Vision-Language Models under Cultural and Inclusive Considerations

Antonia Karamolegkou, Phillip Rust, Yong Cao,
Ruixiang Cui, Anders Søgaard, Daniel Hershcovich

Department of Computer Science, University of Copenhagen
Correspondence: antka@di.ku.dk

Abstract

Large vision-language models (VLMs) can assist visually impaired people by describing images from their daily lives. Current evaluation datasets may not reflect diverse cultural user backgrounds or the situational context of this use case. To address this problem, we create a survey to determine caption preferences and propose a culture-centric evaluation benchmark by filtering VizWiz, an existing dataset with images taken by people who are blind. We then evaluate several VLMs, investigating their reliability as visual assistants in a culturally diverse setting. While our results for state-of-the-art models are promising, we identify challenges such as hallucination and misalignment of automatic evaluation metrics with human judgment. We make our survey, data, code, and model outputs publicly available.

coastalcph/vizwiz-culture

1 Introduction

With the increasing integration of AI applications into our lives, it is important to consider human-centered use cases when evaluating such systems. Large multimodal language models are now used as visual assistants for blind and visually impaired individuals. Given that people across different cultures use such applications, it is essential to ensure not only their accuracy and faithfulness (Brady et al., 2013; Gonzalez et al., 2024) but also their cultural representation and inclusion (Hershcovich et al., 2022; Shi et al., 2024).

Existing evaluation benchmarks for VLMs focus primarily on English with few, implicit mutlicultural references. Although multicultural evaluation datasets like MaRVL (Liu et al., 2021) and XM3600 (Thapliyal et al., 2022) include culture-specific images (e.g., traditional wedding costumes), they also contain images with minimal cultural significance (e.g., a bag of carrots). Consequently, these datasets may not accurately measure the cultural knowledge of VLMs, despite being useful for assessing their multilingual capabilities. Additionally, evaluating these systems as visual assistants presents further challenges due to varying photo quality, user goals, and photo content (Chiu et al., 2020; Jung et al., 2022). Recently, Gonzalez et al. (2024) conducted a diary study with blind and low-vision individuals using an AI-powered scene description application, revealing that significant improvements are still needed for satisfying and trustworthy user experiences.

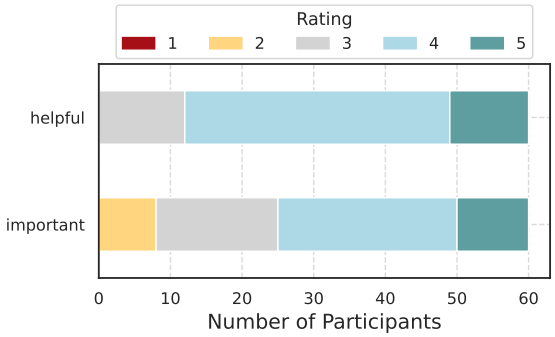


Figure 1: Survey results from people with visual impairments rating *importance* and *helpfulness* of cultural information in image captions. We use a Likert scale from 1 (not important/helpful) to 5 (very important/helpful).

To address both cultural and visual challenges, we first surveyed visually impaired individuals to gather their caption preferences and determine if cultural details are necessary. Then, we filtered an existing dataset with images taken from people who are blind, identifying implicit cultural concepts. This is used as a challenging benchmark to evaluate image captioning performance on cultural images of state-of-the-art models across different prompt settings. With these experiments, we investigate how AI applications, such as image captioning, can foster a more inclusive and culture-aware experience for all.

Proceedings of the 1st Human-Centered Large Language Modeling Workshop, pages 53–66

Background Current models are trained without consideration for the subjective perspectives and cultural influences of those who provided the image descriptions (Ye et al., 2023). This raises the need for carefully curated sources of data and annotation paradigms that are more culturally aware and inclusive (Arora et al., 2023; Cao et al., 2023). Lately, there has been a growing body of work releasing multicultural multimodal datasets for visiolinguistic reasoning (Liu et al., 2021), text to image generation (Liu et al., 2023b; Ventura et al., 2023), and image captioning (Thapliyal et al., 2022). Beyond the focus on the multilingualism of the captions, concurrent work also addresses the cultural concepts depicted in the images (Cao et al., 2024; Burda-Lassen et al., 2024; et al., 2024a; Mukherjee et al., 2024; Bhatia et al., 2024). However, they still do not take into account specific use cases, such as visual assistance. Gurari et al. (2020) released the first image-captioning dataset with photos from people who are blind, and a series of challenges for multimodal systems across different tasks (Gurari et al., 2018). After this initiative, there have been many works trying to improve current models for a specific use-case, to assist people with visual disabilities (Dognin et al., 2022; Ahsan et al., 2021; Delloul and Larabi, 2023). There has also been research in human-computer interaction (HCI) and accessibility on designing image descriptions for visually impaired individuals, primarily focusing on screen readers and functional descriptions of online, publicly available images (Morris et al., 2018; Bennett et al., 2021; Schaadhardt et al., 2021). Despite these efforts, there still seems to be a lack of focus on image captioning for the visually impaired (Ghandi et al., 2023), especially in multi-cultural settings.

