Benchmarking Vision Language Models for Cultural Understanding

Anonymous ACL submission

Abstract

Foundation models and vision-language pretraining have notably advanced Vision Language Models (VLMs), enabling multimodal processing of visual and linguistic data. However, their performance has been typically assessed on general scene understanding – recognizing objects, attributes, and actions – rather than cultural comprehension. This study introduces CULTURALVQA, a visual question-answering benchmark aimed at assessing VLM's geo-diverse cultural understanding. We curate a diverse collection of 2,378 image - question pairs with 1-5 answers per question representing cultures from 11 countries across 5 continents. The questions probe understanding of various facets of culture such as clothing, food, drinks, rituals, and traditions. Benchmarking VLMs on CULTURALVQA, including GPT-4V and Gemini, reveals disparity in their level of cultural understanding across regions, with strong cultural understanding capabilities for North America while significantly weaker capabilities for Africa. We observe disparity in their performance across cultural facets too, with clothing, rituals, and traditions seeing higher performances than food and drink. These disparities help us identify areas where VLMs lack cultural understanding and demonstrate the potential of CULTURALVQA as a comprehensive evaluation set for gauging VLM progress in understanding diverse cultures.

1 Introduction

Recent multimodal vision-language models (VLMs) (Radford et al., 2021; Liu et al., 2023; Peng et al., 2023; Chen et al., 2024; Lu et al., 2024) have shown impressive performance in tasks such as image-to-text generation (Li et al., 2019), visual question answering (Antol et al., 2015; Goyal et al., 2017), and image captioning (Lin et al., 2014; Vinyals et al., 2015). These tasks predominantly focus on general scene understanding capabilities such as recognizing objects, attributes, and actions

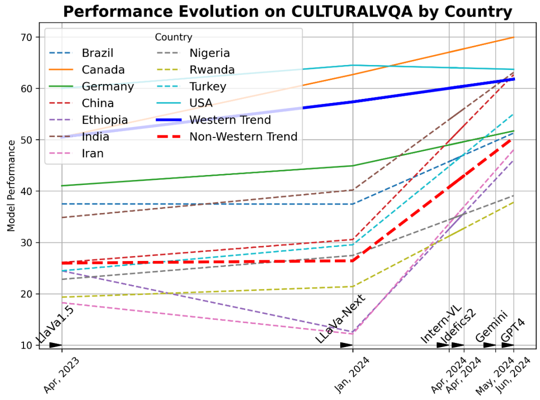


Figure 1: (Top): Samples from CULTURALVQA. (Bottom) The performance of VLMs over time, segmented by non-Western (red) and Western (blue) countries, with model release dates annotated. Dashed and solid lines differentiate trends for non-Western and Western countries, respectively. VLMs' understanding of non-Western cultures has been in a steep upward trend since Jan '24, LLAVA-NEXT (Liu et al., 2024) release.

in scenes containing objects in their common context (Lin et al., 2014). However, given the advancing capabilities of VLMs, we believe the time is now ripe to hold our VLMs to higher standards. We believe that to support increasingly *global* digital interactions, VLMs must also be capable of understanding the *cultural values* (Liu et al., 2021) such as beliefs, rituals, and traditions, for a *variety* of cultures in the world.

In order to adequately assess whether the current state-of-the-art VLMs – including proprietary models such as GPT-4V (OpenAI, 2023) and GEMINI (Gemini Team et al., 2023) – encode cultural knowledge, we need systematic benchmarks. However, evaluating cultural understanding is a challenging task since culture is a multifaceted concept consisting of both tangible (e.g., clothing, and food) as well as intangible elements (e.g., ritual practices). Current benchmarks in this domain, including MaRVL (Liu et al., 2021) and GD-VCR (Yin et al., 2021), while offering foundational insights, have critical shortcomings. MaRVL primarily focuses on visual reasoning tasks (e.g., counting, spatial reasoning) on top of images sourced from various cultures, and lacks probing cultural common sense – the knowledge bank shared by the members of a cultural group (see § 3). GD-VCR although explores commonsense, it is limited by its reliance on movie scenes, which do not encompass the broader spectrum of everyday cultural contexts.

In response to the above challenges, we propose CULTURALVQA, a novel benchmark specifically designed to assess cultural understanding of VLMs. CULTURALVQA is based on Visual Question Answering (VQA), requiring models to integrate both visual and textual information, which permits the formulation of diverse questions, thereby enabling the evaluation of a model's understanding of complex cultural nuances. The CULTURALVQA benchmark extends the language-only CANDLE dataset (Nguyen et al., 2023), which provides a comprehensive collection of cultural commonsense knowledge assertions. We expanded this dataset by automatically collecting images that depict the cultural concept described by the assertions. On top of these images, we collect questions and answers by employing annotators from different cultures who would be familiar with the different cultural concepts depicted in the images. See Fig. 1 (top) for some examples of questions and answers. Our benchmark consists of 2,378 questions collected on top of 2,328 unique images with 1-5 answers per question (total 7,095 answers) from 11 countries.[1]. We also present several analyses to better understand the nature of questions and answers in our benchmark.

Further, we systematically evaluate several state-of-the-art VLMs on CULTURALVQA. Our evaluation reveals a distinct performance gap between proprietary and open-source models, with open-source models significantly underperforming in comparison (a gap of 11.71% between the best-performing open-source and worst-performing closed-source model). Additionally, we observe a significant disparity in model performance across countries. For instance, the highest-performing proprietary model, GPT-4, achieves about 67% accuracy for North American cultural concepts while only 44.15% accuracy on concepts from Africa. VLMs also show varying degrees of proficiency across cultural facets, with closed-source VLMs performing better on questions about rituals and traditions while scoring worse on those related to clothing, food, and drink. We develop CULTURALVQA as a comprehensive evaluation set for gauging VLM progress in understanding diverse cultures and highlighting areas where VLMs lack cultural understanding, with the hope that our benchmark will contribute to accelerating the advancements of VLMs in their cultural understanding, as illustrated in Fig. 1.

