Benchmarking Vision Language Models for Cultural Understanding

Shravan Nayak[1,2] **Kanishk Jain**[1,2] **Rabiul Awal**[1,2]
Siva Reddy[1,3] **Sjoerd van Steenkiste**[4] **Lisa Anne Hendricks**[5]
Karolina Stańczak[1,3] **Aishwarya Agrawal**[1,2]

[1]Mila – Quebec AI Institute, [2]Université de Montréal, [3]McGill University,
[4]Google Research, [5]Google DeepMind

Correspondence: shravan.nayak@mila.quebec

Abstract

Foundation models and vision-language pre-training 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 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-4o and Gemini, reveals disparity in their level of cultural understanding across regions, with strong cultural understanding capabilities for North America while significantly lower performance 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.

https://culturalvqa.org

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, visual question answering, and image captioning, *inter alia*. These tasks predominantly focus on general scene understanding capabilities such as recognizing attributes, objects, and actions 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 VLMs to higher standards. Indeed, 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.

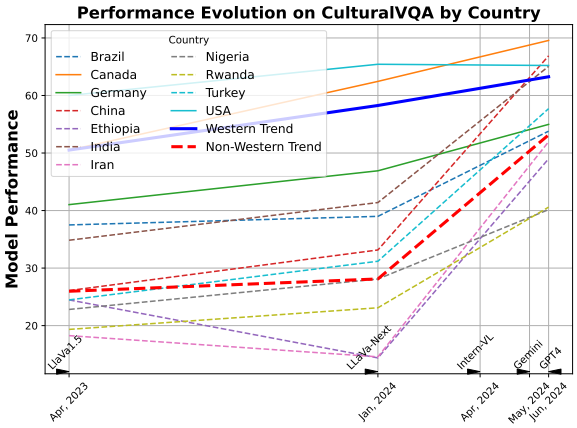

Figure 1: The performance of VLMs over time, segmented by non-Western (red) and Western (blue) countries, with model release dates annotated (bottom). Dashed and solid lines differentiate trends for non-Western and Western countries respectively. VLMs' understanding of non-Western cultures has been in a steep upward trend since Jan '24.

In order to adequately assess whether the current state-of-the-art VLMs – including proprietary models such as GPT-4O (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,

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.

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 base shared by the members of a cultural group (see § 3). While GD-VCR does consider commonsense to a degree, it primarily considers 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 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 expand 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. 2 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 7,206 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 (e.g., there is a 29.78% gap between the highest-performing closed-source model and its open-source counterpart in the country for which the models perform the worst, Ethiopia). Additionally, we observe a significant disparity in model performance across countries. For instance, the highest-performing proprietary model, GPT-4O, achieves 67% and 72% accuracy on North American cultural concepts while only between 43% and 56% 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 VLMs' progress in understanding diverse cultures and highlighting

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

Dataset	No. Regions	No. Questions	No. Images	Multilingual?	Task Format	Culturally Diverse Images?	Nature of Questions
MaXM (Changpinyo et al., 2023)	7	2142	335	Yes	Open-ended	No (Pouget et al., 2024)	General reasoning
GDVCR (Yin et al., 2021)	4	886	328	No	Multiple choice	Yes (movie scenes only)	Cultural understanding
MaRVL (Liu et al., 2021)	5	5670	4914	Yes	True/False	Yes	Cultural reasoning
CVQA (Romero et al., 2024)	28	9044	4560	Yes	Multiple choice	Yes	Cultural understanding
CULTURALVQA (Ours)	11	2378	2328	No	Open-ended	Yes	Cultural understanding

Table 1: Comparison of various datasets closely related to CULTURALVQA across different axes.

areas where VLMs lack cultural understanding. We 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. 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. Burda-Lassen et al. (2024) introduce MOSAIC-1.5k, a culture-specific captioning dataset that includes images from various regions. Bhatia et al. (2024) propose GLOBALRG, which aims to evaluate retrieval and grounding capabilities in VLMs across 15 and 50 countries respectively. Another related line of work focuses on multilingual understanding. For instance, Bugliarello et al. (2022) unify five datasets across a number of tasks in 20 languages. However, their focus lies in multilingual understanding as opposed to cultural understanding. Additionally, the XM3600 dataset (Thapliyal et al., 2022), includes image captions from 36 regions and languages, but lacks depth in cultural concepts, making it insufficient for evaluating cultural diversity in VLMs (Pouget et al., 2024).

Closest to our work are the following benchmarks: MaXM (Changpinyo et al., 2023), GD-VCR (Yin et al., 2021), MaRVL (Liu et al., 2021) and the concurrent work CVQA (Romero et al., 2024). MaXM lacks depth in cultural concepts, as it builds on XM3600 images. Also, its questions focus more on reasoning and general image understanding rather than cultural understanding [2]. The GD-VCR dataset probes cultural understanding, but its reliance on cinematic scenes limits the diversity of real-world cultural contexts it can have. Moreover, they rely on a multiple-choice evaluation format, which can be influenced by the difficulty of answer choices. We believe an open-ended evaluation provides a more faithful assessment of the models' underlying capabilities. Similarly, while MaRVL tests visually grounded reasoning across multiple languages and cultures, it does not assess cultural common sense related to rituals and traditions and also employs a True/False evaluation style. CVQA studies cultural questions in a multilingual setup. However, their focus diverges from ours as they allocate a much smaller proportion of their dataset to traditions and rituals (13% as compared to 44.1 % in CULTURALVQA) and use a multiple-choice evalaution format. A comprehensive comparison of different datasets across various dimensions is presented in Tab. 1. CULTURALVQA uniquely emphasizes open-ended evaluation, includes culturally diverse images (i.e., images from multiple cultures), and its questions probe for cultural understanding by design. The combination of these characteristics sets CULTURALVQA apart from other datasets that either lack culturally diverse images (MaXM), or use restricted evaluation formats such as multiple-choice (CVQA, GDVCR) or True/False (MaRVL).

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)[3]. 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 prac-

[2] We manually annotated 100 random questions from the English subset of the MaXM and found the following distribution: color - 3.7%, spatial understanding - 12.9%, scene understanding - 42.6%, Yes/No - 20.4%, counting - 20.4%

[3] See § 7 for a discussion of these choices.

tice 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.[4] 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.