2 Methodology

We first created a survey seeking to understand the preferences of visually impaired individuals for image captions, focusing on the inclusion of cultural information and the desired level of detail (see Appendix A). We aggregate the participants' assessments of the helpfulness and importance of cultural information in Figure 1.

We then focused on two lines of contribution: **(1)** We filtered the VizWiz dataset for implicit cultural concepts. VizWiz is a widely used visual question answering and image captioning dataset representing a real-world use case, where examples consist of images and questions submitted by people who are blind, together with crowdsourced answers and image captions (Gurari et al., 2020). The selection of this dataset serves two main purposes. Firstly, it is a challenging dataset specifically tailored to real-world challenges faced by people seeking to access visual information. Secondly, VizWiz might contain implicit cultural references that are currently not captured due to the lack of culture-specific captions. **(2)** We evaluated the image captioning performance of state-of-the-art close-sourced and open-sourced models in a culturally diverse setting using our filtered VizWiz dataset. We performed both an automatic scoring of model-generated captions against two sets of annotations using the COCO evaluation package[1] and a human evaluation.

2.1 Data Filtering

To filter the data we hired a total of 165 annotators through the Prolific platform.[2] We first asked participants to specify their country of origin, location, and their cultural background. Then, we asked them to retrieve images from the VizWiz dataset visualizer[3] related to their cultural background, provide the image name, the reason they think the image is culture-related, and their preferred caption from the dataset (VizWiz provides five different image captions per image). We also gave them the option to suggest a better caption that includes cultural aspects. After collecting all the culture-specific candidate images, we proceeded to a second step of verification. In this step, we retained only those images that had received consensus agreement from at least two individuals. We collected a total of 324 images and 648 captions spanning 60 different identified cultures. It should also be noted that more than 96% of the annotators suggested a cultural revision of the original captions. We refer to Appendix B for further information about the annotation guidelines and data filtering approach and results.

2.2 Models and evaluation

We conducted experiments on the image captioning task in the zero-shot setting, in which a pretrained model is queried to produce a textual description for an image without finetuning on the same dataset. We relied on four commonly used open-access

[1] https://github.com/tylin/coco-caption
[2] https://www.prolific.com/
[3] https://vizwiz.cs.colorado.edu/VizWiz_visualization/view_dataset.php

Model	BLEU-4				METEOR				CIDEr				SPICE			
Prompt	*Default*		*Cultural*		*Default*		*Cultural*		*Default*		*Cultural*		*Default*		*Cultural*	
Annotation	*Original*	*Cultural*	*Original*	*Cultural*	*Original*	*Cultural*	*Original*	*Cultural*	*Original*	*Cultural*	*Original*	*Cultural*	*Original*	*Cultural*	*Original*	*Cultural*
BLIP-2	8.0± 0.4	4.8	7.0± 0.4	4.6	12.6± 0.2	10.2	12.3± 0.3	10.3	51.3± 3.2	39.9	44.0± 3.0	36.7	13.8± 0.4	12.5	12.8± 0.5	11.5
InstructBLIP	14.0± 0.5	8.7	14.1± 0.4	9.0	17.3± 0.3	13.2	17.7± 0.3	13.3	77.1± 3.4	60.0	78.8± 3.2	60.2	18.5± 0.4	15.6	18.2± 0.5	14.9
Idefics2	12.0± 0.5	10.1	9.8± 0.5	10.7	18.1± 0.3	15.1	18.9± 0.3	17.1	80.2± 1.9	78.4	74.1± 2.2	78.2	18.0± 0.5	16.7	18.8± 0.2	17.8
LLaVA-1.6	10.0± 0.5	11.4	6.7± 0.3	7.7	18.9± 0.4	17.3	18.4± 0.3	17.0	60.2± 2.3	75.2	40.3± 1.7	56.3	16.3± 0.6	16.5	15.8± 0.5	15.4
Gemini-1.5-Pro	10.8± 0.3	14.1	5.8± 0.1	8.7	20.8± 0.4	21.3	18.2± 0.1	21.0	71.5± 2.1	88.8	14.8± 0.5	34.1	19.6± 0.4	21.6	14.9± 0.3	17.7
GPT-4o	11.9± 0.6	**16.4**	8.1± 0.3	12.2	22.4± 0.4	**23.4**	19.9± 0.3	22.6	66.8± 2.8	**99.8**	40.4± 1.0	72.8	19.1± 0.4	**21.8**	16.6± 0.3	20.1