2 Related work

Cultural understanding is closely related to geo-diverse understanding. Existing geo-diverse datasets, for instance, the Dollar Street dataset (Gaviria Rojas et al., 2022) includes 38,479 images of everyday household items from homes around the world, while the GLDv2 dataset (Weyand et al., 2020) contains 5 million images and 200k distinct instance labels of natural and human-made landmarks, but both only test recognition capabilities as opposed to cultural understanding. The GD-VCR dataset (Yin et al., 2021) probes cultural understanding, but its reliance on cinematic scenes limits the diversity of real-world cultural contexts it can have. Another related line of work focuses on multilingual understanding. For instance, Bugliarello et al. (2022) bring together five datasets across a number of tasks in 20 languages. However, their focus lies in multilingual understanding (as opposed to cultural understanding). Another multilingual dataset, MaRVL (Liu et al., 2021), tests visually grounded reasoning across multiple languages and cultures. However, MaRVL does not explore the cultural common sense of rituals and traditions. Additionally, the XM3600 dataset (Thapliyal et al., 2022), includes image captions from 36 regions and languages, thus providing a broad geographical coverage but nonetheless contains mostly Western content and lacks depth in the included cultural concepts (Pouget et al., 2024). Closest to our work, the MaXM benchmark (Changpinyo et al., 2023), building on the XM3600 dataset, and the concurrent study by Romero et al.

[1]We provide a data statement in App. A

Figure 2: Samples from CULTURALVQA. Our dataset is comprised of images presenting cultural concepts from 11 countries across five facets: traditions, rituals, food, drink, and clothing. It further includes questions probing cultural understanding of the concepts presented in the images and answers to these questions.

(2024) both utilize the VQA format to explore regional and cultural understanding. MaXM focuses primarily on the ability to process images from varied regions rather than on nuanced cultural understanding. Romero et al. (2024) study cultural questions in a multilingual setup. However, their focus diverges from ours as, like MaRVL, they allocate a much smaller proportion of their dataset to traditions and rituals.

3 CULTURALVQA: Dataset Creation

Cultural Taxonomy Culture is a multifaceted concept that describes the way of life of a collective group of people, distinguishing them from other groups with different cultures (Hofstede et al., 2010; Hershcovich et al., 2022). In this paper, we use the concept of a country as a proxy for a cultural group (Adilazuarda et al., 2024).[2] Our work assumes common ground within a cultural group by probing *culturally relevant concepts* that are collectively understood, as well as shared *cultural common sense* employed in reasoning (Hershcovich et al., 2022). For instance, *lavash* – a traditional Persian bread (see Fig. 2) – is an example of a culturally relevant concept, while the common practice of waltzing at weddings exemplifies the cultural common sense among Germans.

Building on these definitions, we introduce a benchmark that evaluates both the tangible aspects of culture through culturally relevant concepts, such as food, drink, and clothing, as well as the intangible facets via shared common sense embedded in rituals and traditions.[3] We frame this evaluation as a VQA task assessing models' cultural understanding. Starting with a pool of countries, we collect images and use culturally knowledgeable annotators to frame questions. Finally, we collect the ground truth answers.

Selection of Countries To build a benchmark that reflects cultural diversity, we aimed to achieve broad geographical coverage. Our final dataset spans 11 countries and 5 continents. These countries were specifically selected to cover different cultural categories from the World Values Survey (Haerpfer et al., 2022) and include Confucian (China), African-Islamic (Turkey, Iran, Ethiopia, Nigeria, Rwanda), Protestant Europe (Germany), English-speaking (USA, Canada), Latin America (Brazil), and South Asian (India) cultures. We opt for an intentional overrepresentation of African-Islamic countries to address their typical scarcity in geo-diverse datasets.

[2]See § 7 for a discussion of this choice.

[3]Herein, the term 'concepts' is used to encompass both cultural concepts and common sense.

Selection of Images The image selection begins with the CANDLE dataset (Nguyen et al., 2023), which provides a rich collection of Cultural Commonsense Knowledge (CCSK). Each of the 1.1 million entries includes URLs to webpages with relevant CCSK data from the C4 corpus (Raffel et al., 2020). Inspired by findings from (Zhu et al., 2023), which highlighted that 80% of webpages in the C4 corpus contain relevant images, we scrape images from these URLs, focusing particularly on CCSK data from the geography and religion domains of our selected countries.

To refine the image dataset derived from web scraping, we applied filters for aspect ratio, size, and specific keywords, and used CLIP similarity (Hessel et al., 2021) to rank images for cultural relevance. Images with low CLIP scores were discarded, and we sampled the remaining images based on their scores, with higher scores having a higher probability of selection. Details of the image filtering process can be found in App. B.

Question Collection Following the conceptual culture framework by Hofstede et al. (2010), we directed annotators to create questions that are easily answerable by someone from their own culture but challenging for outsiders. To elicit such questions, annotators were guided by the instructions shown in App. C and were provided with images and additional context to cultural concepts presented in the image (retrieved from CANDLE). We encouraged them to create questions based on their cultural knowledge, using the additional context (accessible behind a click-to-expand box) only when absolutely necessary. Annotators were also advised to skip images if they found them culturally irrelevant or were unfamiliar with the depicted content.

Initially, for this task, we attempted to engage professional annotators from the Amazon Mechanical Turk (MTurk) platform. However, we encountered challenges in finding sufficient presence of annotators from some of the targeted countries. Therefore, we expanded our search to other communities with a broad cultural representation, an African NLP organization, and an international academic AI research institute.[4] Employing annotators from these sources, we conducted pilot studies to iterate over the task instructions and to pre-select high-quality participants.

[4]We are not disclosing the names of these organizations to maintain anonymity in the reviewing process.

Answer Collection Next, we asked the annotators to write answers to the questions created in the previous step, ensuring that the answers reflected common agreement within their culture (see instructions in App. D). We prompt them to use English for universal concepts like *cats* or *apples* and use widely recognized and agreed upon local terms for concepts like beliefs, festivals, or local cuisine, rather than translating these terms into English. For example, the annotators should use *naan* instead of *Indian bread*. This approach preserves the cultural specificity of the collected answers. Further, we instructed annotators to be as precise as possible in their answers (e.g., *sushi* instead of *food* and *Oolong tea* instead of *tea*) and to keep their responses concise, ideally between one to three words.