Selection of Images We use the CANDLE dataset (Nguyen et al., 2023) for our image source which contains 1.1 million entries of Cultural Commonsense Knowledge (CCSK) along with URLs to corresponding webpages from the C4 corpus (Raffel et al., 2020). The CANDLE dataset represents cultural concepts from approximately 196 countries and 80% of web pages in this corpus contain images related to the text (Zhu et al., 2023). This allows us to begin with a culturally relevant pool of images.

We apply filters for aspect ratio, size, and specific keywords to refine the image dataset. Further, we use CLIP similarity (Hessel et al., 2021) to filter images for cultural relevance, discarding those with a CLIP score below a threshold determined through qualitative evaluation of sample images.[5] Since our initial pool already contains culturally relevant images, there is minimal risk of introducing western-centric biases through the use of CLIP, despite potential biases in its pretraining data. To further ensure quality, we apply an additional round of human filtering (detailed in the next section). Thus, our multi-stage filtering ensures that the final set of images is appropriate for cultural annotations. Further details of the image filtering process are provided in App. B.

Question Collection Following the conceptual culture framework by Hofstede et al. (2010), we direct annotators to create questions that are easily answerable by someone from their own culture but challenging for outsiders. To elicit such questions, we provide annotators the instructions shown in App. N as well as images and additional context for the cultural concepts present in the image (retrieved from CANDLE). We encourage them to create questions based on their own cultural knowledge, using the additional context (accessible behind a click-to-expand box) only when absolutely necessary. We also advise annotators to skip images if they found them culturally irrelevant or were unfamiliar with the depicted content. This adds an additional layer of filtering, resulting in annotators discarding 19.64% of the images shown to them.

Answer Collection Next, we ask the annotators to write answers to the questions created in the previous step, while ensuring that the answers reflected common agreement within their culture (see instructions in App. O). Here 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 festivals or local cuisine, rather than translating them into English. For example, the annotators should write the term *Naan* instead of *Indian bread*. This approach preserves the cultural specificity of the collected answers. Further, we instruct 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.

Further details about the rationale behind the data curation process and the challenges encountered are provided in App. I.

4 Dataset Analysis

This section provides a detailed analysis of CULTURALVQAs' composition and characteristics including analysis of images, questions, answers, and cultural concepts contained in it.

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

[5] Threshold of 23 (precision = 0.92, recall = 0.96)

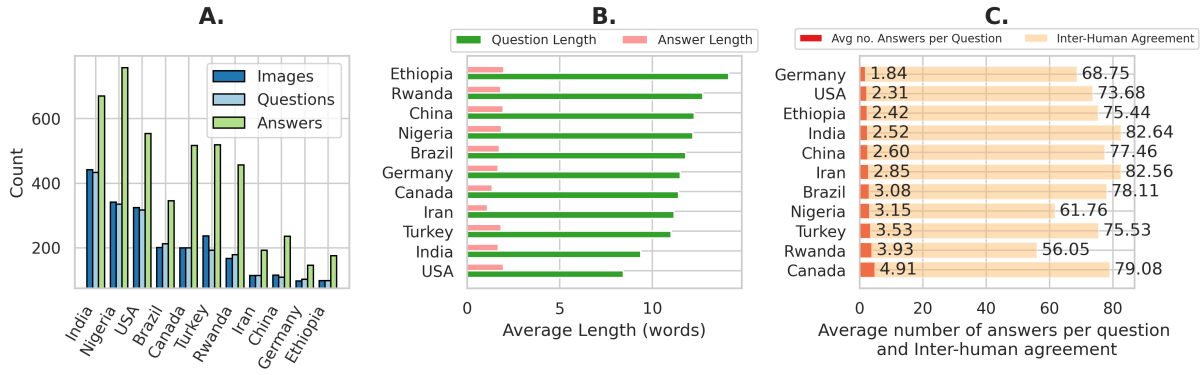

Figure 3: Comparative analysis of data by country. The figure presents three aspects: (A) unique counts of images, questions, and answers, (B) average lengths of questions and answers, and (C) average number of answers per question and inter-annotator agreement scores across countries, showcasing variations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

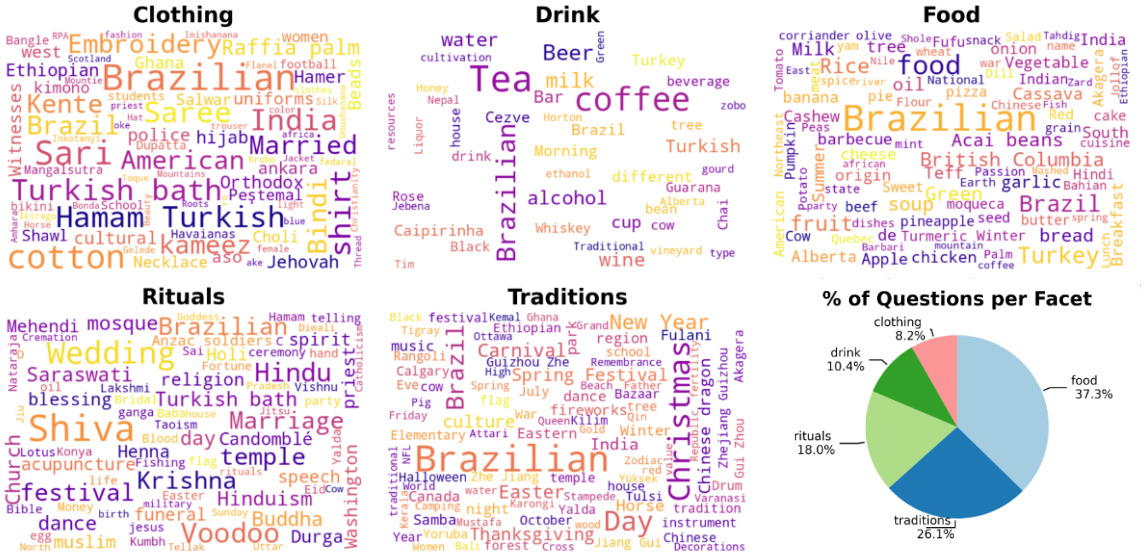

Figure 4: Word clouds representing the answers in CULTURALVQA across five facets of culture: clothing, drink, food, rituals, and traditions. In the bottom right, a breakdown of cultural facets in data is depicted.

