JEEM: Vision-Language Understanding in Four Arabic Dialects

Karima Kadaoui[1*] Hanin Atwany[1*] Hamdan Al-Ali[1*]
Abdelrahman Mohamed[1] Ali Mekky[1] Sergei Tilga[2] Natalia Fedorova[2]
Ekaterina Artemova[2] Hanan Aldarmaki[1] Yova Kementchedjhieva[1]

[1] MBZUAI [2] Toloka AI

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. We find that an impediment to this goal is the lack of reliable evaluation metrics.*

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 [3, 16, 17]. However, these models often struggle to generalize across culturally diverse and dialect-rich environments due to the over-representation of specific geographic regions [7, 9] and standardized language varieties in their training datasets [20]. Similarly, existing evaluation datasets predominantly feature Western-centric images and English text [15, 27], while their non-English counterparts are often derived from the former, either through translation or relabeling of the same images [5]. This results in biased evaluation, which conceals the suboptimal performance of VLMs in geographically and dialectally diverse settings [4].

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 [5, 15, 19, *inter alia*]. Among these, Arabic is rarely included, and when it is, it appears either in its standardized form (Modern Standard Arabic) [23] or a single dialect, such as Egyptian [22]. This approach overlooks the cultural and dialectal diversity found among the ∼400 million speakers of this language.

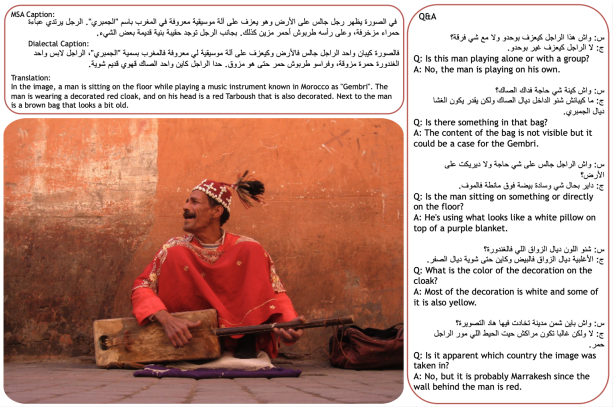

Figure 1. A sample from JEEM (Moroccan set). The image is annotated with an MSA and dialectal captions and five VQA pairs.

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 [10]: **J**ordanian (Leventine), **E**gyptian, **E**mirati (Khaleeji), and **M**oroccan (Maghrebi). JEEM comprises two core tasks: image captioning and VQA. These tasks enable the evaluation of VLMs in terms of their abil-

*Equal contribution.
Correspondence: Karima.Kadaoui@mbzuai.ac.ae
[1]Our data is available at hf.co/datasets/toloka/JEEM.

ity to recognize and appropriately reason about cultural elements, such as traditional clothing, local artifacts, and social settings, while utilizing dialectal language.

We benchmark open- and closed-source VLMs on JEEM and measure performance in terms of standard count-based metrics, LLM-as-a-judge evaluation, and human evaluation. This comprehensive evaluation protocol reveals that automatic metrics exhibit weak correlation with human judgments. We outline next steps for addressing this issue, critical to the effectivebenchmarking of VLMs on image captioning and visual-question answering in Dialectal Arabic.

2. 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. Their sociodemographic statistics, collected through a voluntary survey, can be found in Table 4, 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. TThe process is visualized in Figure 2, Appendix A.

Annotators in each dialect group were assigned images specific to their region. In addition, a set of 25 images per dialect was manually selected to form a *shared* pool of 100 images. These images were annotated in all four dialects to enable the exploration of cross-cultural perspectives.

Country	Image Count	Average Length		Unique Words	
		DA	MSA	DA	MSA
Jordan	606	46	52	8,933	9,751
Emirates	132	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 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.

Step 2: Image Captioning The task is to write a description of the given image in both Modern Standard Arabic (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.

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, providing suggestions for improvements and exchanging feedback in a group chat.

