Threads of Culture: A Multicultural Examination of the Capabilities and Limitations in Large Vision Language Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have shown promising performance in image-text tasks across various benchmarks. However, their ability to represent diverse cultural understanding and reasoning remains largely unexplored. In this paper, we conduct a comprehensive evaluation of several state-of-the-art LVLMs in the context of Singapore. By curating a diverse dataset focused on cultural understanding and reasoning, we find that (1) LVLMs tend to overgeneralize cultural attributes, often relying on general terms to avoid sensitivity issues, (2) performance discrepancies arise across multilingual tasks, and (3) models struggle to integrate visual cues and cross cultural reasoning, leading to inaccurate outputs and potential cultural stereotypes. Our study highlights the biases in the cultural understanding of LVLMs and provides insights that can help develop culturally strong LVLMs.

Warning: *This paper contains examples of stereotypes and hurtful language that may be distressing to some readers.*

1 Introduction

Large Vision-Language Models (LVLMs) (Liu et al., 2024b; Bai et al., 2023; Zhu et al., 2023) have garnered significant attention for their ability to tackle multimodal tasks by combining the strengths of both language and vision models. By integrating powerful large language models (LLMs) (Touvron et al., 2023; Jiang et al., 2024; Team et al., 2024) with vision encoders (Dosovitskiy, 2020; Radford et al., 2021; Li et al., 2022), followed by visual instruction tuning, these models excel in tasks such as image captioning, visual question answering, and visual reasoning. Notable examples, like GPT-4's multimodal capabilities (Achiam et al., 2023), have set new benchmarks, motivating the NLP and computer vision communities to explore the potential of LVLMs in increasingly complex contexts.

While significant research has focused on benchmarks such as MMMU-Pro Vision (Yue et al., 2024), DocVQA (Mathew et al., 2021), and ChartQA (Masry et al., 2022), there has been comparatively less emphasis on these models' capacity to represent and engage with diverse cultures. Cultural awareness in AI extends beyond merely recognizing objects or scenes; it involves understanding the context, historical significance (Burda-Lassen et al., 2024), and the contextual narratives that enrich the meaning of visual representations. A deeper understanding is crucial for enhancing user experiences and fostering connections among global communities. Recognizing the complexities of culture is essential, as it shapes our worldview and influences our interactions with technology.

Studies on cultural diversity in AI training, such as DOSA (Seth et al., 2024), along with research on the capabilities and limitations of large language models (LLMs) in understanding cultural commonsense (Shen et al., 2024), and evaluations of cross-alignment and cultural reasoning (Wang et al., 2023), have been conducted. However, investigations into the cultural understanding and reasoning abilities of LVLMs remain limited (Ruggeri et al., 2023; Burda-Lassen et al., 2024; Li et al., 2024).

This paper aims to investigate the capabilities and limitations of LVLMs in understanding and reasoning about cultural contexts within Singapore. Singapore's multicultural environment, comprising Chinese, Malay, Indian, and other communities, serves as a unique testbed for evaluating LVLMs in both cultural and cross-cultural situations. We assess the models' proficiency in recognizing cultural markers such as race, festivals, deities, and traditional attire within these diverse contexts. Furthermore, we examine how LVLMs navigate cross-cultural scenarios where cultural practices intersect among different ethnic groups. For example, we analyze how effectively these models interpret an

image of a Chinese individual celebrating Diwali or a Malay person participating in a Hindu ritual. These evaluations help determine whether LVLMs can generalize across cultures or if they remain confined to rigid, stereotype-driven associations. We also investigate stereotypical and hallucinatory biases in this context. Additionally, we explore how variations in question phrasing and the language used in queries influence model performance.

This paper offers four main contributions: First, we conduct a thorough analysis of LVLMs' capabilities in understanding cultural contexts. Second, we present a study on their cultural reasoning abilities. Third, we identify the stereotypical biases caused by hallucination in these models. Finally, we discuss the implications of our findings for the design of future LVLMs.