4 Dataset Analysis

This section provides a detailed analysis of our dataset's composition and characteristics. In particular, we offer an analysis of images, questions, answers, and cultural concepts included in the CULTURALVQA dataset.

Images Our dataset comprises of 2,328 unique images. In Fig. 2, we show representative samples showcasing the images and cultural concepts within our dataset. The concepts depicted in the images are sourced from 11 countries, selected through a strategic process to ensure extensive cultural representation. The distribution of unique image count per country is detailed in Fig. 3.

Questions We collected 2,378 questions in total. In Fig. 3, we present the number of unique questions per country. The questions have an average length of 10.98 words (see Fig. 3 for country-wise breakdown). Most frequent question types include 'What'(51.3%), 'Which'(11.2%), 'In' (5.6%), 'Why' (3.4%),'Where'(3.1%) 'Identify'(3.0%), and 'How' (2.7%) questions. For example, 'What' questions often relate to identifying cultural entities like *saree* or *Dirndl* (traditional Indian and German dresses, respectively) in the clothing category, or festivals like *Ramadan* (observed e.g., in Nigeria) and *Spring Festival* (celebrated in China) among rituals. 'Where' questions inquire about locations significant to specific foods, such as the origins of *Quebec chicken*. Finally, we analyzed whether the collected questions contain stereotypes and found that they are largely absent (see App. E).

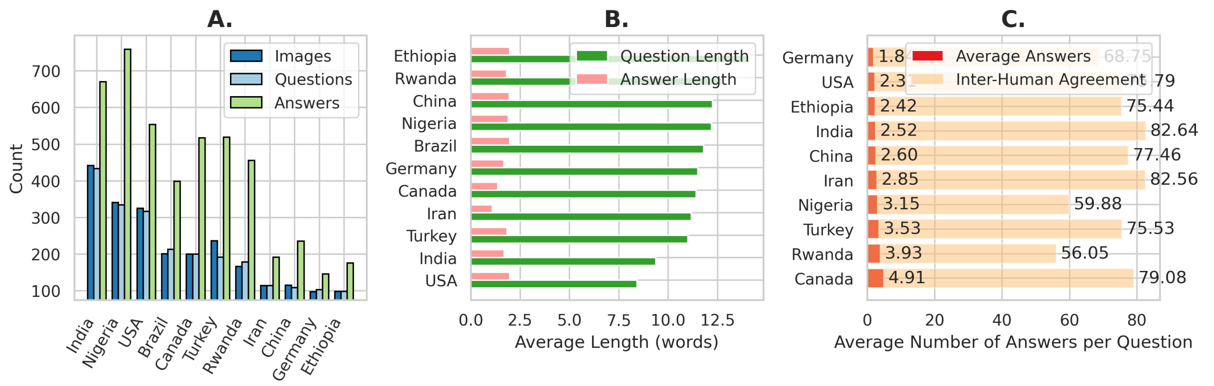


Figure 3: Comparative analysis of data by country. The figure presents three aspects: (A) unique counts of images, questions, and answers, (B) average lengths of questions and answers, and (C) average confidence scores across countries, showcasing variations and trends in CULTURALVQA.

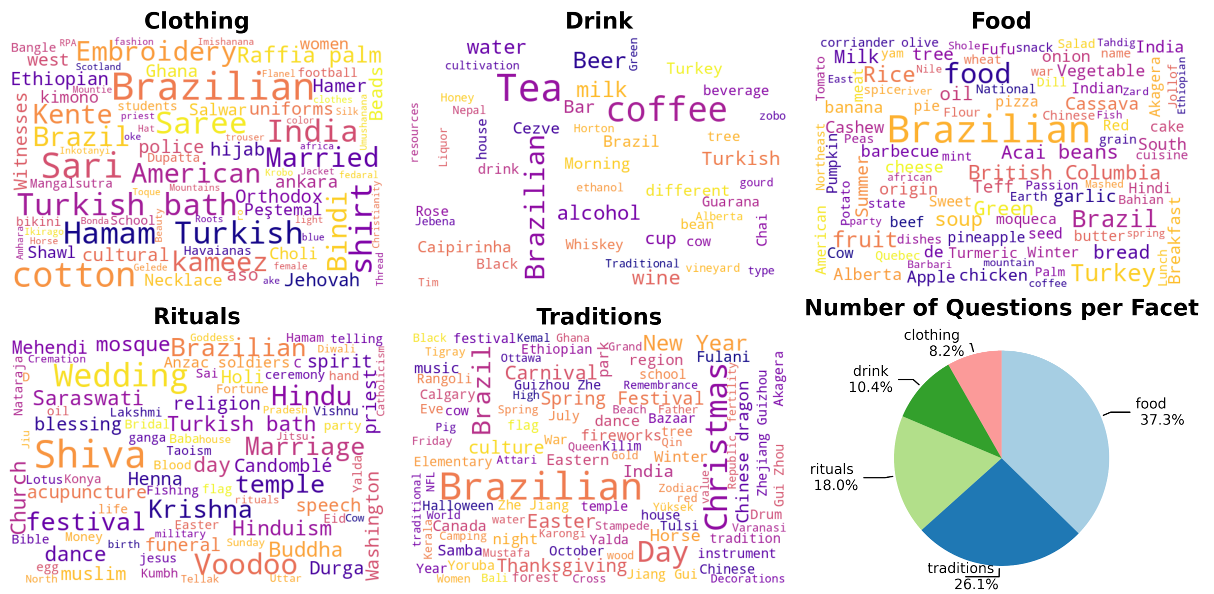


Figure 4: Word clouds representing the answers in CULTURALVQA across five facets of culture: clothing, drink, food, rituals, and traditions. In the bottom right, a breakdown of cultural facets in data is depicted.

Answers CULTURALVQA consists of 7,095 manually curated answers in total.[5] The average answer length is 1.75 words (see Fig. 3 for country wise breakdown). We assess whether answers predominantly feature terms from local languages. To this end, we verified how many answers have corresponding English Wikipedia titles; for 80% of the answers at least one of the answer words is contained in at least one Wikipedia title. Thus our benchmark is still suitable for English VLMs.

