

# WHEN TOM EATS KIMCHI: Evaluating Cultural Bias of Multimodal Large Language Models in Cultural Mixture Contexts

Jun Seong Kim<sup>1,\*</sup>, Kyaw Ye Thu<sup>1,\*</sup>, Javad Ismayilzada<sup>1</sup>, Junyeong Park<sup>1</sup>, Eunsu Kim<sup>1</sup>,  
Huzama Ahmad<sup>2</sup>, Na Min An<sup>2</sup>, James Thorne<sup>2</sup>, Alice Oh<sup>1</sup>

<sup>1</sup>School of Computing KAIST, <sup>2</sup>Graduate School of AI KAIST

## Abstract

In a highly globalized world, it is important for multi-modal large language models (MLLMs) to recognize and respond correctly to mixed-cultural inputs. For example, a model should correctly identify kimchi (Korean food) in an image both when an Asian woman is eating it, as well as an African man is eating it. However, current MLLMs show an over-reliance on the visual features of the person, leading to misclassification of the entities. To examine the robustness of MLLMs to different ethnicity, we introduce MIXCUBE, a cross-cultural bias benchmark, and study elements from five countries and four ethnicities. Our findings reveal that MLLMs achieve both higher accuracy and lower sensitivity to such perturbation for high-resource cultures, but not for low-resource cultures. GPT-4o, the best-performing model overall, shows up to 58% difference in accuracy between the original and perturbed cultural settings in low-resource cultures. Our dataset is publicly available at: <https://huggingface.co/datasets/kyawyeethu/MixCuBe>.

## 1 Introduction

Globalization has brought diverse cultural elements into co-existence within the same time and space. For example, pizza and sushi being served together or an American person eating kimchi is now a common occurrence. Recently, the cultural awareness of multi-modal large language models (MLLMs) has been evaluated using culture-specific (Wang et al., 2024; Baek et al., 2024) and multicultural (Nayak et al., 2024; Liu et al., 2025; Winata et al., 2024; Romero et al., 2024) VQA benchmarks. Also, there are studies such as (Hirota et al., 2022; Howard et al., 2024; Fraser and Kiritchenko, 2024), which examine racial bias in vision models with various approaches, including the

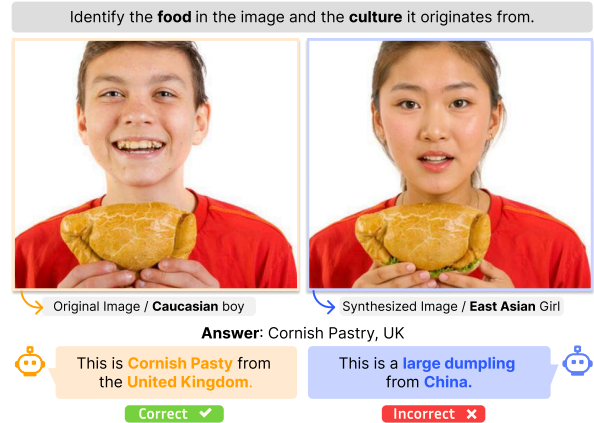


Figure 1: An example of the experiment where a MLLM is tested on both the original image and a synthesized image where the ethnicity of a person is altered.

use of counterfactual images. However, the evaluation of MLLMs’ cultural bias in mixed cultural settings—their ability to recognize certain cultural elements when engaged with people of different ethnicities—remains largely unexplored.

In this study, we examine the cultural bias of MLLMs in the cultural mixture context. Specifically, we focus on cultural markers and people’s ethnic phenotypes as proxies of culture (semantic and demographic proxies as studied in (Adilazuarda et al., 2024)). For instance, while MLLMs may correctly identify kimchi in an image, does that change when the person eating it is of black African background? Specifically, we address the following key research questions:

**RQ 1.** Does replacing the person in an image with a person of a different ethnicity introduce cultural bias in MLLMs?

**RQ 2.** How does this bias differ depending on whether the cultural marker belongs to a low-resource or high-resource culture?

To explore these questions, we introduce MIXCUBE, a **Mixed Culture Benchmark** dataset of 2.5k images of *food*, *festivals*, and *clothing*, labeled

\*Equal contribution.

\*Co-first authors:

09jkim@kaist.ac.kr, kyawyeethu@kaist.ac.kr

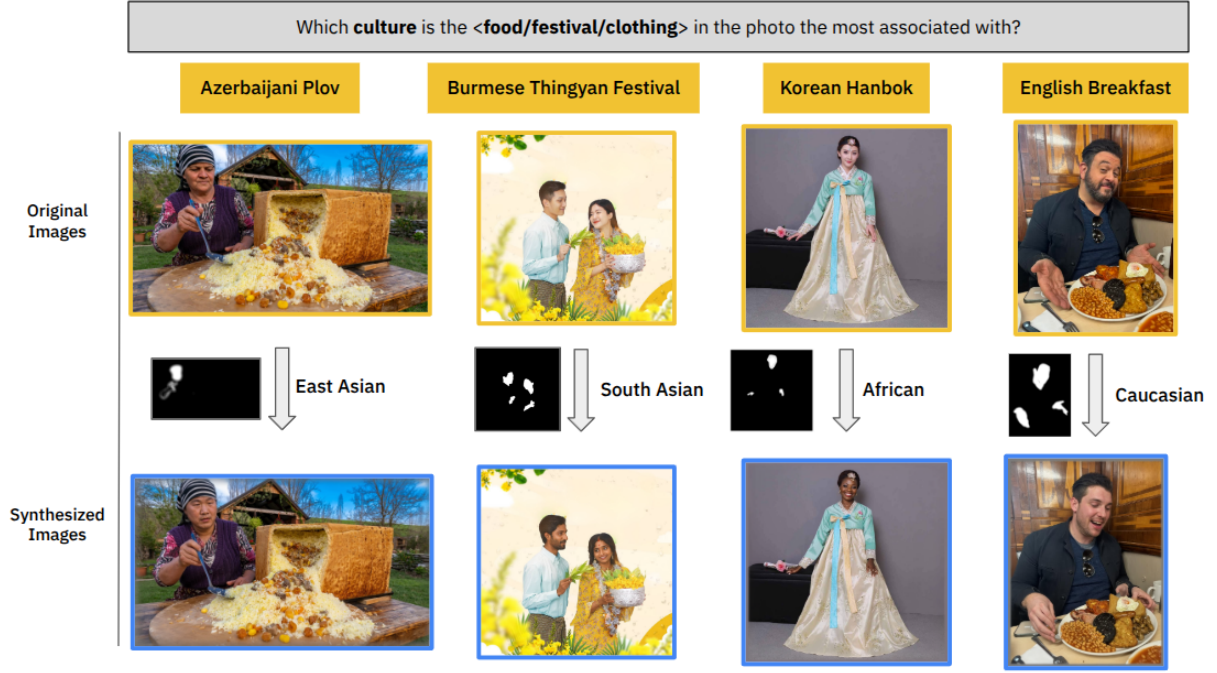


Figure 2: Image synthesis process with sample pairs of original and synthesized images alongside their corresponding masks

with the culture of origin, with food images also labeled with food names. Each image also contains at least one person, and with that original image, we synthesize four additional images in which we replace the person with someone of a different ethnicity (see Fig 1 for an example). We choose four terms to describe broad ethnic phenotypes: *African*, *Caucasian*, *East Asian*, and *South Asian* as they represent geographically diverse populations that can yield distinct phenotypic facial features when inputted, as part of the prompt, into the inpainting model used for synthesis. For original images, we examined five cultures: *Azerbaijan*, *Myanmar*, *South Korea*, *the UK*, and *the US*, representing low-, medium-, and high-resource cultures respectively. Using the dataset, we ask MLLMs to identify the source country and the cultural markers present in each image.

Our results indicate that replacing the person in an image with a person of a different ethnicity degrades MLLM performance, with a larger drop in accuracy for low-resource cultures, Myanmar and Azerbaijan. Models exhibit biases in cultural recognition, showing stable performance across ethnicities for high-resource cultures and large variance for low-resource cultures.

## 2 MIXCUBE: Mixed Culture Benchmark to evaluate cultural bias in MLLMs

MIXCUBE consists of 2.5k labeled images spanning five cultures and three categories of cultural markers (food, festival, clothing). Figure 2 shows the synthesis process of the image set, and Figure 6 illustrates the overall construction pipeline.

**Image Collection.** The seed images are collected using an automatic web scraping tool\* and following a manual web search procedure. During image collection, we followed select criteria, detailed in Appendix A.3, dictating the choice of cultural markers to ensure consistency in the collected data. These criteria aim to reduce misrepresentation of collected cultural data and also ensures variety within each category.

**Image Synthesis.** In preparation for image synthesis, we automatically generate masks of the facial features of each person with the Segment Anything Model (SAM) from Meta (Ravi et al., 2024). Then, we conduct image synthesis via inpainting (Esser et al., 2024), with Stability REST v2 beta API\*. Using the original image, generated mask, and target ethnicity as input, the model generates

\*<https://github.com/ostrolucky/Bulk-Bing-Image-downloader>

\*<https://platform.stability.ai/docs/api-reference>

a synthetic image that replaces the human subject with another of a target ethnicity while closely resembling the original. Image synthesis entails replacing the human subject in each original image with another individual with phenotypic traits that align with the prompts used to guide the synthesis, which are provided in Appendix B.1.

**Quality Assurance.** All generated images were vetted by automated flagging and manual human evaluation to minimize artifacts and misrepresentation of a culture. As an automated filter, we use the combination of BRISQUE (Mittal et al., 2011) and CLIP similarity (Radford et al., 2021) as detailed in Appendix A.6. After filtering out automatically flagged images, each generated image is manually inspected by a human to ensure that the cultural markers remain visually intact, still representing the culture. In cases where artifacts persisted, adjustments were made by manually modifying the mask or further by substituting the original image-mask pair entirely with one that was more suitable for synthesis.

Additional details such as the composition of the dataset, the masking procedure, and labeling are described in Appendix A.

### 3 Evaluating MLLMs with MIXCUBE

We evaluate the cultural bias of MLLMs through two tasks: Country Identification and Cultural Marker Identification. *Country Identification* is the task of identifying the country of origin or culture of a given cultural marker. *Cultural Marker Identification* is the task of identifying the name of a cultural marker. For this task, we focus only on the Food category as foods are the most diverse and distinguishable in terms of their labels. Other categories, Clothes and Festivals, often lack specific names or have only one widely recognized label, presenting a difficulty for even native annotators to identify.