Table 1: Performance of various VLMs on our filtered VizWiz dataset across captioning prompts (default & culture-specific) and annotations (original & culture-specific). We use 2 reference annotations per image. Since the original VizWiz has 5 annotations per image, we report the mean and standard deviation over all 10 combinations with two references. We underline the best result for each model and display the top result for each metric in **bold**.

models:[4] BLIP-2 6.7B (Li et al., 2023a) with OPT as LLM backbone (Zhang et al., 2022), InstructBLIP 7B (Dai et al., 2023) with Vicuna backbone (Chiang et al., 2023), Idefics2 8B (Laurençon et al., 2024), and LLaVa-1.6 7B (Liu et al., 2023a) with Mistral backbone (Jiang et al., 2023). We also used two state-of-the-art closed-access models: GPT-4o (OpenAI, 2024) and Gemini Pro 1.5 (et al., 2024b). For all of these models, we experimented with two different prompt types including a culture-specific prompt following Shi et al. (2024) and a default captioning prompt taken from Dai et al. (2023). The exact prompts can be found in App D. We evaluated the model-generated captions in two ways: (1) via the COCO evaluation suite and (2) through human evaluation. The COCO evaluation suite was first introduced by (Chen et al., 2015) as a framework to assess image captions using numerous automatic metrics, including BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Denkowski and Lavie, 2014), and SPICE (Anderson et al., 2016). For consistency with our culture-specific re-annotations (two captions per image), we also used two reference captions per image to score models on the original annotations. Since each image has five original captions, we report aggregate results over all ten two-caption combinations. Our human evaluation had two stages. In the first stage, we asked 60 participants to determine if a caption is accurate (on a binary scale) given the corresponding image. In the second stage, we asked the same participants to rank all captions (human-generated, and model-generated) according to their preference. We did not make the annotators aware that one caption was model-generated to minimize bias. We provide further details on the human evaluation in Appendix F.

[4]We used implementations and model weights from HuggingFace (Wolf et al., 2020).

3 Results

Automatic evaluation We present the results of our automatic evaluation of model-generated captions in Table 1. Note that due to using two reference captions per image, results for the original annotations are slightly different than when using all five at once; we report the latter in Appendix E for completeness.

As expected, the closed-access models (Gemini and GPT-4o) score best overall. Slightly lower performance is achieved by the instruction-tuned open-access VLMs (LLaVa, Idefics2, and InstructBLIP). BLIP-2, which has not been instruction-tuned, is lagging behind across all metrics. Since VizWiz is naturally noisy due to the high ratio of low-quality, blurry images, the increased scale and overall multimodal reasoning capabilities of the closed-source models appear to give a significant advantage.

Strikingly, Gemini and GPT-4o achieve much better performance on our newly annotated captions that include cultural information than on the original captions (e.g., 11.9 vs. 16.4 BLEU-4 and 66.8 vs. 99.8 CIDEr for GPT-4o with the default prompt), while we observe the opposite for the open-access models (e.g, 14.0 vs. 8.7 BLEU-4 and 77.1 vs. 60.0 CIDEr for InstructBLIP with the default prompt). One possible explanation is that the closed-source models have been tuned to generate more descriptive captions that are aligned better with human preferences and our cultural caption annotations, whereas the open-access models have been tuned to generate slightly more concise captions that align well with benchmark datasets like COCO Captions. Our new cultural annotations are also guaranteed to not have leaked into the VLMs' training data, thus favoring more objectively capable models such as GPT-4o.

Next, while individual models (Idefics2 and InstructBLIP in particular) seem amenable to cul-

Figure 2: Examples of various images from the filtered VizWiz dataset with the original (🌐) and culture-specific (🌍) annotations, and generated captions from Gemini-1.5-Pro, GPT-4o, InstructBLIP, and LLaVA-1.6 with default (🌐) and culture-specific (🌍) prompting.