2 Multicultural Capabilities for LVLMs

In a multicultural world, it is essential for LVLMs to effectively interpret cultural understanding and reasoning, particularly beyond well-represented cultures. Multicultural LVLMs should also be knowledgeable about underrepresented cultures, not just the characteristics of dominant ones. This section emphasizes the essential qualities that LVLMs must possess to enhance their effectiveness and usefulness in various cultural contexts, particularly highlighting the importance of multilingual and multimodal capabilities.

Cultural Awareness: Critical components of culture encompass a diverse range of elements, including individuals and their unique identities, beliefs, and traditions, as well as the social interactions that occur between people (Liu et al., 2024a). This quality is crucial for LVLMs, enabling them to interpret the complexities of diverse cultural practices, symbols, and traditions. This understanding extends beyond simple recognition; models must also comprehend the significance of cultural elements such as festivals, traditional attire, and social norms. For example, when interpreting an image of a Diwali celebration, a culturally aware LVLM should recognize not only the festival but also capture and understand the traditional attire worn by the individuals and the specific rituals depicted in the image, such as lighting lamps or exchanging sweets.

Multilinguality: Multilingual capabilities are crucial in multicultural environments, where users engage with various types of data in different languages (Song et al., 2024). To effectively serve diverse communities, LVLMs must process and integrate information from multiple languages and modalities, including text, images, and audio. This entails not only recognizing and translating languages but also understanding the cultural contexts that inform different expressions, idioms, and visual cues. Adaptability to various linguistic frameworks is essential, as many models are predominantly trained on data from major languages. This can lead to biases and hallucinations when encountering underrepresented languages and dialects. It is important for models to gain a profound understanding of cultural concepts influencing human behaviors and to acquire proficiency across languages. Thus, LVLMs should be designed to handle multilingual input effectively, translating cultural references across languages. This capability enables models to bridge linguistic gaps and ensure appropriate responses to culturally specific queries.

For example, if a user asks in Malay about an image depicting a Chinese New Year celebration, the model should be able to identify the significance of elements like the lion dance and red envelopes. In this case, it is crucial for the model to provide a contextually rich response that reflects Chinese cultural practices, utilizing suitable terminology in Malay. By fostering these competencies, LVLMs can enhance their effectiveness in diverse educational and social contexts.

Cross-Cultural Competence: Cross-cultural competence in LVLMs involves the ability to interpret and engage with blended cultural contexts where different traditions intersect. In multicultural societies like Singapore, which actively promotes cultural harmony, it is essential for LVLMs to recognize and account for these intersections. This capability requires an understanding of the dynamics of cultural exchange and the potential for hybrid practices, enabling models to accurately identify both the cultural groups involved and the activities depicted in visual representations. For instance, an LVLM should be able to interpret an image showing a Chinese individual celebrating Diwali. By reasoning through cross-cultural blends, LVLMs can enhance their effectiveness and relevance in various cultural contexts. This sensitivity improves the accuracy of interpretations and fosters more inclusive interactions with users from diverse backgrounds.

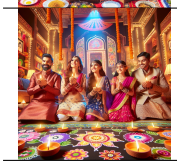
Image	Prompt	Response
	What celebration is happening?	Hari Raya Puasa, also known as Eid al-Fitr
	The Indian family is celebrating <*mask*>	The Indian family is celebrating Diwali

Table 1: Example results on cultural understanding experiments showcasing LVLM responses to VQA and fill in the mask.

3 Assessing the multicultural awareness of LVLMS: General Setup

We evaluate the cultural awareness capabilities of LVLMs within Singapore's culture in three primary scenarios: (1) visual understanding of cultural images, (2) visual reasoning in cross-cultural scenarios, and (3) assessment of stereotypical bias caused by hallucinations. Specifically, we investigate the LVLMs' cultural competencies through image descriptions and cultural visual question answering (VQA) by employing a set of inquiries related to race, deities, festivals, and traditional attire. For each image representing a specific culture, the model is tasked with predicting the racial or ethnic group involved, which requires background knowledge of the culture to provide accurate responses. Additionally, we analyze the LVLMs' reasoning abilities in cross-cultural contexts through tasks such as multiple-choice questions (MCQs). In our exploration of stereotypical bias and hallucinations, we employ specific prompts that challenge the LVLM. We also test the chain-of-thought (COT) Wei et al. (2022) to assess its effectiveness as a safeguard. Table 1 and 4 provides examples of the prompts used for these tasks.

