

Exploring Visual Culture Awareness in GPT-4V: A Comprehensive Probing

Anonymous ACL submission

Abstract

Pretrained large Vision-Language models have drawn considerable interest in recent years due to their remarkable performance. Despite considerable efforts to assess these models from diverse perspectives, the extent of visual cultural awareness in the state-of-the-art GPT-4V model remains unexplored. To tackle this gap, we extensively probed GPT-4V using the MaRVL benchmark dataset, aiming to investigate its capabilities and limitations in visual understanding with a focus on cultural aspects. Specifically, we introduced three visual related tasks, i.e. caption classification, pairwise captioning, and culture tag selection, to systematically delve into fine-grained visual cultural evaluation. Experimental results indicate that GPT-4V excels at identifying cultural concepts but still exhibits weaker performance in low-resource languages, such as Tamil and Swahili. Notably, through human evaluation, GPT-4V proves to be more culturally relevant in image captioning tasks than the original MaRVL human annotations, suggesting a promising solution for future visual cultural benchmark construction.

1 Introduction

Culture serves as the representative symbol reflecting the internal norms and values of diverse human groups (Hofstede, 1984; Mora, 2013; Tomlinson et al., 2014; Gabriel, 2020), including various domains like language, cuisine, common sense, architecture, and activities, etc. The development and growth of artificial intelligence techniques enable us to tackle more complicated tasks, prompting a growing interest in exploring cultural aspects within current models (Arora et al., 2022; Li et al., 2023b; Yang et al., 2023; Huang and Yang, 2023).

Recently, cultural probing in language domains has gained widespread attention, such as probing Delphi model (Talat et al., 2022), GPT-3 (Johnson et al., 2022) and ChatGPT (Cao et al., 2023), etc. While pretrained large Vision-Language models



Caption: Picha ya upande wa kulia ina waumini wameshika kitabu cha dini wakati wa Ibada ya **Ijumaa Kuu** na picha ya upande wa kushoto ina mtu aliyebeba msalaba akifuatwa na watu wengine nyuma wakiadhimisha **Ijumaa Kuu**. (The picture on the right has believers holding a religious book during the **Good Friday** Service and the picture on the left has a person carrying a cross followed by other people in the background celebrating **Good Friday**.) **Label:** True

(a) An example of MaRVL in Swahili.



MaRVL: 兩張圖中的飯廳裡都沒有人。(There is no one in the **dining room** in either picture.)

GPT-4V: 第一张图片展示了一个**现代风格**的饭厅，有着城市景观；第二张则是**传统中式风格**。(The first image shows a **modern style dining room** with a city view; the second one is in a **traditional Chinese style**.)

(b) Caption Comparison between MaRVL and GPT-4V.

Figure 1: Examples of captions from human annotation and GPT-4V where description of GPT-4V is more culturally related than original human annotation, excelling not only in grasping **cultural concept** but also in capturing **fine-grained cultural aspects**.

have demonstrated remarkable performance in visual tasks (Li et al., 2023a; Geigle et al., 2023; Fan et al., 2023; Zhu et al., 2023; Khan and Fu, 2023), there is a pressing need to explore visual cultural assessment in the current landscape. Yang et al. (2023) performed some preliminary experiments exploring multicultural understanding using GPT-4V¹ merely by investigating cultural case studies rather than extensively assessing its performance through quantified metric scores. The extent of cross-cultural awareness in GPT-4V, specifically designed for visual tasks, still remains unexplored

¹We utilize GPT-4-vision-preview model (abbr. GPT-4V) for visual tasks: <https://openai.com/gpt-4>.



Figure 2: Our visual cultural probing framework with MaRVL benchmark dataset, includes three tasks: caption classification, pairwise captioning, and culture tag selection.

(Chang et al., 2023). Therefore, we propose a systematic probing framework to scrutinize the cultural awareness of GPT-4V step by step (Figure 2), aiming to address the following three questions:

RQ1: Can the Language and Visual understanding of GPT-4V be effectively aligned? To address this question, we employ the caption classification task. In this task, we check if the model is able to tell whether the description and images are aligned related to a given concept. Our experimental findings indicate that GPT-4V demonstrates excellent performance in this regard, although it exhibits weaknesses in low-resource languages such as Tamil, Swahili, and Indonesian.