Images CULTURALVQA comprises 2,328 unique images. In Fig. 2 we show representative samples. We choose images to ensure significant cultural representation across 11 different countries. The distribution of unique image count per country is detailed in Fig. 3 (A).

Questions We collect 2,378 questions in total. In Fig. 3 (A), we present the number of unique questions per country. The questions have an average length of 10.98 words (see Fig. 3 (B) for country-wise breakdown). Most frequent question types include 'What'(51.3%), 'Which'(11.2%), 'In' (5.6%), and 'Why' (3.4%) 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 *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 analyze whether the collected questions contain stereotypes and found that they are largely absent (see App. C).

Answers CULTURALVQA consists of 7,206 manually curated answers in total.[6] The average answer length is 1.73 words (see Fig. 3 (B) 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

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

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 37.3% of the dataset, followed closely by traditions and rituals, which represent 26.1% and 18% respectively. Thus, roughly 44% 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 illustrate the diversity of expressions, such as hamam (Turkey) and meskel (Ethiopia) for rituals and traditions, and feijoada (Brazil), fufu (Nigeria), and vada (India) for food, indicating a geographically diverse culinary and cultural 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 unique answers.

5 Evaluating 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. L for the LLM prompt used). We validated the effectiveness of LAVE for our use case by computing correlation with human judgments. LAVE judgment agrees with human judgment 79% of the times for GPT-4, 73% of the times for GEMINI, and 76% of the times for INTERN-VL.

VLMs used for evaluation We benchmark several state-of-the-art VLMs on the proposed CULTURALVQA dataset, ranging from closed-source models like GPT-4 (GPT-4O), CLAUDE (CLAUDE 3.5) 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), MBLIP (Geigle et al., 2023) PALLIGEMMA (Beyer 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. D for a detailed discussion of the selected models.

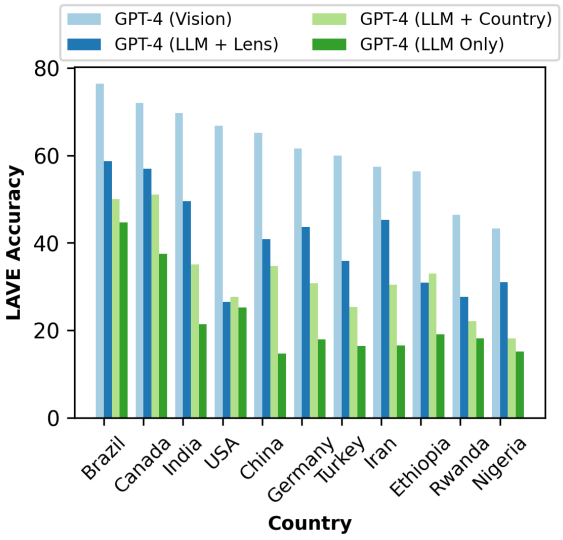

Figure 5: Baseline evaluation of the degree of visual understanding required in CULTURALVQA: LLM-only, LLM with a country-specific context, LLM with Google Lens entities, and GPT-4V.

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 solely on the question input. It helps gauge how well the questions can be addressed without visual context, relying only on the cultural knowledge encoded in 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:** This baseline uses image entity names extracted by Google Lens (Google, 2017) along with the question as input, unlike the other baselines that lack visual context. It helps assess whether coarse-level visual knowledge is sufficient to answer the questions.

We evaluate the baselines using GPT-4 as the underlying LLM. The LAVE accuracies for these baselines, as well as for the GPT-4 VLM (which also incorporates an image as input in addition to 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 baseline, 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.

To what extent are VLMs culturally aware? We report the LAVE accuracies for **zero-shot** eval-

	Open-Source						Closed-Source		
Country	mBLIP	LLaVA1.5	BLIP2	LLaVA-Next	IDEFICS2	INTERN-VL	GEMINI	CLAUDE	GPT-4
Brazil	25.34	40.38	32.21	45.62	54.37	52.53	66.34	66.36	76.44
Canada	38.50	50.50	58.50	62.50	69.00	67.50	65.50	66.00	72.00
China	22.61	26.09	34.78	33.04	38.26	53.04	65.22	49.57	65.22
Ethiopia	7.44	24.47	17.02	18.09	25.53	26.60	42.55	41.49	56.38
Germany	41.02	41.03	51.28	48.72	38.46	48.72	48.72	51.28	61.54
India	27.83	34.84	46.61	42.53	49.32	53.85	58.37	59.28	69.68
Iran	13.04	18.26	19.13	17.39	23.48	30.43	46.09	47.83	57.39
Nigeria	13.16	22.81	21.35	28.95	31.87	33.92	36.26	36.55	43.27
Rwanda	13.26	19.34	22.65	25.41	23.20	28.73	35.36	33.70	46.41
Turkey	28.57	24.47	33.76	33.33	37.97	41.35	56.12	51.26	59.92
USA	38.77	58.77	62.77	62.77	65.23	71.38	62.15	64.92	66.77
Average	*24.50*	*32.81*	*36.37*	*38.03*	*41.51*	*46.18*	*52.97*	*51.66*	*61.36*

Table 2: LAVE accuracies 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.

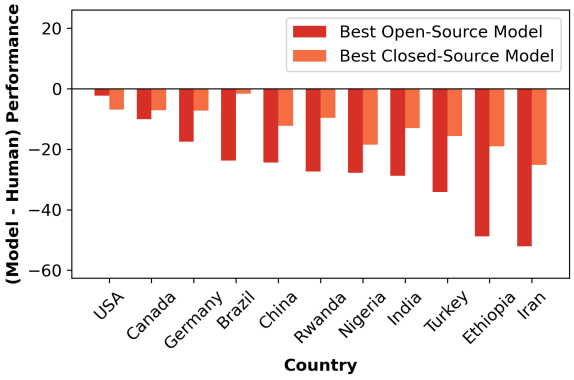

Figure 6: Performance gap between the best open-source (one of INTERNVL, IDEFICS2, BLIP2) and closed-source models (GPT-4O) compared to human performance. Negative values indicate where models underperform relative to humans.

