Uncovering Cultural Representation Disparities in Vision-Language Models

Anonymous EMNLP submission

Abstract

Vision-Language Models (VLMs) have demonstrated impressive capabilities across a range of tasks, yet concerns about their potential biases exist. This work investigates the extent to which prominent VLMs exhibit cultural biases by evaluating their performance on an image-based country identification task at a country level. Utilising the geographically diverse Country211 dataset, we probe several large vision language models (VLMs) under various prompting strategies: open-ended questions, multiple-choice questions (MCQs), including challenging setups like multilingual and adversarial settings. Our analysis aims to uncover disparities in model accuracy across different countries and question formats, providing insights into how training data distribution and evaluation methodologies might influence cultural biases in VLMs. The findings highlight significant variations in performance, suggesting that while VLMs possess considerable visual understanding, they inherit biases from their pre-training data and scale that impact their ability to generalize uniformly across diverse global contexts.

1 Introduction

Vision-Language Models (VLMs) have rapidly advanced, demonstrating exceptional capabilities in integrating visual and textual information for a wide array of tasks, from image captioning to visual question answering (Liu et al., 2024; Alayrac et al., 2022; Wang et al., 2024). These models are increasingly being deployed in diverse applications, impacting areas such as education, healthcare, and public services globally (Zhang et al., 2024).

However, as their influence grows, so do concerns regarding their potential to perpetuate and even amplify societal biases present in their training data (Zhao et al., 2017; Zhou et al., 2022; Weng et al., 2024). Cultural and geographical biases are of particular concern because they can lead to unequal performance and representation across different populations and regions of the world (AlKhamissi et al., 2024; Manvi et al., 2024). Defining "culture" is inherently complex, encompassing a broad spectrum of social norms, values, practices, languages, and historical contexts that shape the lived experiences of individuals and communities (Kroeber et al., 1985). Establishing culture in computational settings presents a persistent challenge due to its multifaceted and dynamic nature. Empirical studies employ tractable proxies such as demographic or geographic proxies to enable systematic analysis (Adilazuarda et al., 2024; Yadav et al., 2025). While nation-level aggregation can mask sub-national heterogeneity, prior work in human–computer interaction and cultural analytics has demonstrated that country labels often serve as a practical proxy for coarse-grained cultural signals when large-scale analyses are required (Obradovich et al., 2022).

In order to quantify cultural disparities in VLMs, we adopt image-based country identification as a concrete proxy task in which a model must infer an image's country of origin solely from visual cues, while also providing a justification. Prior work has shown that geolocation tasks reveal representational imbalances in visual models, as performance often correlates with the prevalence of training data from different regions (Pouget et al., 2024). Using a dataset, with a balanced distribution of images from each country, we ensure that the observed accuracy gaps stem from model bias rather than dataset imbalance. Further, the complex nature of images used could measure VLMs' ability to distinguish similar cultures/countries.

The main contributions of this paper are:

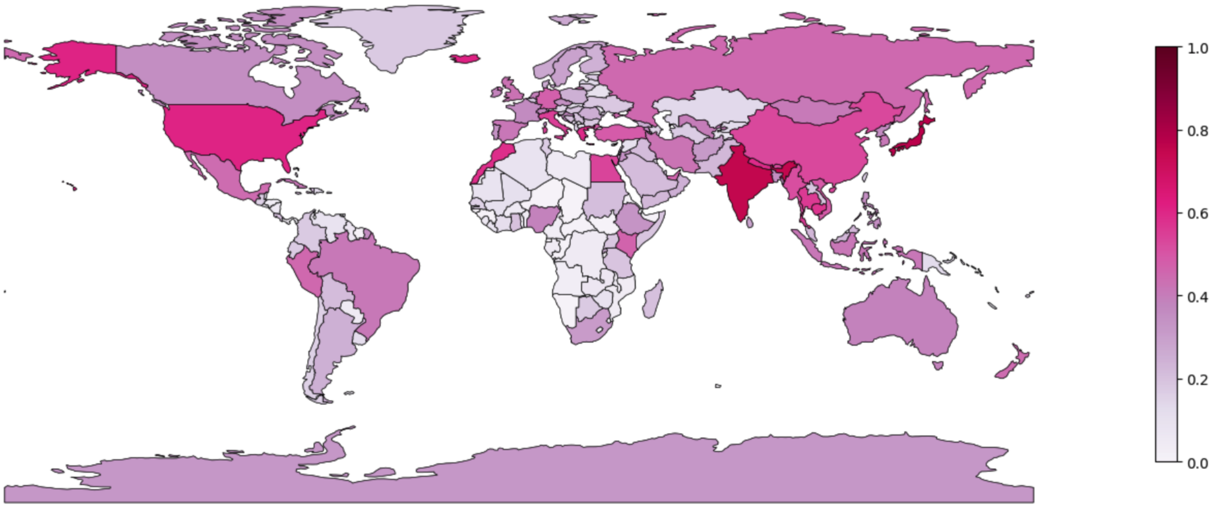

Figure 1: Visualization of the average country-wise recognition accuracy across the VLMs studied in this paper. VLMs perform well at recognizing images from North American and Western European countries, but there are clear disparities in performance for African and Central American countries.

1. We introduce a scalable framework to evaluate cultural biases in VLMs using an image-based country identification task over 211 countries, leveraging the geographically diverse and balanced Country211 dataset.
2. We systematically probe VLMs under varied settings—open-ended and multiple-choice questions (MCQs) with both random and culturally similar distractors—alongside multilingual prompts in five languages, to capture nuanced cultural and linguistic disparities.
3. We examine model robustness to image perturbations and analyse performance across nine image categories (e.g. architecture, landscape, food etc), revealing the influence of image content on cultural bias.
4. Our findings show that VLM biases do not consistently favour Western countries; instead, biases often reflect overrepresentation of certain popular countries (e.g., India, USA) in the training data, suggesting a more complex bias landscape.

2 Related Works

Recent work has increasingly explored the socio-cultural dimensions of Large Language Models (LLMs), including how they encode, express, and respond to culturally specific knowledge. Studies have examined value alignment (Choenni and Shutova, 2024), moral reasoning across languages (Agarwal et al., 2024), and cultural persona (AlKhamissi et al., 2024), while also uncovering strong Western biases in model outputs (Naous et al., 2024) which risk marginalizing cultural diversity if deployed in real world. There have also been efforts to address these concerns, like prompting based on ethnographic fieldwork (AlKhamissi et al., 2024) and fine-tuning culture-specific LLMs (Li et al., 2024a). Similar studies have been extended for Vision Language Models (VLMs) starting from (Liu et al., 2021) over cultural aspects, but in a weaker capacity (Nwatu et al., 2023) showed that CLIP (Radford et al., 2021) struggled in data for poor socio-economic groups worldwide in the Dollar Street dataset (Gaviria Rojas et al., 2022). State-of-the-art off-shelf VLMs score much higher on images depicting Western scenes than equivalent East-Asian scenes for every vision task, such as identification, question-answering, and art emotion classification (Ananthram et al., 2025). Similarly, (Liu et al., 2025; Yadav et al., 2025) reveals that VLMs show stronger performance in Western concepts and weaker results in African and Asian contexts. These findings align with the fact that large pretraining corpora are dominated by high-resource languages and regions. Of the samples that can be geo-located in the OpenImages dataset (Kuznetsova et al., 2020), 32% were from only the United States, and 60% came from only six Western countries (Shankar et al., 2017). Such imbalances translate into a "Western bias" in model behavior (de Vries et al., 2019).

Prior Work	Eval Method	Multilingual?	Adversarial?	Categories	Total Sample Count	Domain
CulturalVQA (Nayak et al., 2024)	Open-Ended	No	No	11 Countries	2,328	5 Categories
WorldCuisines (Winata et al., 2025)	Both	Yes (30 languages)	Yes	189 Countries	6,045	Only Food
Food-500 CAP (Ma et al., 2023)	Open-Ended	No	Yes	7 Regions	24,700	Only Food
MOSAIC-1.5k (Burda-Lassen et al., 2025)	Open-Ended	No	No	N/A	1,500	3 Categories
See It From My Perspective (Ananthram et al., 2025)	Open-Ended	Yes (2 languages)	No	2 Regions	38,479	4 Categories
CVQA (Romero et al., 2024)	MCQ	Yes (31 languages)	Yes	39 Countries	5,239	10 Categories
Ours	**Both**	Yes (5 languages)	**Yes**	**211 Countries**	**63,300**	**9 Categories**

Table 1: Overview of prior datasets used in cultural recognition experiments.