RQ2: Can GPT-4V extract fine-grained visual cultural features and provide high-quality descriptions? To address this, we conduct pairwise captioning task, to assess the ability of the model to capture and summarize the fine-grained cultural-related features exclusively from the image. Extensive evaluation demonstrate that the text generated by GPT-4V surpasses the quality of the initial annotations, indicating its potential to enrich visual cultural datasets.

RQ3: Is GPT-4V proficient in perceiving cross-cultural knowledge? To answer this question, we evaluate how accurately the model can link visual elements presented in given images to broader cultural categories. We formulate this task as multi-culture classification, which is named as culture tag selection. This probes the extent of knowledge that the model possess about different cultures and its ability to categorize cross-culture concepts from only visual clues.

In summary, our key contributions are as follows: 1) We introduce a framework comprising three specific tasks for assessing cultural awareness of GPT-4V; 2) We extensively employ both automated and human evaluations to study the culture understanding of the model; 3) Experimental results prove the outstanding performance of GPT-4V in fine-grained culture understanding as well as highlight challenges in low-resource cultures.

2 Probing Framework

In this section, we elaborate on the probing tasks, human evaluation, and cultural concept filtering.

Rationality. The rationale behind our probing framework lies in its meticulous design, rooted in the principles of progressing from a shallow to a deep understanding and evolving from monocultural to cross-cultural comprehension. Firstly, we input images and captions, requiring the model to demonstrate a coarse-grained understanding of monocultural attributes. Further, we challenge the model to generate cultural interpretations of the images, which are then evaluated by human for its ability to capture and comprehend fine-grained visual cultural aspect. Lastly, we investigate the cross-cultural understanding capabilities of GPT-4V by presenting it with images rooted in Chinese culture and assigning multiple cultural tags.

2.1 Cultural Benchmark

We leverage Multicultural Reasoning over Vision and Language (MaRVL) (Liu et al., 2021) as our probing material, the first and only multimodal

culture reasoning dataset with a diverse set of languages and cultures, including Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish. All the concepts and related images are derived from native speaker annotators.

2.2 Probing Tasks

We designed three tasks to probe the visual culture awareness of GPT-4V, as shown in Figure 2.

Caption Classification Firstly, we replicate the consistent experiment conducted by Liu et al. (2021) and Li and Zhang (2023), namely caption classification. Given that the MaRVL dataset is annotated with captions elucidating the cultural concept depicted in image pairs, we anticipate that GPT-4V possesses the capability to identify these cultural concepts. As shown in Figure 2(a), for each pair of images corresponding to a specific concept, the model is prompted to generate a binary label, i.e. *True* or *False*, indicating whether the provided description accurately aligns with the depicted cultural concept. We also present the evaluation results using open-source models for a more comprehensive comparison.

Pairwise Captioning Illustrated in Figure 2(b), for each pair of images of a given concept, the model is prompted to generate a caption focusing on the differences or the commonalities of the concept presented in the images. We sample a pair of images for each of the concepts in the MaRVL dataset for all five languages. This ended with 429 samples in total. For details of sample statistics and prompts please refer to Appendix A and D.

To better assess the level of cultural awareness in the generated descriptions, which cannot be reflected by conventional captioning metrics, we conduct a human evaluation for this task. It is designed to measure the descriptions of the correctness, culture relevance, and overall quality as the preferred description. More specifically, for each sample with paired images and corresponding descriptions, we request human annotators to assign scores ranging from 1 to 7 for correctness and culture relevance and record their preference by selecting either one of the descriptions or indicating a tie. We ensure that the annotators are native speakers of the corresponding language, and each is assigned five samples after completing the evaluation tutorial and passing attention test. Please refer to Appendix D for more details on human evaluation.

Culture Tag Selection. We construct distinct subsets of concepts within the realm of cultural relevance and cultural neutrality utilizing the MaRVL dataset. In MaRVL, a concept class is represented through dozens of images, with some concepts embodying general objects in the real-world that lack inherent cultural relevance, e.g., "62-胡萝卜" (62-Carrot). To narrow our focus to the cultural domain, we introduce a human evaluation framework involving native speakers tasked with assessing the presence of cultural signals in image pairs from the same concept; see Figure 2 (c). This evaluation contributes to establishing filter criteria for the entire dataset, derived from the consensus among all annotators in each language.