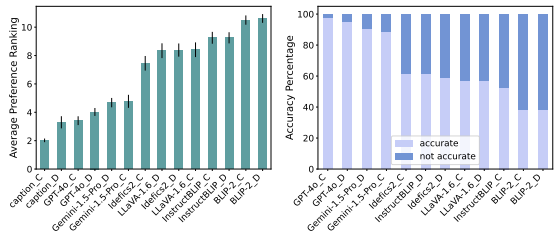


Figure 3: Results of the human evaluation for 100 images and their captions selected at random from the filtered VizWiz dataset. The left plot shows the preference score (participants were asked to rank the captions; lower is better). The right plot shows the accuracy evaluation (participants were asked to assess whether a caption is accurate; higher is better). *'_D'* and *'_C'* denote default and culture-specific prompting, respectively.

tural prompting, leading to improved performance even on the original image captions, the cultural prompting strategy is overall largely ineffective at improving performance on the cultural captions. This result may be due to the models' tendency for sycophantic behavior and them being primed to point out cultural information over other relevant content in the image (Sharma et al., 2023). Alternatively, cultural prompting might elicit more verbose captions that are disfavored by the automatic evaluation metrics, in which case the automatic evaluation results paint an incomplete and potentially misleading picture.

Human evaluation The results of the human evaluation are shown in Figure 3. In line with the automatic metrics, our human annotators tend to prefer the captions produced by closed-access models, GPT-4o and Gemini-Pro, with the BLIP-family models having the lowest ranking. The former are rated as accurate in more than 90% of the cases, while the latter are deemed inaccurate in more than half of the cases. Despite the strong performance of the closed-access models, our preference comparison also shows that the culture-specific human-annotated captions are still preferred over all of the models, suggesting there is ample room for improvement.

In spite of the often stark differences in automatic evaluation scores between cultural and default prompting (with a preference for the latter), human participants prefer the model generations obtained via cultural prompting in 4/6 cases (for both the ranking and the accuracy assessment), supporting our hypothesis that cultural prompting simply elicits an answer format that is disfavored by automatic metrics.

Overall, our results are promising in regard to the reliability of VLMs at zero-shot generating captions that are accurate and useful to users who are blind in culturally diverse scenarios.

4 Further Analysis

To further analyze our results and assess the model-generated captions in a more fine-grained manner, we manually inspected all generated captions for our 324 images filtered VizWiz dataset and provided some examples in Figure 2.

We find that InstructBLIP and BLIP-2 captions tend to be very short, lack a lot of information, and are often irrelevant hallucinations. This is, to an extent, expected as we perform zero-shot captioning, so the models are not necessarily accustomed to the desired captioning style. In this case, few-shot prompting or finetuning the models would likely improve model performance (Brown et al., 2020; Mañas et al., 2023; Ramos et al., 2023). The closed-access models, in contrast, largely provide further or more useful and culture-specific details about the image than given by the human captioners. They also seem to provide more accurate captions compared to the open-sourced models. These points may explain why GPT-4o and Gemini-1.5-Pro and were overall preferred in our human evaluation.

Overall, we observed that the closed-access models can transcribe various language scripts from books, food or beverage packages, giving them an advantage over the smaller models. In most cases, in both culture-specific and default prompts, the models can identify culture-specific beverages like Japanese matcha tea, Chinese jelly grass or lychee juice, and food such as the Indian lijjat papad, Japanese mochi, Tom Kha Gai Thai soup, Korean kimchi, etc. There are also cases where they identify religious or folk items like the Wayang Golek puppets, a jar with traditional Norwegian costumes, or a delft plaque with traditional Dutch costumes.

There is, however, a tendency to generate longer text in the culture-specific prompts by adding generic phrases such as '*hinting at the drink's cultural origin*', '*suggesting a celebration of their shared heritage*', '*highlighting its appeal across age groups in Indian culture*', etc. The most challenging cases for the closed-source models seem to be foreign currencies (especially the Arabic ones), historic figures, and paintings. For example, models seem to confuse Bahraini, Jordan, and Egyptian banknotes, and they do not recognize the Chinese historical figure of Sun Yat-sen, or paintings of Joan Miró or Frederick Morgan. We provide further examples in Appendix G.

5 Discussion

Given the current integration of VLMs as virtual assistants for people who seek sighted support, their performance on culture-specific image captioning seems promising. Examples from our error analysis and case studies highlight some remaining challenges. Measured by automatic evaluation metrics, the performance of the models is overall relatively low compared to results in existing studies evaluating (finetuned) VLMs for image captioning on the full VizWiz and other datasets (Gurari et al., 2020; Chen et al., 2023; Wang et al., 2022). On the other hand, our human evaluation and error analysis show that the generated captions by Gemini-1.5-Pro and GPT-4o are accurate and preferred in many cases. There also seems to be an extended hallucination problem, which remains an existing major challenge not only for VLMs (Li et al., 2023b) but across various language model applications (Bang et al., 2023; Ji et al., 2023).