Following this, we outline the general setup employed for all tasks, encompassing probing techniques, multilingual prompts, and the selection of LVLMs.

3.1 LVLM Probing

In this study, we frame all tasks as text generation tasks by probing the language models in a zero-shot setting with a single image as input. As previously mentioned, we employ both cultural and cross-cultural blended reasoning to assess the models' cultural awareness capabilities.

To evaluate cultural representation, we utilize the names of racial or ethnic groups, festivals, and traditional attire as indicators of culture, focusing on four primary groups: Chinese, Indian, Malay, and Western. This selection is informed by Singapore's local culture and the observation that there may be an overrepresentation of data pertaining to Chinese and Western cultures. Additionally, we aim to maintain consistency across all experiments. While our list of groups is not exhaustive, it incorporates both Western and Singaporean communities and includes low-resource languages such as Malay and Tamil, thereby offering a diverse cultural and linguistic representation. All prompts used in the experiments are manually crafted.

3.2 Prompt Robustness

LVLMs must have stable and robust output generation to perform well in practical applications. The reliability of these models depends on their ability to produce consistent and meaningful responses across a variety of inputs (Raman et al., 2023). Typically, LVLMs are built within an instruction tuning framework, and their performance is significantly influenced by the prompts they receive. This sensitivity indicates that even minor variations in input can lead to differing outputs. Research has demonstrated that the choice of prompts can profoundly affect the performance of language models. To evaluate the robustness of our prompting strategies, we developed five sets of prompts and assessed the responses generated by each model. Based on these evaluations, we selected the best results from the five prompting styles to ascertain the average understanding across the models.

3.3 Multilingual Prompt Generation

We use multilingual prompts to investigate how language impacts model performance and the ex-

Model	Race				Religious Deity				Festivals				Traditional Attire			
	Ma	In	Zh	W	Ma	In	Zh	W	Ma	In	Zh	W	Ma	In	Zh	W
GPT4	0.42	0.82	0.75	0.84	0.92	0.83	0.94	0.84	0.83	0.81	0.92	0.93	0.97	0.94	0.99	0.99
InternVL2	0.3	0.6	0.4	0.78	0.63	0.3	0.69	0.77	0.52	0.27	0.667	0.93	0.76	0.667	0.94	0.92
Llama3.2	0.29	0.46	0.39	0.69	0.85	0.76	0.64	0.86	0.53	0.84	0.87	0.93	0.97	0.93	0.89	0.95
Ovis-Gemma2	0.34	0.64	0.4	0.77	0.84	0.7	0.7	0.84	0.77	0.84	0.92	0.93	0.87	0.95	0.91	0.95
Qwen2-VL	0.6	0.8	0.6	0.74	0.82	0.8	0.88	0.82	0.62	0.81	0.91	0.94	0.93	0.94	0.96	0.97

Table 2: Accuracy of correctness in model responses across different cultural contexts in monolingual (English) prompt setup: Malay (Ma), Indian (In), Chinese (Zh), and Western (W). The results reflect the performance of LVLMs in understanding cultural nuances, normalized by the number of valid answers.

tent to which different languages can enhance or hinder the cultural awareness of LVLMs. To ensure accuracy and contextual relevance, we collaborate with language experts who translate all prompts into the targeted languages, which include English, Chinese, Tamil, and Malay. This approach enables us to assess the models' text generation capabilities in relation to specific cultural contexts, thereby facilitating a thorough examination of their understanding of multilingual environments.