Dialectness Each annotator could complete a limited number of tasks per day to avoid having a small number of annotators dominating the annotations. See Figure 3, Appendix A for the distribution of tasks completed by each annotator. To ensure an appropriate level of dialectness, 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 detailed annotation guidelines made available to the annotators can be found in Appendix A, and the final dataset statistics can be found in Table 1.

3. Data Analysis

Cultural Aspects in Image Captioning 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 involves an image of Omani Halwa, a traditional Gulf dessert with a brown color, made from margarine, sugar, rose water, and semolina. Among the four dialectal captions, only the Emirati one correctly identifies the dessert as Omani Halwa, while the Jordanian one misidentifies it as a visually similar dessert, Karawya, and both Moroccan and Egyptian captions mistakenly describe it as a chocolate dessert, showcasing the diverse regional influences on object recognition. The full example can be found in Figure 4, Appendix B.

Type of Questions 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. 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 Figure 5, Appendix B.) 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 the total, followed by Yes/No questions at 26.42%, Categorical questions at 18.83%, and Quantitative questions at 8.83%.

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.

4. Benchmarking Models on Image Captioning

We benchmark five open-source Arabic-capable VLMs: Maya [1], PALO [21], Peacock [2], AIN [11], and AyaV [6], as well as GPT-4o [17] on the image captions in JEEM. Details about model prompts can be found in Appendix C. We measure image captioning performance using three count-based metrics—CIDEr (C) [25], ROUGE-L (R) [14], and BLEU-4 (B) [18]—one embedding-based metric, BertScore[28],[2], and a LLM-as-a-judge method based on GPT-4 Turbo. The criteria for the latter are as follows: *i*) **Consistency, Con** – alignment with the visual content; *ii*) **Relevance, Rel** – focus on the most important elements; *iii*) **Fluency, Flu** – naturalness and clarity of expression; *iv*) **Dialect authenticity, Auth** – use of appropriate regional language. Each criterion was rated on a five-point Likert scale (1 = poor, 5 = excellent). Evaluation was conducted in two settings: with both the image and reference caption (**Image + Ref**), and with the reference alone (**Ref Only**) [24], allowing us to isolate the impact of visual grounding. See Appendix C for more details. The cost of LLM-based evaluation was $16 for image captioning.

Human Evaluation We further conduct a human evaluation using the same four criteria. For each dialect, 50 images were sampled, and both model outputs and reference captions were rated by native speakers, blind to the source. See Appendix D for evaluation guidelines. Following prior work [12, 26], we use Kendall's Tau-C to assess alignment between automatic metrics and human judgments.

5. Image Captioning Results

Automatic Metrics In Table 3 we show results for the two top-performing models (according to human evaluation) on the four dialectal subsets of JEEM, and in Table 5, Appendix E we include the complete results. We find that count and embedding-based metrics show lower scores than typically observed in English, likely due to the morpholog-

[2]Computed with a CamelBERT [13] backbone. Following [28], we report recall instead of the F1 score.

	Model	Traditional Metrics				GPT Eval (Image + Ref)*				GPT Eval (Ref Only)*				Human Eval*			
		B	C	R	BSc	Con	Rel	Flu	Auth	Con	Rel	Flu	Auth	Con	Rel	Flu	Auth
	AyaV	2.68	0.83	7.59	89.34	3.04	3.16	4.30	2.82	1.78	2.04	4.08	3.06	4.26	4.38	4.04	2.26
	GPT-4o	**5.23**	**6.91**	**9.66**	**90.72**	**3.84**	**4.00**	**4.72**	**3.32**	**2.36**	**2.82**	**4.58**	**3.70**	**4.84**	**4.88**	**4.66**	**3.56**
	AyaV	1.69	0.88	5.80	**90.24**	2.94	3.20	4.34	2.22	1.70	1.76	4.18	2.34	**3.86**	**4.16**	4.74	1.14
	GPT-4o	**3.19**	**2.73**	**7.21**	89.03	**3.58**	**3.84**	**4.74**	**2.58**	**1.96**	**2.34**	**4.54**	**2.62**	3.24	3.32	**4.88**	**2.20**
	AyaV	2.87	0.83	7.85	89.82	2.76	2.80	4.26	3.78	1.74	1.94	4.04	4.02	3.58	4.18	**3.70**	3.44
	GPT-4o	**4.09**	**8.41**	**8.56**	**90.64**	**3.16**	**3.40**	**4.64**	**3.88**	**1.88**	**2.16**	**4.38**	**4.08**	**4.44**	**4.36**	**3.70**	**4.34**
	AyaV	2.21	0.53	6.55	88.33	2.64	2.98	3.98	3.56	1.78	1.92	3.92	3.94	3.72	3.64	3.28	3.04
	GPT-4o	**4.73**	**6.70**	**9.00**	**89.98**	**3.48**	**3.50**	**4.70**	**4.12**	**2.04**	**2.48**	**4.38**	**4.28**	**4.64**	**4.59**	**4.55**	**4.38**
	τ_c	19.79	11.78	15.69	10.62	31.73	34.31	15.42	34.61	25.95	31.64	14.75	34.45	-	-	-	-