6 Conclusion

We evaluated the cultural performance of various models on image captioning using a multicultural dataset tailored to a real-world use case. Although the performance of state-of-the-art closed-source models is promising, there is plenty room for improvement. Examples from our error analysis provide insights into the models' performance, helping us identify some of their weak spots. In our use case, we find that automatic evaluation metrics might not be fully representative of model performance, and therefore encourage researchers to reconsider a more comprehensive assessment framework. For future work, we aim to extend our small filtered cultural dataset by including question-answering tasks with POV cultural questions.

Limitations

Our work focuses primarily on data curation and empirical analysis of large multimodal language models. Our survey, while aimed at determining caption preferences, may not capture the full range of needs and preferences of all people with visual impairment. Further, through our analysis, we gained insights into some weak spots with respect to what cultures and cultural concepts are well recognized by the models. However, since we use a finite amount of data, there might be a data bias in identifying particular cultures or cultural

concepts as problematic. Lastly, cultural complexities and variations make it difficult to develop a standardized approach to cultural inclusion in AI. We do, however, hope that our culture-centric approach in the data filtering and annotation process can serve as an initial step towards evaluating and understanding the cultural awareness and abilities of vision-language models for real-world uses.

Ethics Statement

The motivation behind this study is that large vision-language models have rapidly become mainstream and are used even by those who seek sighted support and cannot easily assess model hallucinations or inaccuracies. The primary purpose of our experiments is to assess the performance of vision-language models in the task of image captioning using a multicultural dataset of images taken from people who are blind. However, it is crucial to recognize that results from our current filtered dataset may not be representative of model performance across cultures. Furthermore, our refined dataset might retain biases present in the original source dataset.

We find it improbable that our experiments and the filtered dataset will meaningfully benefit those intending to create deceptive models for malicious purposes. Additionally, the VizWiz dataset may lack coverage of highly specific subjects, offering only a general overview of factual topics. People who intend to use our resources, however, should state their purpose of usage and be accountable for their own work.

References

Hiba Ahsan, Daivat Bhatt, Kaivan Shah, and Nikita Bhalla. 2021. Multi-modal image captioning for the visually impaired. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 53–60, Online. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. "it's complicated": Negotiating accessibility and (mis)representation in image descriptions of race, gender, and disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

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. *arXiv preprint*.

Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2117–2126, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. 2024. Exploring visual culture awareness in GPT-4V: A comprehensive probing. *arXiv preprint*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing

cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. blog post.

Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. 2020. Assessing image quality issues for real-world problems. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3643–3653.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Khadidja Delloul and Slimane Larabi. 2023. Towards real time egocentric segment captioning for the blind and visually impaired in rgb-d theatre images. *arXiv preprint*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2022. Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge. *J. Artif. Int. Res.*, 73.

David Romero et al. 2024a. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint*.

Gemini Team et al. 2024b. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.*, 56(3).

Ricardo Gonzalez, Jazmin Collins, Shiri Azenkot, and Cynthia Bennett. 2024. Investigating use cases of ai-powered scene description applications for blind and low vision people. *arXiv preprint*.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision – ECCV 2020*, pages 417–434, Cham. Springer International Publishing.

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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint*.

Ju Yeon Jung, Tom Steinberger, Junbeom Kim, and Mark S. Ackerman. 2022. "so what? what's that to do with me?" expectations of people with visual impairments for image descriptions in their personal photo activities. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, DIS '22, page 1893–1906, New York, NY, USA. Association for Computing Machinery.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

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, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. 2023b. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (ccub) dataset. *arXiv preprint*.

Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2023. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2523–2548, Dubrovnik, Croatia. Association for Computational Linguistics.

Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich representations of visual content for screen reader users. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Anjishnu Mukherjee, Ziwei Zhu, and Antonios Anastasopoulos. 2024. Crossroads of continents: Automated artifact extraction for cultural adaptation with large multimodal models. *arXiv preprint*.

OpenAI. 2024. Hello gpt-4o. blog post.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2840–2849.

Anastasia Schaadhardt, Alexis Hiniker, and Jacob O. Wobbrock. 2021. Understanding blind screen-reader users' experiences of digital artboards. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *arXiv preprint*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint*.

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.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2023. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *arXiv preprint*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna. 2023. Cultural and linguistic diversity improves visual representations. *arXiv preprint*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint*.

A Survey on Caption Preferences

We created a survey aiming to understand the preferences of individuals who seek sighted support regarding image captioning. Our interest was particularly focused on whether they prefer image captions to include cultural information, and how detailed they prefer the descriptions to be. We published our survey through the Prolific platform, by choosing 60 participants with an equal gender sample of and representative across countries compensated with 18$ per hour. We also added a screener and selected participants without corrected/normal vision. Overall, the participants were positive regarding the helpfulness and importance of cultural information in the captions with average ratings of 4.1 and 3.9 respectfully.[5] Participants also tended to prefer short captions compared to longer ones. After the anonymity period, we are going to release our survey link and full results.