Cultural Concepts According to the pie chart in Fig. 4, food-related questions are most prevalent, accounting for 31.6% of the dataset, followed closely by traditions and rituals, which represent 28.6% and 22.6% respectively. Thus, roughly 50% of the questions in our dataset probe for cultural understanding of the intangible aspects of culture (rituals and traditions)!

The word clouds generated from the collected answers in Fig. 4 reveal diverse expressions of rituals and traditions represented by terms like *hamam* (Turkey) and *meskel* (Ethiopia). Further, the food category includes diverse items such as *feijoada* (Brazil), *fufu* (Nigeria), and *vada* (India) indicating a geo-diverse culinary scope. While the clothing category is the least prevalent in the dataset, it shows the highest variety in terms of collected answers. The drink category is notably one of the smallest, both in terms of the size and number of

[5]We collected 1-5 answers per question, depending on the availability of annotators.

unique answers.

5 Benchmarking VLMs on CULTURALVQA

Evaluation Metric Evaluating open-ended VQA is challenging. Traditionally, string matching has been used but it is known to underestimate model performance. Based on findings from Mañas et al. (2024), which demonstrate the effectiveness of reference-based LLM evaluation for open-ended VQA tasks, we adopt LAVE, their proposed metric, as our evaluation metric with GPT-4 as the LLM (see App. F for the LLM prompt used). We validated the effectiveness of LAVE for our use case by computing correlation with human judgements.

VLMs used for benchmarking We benchmark several state-of-the-art VLMs on the proposed CULTURALVQA dataset, ranging from closed-source models like GPT-4 (GPT-4O) and GEMINI PRO (GEMINI-PRO-VISION 1.0) to a wide variety of open-source models, ranging from 7 to 25 billion parameter count: BLIP2 (Li et al., 2023), INSTRUCTBLIP (Dai et al., 2024), LLAVA1.5 (Liu et al., 2023), LLAVA_NEXT (Liu et al., 2024), IDEFICS2 (Laurençon et al., 2024), and INTERN-VL 1.5 (Chen et al., 2024). See App. G for detailed discussions on these models.

What degree of visual understanding is required to answer the questions in CULTURALVQA? To investigate this, we employ the following baselines. **LLM-only:** This baseline uses an LLM to answer questions based on solely the question input. It helps gauge the extent to which the questions in our dataset can be addressed without any visual context, solely relying on the language-only cultural information encoded in the parameters of the LLM. **LLM + Country:** It introduces country-specific context into the LLM prompts to determine if knowing the country along with the question can already elicit the correct answer! **LLM + Lens:** Unlike the other two baselines, which do not rely on visual context, this baseline takes as input the image entity names extracted by Google Lens, along with the question. Thus it helps gauge whether the questions in our dataset can be answered with only coarse-level knowledge of the visual context.

We evaluate the baselines using GPT-4 as the underlying LLM. The LAVE accuracies of these baselines, along with that of the GPT-4 VLM (that

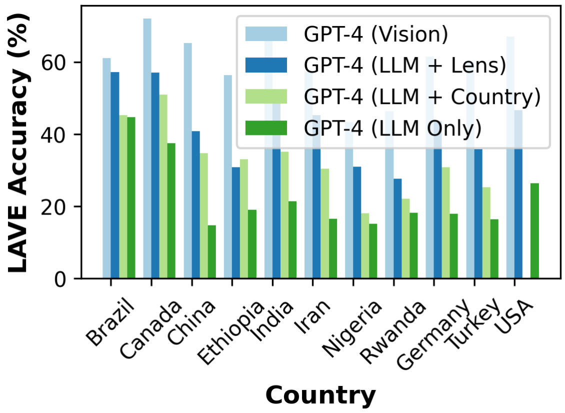


Figure 5: Baseline evaluation of the degree of visual understanding required in CULTURALVQA: LLM-only, LLM with a country-specific context, LLM with Google Lens entities, and GPT-4V.

takes an image also as the input along with the question) are presented in Fig. 5. We see that although the country information and the coarse visual entities help improve the performance on top of the LLM-only, the performance of the strongest baseline (LLM + Lens) is still far from that of the VLM. This verifies that the questions in our dataset require sufficient visual understanding to answer them accurately.

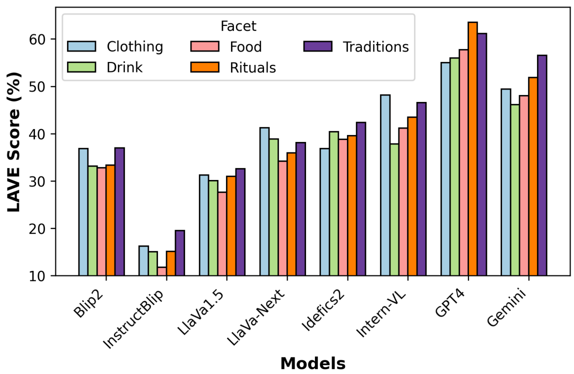


Figure 6: VLM performance across facets as measured using LAVE scores.

To what extent are VLMs culturally aware? We report the LAVE scores of open-source and closed-source vision-language models on the proposed CULTURALVQA benchmark in Tab. 1, which range across countries from 43% to 72% for GPT-4, the best-performing model. The results indicate a significant performance gap between closed-source models and the best-performing open-source models (INTERN-VL for most cases), with an average difference of 11.71% points. This

Country	Open-Source						Closed-Source	
	INSTRUCTBLIP	LLAVA1.5	BLIP2	LLAVA-NEXT	IDEFICS2	INTERN-VL	GEMINI	GPT4
Brazil	11.06	37.50	30.29	43.75	38.46	36.06	51.92	61.06
Canada	17.00	50.50	58.50	62.50	69.00	67.50	65.50	72.00
China	16.52	26.09	34.78	33.04	38.26	53.04	65.22	65.22
Ethiopia	3.19	24.47	17.02	18.09	25.53	26.60	42.55	56.38
Germany	30.77	41.03	51.28	48.72	38.46	48.72	48.72	61.54
India	19.91	34.84	46.61	42.53	49.32	53.85	58.37	69.68
Iran	11.30	18.26	19.13	17.39	23.48	30.43	46.09	57.39
Nigeria	13.74	22.81	21.35	28.95	31.87	33.92	36.26	43.27
Rwanda	4.97	19.34	22.65	25.41	23.20	28.73	35.36	46.41
Turkey	21.52	24.47	33.76	33.33	37.97	41.35	56.12	59.92
USA	58.82	60.0	47.06	64.70	58.82	68.24	61.18	67.06

Table 1: LAVE scores of open- and closed-source models on CULTURALVQA. Best-performing results per country are highlighted in green, and best-performing results among open-source models are highlighted in blue.