Figure 2: Examples of the Country211 dataset, alongside automatically-predicted categories for each image, showcasing the visual diversity of the examples to be classified.

Datasets & Benchmarks : To probe these biases, a growing body of work has constructed specialized datasets and benchmarks with cross-cultural content, such as MOSAIC-1.5k (Burda-Lassen et al., 2025), CULTURAL-VQA (Nayak et al., 2024), and GlobalRG (Bhatia et al., 2024). Many works also opt for probing specific aspects of culture, such as food (Li et al., 2024b), race (tse Huang et al., 2025), art (Mohamed et al., 2024), etc., instead of providing an overall view for bias study. (Winata et al., 2025) introduced WorldCuisines for Food Vision Question Answering and country identification and found that VLMs often fail on adversarially misleading contexts or less-common cuisines. (Ma et al., 2023) introduced the Food-500 CAP dataset and observed that most models exhibited geographical culinary biases. Several studies have also treated country-of-origin or geolocation as a proxy for cultural provenance. WorldCuisines includes a country identification task to reveal failures on uncommon or misleading contexts (Winata et al., 2025), and Food-500 CAP finds systematic mismatches between predicted and actual countries of culinary images (Ma et al., 2023). Even in datasets like Dollar Street (Gaviria Rojas et al., 2022) or OpenImages (Kuznetsova et al., 2020), geographic metadata has been used to analyze representational imbalances across regions (Nwatu et al., 2023; Shankar et al., 2017), demonstrating that country-level annotations provide a practical signal for probing cultural and geographic bias in VLMs.

Impact of Evaluation: The format of evaluation also impacts bias measurement. Many of the above benchmarks use multiple-choice or binary questions, which can mask a model's true understanding. Since language choice can influence bias, benchmarks are often performed across multiple languages. (Romero et al., 2024) showed that the performance of LLaVA-1.5-7B dropped by 19.6% when prompted without multiple choices for CVQA. Models also showed lower performance when prompted in native language of the image's country of origin. However, (Ananthram et al., 2025) observed that prompting in a culturally closer language can reduce Western bias in some VLMs. It was also observed that people of different cultures are capable of differently capable of describing what they see in an image (van Miltenburg et al., 2017). We build on these insights by comparing open-ended vs. multiple-choice prompts (including "hard" questions with challenging distractors) and by evaluating in both English and native languages, to see how the prompting strategy affects cultural bias in VLMs.

3 Dataset Used

The primary dataset used for the experiments is the Country211 (Radford et al., 2021) dataset which was a subset of images from YFCC100M (Thomee et al., 2016) having GPS coordinates associated with them. The images cover several domains including but not limited to - exterior architecture, interior architecture, landscape (vegetation, nature, skyview), people's appearance, attires, scripts, texts, posters, etc. The GPS coordinates associated with the images were then used to map them to individual countries. ISO-3166 [1] codes representing each country were used as labels for each image. ISO labels were used for consistency as country names used by the VLMs were not deterministic i.e Britain was also used simultaneously in place of Great Britain or UK or its constituents, proving the list of tags and corresponding country names led to the models responding consistently with no observable difference in performance. For our experiments, we utilized this dataset, which consists of 21.1 K images, i.e 100 images each from 211 countries.

Key Differences: Existing benchmarks highlight cultural blind spots in VLMs, but they generally either cover fewer categories or countries or are restricted to specialized domains. Our work differs by using an image-based country-identification task over 211 countries, providing much broader geographic coverage and adversarial probing. Further, the datasets utilized in the prior works utilize a images that might be easier to classify, including but not limited to close up shots of food items, popular monuments being the primary object in an image etc... The dataset we utilized introduces a lot of noise and randomness in a majority of images as seen in Figure 2. For instance, the examples from Norway, India and Egypt might be easy to classify, but the examples from Afghanistan and Kuwait require grasping certain features and their associated knowledge i.e the headgear pattern of the Kuwait image and how it is different from other countries in the region. Th example from Afghanistan requires noticing the afghan flag, while the appearance of the person in the left may try to mislead the VLM due to an appearance of different ethnicity.