Accuracy is used to quantify the ability of the MLLMs on both tasks. For *Country Identification*, a model’s output that did not include the exact ground-truth country or culture verbatim was marked as incorrect. For *Cultural Marker Identification* accuracy, a secondary LLM, GLM-4-Plus\*, is used as an evaluator model, given access to ground-truth labels, to assess whether a response sufficiently identified the food.\*

\*<https://bigmodel.cn/dev/howuse/glm-4>

\*Examples of ground-truth labels are provided in Ap-

**Does replacing the person in an image with a person of a different ethnicity introduce cultural bias in MLLMs?** Country Identification accuracy for original images are generally higher than those of synthesized ones as apparent in Figure 3 by an average of 7.64% across all models. However, synthesized ethnicities that closely resemble the demographic of the original culture typically perform better than other ethnicities, achieving an accuracy drop from the original images of just 2.04% and even occasionally matching or outperforming the original images. For example, images synthesized with East Asian ethnicity demonstrate minimal accuracy drops for Korea and Myanmar compared to alterations into other ethnicities. Similarly, UK images synthesized with Caucasian ethnicity show low sensitivity to alterations, achieving accuracy levels close to those of the originals. The alignment is expected, given that Korea and Myanmar belong to East and Southeast Asia respectively, where visual changes made by the diffusion model for East Asian subjects are minimal. Likewise, since the majority of the UK’s population is White British, Caucasian synthesis introduces only trivial visual modifications. In contrast, significant accuracy drops are observed when images are altered to African or South Asian ethnicities, where visual differences are, in general, significant for predominant population of Korea, Myanmar, and the UK. These drops are relative to other synthesized ethnicities within the same country and category.

As can be inferred from Figure 4, a common trend in robustness among the three MLLMs is that their accuracy is barely affected by ethnicity alteration in the UK and the US with drops in accuracy less than 15% across all categories and ethnicities (except InternVL-UK-festival case). Also, all three models show significant accuracy drops in Korean Festival and Korean Food for South Asian while being fairly robust in other ethnicities.

Evaluating Myanmar and Azerbaijan, notable sensitivity is observed in GPT-4o and GLM-4v. GPT-4o shows the highest sensitivity (eg. >40% differences are observed in Azerbaijan) although its absolute accuracy is always higher than the other two models. GLM-4v also exhibits sensitivity but in fewer categories and less intensity than GPT-4o does. Although InternVL is the least sensitive overall, its consistency is partly because of its equally underwhelming accuracy (less than 20%) across

pendix A.5.

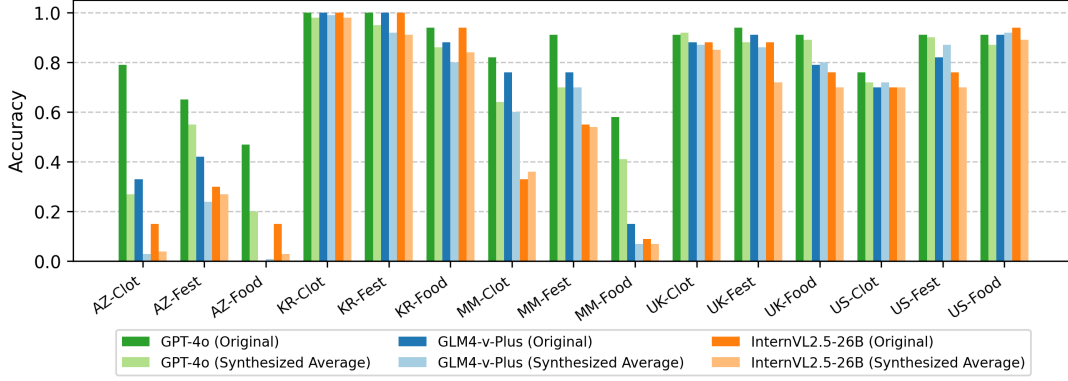


Figure 3: Country Identification accuracy on original images and the average over corresponding synthesized images of four ethnicities (colored in pale) for each country-category pair.

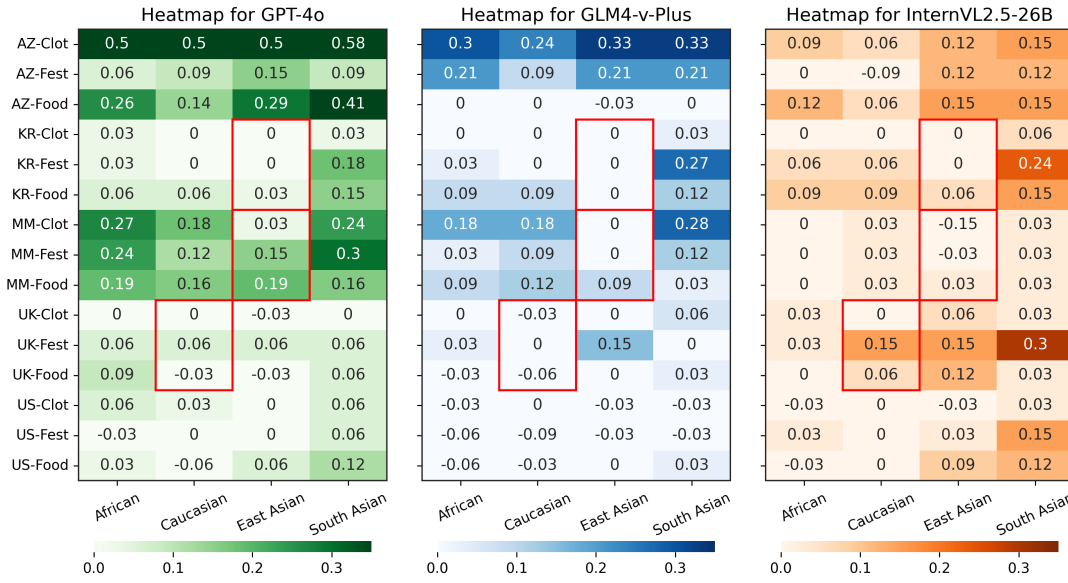


Figure 4: Heatmap of Country Identification accuracy difference. The value in each cell is the difference in Country Identification accuracy between the original and that of synthesized ethnicity. The red boxes highlight the pairs where the synthesized ethnicity by the inpainting model closely resembles a culture demographic.

ethnicities in some categories such as AZ-Food, AZ-Clothes and MM-Food.