Subsequently, we employ these criteria to filter data from MaRVL, retaining image pairs that exhibit unanimous decisions among annotators. Given these remaining image pairs, GPT-4V is prompted to select a culture tag that is most relevant from the five given options. The performance of GPT-4V is evaluated through metrics such as Recall and False Positive Rate, providing insights into its visual cultural awareness.

3 Experiments

3.1 Experimental Setup

For all the probing tasks, we utilize 5GB CPUs to conduct interactive processes with GPT-4V. For the pairwise captioning task we leverage the Prolific crowdsourcing platform² to carry out human evaluation. We ensure that all the annotators belong to the cultural background with the language being evaluated. To ensure a diverse and comprehensive assessment, we enlist a minimum of 10 annotators for each task. For the cultural tag selection task, we invite three native Chinese speakers to filter concepts by presenting image pairs and filtering non-cultural images in a Chinese subset. More technical details are presented in Appendix C.

3.2 Results

Caption Classification. As shown in Table 1, we present the accuracy scores for each language. A notable observation is that GPT-4V demonstrates the highest performance across all languages, showcasing a substantial improvement from 58.9 to 67.4 compared to other models, which answers **RQ1**. Besides, we observe that GPT-4V outperforms in Chinese and Turkish languages but exhibits lower

²<https://app.prolific.com/>

Model	Method	ZH	TA	SW	ID	TR	avg.
mUNITER	MaRVL*	56.8	52.2	51.5	55.0	54.7	54.0
	IGLUE [◊]	55.3	52.7	51.2	54.8	54.7	53.7
	Mixup [†]	60.4	58.9	54.9	55.7	59.5	57.9
xUNITER	MaRVL*	55.0	55.1	55.5	57.1	58.0	56.1
	IGLUE [◊]	53.1	53.1	55.5	55.1	56.2	54.6
	Mixup [†]	57.9	58.0	60.0	58.8	59.7	58.9
GPT-4V	MaRVL	71.8	63.0	64.8	65.2	69.1	67.4

Table 1: F1 score in zero-shot caption classification, where [◊] from Liu et al. (2021), [◊] from Bugliarello et al. (2022), and [†] from Li and Zhang (2023), demonstrating superior performance of GPT-4V across all languages.

performance in Tamil, Swahili, and Indonesian languages, diverging from the patterns observed in mUNITER and xUNITER models.

Caption Generation To answer **RQ2**, we compare the Bertscore (Zhang et al., 2019), distinct-1 scores (Li et al., 2016) of GPT-4V outcomes and MaRVL annotations, as detailed in Figure 3. To further evaluate the cultural impact of the concerned models, we conduct human evaluation experiments shown in Table 2. Notably, both the original MaRVL caption annotations (abbr. MaRVL’s captions) and the generated captions GPT-4V receive high correctness scores, averaging around 5.0 out of 7, indicating precision in concept descriptions. Nevertheless, in terms of cultural relevance, GPT-4V exhibits a significant improvement compared to MaRVL among all languages, albeit still weaker in Indonesian and Turkish.

Interestingly, when voting for better captions, approximately 70% of annotators deem GPT-4V more acceptable than MaRVL, with less than 30% of cases in MaRVL deemed of good quality. Upon closer examination of MaRVL, it becomes apparent that many cases merely describe differences in image pairs rather than cultural relevance, raising questions about the reliability of the human annotation process. Additionally, this experiment highlights the promising capability of GPT-4V in annotating data for cultural-related tasks. For a comprehensive overview of intuitive case studies, please refer to Appendix E, where comparisons of captions across all languages are summarized.