[1] https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

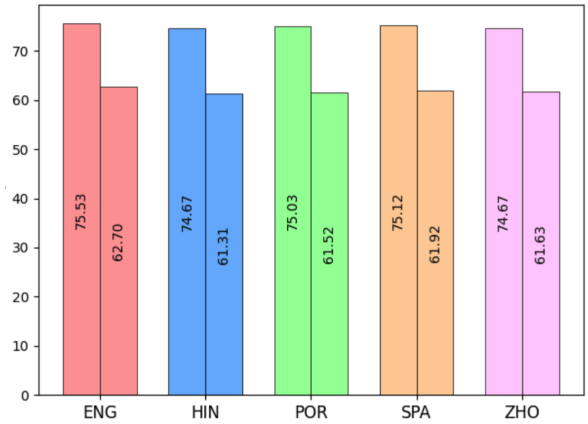

Figure 3: Model-wise averaged accuracy when varying the prompt language or selection of MCQA alternatives (left: random; right: similar). Performance is consistent across conditions.

4 Experiments

Prompt Variations: We probed each VLM under three complementary prompting paradigms.

1. open-ended questions
2. multiple-choice questions (with random distractors)
3. multiple-choice questions (with similar distractors)

Image perturbations: The open-ended experiments were re-done with following adversarial changes:

1. Rotation by 90° clockwise,
2. Rotation by 90° anti-clockwise
3. Flipping the image
4. Gray-scaling the image

However, open ended experiments pose challenges for objective scoring due to semantic variability. Second, multiple-choice questions (MCQs) with random distractors yield correctness metrics yet may understate subtle biases if distractors are easily ruled out. Third, challenging MCQs with similar distractors force models to discriminate between culturally proximate options, thus exposing fine-grained bias patterns. The MCQs are designed as part of discriminative probing and to assess the disparity in the model's cultural knowledge.

Linguistic Variations : We further extend discriminative proving to a multilingual setting, prompting models in five languages : (English, Hindi, Chinese, Portuguese, Spanish) to assess the intersection of cultural and linguistic biases.

Model Variations : A diverse set of VLMs were tested including both proprietary and open-weight models of varying sizes: Gemini-2.5-Flash, Gemma-3-27B (Team et al., 2025), Aya-Vision-8B, Aya-Vision-32B (Dash et al., 2025), GPT-4o-Mini (OpenAI et al., 2024), (etal, 2025).

The experiments being repeated with each permutation of features lead to a total of 168.8 K samples tested. Inference was done in JSON format with the default hyperparameters for each of the models tested through Cohere [2] and OpenRouter's API [3]. More on the JSON formatting and prompts used can be found in Appendix D.

4.1 Open-Ended Evaluation

For the open-ended experiments, we asked each model to provide information on 4 areas: (1) name of the country, (2) country selection rationale in a few sentences, (3) a score from 0 to 100 representing the confidence in the classification, (4) and up to 6 features from the image as a list that had influence in the decision. The accuracies of each country obtained using each of the VLMs used can be seen in Figure 17. The accuracies of many countries were far lower especially in Eastern Europe, South America, Africa and Central Asia. This gap between country level accuracies was far higher in open ended experiments compared to the multiple-choice experiments .

4.2 Evaluation through random distractors in multiple-choice questions

For these experiments, we asked each model to provide information on 4 areas: (1) name of the country, (2) label of the chosen country from the choices provided (3) country selection rationale in a few sentences, and (4) a score from 0 to 100 representing the confidence in the classification. For these experiments, 4 countries were chosen at random from among the other 210 countries for each sample as distractors. The order of options were then shuffled such that the distribution of correct answer's location is made uniform. Compared to other settings, this setting led to the highest average of accuracies obtained due to the clearly contrasting nature of distractors used. However, many central African nations still face a recognition bias likely due to low representation in training data. This was observed across all VLMs that were tested as seen in Figure 18.