Cultural Marker Identification accuracies, shown in Figure 5, exhibit similar sensitivity trends to ethnicity changes. Models like GPT-4o and InternVL drop up to 24% in accuracy for Korean and Azerbaijani food when images are synthesized with South Asian ethnicity. GLM-4v-Plus retains a stable sensitivity across cultures. However, we may still observe for all models that Cultural Marker Identification accuracy values tend to drop for synthesized images, and more so for those that deviate further from the original country’s demographic.

**How does this bias differ depending on whether the cultural marker belongs to a low-resource or**

**high-resource culture?** The Country Identification accuracy across different countries serves as a quantitative measure of the cultural resource levels embedded within various MLLMs. Azerbaijan and Myanmar have consistently lower accuracy, compared to the UK, the US and South Korea, which have accuracy within (80%-100%) in general. This further validates the current literature (Gustafson et al., 2023; Pouget et al., 2024) that vision models tend to possess less robust knowledge of under-represented cultures, highlighting the need to train with more culturally diverse data.

Synthesized images play a crucial role in this analysis by normalizing the distribution of ethnicities across all cultural image sets. This mitigates the unexpected factors introduced by uneven



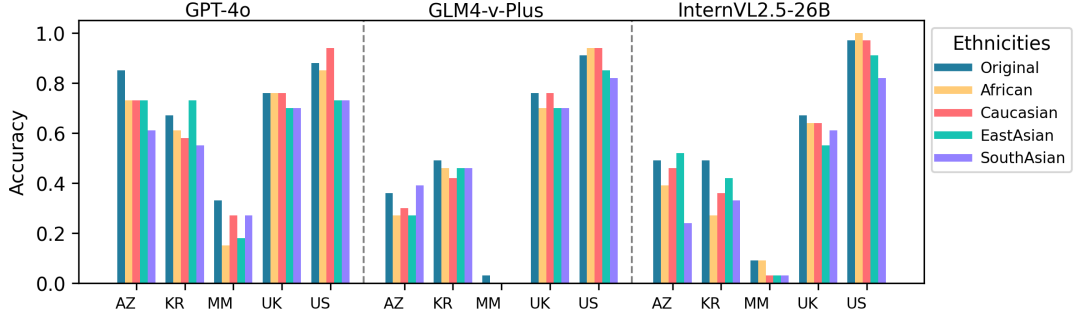


Figure 5: Cultural Marker Identification accuracy evaluated on *Food* images.

dataset representation, ensuring that accuracy differences are primarily attributed to a model’s cultural awareness rather than its familiarity with specific ethnic groups.

#### 4 Discussion

The discrepancy between Food Identification accuracy and Country Identification accuracy in cultures like Azerbaijan and Korea, underscores the MLLMs’ limitation in contextualizing entities into specific cultural frameworks. In Azerbaijan, Food Identification accuracy was significantly higher than the Country Identification accuracy (AZ-food) across all models. This may be attributed to the fact that many Azerbaijani food images in the dataset are visually similar to dishes from neighboring regions or adjacent cultures, such as the Caucasus, the Middle East, and Turkey. Therefore, it can be easier for MLLMs to identify the generic name of the foods (eg. Plov, Kebabs) than to identify the exact country (eg. Azerbaijan) they associated with when the foods are shared by several cultures, albeit with nuanced visual differences. In such cases, models may recognize the food based on its high-level similarities among its variants from similar cultures, rather than its nuanced distinct cultural attributes.

#### 5 Conclusion

In this study, we introduce MIXCUBE to evaluate the robustness of multi-modal large language models (MLLMs) and their cultural awareness and bias with cross-cultural perturbed data across five cultures (Azerbaijan, Myanmar, South Korea, the UK, and the US) and three categories (Food, Festivals, and Clothes). Our results reveal that MLLMs disproportionately favor high-resource cultures while exhibiting both uncertainty and inconsistency in their awareness in underrepresented cultures. Our

findings highlight the need for more diverse, representative data to improve cultural awareness in AI.

#### Limitations and Future Work

The performance drop for synthesized ethnicities may partly stem from minor inpainting artifacts and subtle distortions of cultural markers still persisted despite the quality control, rather than solely from inherent model biases. Furthermore, data contamination — where original images in pretraining datasets inflate Cultural Identification Accuracy — may cause synthesized images to have lower scores due to their novelty. We also acknowledge that the ethnicity alteration that the inpainting model is prompted for is highly generic. For example, ‘South Asian’ encompasses multitudes of ethnicities. Therefore, the synthesized visual appearance is, by no means, intended to be representative of South Asian, but rather a typical sample generated based on the patterns inherently learned by the inpainting model.

Since our study is limited to evaluating three MLLMs on five cultures with four generalized ethnic depictions across three categories of cultural markers, future research will expand along these dimensions — the number of MLLMs, the range of cultures, synthesized ethnic depictions, and categories of cultural markers. By increasing the number of original images and employing multiple inpainting tools to average outputs, technical uncertainties can also be mitigated. This will enable more robust, statistically significant conclusions about changes in model-driven cultural awareness and expand the scope of the analysis.