Culture Tag Selection. Unlike the previous tasks, wherein image pairs were drawn from a composite of cultural relevance and cultural neutrality subsets, we prompt GPT-4V in English, as opposed to any of the languages within the selection, with the aim of mitigating induction biases. In this

	ZH	TA	SW	ID	TR	avg.
MaRVL _{Cor}	5.024	5.000	5.067	5.915	4.189	5.039
GPT4V _{Cor}	5.683	5.577	5.371	5.623	5.552	5.561
MaRVL _{Cul}	2.706	3.352	3.730	2.717	2.378	2.977
GPT4V _{Cul}	4.238	4.183	4.169	3.283	3.371	3.849
All bad	9.5%	7.0%	2.2%	20.8%	12.6%	10.4%
All good	20.6%	29.6%	22.5%	19.8%	12.6%	21.0%
MaRVL	10.3%	15.5%	20.2%	17.9%	15.4%	15.9%
GPT-4V	59.5%	47.9%	55.1%	41.5%	59.4%	52.7%

Table 2: Human evaluation of the pairwise captioning task, indicating that cultural aspect in GPT-4V outcomes surpasses the human annotations of MaRVL dataset.

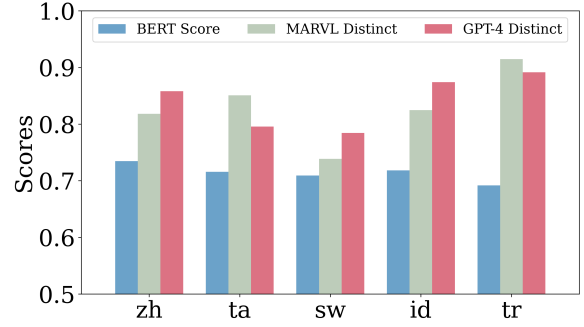


Figure 3: Reference-based evaluation comparison in the cultural pairwise captioning task.

way, we can evaluate cultural awareness in GPT-4V purely based on images. In the specified language, e.g., Chinese, GPT-4V achieves an accuracy of 84.3%, a recall rate of 87.5%, and a false positive rate of 20.7%. Evidently, it can be deduced that GPT-4V demonstrates proficiency in recognizing visual concepts specific to the Chinese language. GPT-4V has been trained using multi-cultural images, and its low false positive rate proves its excellent cross-cultural understanding capability. This finding answers **RQ3** effectively.

4 Conclusion

We systematically conduct probing experiments to assess the visual cultural understanding capabilities of GPT-4V. To achieve that, we investigated three culture-aware tasks, built upon a multimodal cultural reasoning dataset named MaRVL. The experiments demonstrate significant enhancements in cultural understanding of GPT-4V, compared with open-source LLMs. Besides, human evaluation proves the reasonable ability of GPT-4V in terms of cultural understanding and annotating cultural captions, suggesting a promising solutions to construct cultural benchmarks. In future, we will explore the problem of constructing cultural benchmarks that incorporate enhanced cultural awareness.

5 Limitations

While our investigation into GPT-4V as the current state-of-the-art Vision-Language model offers valuable insights, there are still several limitations. Firstly, as the training corpus of GPT-4V remains undisclosed, we cannot guarantee whether the MaRVL dataset we employed overlaps with its training data. However, our caption generation task revealed significant differences between generated captions of GPT-4V and MaRVL, indicating minimal dataset leakage. Secondly, our probing framework assumes language as a proxy for expressing diverse cultural backgrounds. This premise, however, is not entirely consistent. For instance, Swahili is spoken in Tanzania, Kenya, and the Democratic Republic of the Congo, yet it represents a shared cultural heritage.

Despite these limitations, our work offers a unique perspective on exploring potential biases related to visual culture in models. Our investigation represents a pioneering effort in assessing the cross-cultural understanding capabilities of visual models, paving the way for future research in this domain. Future directions could involve the development of more comprehensive datasets reflecting diverse cultural backgrounds, as well as exploring additional methods to evaluate cross-cultural understanding in visual models.

References

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. [Probing pre-trained language models for cross-cultural differences in values](#). [arXiv preprint arXiv:2203.13722](#).

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. In [International Conference on Machine Learning](#), pages 2370–2392. PMLR.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. [Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study](#). In [Proceedings of the First Workshop on Cross-Cultural Considerations in NLP \(C3NLP\)](#), pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. [arXiv preprint arXiv:2307.03109](#).

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Uniter: Universal image-text representation learning](#). In [European conference on computer vision](#), pages 104–120. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In [Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics](#), pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In [Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 \(Long and Short Papers\)](#), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. [Improving CLIP training with language rewrites](#). In [Thirty-seventh Conference on Neural Information Processing Systems](#).