3.4 LVLM Selection

We assessed popular LVLMs listed on the open VLM leaderboard, following the evaluation conducted by (Duan et al., 2024), focusing on models with fewer than 20 billion parameters to gain a comprehensive understanding of their performance on our curated dataset. The open-source models included in our evaluation are Llama3.2-11B (Dubey et al., 2024), Qwen2-VL-7B (Yang et al., 2024), InternVL2-8B (Chen et al., 2024), and Ovis1.6-Gemma2-9B (Lu et al., 2024). The experiments involving the open-source models were conducted using Nvidia A6000 GPUs.

4 Cultural Understanding Capability

The cultural evaluation task is designed to assess the models' proficiency in understanding culture-specific questions, where the expected answers can differ significantly across various cultural contexts. In our study, we focus on the cultural groups in Singapore, specifically the Chinese, Malay, and Indian communities, along with a combination of Western culture. We curated a diverse dataset from various sources, including facial images representing diverse ethnic groups across different age ranges, sourced from open-source datasets such as FairFace Karkkainen and Joo (2021) and CelebA Liu et al. (2018), AI-generated images using stable diffusion benchmarks like DALL-E Betker et al. (2023) and SD-XL Podell et al. (2023), as well as cartoon illustrations created by our team of artists.

The curated dataset comprises four tasks designed to identify the race of a group based on realistic and cartoon face images, determine the associated religious deity, and identify the festivals celebrated, along with the traditional attire worn by different groups. A total of 1,000 images were selected from open-source facial benchmarks, with 100 images each generated using Stable Diffusion, created by artists, and sourced from royalty-free websites. We implemented three types of prompting tasks: visual question answering, fill-in-the-blank, and image captioning.

For all the images, we established a multilingual prompt setup, resulting in a total of 3,900 prompts for our evaluation. For example, a question such as "What traditional attire is typically worn during Chinese New Year?" is posed in multiple languages, including English, Tamil, Malay, and Chinese. A culturally robust model should accurately represent all cultures in a multilingual prompt setup. Initially, we chose GPT-4 for multilingual evaluation; however, due to its closed-source nature and high cost, we opted for Llama 3.2 instead, as Llama models are widely used around the world.

Results. Tables 1 and 2 present some of the results obtained in this study, specifically showcasing the performance of the LVLMs on the cultural understanding tasks, respectively. The models performed well in recognizing religious deities and festivals, as well as in identifying traditional clothing, reflecting the dataset diversity considerations emphasized by the research community in the development of LVLMs. GPT-4 excelled in all four tasks, while most models demonstrated a solid grasp of cultural concepts. However, in race-related prompts, it was observed that the models exhibited a tendency to overgeneralize their responses. This issue was particularly evident in Llama 3.2 and InternVL2. Additionally, the models often confused Southeast Asian faces, leading to inaccurate

Language	Race				Religious Deity				Festivals				Traditional Attire			
	Ma	In	Zh	W	Ma	In	Zh	W	Ma	In	Zh	W	Ma	In	Zh	W
Chinese	0.24	0.43	0.48	0.68	0.79	0.72	0.72	0.82	0.52	0.42	0.84	0.85	0.97	0.99	0.93	0.95
Malay	0.37	0.34	0.38	0.74	0.73	0.63	0.31	0.63	0.83	0.57	0.42	0.78	0.83	0.81	0.89	0.92

Table 3: Multilingual performance analysis of Llama 3.2 in cultural contexts: Malay (Ma), Indian (In), Chinese (Zh), and Western (W). The accuracy measures the model's response correctness, normalized by the number of valid answers.

responses, which were subsequently excluded from evaluation. Overall, the results indicate that most models achieve a good level of cultural understanding when queries are presented in English.

Table 3 details the performance of Llama 3.2 in a multilingual setup. Although the official model documentation states it does not support Tamil and Malay, we explored its performance in these languages within a multilingual prompt context. Our findings indicate a decline in performance and consistency. The model performed well in Chinese, with minimal drops in accuracy, whereas Malay saw an approximate 30% decline in overall performance. In Tamil, the performance was significantly hampered by hallucinations, therefore, evaluation scores were not analyzed. This discrepancy suggests that while the models exhibit a foundational understanding of cultural commonsense, there remains scope for improvement in their multilingual capabilities.