B VizWiz Data Filtering—Human Annotation

As mentioned in the experimental set-up section, to filter the data we created a survey through the Prolific annotation platform. All annotators were compensated with 18$ per hour. We ran this survey 4 times asking for 40 participants each time.

We asked people to identify images from the VizWiz dataset based on their cultural background, provide an original and a corrected caption, and specify the reason they selected the image as culture-specific. We grouped the reasons that the annotators provided for selecting culture-specific images in Figure 4.

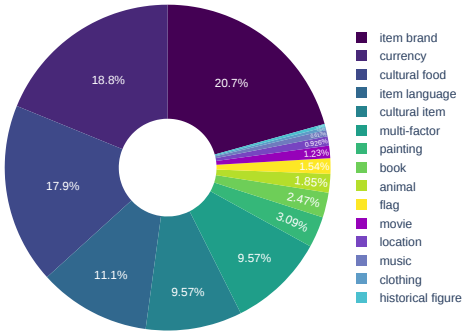


Figure 4: Distribution of the factors/indicators that lead the annotators to select a specific image as culture-related and specify the corresponding culture.

[5]The scale is from 1. Not important/helpful at all to 5. Very important/helpful)

The cultural concepts identified by our annotators can be found in Figure 5.

Figure 5: Distribution of the cultural concepts identified in the VizWiz dataset by the annotators.

The full annotation guidelines were the following:

> Creating datasets that reflect a variety of cultures is a challenging task. This is why we will try to filter an existing dataset. Your task is to find culture-related images from a dataset called VizWiz. You need to:
> - Visit the dataset website[link]. - Browse the dataset or use the search bars on the left side of the page and search key-terms related to your culture 'Within visual question', 'Within visual answer' or 'Within captions'. - Try to find an image that is related to your culture/cultural background (i.e. food brand, currency, books, culture-specific locations etc.) - Provide your answers to the 5 following questions.
>
> 1. Copy and paste the image name (VizWiz_train_**number**.jpg).
> 2. Based on your cultural background, specify what culture you think is the image related to.
> 3. Select a caption for the image from the suggested Image Captions.
> 4. Do you have a better suggestion for the image caption? To guide your caption generation, imagine that you are describing the image to a visually impaired friend. The caption should explain the whole image, including all the main objects, activities, and their relationships, and reflect the culture information of the image.
> 5. Provide a reason as to why the image is culture-specific.

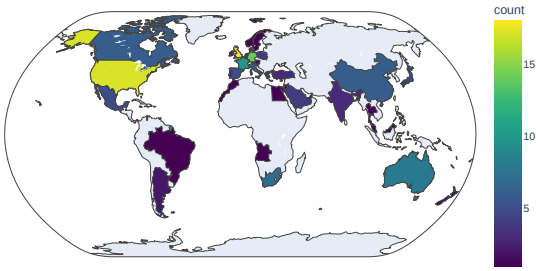
After this, we collected information about the annotators' cultural backgrounds. We asked for both home-country of origin and current country location information since sometimes both can affect our cultural beliefs and practices. The distribution of the annotators counties of origin and location are presented in Figure 6b.

> The last step is to answer some final questions about your cultural background, and age. We do not collect any other personal information. Your answers will only be used for statistical research purposes.
>
> - What is your country of origin that you consider your 'home', influencing your cultural beliefs and other aspects of your identity?
> - Is there a country in which you are currently located for a long period of time?
> - How old are you? Fill in years in numbers.

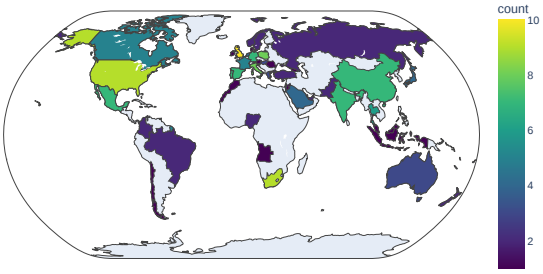
After collecting all the responses, we kept only the images where at least two annotators agreed to select the image as culture-specific. After this extra validation, we resulted in a total of 324 images spanning 60 different identified cultures. We compared the similarity between the suggested captions by the annotators and the original VizWiz captions and the results can be found in Table 2 indicating a high similarity between the new culture-specific suggested captions.

Captions	BLEU-4	ROUGE-L	F1
Culture-specific	37.10	61.90	93.0

Table 2: Results from comparing the culture-specific captions of the two annotators against the five original VizWiz captions.



(a) Distribution of the current country location of the annotators.



(b) Distribution of the country of origin of the annotators.