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. [Minds and machines](#), 30(3):411–437.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. [mbliip: Efficient bootstrapping of multilingual vision-llms](#). [arXiv](#), abs/2307.06930.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. [News summarization and evaluation in the era of gpt-3](#). [ArXiv](#), abs/2209.12356.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural nlp](#). [ArXiv](#), abs/2203.10020.

Geert Hofstede. 1984. [Culture’s consequences: International differences in work-related values](#), volume 5. sage.

Dirk Hovy and Diyi Yang. 2021. [The importance of modeling social factors of language: Theory and practice](#). In [North American Chapter of the Association for Computational Linguistics](#).

Jing Huang and Diyi Yang. 2023. [Culturally aware natural language inference](#). In [Conference on Empirical Methods in Natural Language Processing](#).

- Douglas B Johnson, Rachel S Goodman, J Randall Patrinely, Cosby A Stone, Eli Zimmerman, Rebecca Rigel Donald, Sam S Chang, Sean T Berkowitz, Avni P Finn, Eiman Jahangir, Elizabeth A Scoville, Tyler Reese, Debra E. Friedman, Julie A. Bastarache, Yuri F van der Heijden, Jordan Wright, Nicholas Carter, Matthew R Alexander, Jennifer H Choe, Cody A Chastain, John Zic, Sara Horst, Isik Turker, Rajiv Agarwal, Evan C. Osmundson, Kamran Idrees, Colleen M. Kiernan, Chandrasekhar Padmanabhan, Christina Edwards Bailey, Cameron Schlegel, Lola B. Chambless, Mike Gibson, Travis J. Osterman, and Lee Wheless. 2023. [Assessing the accuracy and reliability of ai-generated medical responses: An evaluation of the chat-gpt model.](#) [Research Square.](#)
- Rebecca L Johnson, Giada Pistilli, Natalia Menéndez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. [The ghost in the machine has an american accent: value conflict in gpt-3.](#) [arXiv preprint arXiv:2203.07785.](#)
- Zaid Khan and Yun Fu. 2023. [Contrastive alignment of vision to language through parameter-efficient transfer learning.](#) In [The Eleventh International Conference on Learning Representations.](#)
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality.](#) In [European Association for Machine Translation Conferences/Workshops.](#)
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models.](#) In [Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies](#), pages 110–119, San Diego, California. Association for Computational Linguistics.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. [Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models.](#) In [Proceedings of the 40th International Conference on Machine Learning, ICML’23.](#) [JMLR.org.](#)
- Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuanjing Huang, and Zhongyu Wei. 2023b. [Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training.](#) In [Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 5939–5958, Toronto, Canada. Association for Computational Linguistics.
- Zhi Li and Yin Zhang. 2023. [Cultural concept adaptation on multimodal reasoning.](#) In [Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing](#), pages 262–276, Singapore. Association for Computational Linguistics.
- Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023a. [On the cultural gap in text-to-image generation.](#)
- 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, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023b. [G-eval: Nlg evaluation using gpt-4 with better human alignment.](#) In [Conference on Empirical Methods in Natural Language Processing.](#)
- Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. 2023c. [Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark \(ccub\) dataset.](#)
- Roberto Menicatti, Barbara Bruno, and Antonio Sgorbissa. 2017. [Modelling the influence of cultural information on vision-based human home activity recognition.](#) In [2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence \(URAI\)](#), pages 32–38.
- Cristina Mora. 2013. [Cultures and organizations: Software of the mind intercultural cooperation and its importance for survival.](#) [Journal of Media Research](#), 6(1):65.
- Zeeraq Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. [On the machine learning of ethical judgments from natural language.](#) In [Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies](#), pages 769–779, Seattle, United States. Association for Computational Linguistics.
- Marc Tomlinson, David Bracewell, and Wayne Krug. 2014. [Capturing cultural differences in expressions of intentions.](#) In [Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers](#), pages 48–57, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. [Is chatgpt a good nlg evaluator? a preliminary study.](#) [ArXiv](#), abs/2303.04048.
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. [The dawn of lmms: Preliminary explorations with gpt-4v\(ision\).](#)

Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna. 2023. [Cultural and linguistic diversity improves visual representations](#).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. 2022. [VLUE: A multi-task multi-dimension benchmark for evaluating vision-language pre-training](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27395–27411. PMLR.

Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. 2023. [Vi-gpt: A generative pre-trained transformer for vision and language understanding and generation](#).

A Prompt setting

In consideration of diverse languages and task requirements, distinct prompts have been tailored. Specifically, for culture checking and cultural pairwise captioning tasks, prompts have been meticulously crafted for each of the five languages. These prompts are designed to guide the model’s focus predominantly towards cultural concepts, endeavoring to closely emulate the data collection methodology of the MaRVL dataset (Liu et al., 2021).

Table 4 showcases the prompts and exemplar responses for the Chinese language. For a comprehensive collection of prompts, kindly consult the configuration file embedded within the codebase. In the context of the culture tag selection task, English has been designated as the language to preclude the introduction of bias into the task.

B Related work

Culture in Multimodal Probing. Modeling cultural factors in language models has received increasing attention in the NLP community (Hovy and Yang, 2021; Hershcovich et al., 2022; Huang and Yang, 2023). In the multimodal domain, previous works have highlighted the need for culture integration within vision-language models (Zhou et al., 2022). Menicatti et al. (2017) find that cultural information can improve the performance of vision systems. Ye et al. (2023) highlight that cultural nuances can affect image annotations and show that cultural/linguistic diversity improves visual representations. Li and Zhang (2023) use multimodal data augmentation to address the data

scarcity problem in low-resource cultures. (Liu et al., 2021) introduce MaRVL, a Multicultural Reasoning over Vision and Language (MaRVL) dataset from 5 different languages and cultures. Liu et al. (2023a) build a Challenging Cross-Cultural dataset of textual prompts to generate images in Chinese cultural style. Liu et al. (2023c) introduce a Cross-Cultural Understanding Benchmark for text-to-image synthesis representing eight distinct cultural backgrounds. These datasets primarily serve as a means to gauge the generalization performance of models without using any visual culture-specific tasks.

LLMs for Evaluation. As LLMs emerge as powerful tools in NLP, their potential as evaluators in various contexts is gaining recognition (Johnson et al., 2023; Goyal et al., 2022). For example, Liu et al. (2023b) present G-Eval, a framework of using LLMs to assess the quality of NLG outputs. Furthermore, Wang et al. (2023) undertake an initial survey to explore the utilization of ChatGPT as an evaluator for NLG. In machine translation, Kocmi and Federmann (2023) propose to use GPT models for translation evaluation. Most relevant to our work, Yang et al. (2023) performed preliminary experiments in multicultural understanding with GPT-4V with a few cases from wedding scenarios.

C Experiment Setting

In Table 3, we present the sample data sizes for each task in our probing experiments. For the culture checking task, we ensured an equivalent data scale with MaRVL to facilitate a fair comparison with results from other zero-shot baseline models. For Pairwise captioning, we randomly sampled from all samples where labels were True for each cultural concept, creating a Probing subset. Our Probing experiments were conducted on a 4-core Linux system using OpenAI library³.

Baseline. The UNITER (Chen et al., 2020) is the popular vision-language model, similar to BERT_{BASE} framework. In our paper, we select two representative baselines, including mUNITER and xUNITER, where mUNITER is initialized by UNITER with mBERT (Devlin et al., 2019) and xUNITER is initialized by UNITER with XLM-R_{BASE} (Conneau et al., 2020).

³Python Package: <https://github.com/openai/openai-python>

文本生成质量评估 (图像→文本)

请阅读下面的两张图片和两句话描述, 并对下面的问题进行评分。

左图



右图



句子a: 左图中只有一支毛笔, 右图则有数支。
句子b: 第一张图片中的毛笔头部尖细, 而第二张中的毛笔头部更为饱满。

问题

正确性(句子a)- 问题1: 句子a能够正确描述两张图片的内容。
(1表示完全错误, 7表示完全正确。
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

蕴含文化(句子a)- 问题2: 句子a的描述内容和文化属性相关。
(1分代表和文化完全没关系, 7分代表和文化强相关)。
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

正确性(句子b)- 问题3: 句子b能够正确描述两张图片的内容。
(1表示完全错误, 7表示完全正确。
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

蕴含文化(句子b)- 问题4: 句子b的描述内容和文化属性相关。
(1分代表和文化完全没关系, 7分代表和文化强相关)。
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

文本对比 - 问题5: 句子a和句子b, 哪个更适合作为两张图片的描述?
☐ 1. 句子a ☐ 2. 句子b
☐ 3. 平局 (质量相同, 都很好) ☐ 4. 平局 (质量相同, 都很差)

* 如有疑问, 请留下您的建议。

提交

Progress: 1/5
User: undefined

Figure 4: Our proposed platform for human evaluation in cultural pairwise captioning task.