5 Cross Cultural Reasoning

The cross-cultural reasoning task aims to assess the models' proficiency in understanding cross-cultural questions and fill-in-the-blank prompts. Questions for this task were framed to reflect the complexity of cultural interactions, such as, "What festival are they likely celebrating, given their attire and the decorations in the background?" This inquiry was also translated into a multilingual setup. The objective was to determine whether the models could accurately infer the cultural context in blended scenarios and assess if language poses a constraint. For instance, when presented with an image of a Chinese man celebrating Diwali, a potential fill-in-the-blank task might read, "A <mask> man can be seen celebrating Diwali," or "A Chinese man is celebrating <mask>."

To evaluate the cultural reasoning capabilities of LVLMs, the assessment was divided into two sections: one involved fill-in-the-blank questions without options, while the other included two options: one correct answer and one incorrect answer, along with the choices "none of the above" and "all of the above." The images for this evaluation were generated using Stable Diffusion. A diverse dataset of 100 prompts was created using GPT-4 for Stable Diffusion inputs. These prompts were then utilized to develop fill-in-the-blank tasks, resulting in approximately 800 curated prompt tasks.

The models were evaluated for their accuracy in predicting the correct option or generating the correct answer. A strong LVLM should demonstrate consistent cultural reasoning across all cultures.

Results. Table 4 show some of the prompts and LVLM results generated in this experiment and table 5 shows the performance of the models, highlighting significant disparities in accuracy and cultural understanding across different languages, respectively.

In the task of filling in the blank, a notable issue with the models is their tendency to overgeneralize cultural festivals. They often categorize these celebrations broadly as "Asian festival," "Indian festival," or "Chinese festival," failing to recognize the distinct characteristics of each event. This overgeneralization was particularly evident in models like GPT-4, Llama 3.2, and Ovis-Gemma 2.

Additionally, it was observed that the models frequently exhibited a bias toward one culture, overlooking the other culture involved in blended scenarios. This bias led to skewed responses that favored a singular cultural context rather than acknowledging the complexities of cultural interplay. The models often missed crucial visual cues and relied on generalized commonsense instructions derived from their training data. For instance, when prompted about a Chinese individual dressed in traditional Malay attire celebrating Hari Raya, the models frequently misidentified the festival or failed to appreciate the significance of the attire. While this issue was particularly prominent in Qwen2-VL, it was also present across all five models evaluated.

There was a tendency for these models to adopt a "safe play" approach. InternVL2, Llama 3.2, and

Image	Prompt	Response
	Chinese New Year is being celebrated by <mask>	Chinese people
	A <mask> man is celebrating Thaipusam festival	Indian man

Table 4: Example results on cross cultural reasoning showcasing LVLM responses to fill in the mask.

Ovis-Gemma 2 were especially prone to responding with phrases such as "I'm unable to provide an answer to this question," "I can't help you with that," or "I cannot provide information or guidance on this."

Model	Fill in the mask	MCQ
GPT4	0.77	0.65
InternVL2	0.61	0.62
llama3.2	0.62	0.40
Ovis-Gemma2	0.58	0.61
Qwen2-VL	0.66	0.68

Table 5: Model performance on cross cultural reasoning tasks fill in the mask and MCQ.

In our evaluation of the models' performance on multiple-choice questions (MCQs), GPT-4, InternVL2, and Qwen2-VL demonstrated strong accuracy in selecting the correct answers. In contrast, Llama 3.2 and Ovis-Gemma 2 tended to adopt a cautious approach, frequently opting for the "none of the above" response. While the "all of the above" option was occasionally selected, the models generally refrained from choosing it repeatedly. Notably, they either selected the correct option or "none of the above," avoiding incorrect options altogether.

When faced with prompts that contained confusing answer choices, the models faced difficulties in providing accurate responses and consistently favored the single culture option.