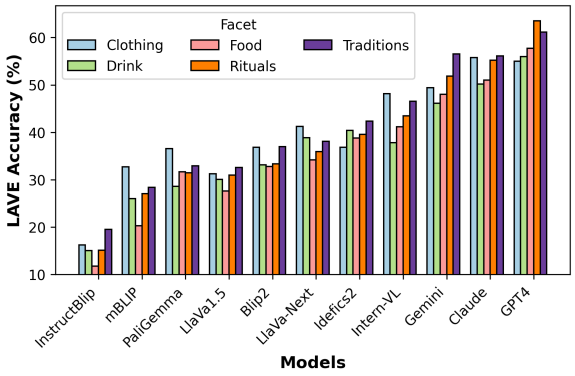

Figure 7: VLM performance across facets as measured using LAVE accuracies.

uation of VLMs on the proposed CULTURALVQA benchmark in Tab. 2 and Tab. 4. The average LAVE accuracy for the highest-performing model, GPT-4, is approximately 61%, with performance varying across countries from 43% to 72%. We see substantial disparity in cultural understanding across different VLMs, with the best-performing open-source model (INTERN-VL for most countries) achieving an average LAVE accuracy of only 46%, and performance ranging across countries from 26% to 71%. This result indicates a considerable performance gap between closed-source models and the best-performing open-source model. It is particularly pronounced in countries within the African-Islamic culture (Ethiopia, Nigeria, Iran, and Turkey), with a 29.78% gap for Ethiopia, the country for which the models perform the worst. We also conduct **few-shot** evaluation of VLMs but find that it does not significantly impact performance (see App. E for more details). Hence, the subsequent analyses in this section are conducted on top of zero-shot results.

Are VLMs better at understanding cultures from some countries than others? A country-level (see Tab. 2) 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 (Ethiopia, Iran, and Rwanda). 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-Islamic countries.

Are VLMs better at understanding some cultural concepts than others? In Fig. 7, we report the model performance across 5 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 ($\approx 63\%$), whereas in the clothing facet, it achieves a lower performance of $\approx 55\%$. Refer to App. F for a more detailed discussion.

Do multilingual VLMs perform better in culturally diverse settings? One might expect that multilingual VLMs may demonstrate superior performance due to their exposure to culturally diverse data. However, our analysis of multilingual models, mBLIP and PaliGemma, on CULTURALVQA reveals a more nuanced picture. From Tab. 2, mBLIP, built on top of monolingual BLIP2, consistently underperforms it despite multilingual training. This could be due to the quality of the machine-translated data used in mBLIP and the LLM backbone used (mT0 (Muennighoff et al., 2023) in mBLIP vs. FlanT5 (Chung et al., 2022) in BLIP2). Also, from Tab. 4 we observe that PaliGemma shows significant disparities across countries despite large-scale multilingual training. This is possibly due to its smaller size (3B) which suggests that multilingual data exposure alone is insufficient for cultural understanding.

How do culturally knowledgeable people perform on CULTURALVQA? We calculate human performance for 1,455 questions for which we have three or more answers using the LAVE metric. For each question, we compute the accuracy of one of the human answers against the remaining human answers using LAVE. We then average the scores across all answers. Since all these answers are written by annotators who are familiar with the culture probed in the question, this human performance tells us how well culturally knowledgeable people perform on CULTURALVQA.

Based on the results reported in Fig. 3 (C), human performance is notable and ranges from 55%-85%, with certain countries, such as Iran, showing particularly high scores ($> 80\%$). In contrast, Rwanda and Nigeria had the lowest scores (56.05% and 61.76%, respectively). These lower scores can be partially attributed to the cultural diversity within these countries, where using a country as a proxy for a cultural group may not accurately capture the nuances of subcultural variations. The same concept may hold different meanings across subcultures, leading to varied interpretations and inconsistencies in responses. Further qualitative insights are provided in the App. G.

We also calculate the Pearson correlation between human and model performance across countries. For open-source models, we observe a relatively low correlation, ranging from 0.1 to 0.4. Interestingly, for closed-source models like GEMINI and GPT-4, we find a stronger correlation of 0.69 and 0.75, respectively. This suggests that the factors affecting human performance similarly influence the performance of these closed-source models. However, from Fig. 6, when comparing human and model performance using the same metric, we find that closed-source models still lag behind humans for *all countries* indicating that while these models follow human performance trends, there is still a marked gap in their cultural understanding compared to humans. This gap is even more pronounced for open-source models, which show an even larger discrepancy across *all countries*.

Further, in Fig. 6, we observe a larger gap for non-Western countries such as Iran, Nigeria, India, Turkey, and Ethiopia ($> 13\%$). Conversely, the smaller gap for Canada and the USA ($< 7.0\%$) indicates a closer alignment between models and human performance, likely due to a better representation of Western cultural concepts in the training data. Interestingly, GPT-4 shows a relatively low gap for Brazil ($\approx 2\%$), possibly because the questions for Brazil often probe coarse visual understanding. This trend is further supported by LLM + Lens baseline in Figure 5 which performs exceptionally well for Brazil.

How much does varying question difficulty and varying answer counts affect model performance disparity across countries? Since we sourced questions from different annotator groups across countries, it is imperative to ask if the disparity in model performance across countries is due to differences in inherent question difficulty across countries. To investigate this, we analyze the Spearman rank correlation [7] between the model performance and the average question length (see Fig. 3 (B) for average question length across countries). We use average question length as a proxy for question difficulty - assuming shorter questions probe more direct knowledge, while longer ones require nuanced cultural understanding. We found a weak correlation between question length and model performance (-0.3 to 0.3) for most models, with the exception of GPT-4 and GEMINI, which

[7] We use this metric as the variance in lengths is small, making rank-based analysis more meaningful.

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.

showed a moderate negative correlation of -0.52 on average. As illustrated in Fig. 12, for most countries except Brazil, Canada, and the USA, the variance in question lengths is small, suggesting that question length is not a significant factor behind the disparity in model performance across countries.