Table 3. Image captioning evaluation. * Metrics computed on the same 200-image sample (50 per dialect).

ical richness of Arabic. All four metrics exhibit low correlation with human judgments, which renders them unfit for the task. The average scores obtained with GPT-4 Turbo show stronger alignment with human judgments across Consistency, Relevance, and Fluency when both the image and reference caption are provided. In the image-plus-reference configuration, Kendall's τ_c reaches 34.3 for Relevance and 31.7 for Consistency, compared to 31.6 and 26.0, respectively, in the reference-only setup. This trend aligns with findings from [24], highlighting that including visual context improves the reliability of GPT-based evaluations. While Dialect Authenticity shows relatively stable correlation across both settings, the observed gains in other criteria suggest that visual information helps GPT-based evaluators make more grounded and accurate judgments.

Human Evaluation Based on the human evaluation scores in Table 3, we observe that GPT-4o outperforms AyaV on nearly all criteria and dialects, often with a considerable margin. A general issue observed with open-source models is their inability to produce text in Dialectal Arabic. This problem is especially pronounced in the low-resource Emirati dialect, where even GPT-4o scores considerably lower on the Auth metric, compared to other dialects. Both models show the best dialect authenticity for Egyptian Arabic, known to be relatively high in resources compared to the rest. For the other three criteria, the gap is smaller between AyaV and GPT-4o, since AyaV can get credit even with a caption in MSA, which does not strictly meet the task requirements but is reflective of the general vision-language capabilities of the model. Here, AyaV outperforms GPT-4o on consistency and relevance in Emirati Arabic, but lags behind with up to a point for the remaining dialects on these two criteria, and on fluency for all dialect.

6. Conclusion

We present JEEM, a culturally-representative benchmark for four Arabic dialects—Jordanian, Emirati, Egyptian, and Moroccan—across image captioning and VQA tasks. We evaluated open- and closed-source VLMs using automatic metrics and human judges, finding that the former only weakly correlate with the latter. Results show that low-resource varieties like Emirati Arabic are particularly challenging, but even higher-resource varieties like Egyptian Arabic leave room for improvement. In a trend common to various vision-language evaluation context beyond Dialectal Arabic, GPT-4o proves superior to open-source VLMs.

Future Directions A key challenge we identified relates to the automatic evaluation of dialectal image captions, which hinders immediate progress in image captioning in Dialectal Arabic. We continue to experiment with alternative recent metrics which take on a more structured approach of image and text decomposition [8]. As these are primarily developed for and tested on English, it remains to see how viable their adaptation and application to Dialectal Arabic will prove. In parallel, we are carrying out more extensive human evaluation of the image captioning capabilities of VLMs, and also of their visual question-answering capabilities. The automatic evaluation of visual question-answers is also non-trivial as even seemingly simple question types, such as yes/no questions, can pose a challenge when the answer is not a simple 'yes' or 'no', but a more nuanced elaboration on the contents of the image. This is the true nature of in-the-wild question answering, which arises organically from the image-blind setup for answer writing that we used. Yet, existing automatic metrics fail in this setting of open-ended question-answering, much like they do in image captioning.