Country	GPT-4	Human	Δ (%)
Iran	57.39	82.56	43.86%
Nigeria	43.27	59.88	38.39%
Ethiopia	56.38	75.44	33.81%
Turkey	59.92	75.53	26.07%
Rwanda	46.41	56.05	20.77%
India	69.68	82.64	18.58%
China	65.22	77.46	18.77%
Germany	61.54	68.75	11.73%
Canada	72.00	79.08	9.83%

Table 2: Comparison of GPT-4 performance against human performance across countries, ordered by decreasing percentage difference (**Δ (%)**) between them.

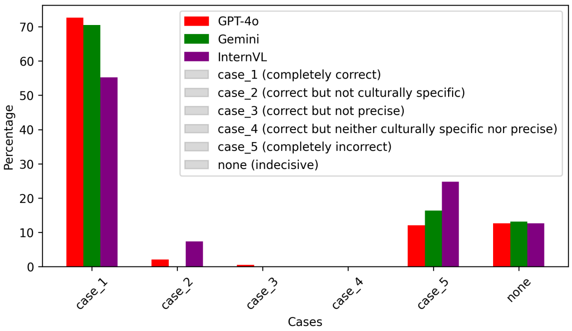


Figure 7: Distribution of human judgments for model answers in India across different models (GPT-4O, GEMINI, INTERNVL). GPT-4O and Gemini show the highest percentage of completely correct answers (case_1), while INTERN-VL has a significant percentage of completely incorrect answers (case_5).

gap is particularly pronounced in countries from Africa (Ethiopia, Nigeria) and the Middle East (Iran, Turkey).

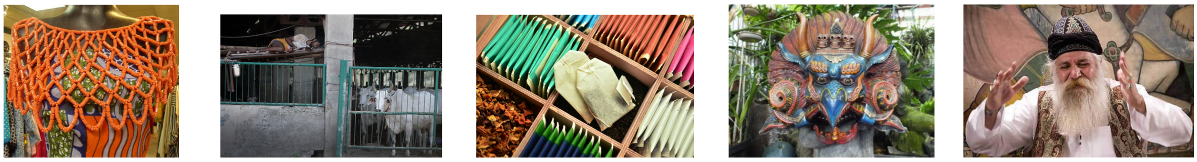
Are VLMs better at understanding cultures from some countries than others? A country-level (see Tab. 1) analysis of the models reveals stark variance in performance across different regions. Generally, open-source models perform well for high-resource countries such as the USA, Canada, Brazil, and India while achieving inferior performance in underrepresented countries. This trend holds true even for open-source models with large parameter sizes, such as INTERN-VL, indicating that data diversity is more crucial for cultural understanding than model size. Although closed-source models showcase less drastic performance discrepancies across countries, their performance also degrades significantly for African countries.

Are VLMs better at understanding some cultural concepts than others? In Fig. 6, we report the model performance across five cultural facets. Generally, we find that proprietary models tend to perform better on intangible concepts – rituals, and traditions, compared to drink and food. Indeed, the highest performance of GPT-4 is achieved in the rituals facet ($> 60\%$), whereas in the clothing facet, it achieves a lower performance of $\approx 53\%$.

How do culturally knowledgeable people perform on CULTURALVQA? We calculate human performance for 1,325 questions for which we have three or more answers using the LAVE metric.[6] For each question, we compute the accuracy of one of the human answers against the remaining human answers using LAVE. We do this for each human answer and average the scores across all answers. Since all these answers are written by annotators who are familiar with the culture being probed in the question, this human performance tells us how well culturally knowledgeable people can perform on CULTURALVQA.

Based on the results in Tab. 2 (also reported in Fig. 3), human performance is notable and ranges from 55%-85%, with certain countries, such as Iran,

[6]Brazil is currently not included in this study as the collection of multiple answers is still in progress.



Q: This image depict or give the sign of what in Nigerian culture?
Human: sign or symbol of royalty
GPT-4: Coral Beads

Q: Which planet is the above animal compared with?
Human: Earth
GPT-4: Jupiter

Q: In which type of glass do Turkish people consume the thing depicted in the image?
Human: Slim-waisted glass
GPT-4: Tulip glass

Q: What does the animal in the image depict?
Human: Garuda
GPT-4: Tibetan Snow Lion

Q: What is the traditional occupation name of this person?
Human: Naghali
GPT-4: Dervish

Figure 8: Qualitative failure examples of GPT-4 predictions.

showing particularly high scores ($> 80\%$). Further, we find a major gap between human performance and the best-performing model, GPT-4, with larger differences observed for non-Western countries such as Iran, Nigeria, and Ethiopia ($> 33\%$). Conversely, the smaller gap for Canada (9.83%) indicates a closer alignment between GPT-4 and human performance, likely due to a better representation of Western cultural concepts in the training data.

Human judgment of model performance We evaluate responses from the GPT-4, GEMINI, and INTERNVL models for questions from India, with each answer rated by 5 humans on a scale of 1 to 5, from completely correct to completely incorrect. See App. J for details on the human evaluation study. Fig. 7 shows the percentage of questions that fall into each of the five scales.

The results indicate that the GPT-4's and INTERNVL's scores closely align with human judgments for case 1 scores, suggesting that our metric predicts answers to be correct only if they are both precise and culturally specific. We note that humans tend to rate model predictions higher than the LAVE metric. Finally, the evaluation shows that humans very often choose the extreme ratings, considering most model responses as either fully accurate or entirely wrong.