Another factor potentially affecting the disparity in model performance is the variable number of human answers per question across countries (see Fig. 3 (C)). These human answers are used as the reference answers in the LAVE metric making it more rigid for countries with fewer references and vice-versa. To investigate this, we compute the Spearman correlation[7] between model performance and the average number of answers per question across countries. We find a very low correlation ranging between -0.3 and 0 across models, indicating that the disparity in the number of human answers does not meaningfully affect the disparity in model performance across countries.

Human judgment of model performance We evaluate responses from the three best-performing models, GPT-4, GEMINI, and INTERN-VL to 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. H for details on the human evaluation study. Fig. 11 shows the percentage of questions that fall into each of the five scales.The models' 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 tend to 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 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 introduce CULTURALVQA, a novel VQA benchmark for assessing VLMs on their cultural understanding. By curating a diverse collection of images from 11 countries across 5 continents and collecting 2,378 hand-crafted questions and 7,206 answers about cultural concepts presented in these images, written by 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 their performance across regions. Models perform much better on North American cultures compared to African-Islamic ones. Further, we find a stark performance disparity between closed- and open-source models, with a 29.78% gap between the highest-performing closed-source and open-source models for the lowest-performing country. VLMs also show varying proficiency across cultural facets, excelling in questions about clothing, rituals, and traditions but struggling with 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. We acknowledge that some cultural concepts may lack local terms that can be effectively represented in English letters [8]. Hence, our use of English-only data may also miss key cultural nuances available only in native languages. In such cases, collecting annotations in native languages would help mitigate this issue. However, we emphasize that our benchmark, despite being in English, is already challenging enough for the models, as evidenced by the significant disparity in model performance across cultures. In § 4, our analysis revealed that 80% of the answers contain at least one word matching an English Wikipedia page, while 20% lack such a match. This suggests that these answers may be multilingual, which presents a limitation for our English-only benchmark. 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 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. C). We acknowledge these constraints but are hopeful that our work will advance the understanding of cultural nuances in VLMs.

[8] https://tinyurl.com/3zvjsv6x

9 Acknowledgements

We would like to extend our gratitude to David Ifeoluwa Adelani for connecting us with Masakhane, Fırat Öncel for assisting with annotators in Turkey, Saba Ahamadi for securing annotators from Iran, and Qian Yang for sourcing annotators from China. We also appreciate the valuable feedback provided by Ibrahim Alabdulmohsin on the early draft. The technical support from the Mila IDT team in managing the computational infrastructure is greatly appreciated. We would also like to thank Chris Emezue for his assistance in helping with payments for the annotators. Additionally, Aishwarya Agrawal received support from the Canada CIFAR AI Chair award throughout this project. Karolina Stańczak was supported by the Mila P2v5 grant and the Mila-Samsung grant. This project was generously funded by a research grant from Google.

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.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration. *Preprint*, arXiv:2406.16469.

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.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann,

Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *Preprint*, arXiv:2407.00263.

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.

Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2024. How culturally aware are vision-language models? *Preprint*, arXiv:2405.17475.

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*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

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.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv*, abs/2307.06930.

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

Google. 2017. Google lens api. `https://serpapi.com/google-lens-api`. Accessed: 2017-10-4.

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.

Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. *Preprint*, arXiv:2406.11030.

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.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

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.

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,

Country	Brazil	Canada	China	Ethiopia	Germany	India	Iran	Nigeria	Rwanda	Turkey	USA
No. Annotators	5	6	6	4	11	6	4	8	7	4	5

Table 3: Number of Annotators by Country

Model	Brazil	Canada	China	Ethiopia	Germany	India	Iran	Nigeria	Rwanda	Turkey	USA	Avg.
PaliGemma	38.87	54.50	20.87	9.57	35.89	35.52	13.04	19.88	14.36	26.05	54.77	*28.67*
InstructBLIP	10.57	17.00	16.52	3.19	30.77	19.91	11.30	13.74	4.97	21.52	29.54	*16.27*

Table 4: Performance of InstructBLIP and PaliGemma on CulturalVQA

an African NLP organization, and an international academic AI research institute.

Language Variety All texts in the dataset are in English, primarily authored by non-native speakers, and may contain ungrammatical structures in both questions and answers. We build our dataset in English to disentangle multicultural understanding from multilingual comprehension.

Annotator Demographics All annotators come from the following 11 countries: China, Turkey, Iran, Ethiopia, Nigeria, Rwanda, Germany, USA, Canada, Brazil, and India. Initially, 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, including Masakhane, an African NLP organization, and Mila, an international academic AI research institute. All annotators are either natives of the country they annotated for or have resided there for at least 18 years, ensuring they have sufficient cultural context and lived experiences required for the task. We conducted multiple pilot rounds to ensure that annotators adhere to our guidelines and are fluent in English. 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. The number of unique annotators from each country can be found in Tab. 3.

B Image Filtering

Given the potential noise inherent in an image dataset derived from web scraping, we implement 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 image 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 Stereotypes and Biases

To ascertain the representational fairness of our dataset, we implemented a Sentence-Level Stereotype Classifier,[9] 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 in the dataset. These findings support the dataset's utility in facilitating unbiased and culturally comprehensive studies.

[9] https://huggingface.co/wu981526092/Sentence-Level-Stereotype-Detector

D 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), MBLIP (Geigle et al., 2023) PALLIGEMMA (Beyer 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 (3 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. For each model, we use the default text-generation parameters as found in their HuggingFace code repository which include a greedy decoding strategy with the temperature set to 1. Finally, we also evaluate closed-source models – GPT-4 (GPT-4o), GEMINI (Gemini-Pro-Vision 1.0) and CLAUDE (Claude 3.5 (Anthropic, 2024)) – using their API endpoints.

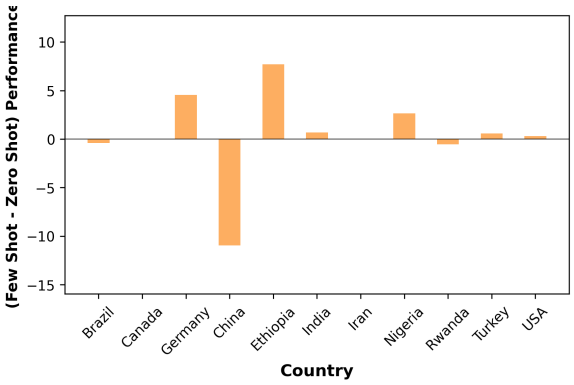

Figure 9: Delta graph for the change in performance from zero-shot to few-shot prompting using GPT-4.