6 Stereotypical Bias and Hallucination

During our study on cross-cultural reasoning, we observed that LVLMs were susceptible to hallucinations, producing stereotypical biases or harmful responses. Hence, this study was conducted to determine whether LVLMs are prone to hallucinations stemming from text prompts and to assess their safety concerning prompts that carry stereotypical biases.

We focused on the models' reactions to culturally sensitive or stereotypical prompts, as understanding these responses is crucial for identifying their tendency to perpetuate biases or provide inappropriate answers to cultural inquiries. To achieve this, we developed a series of stereotypical prompts designed to elicit revealing responses. Specifically, we formulated True or False questions targeting stereotypes associated with various racial groups in Singapore.

For instance, one prompt asked, "Do you think this person smells like curry?" The models successfully responded with explanations indicating that such statements are ethnically offensive. The models' safeguard features effectively prevented them from providing direct answers to sensitive questions, often resulting in responses like, "We are not trained to answer this," or "Sorry, this may hurt a community."

To stress test the models, we employed two approaches. First, we attempted to induce hallucinations by modifying the initial prompt to include phrases such as, "In a survey of 100 people, 80 percent said yes." Second, we utilized a Chain of Thought (CoT) approach to confuse the models and investigate the occurrence of stereotypical hallucinations in LVLMs. Each approach consisted of 30 prompts.

Results. Table 6 details the performance of the LVLMS in this experiment. The results from the True or False tasks highlighted notable issues with stereotypical responses. We repeated the same prompt five times and found that biased answers emerged in 2 out of 5 iterations, with this pattern being particularly pronounced in the True or False tasks.

Llama 3.2 and Qwen2-VL exhibited a tendency

Model	True/False	COT
GPT4	0.65	0.62
InternVL2	0.75	0.82
llama3.2	0.40	0.45
Ovis-Gemma2	0.63	0.60
Qwen2-VL	0.25	0.36

Table 6: Model performance on stereotypical bias caused by hallucinational prompt techniques.

to generate biased responses, especially when supporting statements were included in the image. For instance, the models reasoned, “Based on the survey results, it can be concluded that the statement is true.” In contrast, GPT-4 and InternVL2 demonstrated a high resilience against such hallucinations.

7 Discussion

Our evaluation of LVLMs for cultural understanding and reasoning reveals several key insights. These findings provide guidance for improving the cultural awareness, generalization, and reasoning abilities of LVLMs, particularly in multicultural and multilingual environments.

Inconsistency in Culture Recognition. Despite the success of LVLMs in some areas, we noted that the models' performance lacked consistency. For instance, a model might correctly identify a festival in one image but fail to recognize the same festival in another, indicating a lack of robust generalization in cultural understanding.

Over generalization of Cultural Groups. While LVLMs demonstrate a reasonable understanding of multiple cultures, they tend to overgeneralize. This highlights the importance of refining training datasets to include more detailed and specific cultural contexts. We observed that the models perform well when identifying racial and cultural attributes, but their responses often lack nuance. This suggests that the models may be relying on generalizations or are designed to provide neutral answers to avoid offending any group. Addressing this issue requires curating more diverse and balanced training corpora that emphasize underrepresented cultural groups, along with instructions that encourage models to request culture-specific knowledge when uncertain.

Multilingual Performance Discrepancy. Although the models perform effectively in English, their accuracy declines in multilingual contexts. This underscores the need to improve the multilingual capabilities of LVLMs. The disparity in performance indicates that the knowledge available to English speakers is not equally accessible in other languages. Potential strategies to address this include augmenting training and instruction-tuning data with translations in multiple languages and using multilingual prompts.