Qualitative examples of model failures Our qualitative evaluation of the best-performing model, GPT-4, highlights its limitations in recognizing and interpreting cultural nuances. For instance, GPT-4 overlooks the cultural significance of intangible cultural concepts like coral beads in Nigeria, which symbolize wealth and heritage but are treated merely as decorative objects, as well as it fails to recognize the symbolic connection between cows and planet Earth in Indian culture (see Fig. 8). Focusing on tangible cultural concepts in Fig. 8, the model's shortcomings are evident as it inaccurately recognizes cultural entities and objects. For instance, it mislabels Naghali, a traditional Iranian storyteller as a Dervish and mistakes a traditional Turkish tea glass for a tulip glass, commonly used for serving beer. These examples reveal how GPT-4's struggles with both tangible and intangible cultural concepts: it has difficulties distinguishing between visually similar but culturally distinct entities and objects, and it lacks a deep understanding of cultural beliefs and symbolic meanings.

6 Conclusions

In this paper, we highlight the significance of evaluating multimodal vision-language models not just on general scene understanding but also on their ability to comprehend diverse cultural contexts. We introduce CULTURALVQA, a novel cultural VQA benchmark for assessing VLMs on their cultural understanding. By curating a diverse collection of images from 11 countries across five continents and collecting 2,378 hand-crafted questions and 7,095 answers about cultural concepts presented in these images, written by professional annotators, we ensured a broad representation of cultural concepts pertinent to diverse cultural groups.

Benchmarking state-of-the-art models on CULTURALVQA reveals notable disparities in the performance of VLMs across regions. Specifically, models demonstrate substantially higher accuracy in answering questions related to North American cultures compared to African and Middle Eastern ones. Further, we find a stark performance disparity between proprietary and open-source models, with an 11.71% difference between the best-performing open-source model and the worst-performing proprietary model. The benchmarked VLMs also showed varying levels of proficiency across cultural facets, performing well on questions about clothing, rituals, and traditions, but less effectively on those concerning food and drink. Our results underscore the current limitations of VLMs in achieving uniform cultural comprehension and pinpoint specific areas that require improvement.

7 Limitations

Our study faces limitations due to our data collection methods, the scope of the CULTURALVQA dataset, and our focus on the English language. We approximated cultural groups using geographical regions for annotator recruitment, potentially oversimplifying cultural identities and conflating culture with nationality due to practical constraints like annotator availability. Our use of English-only data may also miss key cultural nuances available only in native languages. Although our dataset aims for cultural diversity, it does not capture the full spectrum of global cultural diversity. Future work will expand the dataset to represent diverse cultures and regions more broadly and develop multilingual datasets for greater inclusivity.

Challenges in collecting culturally informative data Collecting culturally rich content from diverse annotators proved challenging, particularly because the images and concepts were limited to those available on English-language websites. This restriction likely omits important cultural details. Allowing annotators to skip inadequate images did not fully overcome the drawbacks of limited image quality, impacting the depth of the questions created.

8 Ethical Considerations

Our CULTURALVQA benchmark involves culturally specific questions and answers, developed by professional annotators from the relevant countries. We sought wide cultural representation by engaging with three different communities, compensating annotators at $10-15 per hour for both included and excluded contributions after pilot testing. This reflects our best effort to maintain fairness and inclusivity in our data collection process.

Despite these efforts, we recognize our approach's limitation in equating cultural groups with national borders, potentially overlooking the complex realities of minority and diaspora communities. We urge future research to explore finer distinctions within cultural groups to enhance representation. Although we have rigorously tried to remove biases, some subjective content may persist; however, a substantial portion of the dataset has been verified as unbiased (see App. E). We acknowledge these constraints but are hopeful that our work will advance the understanding of cultural nuances in VLMs.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. *Preprint*, arXiv:2403.15412.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR.

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, Singapore. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems*, volume 35, pages 12979–12990. Curran Associates, Inc.

Gemini Team et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv e-prints*, arXiv:2312.11805.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Marta Lagos, Juan Diez-Medrano, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World Values Survey: Round seven - country-pooled datafile version 3.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

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.

Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival*, 3rd edition. McGraw-Hill, New York; London.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742.

Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. 2019. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Wang, Yejin Choi, and Bill Yuchen Lin. 2024. WildVision Arena: Benchmarking multimodal LLMs in the wild.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic VQA evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179. AAAI.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference*, page 1907–1917.

OpenAI. 2023. Gpt-4v. Retrieved from `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

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. *arXiv preprint arXiv:22405.13777*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:22406.05967*, arXiv:2406.05967.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164. IEEE Computer Society.

T. Weyand, A. Araujo, B. Cao, and J. Sim. 2020. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2572–2581, Los Alamitos, CA, USA. IEEE Computer Society.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

Appendix

A Data Statement

We provide a data statement (Bender and Friedman, 2018) to document the generation and provenance of CULTURALVQA.

Curation Rationale CULTURALVQA benchmark is designed to evaluate VLMs' cultural understanding capacities across various cultures. The images are sourced from the CANDLE dataset (Nguyen et al., 2023), which offers a comprehensive collection of Cultural Commonsense Knowledge (CCSK) from the C4 corpus (Raffel et al., 2020), consisting of 1.1 million entries each linked to relevant CCSK data via URLs to webpages. Annotators writing questions and answers for this project are recruited through the MTurk platform, an African NLP organization, and an international academic AI research institute.

Language Variety All texts included in the dataset are in English, primarily authored by non-native speakers, and may thus contain ungrammatical structures both in questions and answers.

Annotator Demographics All annotators come from the following 11 countries: China, Turkey, Iran, Ethiopia, Nigeria, Rwanda, Germany, USA, Canada, Brazil, and India. Other demographics such as age and gender are unknown. All annotators were compensated at an hourly rate of 10-15$ per hour depending on a task and the number of completed HITs.