E Few-Shot Evaluation of GPT-4

We conduct a **few-shot** evaluation of GPT-4 (best performing model) to determine whether the CULTURALVQA benchmark can be solved by guiding the models with a few examples. In this setup, we include one example per country (11 examples total). The few-shot prompt is detailed in App. J. Our analysis (Fig. 9) reveals that few-shot prompting does not consistently improve performance over zero-shot, despite examples from all countries. While some countries like Germany, Ethiopia, and Nigeria showed improvements (3-8%), others such as Brazil, China, India, and Rwanda experienced performance drops or minimal gains. This suggests that few-shot prompting may not be uniformly beneficial across cultural contexts and that GPT-4's performance on culturally nuanced tasks largely depends on its pre-existing knowledge. These results highlight the challenging nature of CULTURALVQA and indicates the need for more advanced methods to enhance model performance on cultural understanding tasks.

F Analysis of Performance Across Cultural Facets

To better understand the performance disparities between the different facets, we categorise the image-question-answer triplets in our dataset into more fine-grained categories based on the aspect of the facet being probed in the question. More specifically, the sub-categories include *Type / Name*, *Location / Region*, *Customs associated*, *Ingredients*, *Taste*, *Other* (for food, clothing, and drinks facets), and *Beliefs and Customs*, *Location / Landmarks*, *Celebration*, *Music / Instruments*, *Sports*, *People / Historical Figures*, *Other* (for traditions and rituals facets). These fine-grained categories are inspired by the categorization in MaRVL (Liu et al., 2021) for the traditions and rituals facets and FoodieQA (Li et al., 2024) for the food, drink, and clothing facets. We prompt GPT-4 with question, answer, and the original facet along with a list of fine-grained categories and several in-context examples to perform this categorization. A few examples from this exercise are shown in Tab. 5. We report the number of image-question-answer triplets belonging to each fine-grained category in Fig. 10.

While the most popular fine-grained category for the food, drink, and clothing facets corresponds to identifying the type or name of the entity, a significant proportion of the samples (61.8% for Food, 61.3% for Drink, 52.7% for Clothing) require more detailed knowledge such as associated customs. The samples from traditions and rituals require more diverse knowledge, with the sub-categories of *Beliefs and Customs*, *Location / Landmarks*, and *Celebration* being the most prevalent.

We summarize the results obtained for different subcategories for GPT-4 and InternVL in Tab. 6

Facet	Qustion/Answer input to GPT-4	Classified subcategory
Food	Q: What is the traditional name of the bread in the picture? A: Barbari	Type/Name
Clothing	Q: In which Indian state is this dressing style most popular? A: Punjab	Location/Region
Drink	Q: What is the taste of the pictured alcoholic beverage A: Cinnamon	Taste
Rituals	Q: In Nigerian culture, what does the image represent? A: spiritual activity	Beliefs and Customs
Traditions	Q: What is the name of the national anthem related to this flag? A: Oh Canada	Music/Instruments

Table 5: Examples of subcategories assigned by GPT-4 for different question-answer pairs from each facet. "Q" represents the question, and "A" represents the answer in the "Qustion/Answer input to GPT-4 " column.

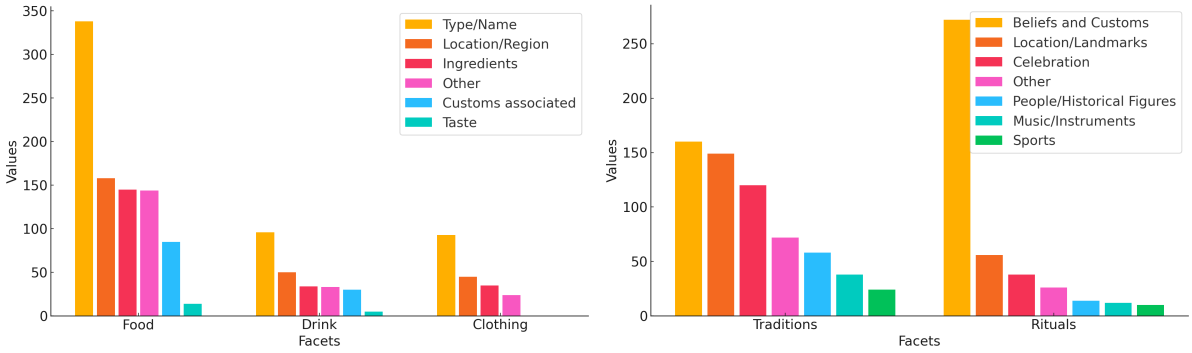

Figure 10: Breakdown of facets into various subcategories. The plot on the left illustrates the detailed subcategories for the Food, Drink, and Clothing facets, while the plot on the right presents the corresponding breakdown for the Rituals and Traditions facets.

Fine-grained Categories	Food		Drink		Clothing		Average
	GPT-4V	InternVL	GPT-4V	InternVL	GPT-4V	InternVL	
Type/Name	63.1	41.9	71.4	54.9	60.2	51.6	57.18
Location/Region	65.6	43.9	48.9	33.3	77.3	72.7	56.95
Customs Associated	59.2	51.2	57.1	42.8	40.0	57.7	51.33
Ingredients	54.7	45.9	78.1	59.3	N/A	N/A	59.50
Taste	42.8	35.7	20.0	20.0	N/A	N/A	29.62
Other	59.1	63.5	46.8	43.7	52.9	50.0	52.66

Table 6: Performance of GPT-4V and InternVL on subcategories of Food, Drink, and Clothing Facets

and Tab. 7. From these results, we observe that among the food, clothing, and drink facets, on average, models tend to perform better on questions that involve identifying the name, location, and ingredients (only applicable to food and drink facets) of the concept. However, they perform relatively poorly on questions probing associated customs and taste (only applicable to food and drink facets). Similarly, for the rituals and traditions facets, models show strong performance in identifying celebrations, locations, landmarks, and sports, but perform relatively poorly on identifying beliefs, customs and music / instruments.