#### Ethics Statement

All studies in this research project were conducted with the approval of KAIST IRB (KAISTIRB-

2025-37). This study evaluates the robustness of MLLMs in cultural awareness to promote transparency, fairness, and inclusivity in artificial intelligence while carefully considering the ethical implications of altering human features such as ethnicity. Our work focuses exclusively on assessing model robustness and biases without endorsement of stereotypes of cultural misrepresentation, using synthetic alterations solely to uncover dependencies on peripheral attributes and foster greater inclusivity in future models. We acknowledge the potential misuse of our methodologies—such as exploiting synthesized data for discriminatory purposes, and thus advocate for the responsible use of the benchmark and related tools within clearly defined ethical and scientific boundaries.

We acknowledge that our positionality as researchers—including our cultural and social backgrounds—may pose an influence on our approach to assessing bias within MLLMs. We remain committed to transparency within our methodology and strive for objectivity. Additionally, we understand the risks involved in the reinforcement of stereotypes that may arise during the image synthesis stage. To minimize this, our research emphasizes that no culturally connected elements were synthesized, with models instead focused solely on altering the ethnic aspects of each image. Furthermore, the focus of our research is conducted in an effort to quantify the potential reliance of MLLMs on stereotypical markers in an effort to reduce such biases in future models.

## Acknowledgements

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW), supervised by the IITP(Institute of Information & communications Technology Planing & Evaluation) in 2024 (2022-0-01092).

## References

- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling "culture" in llms: A survey](#).
- Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. [Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration](#).
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. [Scaling rectified flow transformers for high-resolution image synthesis](#).
- Kathleen C. Fraser and Svetlana Kiritchenko. 2024. [Examining gender and racial bias in large vision-language models using a novel dataset of parallel images](#).
- Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. 2023. [Pinpointing why object recognition performance degrades across income levels and geographies](#).
- Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. [Gender and racial bias in visual question answering datasets](#). In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 1280–1292. ACM.
- Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandiwala, and Svetlana Kiritchenko. 2024. [Uncovering bias in large vision-language models at scale with counterfactuals](#).
- Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. [Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries](#).
- Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. 2011. [Blind/referenceless image spatial quality evaluator](#). In *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 723–727.
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. [Benchmarking vision language models for cultural understanding](#).
- Anisha Pal, Julia Kruk, Mansi Phute, Manogna Bhattaram, Diyi Yang, Duen Horng Chau, and Judy Hoffman. 2024. [Semi-truths: A large-scale dataset of ai-augmented images for evaluating robustness of ai-generated image detectors](#).
- Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. [No filter: Cultural and socioeconomic diversity in contrastive vision-language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#).

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. [Sam 2: Segment anything in images and videos](#).

Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. 2024. [Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines](#).

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

Yuxuan Wang, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, and Wanxiang Che. 2024. [Cvlue: A new benchmark dataset for chinese vision-language understanding evaluation](#).

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Willie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Santosa, Peerat Limkonchotiwat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie Lee, Shogo

## Appendix

### A Details of Dataset Construction

Our dataset is publicly accessible at <https://huggingface.co/datasets/kyawyethu/MixCuBe>, and it includes original images, synthesized images, and masks. Additionally, a list of labels for food item names have been provided within the dataset. The pipeline illustration for the data set construction is provided in Figure 6.

#### A.1 Reference for Country ISO codes and Category Abbreviations

Throughout the paper, we use the two-letter ISO codes for each country and four-letter abbreviations for each category of cultural marker as follows.

Country/Culture	Abbreviation
Azerbaijan	AZ
South Korea	KR
Myanmar	MM
United Kingdom	UK
United States	US
Category	Abbreviation
Clothes	Clot
Festival	Fest
Food	Food

Table 1: Reference for country ISO codes and abbreviations of categories

#### A.2 Composition of the Dataset

MIXCUBE consists of 2.5k labeled images spanning

- 5 cultures: *Azerbaijan, South Korea, Myanmar, the United Kingdom, and the United States*
- 3 identifying categories of cultural markers: *Food, Festival, and Clothing*
- 4 Synthesized Ethnicities: *African, Caucasian, East Asian and South Asian*

For each category of each culture, we collected 33 original images, which were then synthesized by inpainting to generate four sets of synthesized images. The total data is composed of 2475 images.

#### A.3 Image Collection Criteria

1. Annotators for each set of images within a culture must be native for that culture.

Country / Culture	Original Images	Synthesized Images	Categories	Total Images
AZ	33	$33 \times 4$	$\times 3$	$33 \times 5 \times 3 = 495$
KR	33	$33 \times 4$	$\times 3$	$33 \times 5 \times 3 = 495$
MM	33	$33 \times 4$	$\times 3$	$33 \times 5 \times 3 = 495$
UK	33	$33 \times 4$	$\times 3$	$33 \times 5 \times 3 = 495$
US	33	$33 \times 4$	$\times 3$	$33 \times 5 \times 3 = 495$

Table 2: The Composition of the Dataset

2. The cultural marker(s) in each image must be easily identifiable by native annotators. (The cultural marker should be both visually clear and popular enough among their culture.)
3. When choosing images, cultural overlap must be minimized (e.g. American pizza is avoided because while its nuances are specific to the US, pizza, in general, is a very common food eaten worldwide.)
4. The number of types in a category must be at least a fourth of the total number of images in that category. (e.g. 33 food images  $\rightarrow$  8 different types of food.)