Figure 6: Plots as subfigures.

C Model Overview

We list models with their API identifiers in Table 3 below.

Name	Identifier	Reference
BLIP-2	`Salesforce/blip2-opt-6.7b`	Li et al. (2023a)
InstructBLIP	`Salesforce/instructblip-vicuna-7b`	Dai et al. (2023)
Idefics2	`HuggingFaceM4/idefics2-8b`	Laurençon et al. (2024)
LLaVA-1.6	`llava-hf/llava-v1.6-mistral-7b-hf`	Liu et al. (2023a)
Gemini-1.5-Pro	`gemini-1.5-pro-preview-0514`	et al. (2024b)
GPT-4o	`gpt-4o-2024-05-13`	OpenAI (2024)

Table 3: Overview of models used in this study

D Model Prompting

We provide the templates we used to prompt our models. The default templates have been sourced from Dai et al. (2023) and Shi et al. (2024).

E Vizwiz Results – 5 Original References

We report model performance on our filtered VizWiz dataset when using all five original captions per image (rather than combinations of two references at a time) in Figure 5.

F Human Evaluation

To conduct the human evaluation of the model generated responses we created a survey and hired 54 annotators through the Prolific platform compensated with 18$ per hour. We added a screening in the platform for a representative sample of countries and an even distribution of male and female participants. Each annotator evaluated 12 images and their captions and for each image, we assigned two annotators and averaged their scores. We provided the following instructions to the annotators for evaluating the captions:

> This study involves evaluating captions. To guide your ratings, imagine that you are describing the image to a visually impaired friend. Then consider:
> How well does the caption describe the image to this friend? Does it take into account cultural considerations? You will be given two sets of captions describing an image.
>
> 1. Specify which caption you prefer for the given image (1, 2 or both).
> 2. Determine if each caption is accurate and relevant to the given image.
>
> As a general guidance you should consider a caption as bad when it has one or more of the following issues:
> a) Caption misses the main topic of the image. b) Caption has major grammatical errors (such as being incomplete, words in the wrong order, etc). Please ignore the capitalization of words and punctuation. c) Caption includes hallucinations and mentions objects, activities, or relationships that are definitely not in the image. d) Caption is not as informative. e) Caption does not reflect the cultural information depicted in the image.

G Error Analysis II

We provide further examples from currency-related images in Figure 7. We can see that for countries such as US, or Australia, the original VizWiz captions provide culture-specific information, but this is not the case for Japanese or Arabic currencies. Moreover, the models seem robust in western and Asian currencies, but not with all the Arabic ones. The example provided in Figure 7 shows how the models confuse a Jordan currency with Egyptian or Saudi Arabian currencies and how the smaller open-source models are more prone to hallucinations.

Default prompting

```
<Image> A short image description:
```

```
<Image> Write a caption that describes the photo.

Format your response in JSON as follows:
{
  "caption": "Caption for the image"
}
```

```
<Image> A photo of
```

```
<Image> Can you briefly describe the content of the image?
```

```
<Image> Write a caption that describes the photo.
```

Culture-specific prompting

```
<Image> A short, culture-aware image description:
```

```
<Image> Cultural information encompasses content that showcases the distinctive characteristics, artifacts, or manifestations of a
↪  specific group, community, or region.
This includes, but is not limited to, practices, behaviors, norms, values, beliefs, habits, customs, architectural styles,
↪  environmental engagements, and any other elements that are emblematic of a particular cultural setting.
It does not include generic information or widespread practices that are not distinctly tied to a specific cultural identity.

For this task, consider information as "cultural" if:
1. It is associated with or characteristic of a specific identified group (e.g., Americans, Italians, midwestern Americans, etc.).
2. It reveals a unique aspect of that group's way of life, including social conventions, physical creations, or interactions with
↪  their surroundings that are not typically seen in other cultures.
3. It provides insight into the cultural uniqueness, whether through social practices, material culture, or other culturally
↪  significant elements.

Please exclude generic or ubiquitous statements or observations that do not clearly relate to the unique cultural context of a
↪  specific group.

Given this image, do two things:
1. Determine whether the provided example contains cultural information.
2. Write a caption that describes the photo and includes the cultural information extracted.

Format your response in JSON as follows:
{
  "caption": "Caption for the image",
  "is_cultural": true/false,
  "justification": "Why or why not the image contains cultural information"
}
```

```
<Image> Write a caption that describes the photo and includes any cultural information present.
```

Table 4: Image captioning templates used to prompt our models.