Evaluation Metrics. For the purposes of human evaluation in Table 2, we have devised three metrics: correctness (MaRVL_{Cor} , GPT4V_{Cor}), cultural relevance (MaRVL_{Cul} , GPT4V_{Cul}), and voting preference. The correctness and cultural relevance metrics are assigned scores ranging from 1 to 7, while voting preference requires a decision among four choices (All bad, All good, MarRVL is better, GPT-4V is better).

D Human Evaluation

For the cultural pairwise captioning task, we recruited evaluators proficient in the respective languages through Prolific, with the number of evaluators for each language outlined in Table 3. Our evaluation methodology involved the development of a web interface⁴, as depicted in Figure 4. Participants commenced with an attention test following our tutorial, which included a round of guided tasks with prompts to ensure their comprehension of the assignment. This involved answering five questions based on specific options.

Subsequently, participants independently completed the scoring for five rounds of tasks. To en-

⁴Evaluation: https://recipecoastalproject.studio/gpt4_evaluate_cn_1996/

Task	ZH	TA	SW	ID	TR	Total
CC	1012	1242	1108	1128	1180	5670
PC	94	83	78	95	79	429
CTS	188	166	156	190	158	858
NE	25	15	18	21	29	108

Table 3: Statistics pertaining to our probing data samples across the three tasks, including Culture checking (CC), Pairwise Captioning (PC), and Culture Tag Selection (CTS) along with the number of evaluators (NE) sourced from the crowdsourcing platform.

sure the validity of the evaluation, we automatically excluded assessment records with durations under 1 minute. The average evaluation time per user was 6.12 minutes. Notably, the Chinese language had the shortest average time, at only 3.44 minutes, while Swahili exhibited the lengthiest average time, amounting to 9.22 minutes.

E More Case Study

Here, we provide supplementary cases, as shown in Figure 5. It is apparent that, across all languages, the descriptions generated by GPT-4 adeptly capture cultural concepts with precision. Moreover, these descriptions demonstrate a heightened level of granularity in terms of cultural relevance, thereby substantiating the model’s superior quality and enhanced cultural awareness capabilities.

<p style="text-align: center;">Caption Classification</p> <p>Prompt: 任务：我会给你两张图片和一个描述句子，你需要判断这个描述句子是不是正确描述两张图片中的#concept（关于#concept的描述正确即可）。如果给定的句子描述内容为正确的，则输出为[真]，否则输出为[假]。 <i>Task: I will provide you with two images and a descriptive sentence. Your task is to determine whether the descriptive sentence accurately describes the concept (#concept) depicted in the two images. If the given sentence correctly describes the content related to #concept, the output should be [True]; otherwise, the output should be [False].</i> 描述句子:#Caption Sentence: #Caption. 确保严格遵循以下输出格式。 <i>Ensure strict adherence to the following output format.</i> [输出格式]: [Output Format]: 答案: [真/假]。 Answer: [True/False]. 解释: [解释你给出答案的理由]。 Explanation: [Explain the reasoning behind your answer].</p> <p>Output 根据提供的图片，左图显示的是一轮满月而不是月饼，右图展示的是一个月饼，与描述中的悬挂的灯笼不符。因此，描述句子不正确描述了这两张图片。 <i>According to the provided images, the left picture depicts a full moon, not a mooncake, while the right picture shows a mooncake. This is inconsistent with the description of hanging lanterns. Therefore, the descriptive sentence does not accurately describe the content in these two images.</i></p>
<p style="text-align: center;">Pairwise Image Captioning</p> <p>Prompt: 任务：我会给你两张图片，你需要生成一个句子，能够同时描述这两张图片中的#concept (你的描述重点将是关于#concept)。文本内容长度大约10-40个字。 <i>Task: I will provide you with two images, and you need to generate a sentence that can simultaneously describe the #concept in both pictures (with the focus on #concept). The text should be approximately 10-40 words in length.</i> 确保严格遵循以下输出格式。 <i>Ensure strict adherence to the following output format.</i> [输出格式]: [Output Format]: 描述句子: Sentence: [描述内容] [Description]</p> <p>Output 第一张照片展示了室内浪漫的七夕布置，而第二张则是户外放天灯的庆祝活动。 <i>First image showcases romantically decorated indoor setting for Qixi Festival, while the second one captures an outdoor celebration with sky lanterns.</i></p>
<p style="text-align: center;">Culture Tag Selection</p> <p>Prompt: Task: Given two images, select the most appropriate cultural tag from the provided options. Options: Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish." Ensure strict adherence to the specified output format. Output format: Answer: [The option you chose].</p> <p>Output Mandarin Chinese</p>