Additionally, we aim to ensure that the ethnic composition of people in the images for each country in the dataset reflects the demographic makeup of that country.

#### A.4 Masking Procedure

Masking was primarily automated using YuNet<sup>\*</sup> from OpenCV to automatically obtain the coordinates of faces as an input to Segment Anything Model (SAM).

For images containing multiple individuals, we limit the number of faces to detect by filtering out the faces with confidence score lower than 0.65 and selecting at most three faces with top confidence if any. This is to ensure only the central and most prominently visible humans are masked, as current inpainting models tend to show a degradation in performance when required to modify multiple subjects simultaneously. Gaussian blur is applied after mask generation to feather the edges of our mask generally helps provide a better inpainting result.

#### A.5 Labels

Multiple acceptable labels for Azerbaijan, the UK and the US were considered as follows to eliminate false negatives in country identification.

<sup>\*</sup>[https://docs.opencv.org/4.x/df/d20/classcv\\_1\\_1FaceDetectorYN.html](https://docs.opencv.org/4.x/df/d20/classcv_1_1FaceDetectorYN.html)



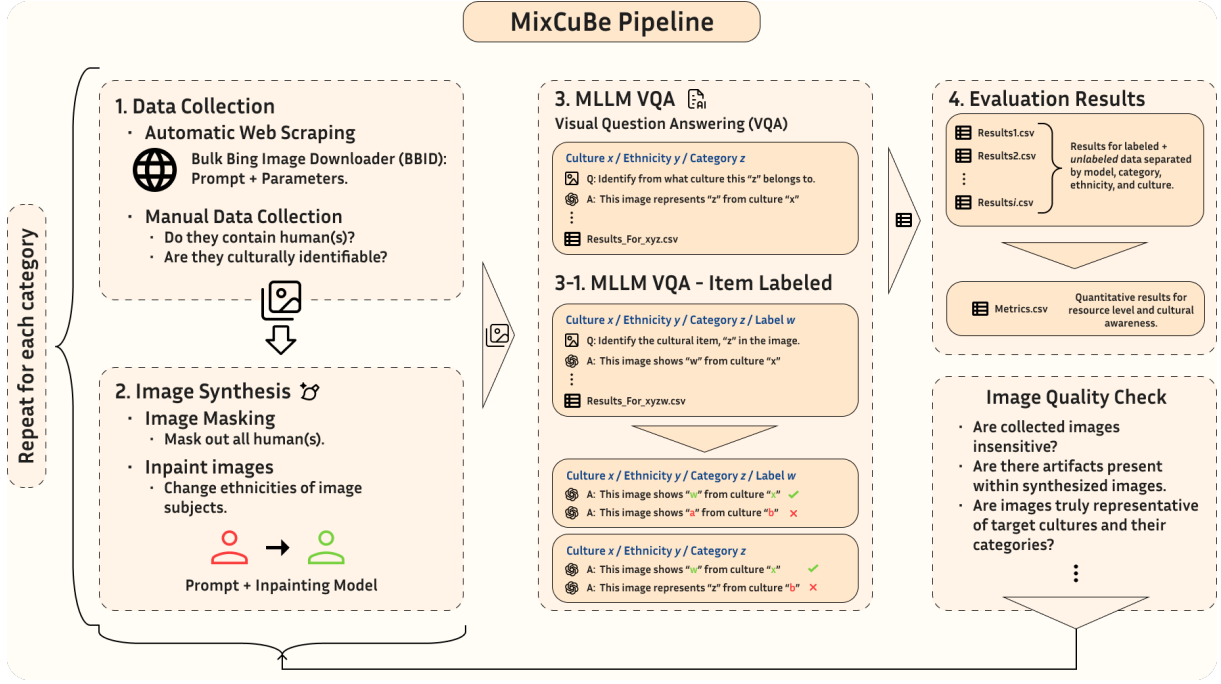


Figure 6: The overall pipeline of the construction of MIXCuBE and the evaluation of cultural awareness

Azerbaijan: "Azerbaijan, Azerbaijani, Azeri"

UK: "UK, United Kingdom, Scotland, Britain, British, Irish, Wales, England, English"

US: "USA, US, the United States of America, the United States, Hawaii, American"

Likewise, we pre-defined multiple acceptable ground-truth labels for food of each *Food* image to aid the evaluator model in its assessment. A label can be either unique or shared across multiple images. One food label for each country is provided as examples in the following.

Azerbaijan: "Azerbaijani Baklava"

Korea: "Jjajangmyeon, Kimchi, Ramyeon, Black Bean Noodles"

Myanmar: "Laphet Thoke, Tea leaf salad"

UK: "Cottage Pie, Shepherds Pie, Shepherd"

US: "Grilled cheese, Toastie"

## A.6 Automated Filtering for Synthesized Images

The automated flagging technique we employed partially follows the procedure of image saliency check from (Pal et al., 2024). BRISQUE, a reference-free metric that quantifies the perceptual quality of an image, is used to detect images

with low structural integrity, indicated by a high BRISQUE score. To ensure that the inpainting model performs enough augmentations on the original while retaining certain resemblance to the original, images beyond the defined range of CLIP similarity are flagged. A synthesized image with low CLIP similarity cannot impose a sufficient visual challenge on MLLMs while extreme visual divergence from the original may potentially distort the cultural marker that the model is supposed to identify. Therefore, images with either BRISQUE score of greater than 80 or CLIP similarity outside the range (65 - 98) are discarded.