Licensing Information The images in JEEM are subject to the underlying licensing terms of Wikimedia Commons[3] and Flickr[4]. The image captions and questions-answer pairs will be distributed under the MIT license[5].

Fair Job Conditions Our team of writers is based in the

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

United Arab Emirates, Jordan, Morocco, and Egypt. Their pay rates exceed the respective hourly minimum wages. Annotations are collected and stored anonymously. Writers are informed in advance about potentially sensitive or harmful content in the images.

References

[1] 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. Maya: An instruction finetuned multilingual multimodal model. *arXiv preprint arXiv:2412.07112*, 2024. 3

[2] Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 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, 2024. Association for Computational Linguistics. 3

[3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1

[4] Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 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, 2024. 1

[5] Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, 2023. 1

[6] Cohere. Aya vision: Expanding the worlds ai can see, 2025. Accessed: March 21, 2025. 3

[7] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59, 2019. 1

[8] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption, 2024. 4

[9] Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. 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*, 2023. 1

[10] Nizar Y. Habash. *Introduction to Arabic natural language processing*. Morgan and Claypool Publishers, 1 edition, 2010. 1

[11] Ahmed Heakl, Sara Ghaboura, Omkar Thawkar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman Khan. Ain: The arabic inclusive large multimodal model. *arXiv preprint arXiv:2502.00094*, 2025. 3

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 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, 2021. Association for Computational Linguistics. 3

[13] Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 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), 2021. Association for Computational Linguistics. 3

[14] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 3

[15] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 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, 2021. Association for Computational Linguistics. 1

[16] LlamaTeam. The llama 3 herd of models, 2024. 1

[17] Team OpenAI. Gpt-4 technical report, 2024. 1, 3

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. Association for Computational Linguistics. 3

[19] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, 2022. 1

[20] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[21] Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. Palo: A polyglot large multimodal model for 5b people. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1745–1754, 2025. 3

[22] 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. Cvqa: Culturally-diverse multilin-

gual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024. 1

[23] 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. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024. 1

[24] Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o, 2024. 3, 4

[25] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[26] Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 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, 2024. 3

[27] Yuxuan Wang, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, and Wanxiang Che. Cvlue: A new benchmark dataset for chinese vision-language understanding evaluation, 2024. 1

[28] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 3

JEEM: Vision-Language Understanding in Four Arabic Dialects

Supplementary Material

A. Data Annotation

A.1. Demographic Profiles

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); Emirates: 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+: 12.5
What is your current employment status?	Not working: 8.2, Self-employed: 25, Part-time: 33.3, Full-time: 33.3

Table 4. Results of the voluntary survey of 24 respondents.

Table 4 presents the demographic and professional background of respondents who participated in our voluntary survey. The survey was conducted via Google Forms and received 24 responses. Participants were not compensated for their time. The aim was to gather demographic information on the linguistic and professional diversity of contributors.

A.2. Image Captioning Instructions

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.

A.3. Question Writing Instructions

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?"