B Image Filtering

Given the potential noise inherent in an image dataset derived from web scraping, we implement a series of heuristic filters to refine our selection. First, we apply aspect ratio filtering, retaining only images with an aspect ratio between 0.5 and 2, effectively removing many banner-like advertisements. Next, we discard any images smaller than 100 pixels due to their inadequate detail for analysis. We also exclude images containing specific keywords such as "logo" and "social," which typically denote non-relevant graphics or branding content.

To guarantee the high quality of images included in our benchmark, we first employed CLIP similarity (Hessel et al., 2021) to rank the remaining images for cultural relevance. Based on a manual annotation of images for 200 CCSK assertions, to assess their relevance to the CCSK, we set a threshold of 23 to ensure culturally relevant images (precision = 0.92, recall = 0.96). Images below this score were discarded. Higher-scoring images were more likely to be selected for question creation.

C Instructions for Human Question Generation

The detailed instructions given to the annotators for writing questions can be found in Fig. 9.

D Instructions for Human Answer Generation

The detailed instructions given to the annotators for collecting answers can be found in Fig. 10.

E Stereotypes and Biases

To ascertain the representational fairness of our dataset, we implemented a Sentence-Level Stereotype Classifier, a transformer-based model, for detecting stereotypical content within the dataset's questions. This model's efficacy in classifying sentences based on the presence of stereotypes or anti-stereotypes was evaluated across various dimensions including race, gender, religion, and profession. The classifier identified relatively few stereotypical instances: 69 cases pertained to race, 44 to gender, 22 to religion, and 8 to profession. In contrast, anti-stereotypical content was more prevalent, with 169 cases for race, 25 for religion, 23 for gender, and 7 for profession. A significant portion of the data, 923 instances, did not correlate with any stereotypical or anti-stereotypical categories, underscoring the minimal presence of biased content

Instructions for Writing Cultural Visual Question and Answer

Thank you for participating in our study. Please start by watching the following video, which contains important information about how to complete the task. Watching the video will help you understand the task and instructions much better. After watching the video, make sure to carefully read the written instructions below, as there are a few more details you need to know.

Click here to watch the instructional video

Instructions:

In this task, you will be shown an image that depicts a cultural concept from your culture such as a practice, tradition, food, or clothing. Your task is to ask a question **about the cultural concept depicted in the image** that someone from your culture will be able to answer easily, **but someone who is not familiar with your culture will not be able to answer.**

IMPORTANT 1: The question must require looking at the image to be able to answer it correctly. The question must not be answerable without looking at the image.

IMPORTANT 2: The question must require an understanding of your culture to be able to answer it correctly.

IMPORTANT 3: The question must elicit a single correct answer. Do not ask questions that are vague or under-specified and may have multiple correct answers.

Please see the examples below to understand the above requirements better. **Your work will be rejected if your questions do not satisfy either of the above requirements.**

Before writing the question for each image, you need to answer the following question:

Are you familiar with the cultural concept depicted in the image?

1. Yes, I am familiar.
2. Yes, I am somewhat familiar.
3. No, I am not familiar.

If you are not familiar with the cultural concept depicted in the image, we provide you with some supporting information to help you understand the cultural concept. You can view this information by clicking on the "Supporting Information (click to expand)" which will expand the dialog box. The supporting information includes the name of the cultural concept and some additional context. **But please use this information only if you are not already familiar with the cultural concept depicted in the image.**

Finally, we also need you to write the answer to the question.

IMPORTANT 1: Your answer must be such that most people from your cultural group would agree on it.

IMPORTANT 2: Your answer must be a brief phrase. It must not be a full sentence. For example,

- "It is a potato." -> "Potato"
- "Yes, it is." -> "Yes"

In addition, the question–answer pair must follow each of the below criteria:

1. **No Stereotypes:** Please frame your question around a fact that is true about your culture. Do not ask a question based on stereotypes i.e., over-simplified beliefs about your cultural group.
2. **Culturally Precise Answer:** Write answers that most people from your culture would agree with. For universal concepts like "cats," "apples," etc., please use English terms. However, for culturally specific concepts like beliefs, festivals, local cuisine, or drinks, use the local name that is widely recognized and agreed upon in your culture.
   - "Kutta" -> "Dog"
   - "Naan" -> "Naan" (instead of "Bread" or "Indian bread").
3. **Answer Specificity:** Please provide precise answers and avoid generic ones. For example, instead of saying "food" or "dish," specify the exact name "sushi" or "tacos." Instead of saying "festival," specify "Diwali" or "Carnival." Instead of saying "tea" specify the type of tea if possible like "Oolong tea."
4. **Use digits for numerical answers:** For numerical answers, please use digits (eg: Write 10 instead of ten)

For a detailed look at the image, please hover over it.

Please write the questions following the instructions the best you can. Careless work will be rejected. Thank you for your careful attention to detail and your valuable contribution!

Figure 9: The instructions given to annotators to write questions and answers for images. To assist with writing, we provide a brief video detailing our task and guidelines. Additionally, we offer multiple examples showcasing both good and poor practices (examples not included here)

Instructions for Writing Culturally aware answers

In this task, you will be provided with an image and a question about the image. Your task is to provide an appropriate answer to the given question.

Nature of the image and the associated question: The provided image depicts a cultural concept from your culture such as a practice, tradition, food, or clothing. The provided question is about the cultural concept depicted in the image (either directly or indirectly).

Your **task** is to provide an appropriate answer to the question. Your answer should satisfy **each** of the following criteria.

1. **The answer should be culturally specific:** Write answers that most people from your culture would agree with. For universal concepts like "cats," "apples," etc., please use English terms. However, for culturally specific concepts like beliefs, festivals, local cuisine, or drinks, use the local name that is widely recognized and agreed upon in your culture.

   Below are examples of universal concepts, so please use English terms for such concepts. The word before "->" denotes the incorrect way of answering whereas the word after "->" denotes the correct way of answering.
   - "Dhaniya patta" -> "Coriander leaves"
   - "Anar daana" -> "Pomegranate seeds"

   Below are some examples of culturally specific concepts, so please use the widely accepted local terms for these concepts. The word before "->" denotes the incorrect way of answering whereas the word after "->" denotes the correct way of answering.
   - "bread" -> "Naan"
   - "dress" -> "Saree"

2. **The answer should be precise:** Please provide precise answers and avoid generic ones. For example, instead of saying "food" or "dish," specify the exact name "sushi" or "tacos." Instead of saying "festival," specify "Diwali" or "Carnival." Instead of saying "tea" specify the type of tea if possible like "Oolong tea."