H Case Study

We illustrate the value of cultural and inclusive VL models via a case study on evaluating GPT-4V as a visual assistant integrated into the 'Be My Eyes' platform. In this case study, we took a random sample of 20 images from the MaRVL dataset (Liu et al., 2021). Here we provide a selection of images we tried in our case study. Each figure includes the target culture behind each image and the GPT-4 Vision output after loading the image in the Be My Eyes application.

Model	BLEU-4		METEOR		CIDEr		SPICE	
Prompt	*Default*	*Cultural*	*Default*	*Cultural*	*Default*	*Cultural*	*Default*	*Cultural*
Annotation	*Original-Full*		*Original-Full*		*Original-Full*		*Original-Full*	
BLIP-2	14.9	12.1	16.1	15.5	51.7	44.3	10.6	9.9
InstructBLIP	**25.3**	24.5	22.0	22.1	77.4	78.9	15.0	15.0
Idefics2	20.8	16.7	22.2	23.3	**82.0**	76.1	15.1	16.6
LLaVA-1.6	17.3	11.8	23.3	22.1	60.9	40.5	15.3	15.3
Gemini-1.5-Pro	18.6	9.9	25.5	21.9	73.0	15.0	18.0	15.3
GPT-4o	20.5	13.8	**27.4**	24.4	67.7	41.0	**18.4**	16.6

Table 5: Performance of various VLMs on our filtered VizWiz dataset across captioning prompts (default & culture-specific) using the five original reference captions per image. We underline the best result for each model and display the top result for each metric in **bold**.

Figure 7: Examples from images related to currency comparing original (🌐) with culture-specific (🌍) annotations and generated captions from Gemini-Pro and GPT-4o with default (🌐) and culture-specific (🌍) prompting.

Culture information: The Agamas (Devanagari: आगम, IAST: *āgama*) (Tamil: ஆகமம், romanized: *ākamam*) are a collection of several Tantric literature and scriptures of Hindu schools.

GPT4: The picture shows a decorative element that appears to be part of a larger structure, possibly a shrine or temple. At the center is a gold-colored figure with multiple arms in a symmetrical pose. The figure is adorned with a detailed halo-like element and is set against a dark, intricately carved background. Below the figure, there are offerings placed in a bowl, which include what looks like a coconut shell and other items wrapped in leaves. The background is a clear blue sky with some blurred tree branches at the top right corner.

Image description provided with Be My Eyes.

Figure 8: A picture extracted from MaRVL depicting a statue of Agamas and the GPT-4V image description provided in BeMyEyes.

Culture information: Chinese Buddhism is a sinicized form of Mahāyāna Buddhism, which draws on the Chinese Buddhist Canon (大藏經, Dàzàngjīng, "Great Storage of Scriptures") as well as numerous Chinese traditions.

GPT4: The picture shows three statues that appear to be of Buddhist deities or figures, each seated on an ornate throne. They are positioned side by side against a backdrop that looks like a temple wall with intricate designs. The central figure is slightly elevated compared to the other two. All figures have a peaceful expression, with eyes closed as if in meditation. The statues are richly decorated with details on their clothing and headpieces. The photo is taken from a slight angle and the image is rotated 90 degrees to the right.
Image description provided with Be My Eyes.

Figure 9: A picture extracted from MaRVL depicting Buddhist statues and the GPT-4V image description provided in BeMyEyes.

Cultural information: Sambar, known as Kuzhambu in Tamil Nadu, is a lentil-based vegetable stew, cooked with pigeon pea and tamarind broth.

GPT4: The picture shows a bowl of soup with a variety of ingredients visible. The soup is in a terracotta bowl with a textured edge. The soup itself appears to be a broth-based one with chunks of different vegetables and possibly meats. Visible ingredients include what looks like red tomato pieces, green herbs or vegetables, and possibly some beans or small pasta. The colors suggest a hearty and possibly spicy dish. The bowl is sitting on a surface with a striped pattern, and the bottom edge of a laptop keyboard is visible in the foreground, indicating the photo was taken from above a desk or table.

Image description provided with Be My Eyes.

Figure 10: A picture extracted from MaRVL depicting sambar, a traditional Tamil dish, and the GPT-4V image description provided in BeMyEyes.

Cultural Information: Doner kebab, also spelled as döner kebab, is a dish of Turkish origin made of meat cooked on a vertical rotisserie

GPT4: The picture shows a person, likely a chef, wearing a white apron and a white hat with a red trim, standing in front of a vertical rotisserie with a large piece of meat cooking on it. The meat is browned on the outside, suggesting it's being roasted. The chef appears to be in a restaurant or a kitchen, as there are other people and what looks like a menu in the background.

Image description provided with Be My Eyes.

Figure 11: A picture extracted from MaRVL depicting a döner, a traditional Turkish dish, and the GPT-4V image description provided in BeMyEyes.