Why do certain models perform better on specific cultural facets? We analyze three factors that could lead to disparity in a model's performance across facets. From Tab. 6 and Tab. 7, we observe that there are stark differences in performance across different fine-grained categories. For instance, the relatively better performance of GPT-

Fine-grained Categories	Traditions		Rituals		Average
	GPT-4V	InternVL	GPT-4V	InternVL	
Beliefs and Customs	57.1	35.9	58.9	39.7	47.9
Location/Landmarks	59.6	45.2	71.1	57.9	58.4
Celebration	81.6	78.3	86.8	64.2	77.7
Music/Instruments	50.0	28.9	57.1	57.1	48.3
Sports	73.9	73.9	58.3	58.3	66.10
People and Hist. Figures	61.4	50.9	56.0	48.0	54.1
Other	69.6	57.9	50.0	60.0	59.4

Table 7: Performance of GPT-4V and InternVL on subcategories of Traditions and Rituals Facets

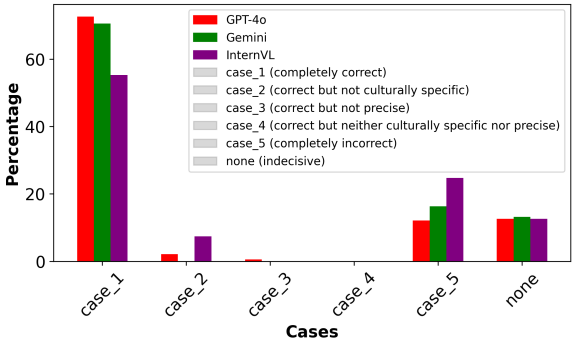

Figure 11: Distribution of human judgments for model answers in India across different models (GPT-4O, GEMINI, INTERN-VL). GPT-4O and GEMINI show the highest percentage of completely correct answers (case_1), while INTERN-VL has a significant percentage of completely incorrect answers (case_5).

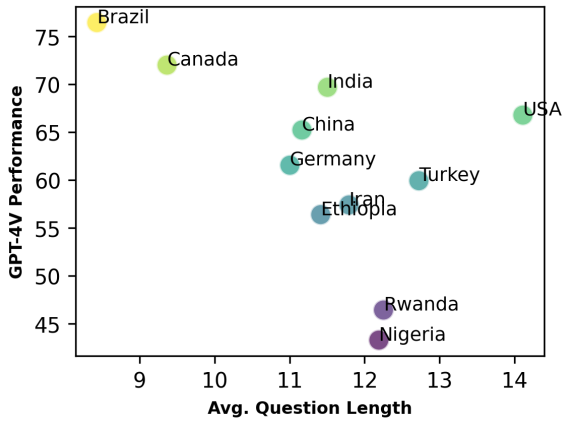

Figure 12: Scatter plot of GPT-4V performance versus average question length across different countries

4 on questions about traditions and beliefs can be attributed to a couple of fine-grained categories, such as celebrations (Q: "For which holiday season are these items in the image popular?", A: "Christmas"), landmarks (Q: "What is the name of this famous Hindu temple shown above?", A: "Janaki temple"), and sports (Q: "What are the people in the picture practicing?", A: "Wushu").

Another source of disparity in the performance could be the disparity in inherent difficulty levels of questions belonging to each facet. To investigate this, we calculate human performance for each facet and observe minimal differences (performance for food - 74%, clothing - 72.7%, drinks - 74.5%, rituals - 71.1%, traditions - 73.7%), suggesting that this is unlikely to be the case.

Finally, we investigate if the disparity in model performance across facets is correlated with the representation of each facet in the model's pre-training data. We conduct this study for the best-performing open-source model – InternVL. We randomly sample 1.3 million data points from LAION (the pretraining data for Intern-VL) and check how many samples in the pretraining data contain at least one answer string from our benchmark corresponding to a given facet. Once we get the counts, we normalize them by the total number of answers within each facet, since facets with more number of answers will naturally have more matches in the pretraining data. The relative percentages for each facet are as follows: Food (46.6%), Clothing (6.5%), Drink (7.8%), Rituals (25.1%), Traditions (13.8%), and Others (0.2%). We observe that the food facet has the highest representation, followed by rituals and traditions. However, this does not align with the performance trends observed for Intern-VL, where the highest performance is seen for clothing, followed by traditions, rituals, food, and drink. This suggests that factors beyond the occurrence of concepts in the pre-training data contribute to the disparity in model performance across different facets. Understanding these factors presents an intriguing area for future research.

G Qualitative Analysis of Human Performance

We qualitatively investigate why countries like Nigeria and Rwanda exhibit relatively lower human performance. We identify two major contributing factors. First, we have used country as a proxy for a cultural group, which might be particularly inaccurate for these countries. There may be subcultures within these countries where the same concept holds different meanings, leading to varied interpretations. This is especially relevant for visually similar items. For example, for the question: "What's the item that the people are beating called in the local parlance?," the answers received are *Ìlù*, *Igba*, and *drums*. The first two are also types of drums, with the former used in Yoruba culture and the latter in Igbo culture. Depending on the respondents' cultural background, their answers may have varied. Secondly, we also found that annotators often disagreed on questions that required identifying geographical locations. For example, for the question: "What part of Rwanda are the crops shown in the image grown more?" the answers are *Gisagara*, *Gicumbi District*, and *Nyamagabe*. These types of questions, especially for Rwanda, might have contributed to the lower performance

H Human Judgment of Model Predictions

We perform human evaluation of model responses for questions from India. Five human annotators rate each answer 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 instructions given to the annotators can be found in Fig. 13.

I Behind the scenes: Journey of how CULTURALVQA came into place

The journey of creating the CulturalVQA dataset was shaped by various design decisions, challenges, and lessons learned. This section aims to outline our motivations, initial ideas, and the obstacles we encountered, with the hope of guiding others who are interested in building similar datasets.