Table 4: Example instances of our probing methodology encompass three tasks in Chinese and English prompts, please refer to our code for other languages. The italicized text has been translated from Chinese to English using Google Translate for a better understanding.

Chinese (zh)		<p>MaRVL: 兩張圖中的麵條都是被放在盤子或碗哩，而且左圖中還有筷子。(The noodles in both pictures are placed on plates or bowls, and there are chopsticks in the left picture.)</p> <p>GPT-4: 第一张图片显示的是一碗带有肉丸和绿叶蔬菜的亚洲风味汤面条，而第二张图片中的面条是意大利风格，撒有肉末和帕尔马干酪。(The first picture shows a bowl of noodles in an <i>Asian-style soup</i> with meatballs and leafy greens, while the noodles in the second picture are <i>Italian-style</i>, sprinkled with <i>minced meat and parmesan cheese</i>.)</p>
		
Tamil (ta)		<p>MaRVL: இடது பக்க படத்தில் புடவை அணிந்த பெண்ணும் ஒரு பாளையம் இருப்பதை நம்மால் காண முடிகிறது. (In the left picture we can see a woman wearing a sari and a pot.)</p> <p>GPT-4: முதல் படத்தில் பச்சை நிற புடவை பிரகாசிக்கிறது, இரண்டாவதில் சிவப்பு நிற புடவை சடங்குக்கு உகந்தது. (In the first image the <i>green saree</i> shines and in the second the <i>red saree</i> is <i>ritualistic</i>.)</p>
		
Swahili (sw)		<p>MaRVL: Picha ya upande wa kulia ina mchicha ulio ndani ya chombo fulani. (The picture on the right shows spinach in a container.)</p> <p>GPT-4: Kwenye picha ya kwanza, mchicha unaonekana ukiwa umepandwa shambani na watu wawili wanauhudumia, huku picha ya pili ikiwaonyesha mchicha uliostawi vizuri karibu na kikapu kilichojazwa mazao ya kijani. (In the first photo, <i>spinach</i> is seen <i>planted in a field</i> and two people are tending to it, while the second photo shows <i>well-developed spinach</i> next to a basket full of green produce.)</p>
		
Indonesian (id)		<p>MaRVL: Kedua gambar adalah gambar piring putih yang bergambar. (Both images are pictures of white plates with images.)</p> <p>GPT-4: Piring dalam gambar pertama dihiasi dengan pola biru di pinggiran dan digunakan untuk sajian makanan, sedangkan piring dalam gambar kedua menampilkan ilustrasi lengkap pemandangan biru putih yang sepenuhnya untuk dekorasi. (The plate in the first image is decorated with a <i>blue pattern</i> on the rim and is used for serving food, while the plate in the second image features a complete <i>illustration of a blue and white scene</i> that is entirely for decoration.)</p>
		
Turkish (tr)		<p>MaRVL: Soldaki resimde basketbol topuyla oynayan sadece bir kişi görülüyor. (The picture on the left shows only one person playing with a basketball.)</p> <p>GPT-4: Her iki resimde de Türkiye milli basketbol takımı oyuncuları maç sırasında topa hakim. (In both pictures, <i>Turkish national</i> basketball team players dominate the ball during the match.)</p>
		

Figure 5: Additional case studies in the cultural pairwise caption generation task across languages, with cultural-relevant descriptions highlighted in blue.