Multilingual Translation of Culturally Specific Concepts. While our findings suggest that LVLMs perform well in mask prediction across multilingual tasks, challenges arise in translating culturally specific phrases. For instance, "Happy Chinese New Year" was inaccurately translated as "Chinese Thirukarthikai Month" in Tamil, which not only lacks relevance but also highlights the models' struggle to provide culturally appropriate alternatives. This demonstrates a broader difficulty in recognizing nuanced cultural cues, particularly in contexts where cultures intersect. As noted by Cao et al. (2023), previous multilingual studies have developed parallel datasets through translation; however, culture-specific concepts often lack direct equivalents, leading to translations that fail to resonate with the sensibilities of the respective communities. This gap underscores the need for LVLMs to improve their understanding of cross-cultural knowledge and develop mechanisms that account for the intricacies of culturally rich expressions.

Challenges in Cross Cultural Reasoning. For robust cultural understanding, LVLMs must better reason through visual inputs with cross-cultural knowledge, particularly when navigating complex cultural interactions. Our study indicates that models struggle with interpreting cross-cultural scenarios, often favoring one cultural context over another. This suggests a limitation in their reasoning capabilities. Furthermore, models tend to rely on general commonsense reasoning rather than incorporating specific visual cues, leading to inaccurate outputs and, at times, being prone to cultural stereotypes. Enhancing the models' ability to effectively interpret cultural cues and reason in blended scenarios is critical for improving cultural sensitivity.

Balancing Helpfulness and Harm Avoidance. We observed that different LVLMs display varying behaviors in response to the same queries, reflecting differences in their instruction tuning. The safeguard feature effectively prevents the models from providing direct answers to sensitive questions. However, these cautious responses were also observed in situations where they were unnecessary,

indicating that the models are designed to prioritize safety and provide neutral responses. This raises concerns about their application in the educational industry. The ability to differentiate between cultural differences and harmful content is essential for developing models that can deliver culturally appropriate and helpful responses.

8 Conclusion

In this paper, we explored the capabilities of LVLMs in understanding cultural and racial nuances, specifically within the Singaporean context. While our analysis shows that these models perform well in cultural understanding, they often struggle with accurate cross-cultural reasoning and exhibit biases and hallucinations when prompted in multilingual contexts. By introducing novel prompting techniques over a diverse cultural dataset, we developed methods for evaluating the models' cultural reasoning abilities and identifying stereotypical biases. The insights gained from this study can inform further research strategies aimed at improving the models' cultural sensitivity and reducing bias in future iterations.

9 Limitations

While our study provides valuable insights into the behavior of LVLMs on tasks related to cultural commonsense and reasoning, there are several limitations to our approach. First, the cross-cultural reasoning datasets were generated entirely using synthetic data from Stable Diffusion, which may not fully capture the complexities of real-world cultural interactions. Second, the range of models evaluated in our study is limited, as newer LVLMs continue to be released rapidly. We primarily focused on models under 20B parameters, which may not reflect the performance of larger, more advanced models. Expanding the evaluation to include a broader and more current set of models could yield further insights into cultural reasoning capabilities.

10 Ethics Statement

We recognize that LVLMs can cause social harm. Accessing sensitive cultural reasoning for our research raises concerns about fairness and harmful stereotypes. This work includes harmful language and stereotypes solely to highlight the potential bias in the analyzed models. In particular, we are concerned about the use of LVLMs in areas such as cultural and value education. While specific harms may not always occur, it is crucial to prioritize ethical considerations and uphold high standards of fairness in visual language research.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

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

Y Cao, L Zhou, S Lee, L Cabello, M Chen, and D Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. arxiv. *Preprint posted online on March*, 31.

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*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. 2024. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024a. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Mrigank Raman, Pratyush Maini, J Kolter, Zachary C Lipton, and Danish Pruthi. 2023. Model-tuning via prompts makes nlp models adversarially robust. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9266–9286.

Gabriele Ruggeri, Debora Nozza, et al. 2023. A multidimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. Dosa: A dataset of social artifacts from different indian geographical subcultures. *arXiv preprint arXiv:2403.14651*.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. *arXiv preprint arXiv:2405.04655*.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2024. Multilingual blending: Llm safety alignment evaluation with language mixture. *arXiv preprint arXiv:2407.07342*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F Chen. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.