Data Annotation Pipeline

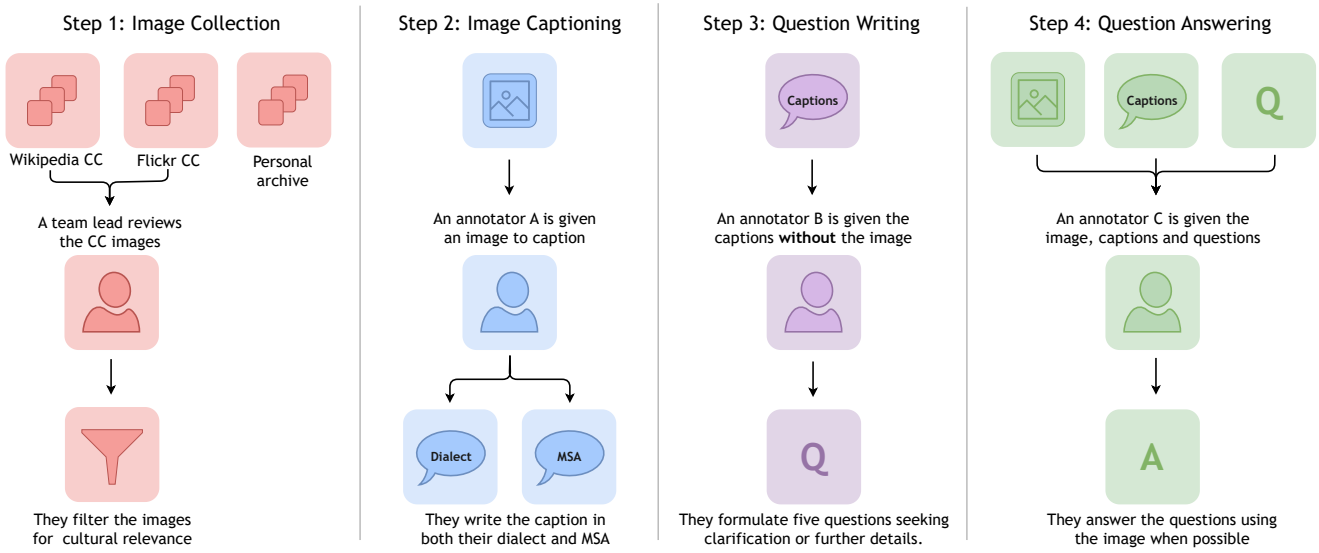

Figure 2. Data annotation pipeline.

Annotator Contributions by Task and Dialect

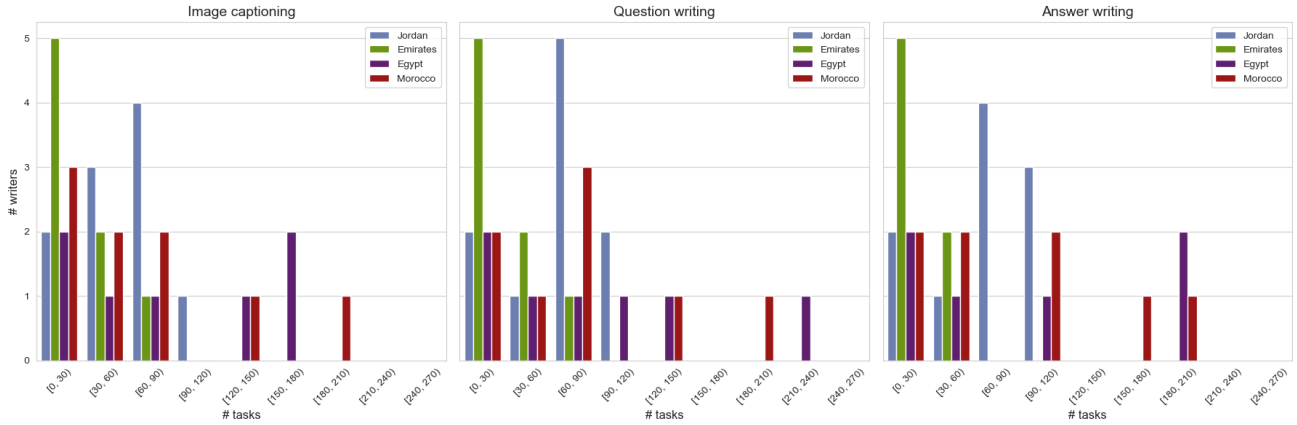

Figure 3. Distribution of annotators based on the number of tasks completed for three tasks: Image Captioning, Question Writing, and Answer Writing. Each bar represents the number of writers contributing within a given range, with colors indicating different dialects. Y-axis: number of unique writers. X-axis: the number of tasks grouped into intervals.

- **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 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 simply ask, "What color is it?"

A.4. Question Answering Instructions

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. Data Analysis

B.1. An Example of Shared Image

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 يديدة 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 4. 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 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: مين شو اسم المطعم فالصورة؟
النوع: تحقق

سؤال: هل البهال 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 5. Question Type Identification Prompt. The task is to determine the type of a given question in dialectal Arabic based on a predefined classification.