3. **The answer should be short:** Your answer should be a brief phrase. It should not be a full sentence.
   - "It is a potato" -> "potato"
   - "People are celebrating Holi" -> "Holi"

4. **The answer should use digits for numerical answers:** For numerical answers, please use digits (eg: Write 10 instead of ten)

If you don't know the answer, provide your best guess. Your answer should be such that most people from your cultural group would agree on it.

In addition to answering the question, please also indicate whether you think you were able to answer the question correctly by answering the following question:

"Do you think you were able to answer the question correctly?"

1. Yes
2. Maybe
3. No

Figure 10: The instructions given to annotators to write answers for questions collected for images. To assist with writing, we provide clear guidelines and offer multiple examples showcasing both good and poor practices.

in the dataset. These findings support the dataset's utility in facilitating unbiased and culturally comprehensive studies.

F System Prompt for the Evaluation Metric

System prompt used for the LAVE evaluation metric

You are an expert cultural anthropologist tasked with evaluating the correctness of candidate answers for cultural visual question answering. Given a question, a set of reference answers by an expert, and a candidate answer by a model, please rate the candidate answer's correctness. Use a scale of 1-2, where 1 indicates an incorrect, irrelevant, or imprecise answer, and 2 indicates a correct and precise answer. Specify the rating in the format 'rating=X', where X is either 1 or 2. Also, provide the rationale for your rating.

G VLMs Used for Benchmarking

We benchmark the following state-of-the-art open-source VLMs on our proposed CULTURALVQA dataset: BLIP2 (Li et al., 2023), INSTRUCTBLIP (Dai et al., 2024), LLAVA1.5 (Liu et al., 2023), LLAVA_NEXT (Liu et al., 2024), IDEFICS2 (Laurençon et al., 2024), and INTERN-VL 1.5 (Chen et al., 2024). These models were selected based on their release year and parameter size (7 to 25 billion) to test how these aspects affect cultural understanding. INSTRUCTBLIP, fine-tuned with instruction tuning, is compared to BLIP2 to see if instruction tuning enhances cultural understanding. IDEFICS2, with 8 billion parameters, is evaluated for its performance on open datasets, surpassing larger models. INTERN-VL 1.5, with 25 billion parameters, bridges the gap between open-source and proprietary models, showing strong multimodal benchmark performance, even outperforming proprietary models on some benchmarks. Finally, we also evaluate closed-source models – GPT-4 (GPT-4o) and GEMINI PRO (Gemini-Pro-Vision 1.0) – using their API endpoints.

H Prompt for VLM Inference

Prompt used to test VLM inference

You will be given an image depicting a cultural concept and a question about the image. Answer the question with a precise, culturally specific response (e.g., 'sushi' instead of 'food', 'Diwali' instead of 'festival') of 1-3 words.

I Inference Using Closed-Source Models

In this section, we provide the sample code used for accessing Gemini-Pro and GPT-4.

For performing inference using Gemini, we leverage the Vertex AI API for Gemini with multimodal prompts. The code snippet for inference is provided below.

```
import google.generativeai as genai

genai.configure(api_key=<api_key>)
model = genai.GenerativeModel('gemini-
    pro-vision')

response = model.generate_content([
    question, image],
stream=False,
request_options={"timeout": 600})
response.resolve()
predicted_answer = [response.text]
```

Listing 1: Code snippet for accessing Gemini using API

J Human Judgment of Model Predictions

We evaluate model responses for questions from India, with each answer rated by 5 humans on a scale of 1 to 5: 1 (completely correct), 2 (correct but not culturally specific), 3 (correct but not precise), 4 (correct but neither culturally specific nor precise), and 5 (completely incorrect). The detailed instructions given to the annotators can be found in Fig. 11.

Instructions

In this task, you will be provided with an image, a question about the image and a response to the question. **Your task is to rate the correctness of the response.**

Nature of the image and the associated question: The provided image depicts a cultural concept from your culture such as a practice, tradition, food, or clothing. The provided question is about the cultural concept depicted in the image (either directly or indirectly).

Your task is to rate the correctness of the response by choosing one of the 5 options:

1. The response is **completely correct.**
2. The response is **correct but not culturally specific.**
3. The response is **correct but not precise.**
4. The response is **correct but neither culturally specific nor precise.**
5. The response is **completely incorrect.**

Please see below to understand what we mean by **culturally specific** and **precise response**.

Culturally specific response: A response is considered to be culturally specific if it uses a term that most people from your culture would agree on. For universal concepts like "cats," "apples," etc. the response should use English terms. However, for culturally specific concepts like beliefs, festivals, local cuisine, or drinks, the response should use the local name that is widely recognized and agreed upon in your culture.

Below are examples of universal concepts, so the response should use English terms for such concepts. The word before "->" denotes an incorrect response whereas the word after "->" denotes a correct response.

1. "Dhaniya patta" -> "Coriander leaves"
2. "Anar daana" -> "Pomegranate seeds"

Below are some examples of culturally specific concepts, so the response should use widely accepted local terms for these concepts. The word before "->" denotes an incorrect response whereas the word after "->" denotes a correct response.

1. "Bread" -> "Naan"
2. "Festival of colors" -> "Holi"

Precise response: The response should be a precise answer to the question, it should not be a generic answer. For example, a response that just says "food" or "dish" is a generic response. A precise response would specify the exact name of the dish such as "sushi" or "tacos". Similarly, a generic response would just say "festival" whereas a precise response would specify the exact name of the festival such as "Diwali" or "Carnival". Just saying "tea" would be a generic response, specifying the type of tea such as "Oolong tea" would be a precise response (if indeed the type of the tea can be identified from the shown image).

Please see the examples to understand this better.

Figure 11: The instructions given to annotators to evaluate answers generated by various models. To assist with writing, we provide clear guidelines and offer multiple examples showcasing both good and poor practices.