## B Prompts

### B.1 Prompts for Diffusion Inpainting Model

For inpainting of images in batch using Stable diffusion, we use a general prompt for each category as follows.

Clothes: "{Ethnicity} person(s) in clothes"

Food: "{Ethnicity} person(s) with food"

Festival: "{Ethnicity} person(s) at an occasion"

The placeholder, {Ethnicity}, is one of African, Caucasian, East Asian, South Asian.

We prompt the model again with a tailored prompt for each image having undesirable result

from the initial batch inpainting. Some examples are

- "South Asian men playing a flute"
- "Caucasian ladies performing a dance"
- "An East Asian couple in Myanmar traditional clothes"

## B.2 Prompts for Evaluating MLLMs

For evaluating MLLMs, we use a dedicated prompt for each category as follows.

Clothes: "Which country is the clothing in the photo the most associated with? Which visual cues did you use to determine it?"

Food: "Which country is the food in the photo the most associated with? Which food is it?"

Festival: "Which country is the celebration/cultural activity/festival in the photo the most associated with? Which visual cues did you use to determine it?"

## B.3 Prompt for Evaluator LLM of Responses from Food Images

For determining whether or not a response from a MLLM sufficiently and correctly identifies the food in an image, we used GLM-4-Plus as an evaluator LLM with the following one-shot prompt.

You'll be provided with a label and a response by a multi-modal LLM that identifies the name of the food in an image. Determine whether the food name contained in the response can be considered as correct given the ground-truth label. under label. Consider it as correct ('Yes') if the names of the food refer to the same food semantically either in native language or in English. Otherwise, 'No'.

- Emphasize on the name instead of the description.

- The names do not need to match exactly.

- If the name provided is wrong, it's 'No' even if the description is close. For example, if the label is "Dote Htoe, Wat Thar Dote Htoe, pork offal skewers, pork skewers" and the response is "This food is mostly associated with Myanmar and is called 'E Kya Kway' or 'Inn Kyaik Kyaw'. It's a popular street food featuring skewers, often with a variety of meats and offal, cooked in a boiling broth.", the answer should be 'No' since the name is completely wrong and the description does not include 'pork'.

- Answer only in 'Yes' or 'No'.

## C Experimental Settings

### C.1 Models

We evaluated the cultural awareness of three MLLMs — GPT-4o (gpt-4o-2024-08-06)\* by OpenAI, GLM-4v (glm-4v-plus)\* by ZhipuAI, and InternVL2.5 (InternVL2.5-26B-AWQ)\* by OpenGVLab.

### C.2 Hyperparameters

The following table provides the values of some key hyperparameters used in the experiments.

Model	Hyperparameter	Value
Stable Diffusion	Diffusion steps	60
	Guidance scale	7.0
GPT-4o, GLM-4v, InternVL2.5	Maximum Token	120
	Temperature	0.3
	Top-p	0.6
GLM-4v	Maximum Token	10
	Temperature	0.2
	Top-p	0.5

Table 3: Hyperparameters used in the experiments.

## D MLLM Evaluation Results

Table 7 and Figure 4 collectively show the results of country identification, presenting, in each cell, the absolute accuracy and the difference in accuracy with respect to that of the original respectively. Likewise, Table 8a and Figure 8b display the results of Cultural Marker Identification.

\*<https://platform.openai.com/docs/models>

\*<https://bigmodel.cn/dev/howuse/glm-4v>

\*[https://huggingface.co/OpenGVLab/InternVL2\\_5-26B-AWQ](https://huggingface.co/OpenGVLab/InternVL2_5-26B-AWQ)

Country/ Culture	Ethnicity	Clothes			Food			Festival		
		GPT-4o	GLM-4v -Plus	InternVL 2.5-26B	GPT-4o	GLM-4v -Plus	InternVL 2.5-26B	GPT-4o	GLM-4v -Plus	InternVL 2.5-26B
Azerbaijan	Original	0.79	0.47	0.65	0.33	0.00	0.42	0.15	0.15	0.30
	African	0.29	0.21	0.59	0.03	0.00	0.21	0.06	0.03	0.30
	Caucasian	0.29	0.33	0.56	0.09	0.00	0.33	0.09	0.09	0.39
	East Asian	0.29	0.18	0.50	0.00	0.03	0.21	0.03	0.00	0.18
	South Asian	0.21	0.06	0.56	0.00	0.00	0.21	0.00	0.00	0.18
	Average (Synthesized)	0.27	0.20	0.55	0.03	0.01	0.24	0.04	0.03	0.27
Korea	Original	1.00	0.94	1.00	1.00	0.88	1.00	1.00	0.94	1.00
	African	0.97	0.88	0.97	1.00	0.79	0.97	1.00	0.85	0.94
	Caucasian	1.00	0.88	1.00	1.00	0.79	1.00	1.00	0.85	0.94
	East Asian	1.00	0.91	1.00	1.00	0.88	1.00	1.00	0.88	1.00
	South Asian	0.97	0.79	0.82	0.97	0.76	0.73	0.94	0.79	0.76
	Average (Synthesized)	0.98	0.86	0.95	0.99	0.80	0.92	0.98	0.84	0.91
Myanmar	Original	0.82	0.58	0.91	0.76	0.15	0.76	0.33	0.09	0.55
	African	0.55	0.39	0.67	0.58	0.06	0.73	0.33	0.09	0.55
	Caucasian	0.64	0.42	0.79	0.58	0.03	0.67	0.30	0.06	0.52
	East Asian	0.79	0.39	0.76	0.76	0.06	0.76	0.48	0.06	0.58
	South Asian	0.58	0.42	0.61	0.48	0.12	0.64	0.30	0.06	0.52
	Average (Synthesized)	0.64	0.41	0.70	0.60	0.07	0.70	0.36	0.07	0.54
UK	Original	0.91	0.91	0.94	0.88	0.79	0.91	0.88	0.76	0.88
	African	0.91	0.82	0.88	0.88	0.82	0.88	0.85	0.76	0.85
	Caucasian	0.91	0.94	0.88	0.91	0.85	0.91	0.88	0.70	0.73
	East Asian	0.94	0.94	0.88	0.88	0.79	0.76	0.82	0.64	0.73
	South Asian	0.91	0.85	0.88	0.82	0.76	0.91	0.85	0.73	0.58
	Average (Synthesized)	0.92	0.89	0.88	0.87	0.80	0.86	0.85	0.70	0.72
US	Original	0.76	0.91	0.91	0.70	0.91	0.82	0.70	0.94	0.76
	African	0.70	0.88	0.94	0.73	0.97	0.88	0.73	0.97	0.73
	Caucasian	0.73	0.97	0.91	0.70	0.94	0.91	0.70	0.94	0.76
	East Asian	0.76	0.85	0.91	0.73	0.91	0.85	0.73	0.85	0.73
	South Asian	0.70	0.79	0.85	0.73	0.88	0.85	0.67	0.82	0.61
	Average (Synthesized)	0.72	0.87	0.90	0.72	0.92	0.87	0.70	0.89	0.70