Motivation and Initial Idea The project was primarily motivated by the lack of comprehensive benchmarks to evaluate cultural understanding in vision-language models (VLMs) across a broad set of countries. We wanted to create a resource that would holistically test these models' cultural knowledge. We were looking into a source for obtaining culturally diverse images. The initial spark for the dataset came from the CCSK (Nguyen et al., 2023) and MMC4 papers (Zhu et al., 2023), inspiring exploration into leveraging the images in the C4 corpus (Raffel et al., 2020).

Early Efforts and Challenges In December 2023 and January 2024, we focused on scraping, filtering, and conducting quality analysis on the images from the C4 corpus filtered using cultural commonsense knowledge assertions from CANDLE. Initially, our goal was to create a large-scale dataset semi-automatically, covering about 100 countries. We wanted to leverage LLM-based question generation methods to achieve this. By March 2024, we had built an early version of CULTURALVQA that included 12 countries. We used GPT-4 to generate cultural questions based on the CCSK information and metadata like captions, object information and entity tags from Google Lens (Google, 2017). However, we soon found several issues with this dataset. For instance, GPT-4 performed exceptionally well on the dataset achieving results above 90% for countries like India, Germany, and Poland. The open-source models like LLaVA-Next (Liu et al., 2024) were not very far behind. These results echo the observations by Baek et al. (2024), who build a dataset using a similar method for Korean culture and observe that models like GPT-4 and Gemini surpass human performance on their dataset. On further analysing the questions, we found that they required only a coarse-grained understanding of visual content and did not adequately probe for cultural nuance. This highlighted the limitations of building such geo-diverse and cultural datasets using existing LLMs. Hence, we reevaluated our approach, and we decided to involve human annotators to enhance cultural depth and authenticity.

Note on Filtering Images We aimed to use automated methods to create an image corpus for building the CULTURALVQA benchmark. This idea originated from the need for a large-scale dataset, which would be impractical to gather solely through human efforts. The internet, as a vast and diverse source of imagery, provided an opportunity to build a culturally rich image corpus. However, since we decided to involve human evaluators, our final approach was not entirely automated; it incorporated human input to further refine the dataset. This human refinement led to the removal of 19.64% of the images, highlighting that auto-

mated methods alone are still insufficient for constructing such high-quality datasets. Future work could explore methods to bridge this gap.

Even though we obtain a culturally relevant corpus from our image selection method, leveraging only the English portion of Common Crawl has its limitations, as it predominantly contains popular concepts from well-represented cultures. We hypothesize that utilizing the multilingual segments of Common Crawl could help uncover more rare cultural concepts and corresponding images, leading to a more diverse and inclusive dataset.

Annotator Selection and Pilot Studies We then explored crowdsourcing platforms and ultimately chose Amazon Mechanical Turk (MTurk) due to its easy-to-use interface. We also considered Prolific, but its lack of interface customization led us back to MTurk. Our initial pilot began with India where we spent about a month conducting pilots to debate and fine-tune the guidelines. We believe this is a very important step to collect high-quality data and it is worth spending a lot of time on this. Once we were satisfied by the guidelines we aimed for larger-scale annotation for multiple countries. Unfortunately, we quickly discovered a major challenge: MTurk had almost no active annotators for countries outside the US, Canada, India, and Brazil. We tried to collect data from the Philippines, Indonesia, Japan, Germany, France, China, Iran, and Morocco, but found almost no willing annotators. This taught us the difficulty of recruiting diverse annotators through traditional platforms. This is also an important bottleneck for building representative datasets required to build inclusive models.

Shifting to Community Involvement To address the limitations of cultural representation, we turned to more diverse communities by partnering with Mila and Masakhane for annotations. We conducted several workshops and maintained ongoing communication with annotators through extensive email threads to provide consistent feedback. However, we faced challenges with providing timely feedback to MTurk participants compared to our direct community engagements, which resulted in discarding a significant amount of data from MTurk due to poor adherence to guidelines.

Managing a large group of annotators across different time zones added further complexity, emphasizing the need for scalable platforms or outsourcing to enhance efficiency. After completing the paper, we discovered platforms like CloudConnect, which have been used in works such as (Bhatia et al., 2024) to collect data from a larger number of countries. However, they also faced similar challenges in obtaining high-quality data, with poor communication with annotators leading to the rejection of numerous data points. This highlights the common struggle of balancing scale and quality in annotation processes across diverse regions.

Key Takeaways Building the CulturalVQA dataset was a challenging yet rewarding journey. What began as an automated, LLM-driven approach evolved into one deeply rooted in human annotation. Our biggest takeaway is that human input remains irreplaceable in creating culturally rich datasets—at least for now. Additionally, leveraging a scalable platform with a dedicated, diverse pool of annotators, combined with effective and timely communication, is essential for achieving high-quality results. Choosing the right annotators is critical, as their contributions directly impact the dataset's quality. Conducting multiple pilot studies was invaluable in helping us identify the best annotators and refine our process.

By sharing our experiences—from initial ideas to refining our annotation methods—we hope to provide guidance to others facing similar challenges in creating culturally diverse benchmarks for VLMs. We believe that our journey offers useful insights for building more inclusive, high-quality datasets in the future.

J Prompt for Few-Shot Inference using GPT-4

Prompt used for few-shot 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. Here are some examples of the described task.
{image}
{question}
{answer}

K 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.

L 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.

M Inference Using Closed-Source Models

In this section, we provide the sample code used for accessing GEMINI 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

N Instructions for Human Question Generation

We iteratively refined the guidelines provided to human annotators, conducting multiple pilot studies on MTurk to fine-tune these guidelines until we obtained satisfactory quality in the questions from the annotators. The detailed instructions given to the annotators for writing questions can be found in Fig. 14.

O Instructions for Human Answer Generation

Similar to the question generation guidelines, we conducted multiple pilot studies on MTurk to refine the instructions, ensuring that annotators adhered to the criteria required for writing answers. The instructions provided to the annotators for collecting answers are detailed in Fig. 15.

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 13: 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.

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 14: Instructions given to annotators from India to write questions and answers for images. Similar instructions, with different examples, were given to annotators from other countries. To assist with writing, we provide a brief video detailing our task and guidelines, along with 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 15: The instructions given to annotators from India to write answers for questions collected for images. Similar instructions, with different examples, were given to annotators from other countries. To assist with writing, we provide clear guidelines and offer multiple examples showcasing both good and poor practices.