C. Prompt Templates Used for Generation

اوصف الصورة

Figure 6. Prompt for generating captions in Modern Standard Arabic (MSA). Translation: *Describe the image.*

اوصف الصورة باللهجة المحلية

Figure 7. Prompt for generating captions in a regional dialect. Translation: *Describe the image in the* {*dialect name*} *dialect,* where {`dialect name`} is replaced with the target dialect (e.g., Egyptian or Moroccan).

D. Evaluation

D.1. Human Evaluation Instructions

You are presented with an image and an image caption. You need to look closely at the image, read its caption, and evaluate the caption according to the following four criteria 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 use?

Note that fluency and dialectal language are not the same. A caption might be non-fluent but still dialectal.

D.2. GPT-4 Turbo Prompts

To evaluate the quality of generated captions, we leverage GPT-4 Turbo in two settings: (1) image + reference caption as in Figure 8, and (2) reference caption only as in Figure 9. In (1) and (2), we assess four criteria from human evaluation: consistency, relevance, fluency, and dialect authenticity. (2) removes the image, focusing solely on text-to-text alignment. This setup isolates the role of visual context: (1) measures grounded caption quality, while (2) evaluates textual fidelity.

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 8. Evaluation prompt using **image and reference caption**. For MSA, Dialect Authenticity was omitted.

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):
Consistency: X/5
Relevance: X/5
Fluency: X/5
Dialect Authenticity: X/5

Figure 9. Evaluation prompt using **reference caption only**. For MSA, Dialect Authenticity was omitted.