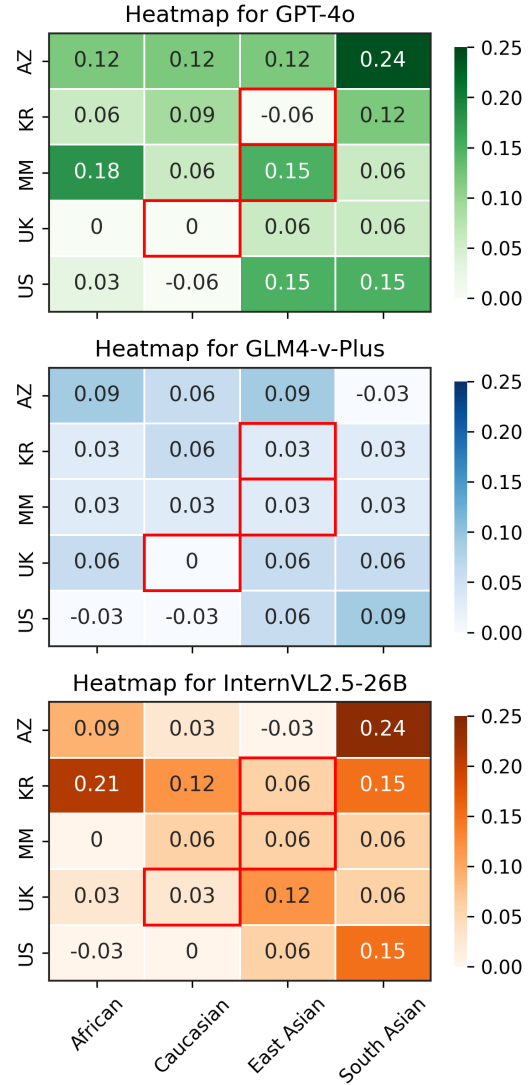
Note: Each cell represents accuracy percentage calculated out of 33 images except cells in the row of Average (Synthesized).

Figure 7: Country Identification Accuracy Data

Country/ Culture	Ethnicity	GPT4-o	GLM4-v-Plus	InternVL2.5-26B
Azerbaijan	Original	0.85	0.36	0.49
	African	0.73	0.27	0.39
	Caucasian	0.73	0.30	0.46
	East Asian	0.73	0.27	0.52
	South Asian	0.61	0.39	0.24
	Average (Synthesized)	0.70	0.31	0.40
Korea	Original	0.67	0.49	0.49
	African	0.61	0.46	0.27
	Caucasian	0.58	0.42	0.36
	East Asian	0.73	0.46	0.42
	South Asian	0.55	0.46	0.33
	Average (Synthesized)	0.61	0.45	0.35
Myanmar	Original	0.33	0.03	0.09
	African	0.15	0.00	0.09
	Caucasian	0.27	0.00	0.03
	East Asian	0.18	0.00	0.03
	South Asian	0.27	0.00	0.03
	Average (Synthesized)	0.22	0.00	0.05
UK	Original	0.76	0.76	0.67
	African	0.76	0.70	0.64
	Caucasian	0.76	0.76	0.64
	East Asian	0.70	0.70	0.55
	South Asian	0.70	0.70	0.61
	Average (Synthesized)	0.73	0.71	0.61
US	Original	0.88	0.91	0.97
	African	0.85	0.94	1.00
	Caucasian	0.94	0.94	0.97
	East Asian	0.73	0.85	0.91
	South Asian	0.73	0.82	0.82
	Average (Synthesized)	0.81	0.89	0.92

Note: Each cell represents accuracy percentage calculated out of 33 images except cells in the row of Average (Synthesized).

(a)



(b)

Figure 8: (a) Cultural Marker Identification Accuracy Data (b) Cultural Marker Identification Accuracy Difference Heatmap. The value in each cell is the difference in Cultural Marker Identification Accuracy between the original and that of synthesized ethnicity. The red boxes highlight the pairs where the synthesized ethnicity by the inpainting model closely resemble to a demographic of the culture.