E. Complete Results

	Model	Traditional Metrics				GPT Eval (Image + Ref)*				GPT Eval (Ref Only)*				Human Eval*			
		B	C	R	BSc	Con	Rel	Flu	Auth	Con	Rel	Flu	Auth	Con	Rel	Flu	Auth
MSA	AIN	4.00	1.05	7.46	80.31	2.62	2.45	4.23	-	1.54	1.58	4.02	-	-	-	-	-
	AyaV	4.10	0.76	9.85	90.36	2.90	3.24	4.50	-	1.72	1.82	4.18	-	-	-	-	-
	Maya	4.25	1.79	9.47	90.35	2.32	2.52	4.00	-	1.54	1.60	3.80	-	-	-	-	-
	PALO	4.26	1.76	9.48	**90.46**	2.33	2.55	3.96	-	1.49	1.57	3.76	-	-	-	-	-
	Peacock	2.08	1.51	7.18	84.24	1.73	1.90	3.82	-	1.16	1.21	3.07	-	-	-	-	-
	GPT-4o	**5.87**	**7.27**	**10.61**	90.35	**3.50**	**3.62**	**4.62**	-	**1.90**	**2.22**	**4.26**	-	-	-	-	-
Jordan	AIN	2.19	0.45	5.57	81.55	2.77	2.79	4.38	2.65	1.63	1.79	4.23	2.63	2.94	3.13	4.25	1.25
	AyaV	2.68	0.83	7.59	89.34	3.04	3.16	4.30	2.82	1.78	2.04	4.08	3.06	4.26	4.38	4.04	2.26
	Maya	1.91	0.49	6.44	90.16	2.50	2.60	4.12	2.60	1.54	1.66	3.94	2.72	3.08	3.42	4.04	2.22
	PALO	2.05	0.68	6.63	**90.73**	2.56	2.52	4.28	2.58	1.64	1.74	4.20	2.64	3.36	3.26	3.74	1.08
	Peacock	1.55	1.88	5.91	83.57	2.32	2.36	3.56	2.52	1.42	1.56	3.48	2.60	2.32	2.28	3.46	1.32
	GPT-4o	**5.23**	**6.91**	**9.66**	90.72	**3.84**	**4.00**	**4.72**	**3.32**	**2.36**	**2.82**	**4.58**	**3.70**	**4.84**	**4.88**	4.66	3.56
	Human	-	-	-	-	-	-	-	-	-	-	-	-	4.74	4.76	**4.78**	**4.82**
Emirates	AIN	1.63	0.52	5.03	81.89	2.59	2.76	4.20	2.07	1.49	1.60	4.15	2.06	3.34	3.62	4.74	1.00
	AyaV	1.69	0.88	5.80	**90.24**	2.94	3.20	4.34	2.22	1.70	1.76	4.18	2.34	3.86	4.16	4.74	1.14
	Maya	1.55	0.36	5.98	89.43	2.44	2.42	4.20	2.04	1.52	1.58	4.16	2.06	1.84	2.20	4.56	2.00
	PALO	1.50	0.29	5.75	89.43	2.53	2.71	4.16	1.88	1.57	1.69	4.12	1.90	2.38	2.78	**5.00**	1.00
	Peacock	1.23	0.95	4.09	79.09	2.04	2.06	3.86	1.82	1.22	1.33	3.84	1.73	2.28	2.84	4.00	1.00
	GPT-4o	**3.19**	**2.73**	**7.21**	89.03	**3.58**	**3.84**	**4.74**	**2.58**	**1.96**	**2.34**	**4.54**	**2.62**	3.24	3.32	4.88	2.20
	Human	-	-	-	-	-	-	-	-	-	-	-	-	**4.48**	**4.72**	4.92	**4.94**
Egypt	AIN	2.08	0.31	5.20	79.60	2.40	2.36	4.23	2.47	1.44	1.58	4.06	2.23	3.67	3.79	4.27	2.21
	AyaV	2.87	0.83	7.85	89.82	2.76	2.80	4.26	3.78	1.74	1.94	4.04	4.02	3.58	4.18	3.70	3.44
	Maya	2.16	0.49	6.67	**90.82**	2.06	2.14	4.00	2.46	1.40	1.48	4.04	2.36	3.76	3.12	2.88	1.64
	PALO	2.05	0.52	6.37	91.10	2.13	2.55	4.18	2.55	1.61	1.78	4.10	2.37	3.48	4.02	4.50	1.74
	Peacock	0.86	0.54	4.50	81.88	2.04	1.98	3.92	2.32	1.28	1.34	3.78	2.24	3.14	2.52	3.40	1.66
	GPT-4o	**4.09**	**8.41**	**8.56**	90.64	**3.16**	**3.40**	**4.64**	**3.88**	**1.88**	**2.16**	**4.38**	**4.08**	4.44	**4.36**	3.70	4.34
	Human	-	-	-	-	-	-	-	-	-	-	-	-	**4.46**	4.16	**4.62**	**4.90**
Morocco	AIN	1.34	0.58	3.40	81.37	2.54	2.75	4.40	1.69	1.44	1.50	4.06	1.27	4.00	3.46	4.35	1.00
	AyaV	2.21	0.53	6.55	88.33	2.64	2.98	3.98	3.56	1.78	1.92	3.92	3.94	3.72	3.64	3.28	3.04
	Maya	1.06	0.37	3.79	88.85	2.26	2.54	4.10	1.72	1.48	1.56	3.98	1.32	3.09	3.28	4.01	2.32
	PALO	1.06	0.46	3.76	89.47	2.66	2.78	4.20	1.74	1.60	1.72	4.18	1.28	4.14	3.80	4.70	1.00
	Peacock	0.51	0.40	2.55	79.88	1.98	1.98	2.72	2.04	1.36	1.42	2.66	2.48	3.56	2.56	3.86	1.00
	GPT-4o	**4.73**	**6.70**	**9.00**	**89.98**	**3.48**	**3.50**	**4.70**	**4.12**	**2.04**	**2.48**	**4.38**	**4.28**	4.64	4.59	4.55	4.38
	Human	-	-	-	-	-	-	-	-	-	-	-	-	**4.84**	**4.70**	**4.87**	**4.97**
	τ_c	19.79	11.78	15.69	10.62	31.73	34.31	15.42	34.61	25.95	31.64	14.75	34.45	-	-	-	-

Table 5. Image captioning evaluation. Human scores are included where available. * Metrics computed on the same 200-image sample (50 per dialect).