[2] https://docs.cohere.com/cohere-documentation
[3] https://openrouter.ai/docs/quickstart

		MCQA	
Region	**Open-Ended**	**Similar**	**Random**
North America	41.9	73.7	80.2
Central America	11.1	69.7	68.0
Caribbean	13.6	50.5	71.4
South America	20.4	70.9	68.7
Oceania	19.0	57.5	68.9
Western Europe	30.9	57.9	77.5
Northern Europe	25.3	60.6	79.4
Eastern Europe	26.6	53.4	75.9
Middle East	29.3	68.4	77.1
Central Asia	26.7	53.5	78.1
East Asia	43.6	71.6	83.8
Southeast Asia	41.7	67.5	81.7
South Asia	49.1	69.0	85.5
North Africa	31.9	54.3	78.9
Central Africa	11.8	57.0	68.2
Southern Africa	20.4	74.2	74.2
Overall	27.7	63.1	76.1

Table 2: Region-wise averaged accuracy across models. There are consistent disparities in performance across different regions, regardless of the prompting method.

4.3 Evaluation through similar distractors in multiple-choice questions

Similar to the prior experiments with MCQs using random distractors, in this setting use similar nations as distractors. These were chosen from among the bordering nations. Any countries with high similarity in culture if any were added manually. (Ex : Spain -> Mexico). This led to the average of accuracies dropping considerably due the challenging nature of the options presented to the models. However, the drops were observed for only a few countries where choosing similar distractors led to these countries' images being classified as belonging to one of their popular neighbors. This can be observed in Figure 18 and Figure 19.

5 Results

The results for experimental setting over countries of each region can be seen in Table 2.

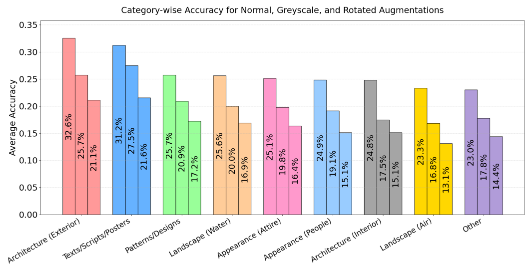

Figure 4: Model-wise averaged accuracy across the nine image categories, as a function of the image perturbations. There is a clear trend of models performing better with the original images (left), compared to the grayscale images (middle) , or rotated images (right).

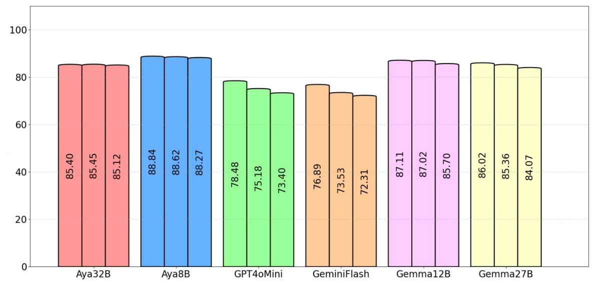

Figure 5: Average model confidence, given the original images (left), grayscale images (middle), and rotated images (right). GPT4o, GeminiFlash, and Gemma27B are most sensitive to image perturbations.

5.1 Effect of Language of Inputs on Results

The average of country level accuracies compared to each language as input can be seen in Figure 3. The language used for inputs had a very little effect i.e <2% for all languages. But at a country level, most countries remained unaffected by language of the prompt to a large extent with change in accuracy less that 0.1%. The only cases with a noticeable change in accuracy are some but not all of the countries that speak the target language predominantly. This can be seen in Figure 3. For example, Changing the input language from English to Spanish improved accuracy for Spain but the change over Latin-American countries was negligible. Similarly, while switching to Portuguese had improved the accuracy for Brazil, it lead to a drop in accuracy for Portugal. Overall, the input language improves performance for some countries primarily associated with the language used. The results also partially contradict prior findings that prompting in culturally similar languages reduces western bias (Ananthram et al., 2025).

5.2 Effect of Image Perturbations on Results

Figure 4 and Figure 5 display the changes in accuracy observed due to gray-scaling and rotating the images compared to the original images. Input image perturbations can have a large impact on the country-level biases in VLMs. Further, It can be assumed that the VLMs tested are not robust enough towards image perturbations, with each country being effected at a different scale between each model/perturbation. The overall averages can also be seen in Figure 8, Figure 9 and Figure 10 respectively.

Figure 18 shows how perturbations affect model performance across different semantic image categories. For all nine categories, models perform best on original (unaltered) images, with decreasing accuracy for gray-scaled and worse for rotated versions. Categories like exterior architecture, text/scripts/posters, and attire/patterns are especially impacted by perturbations. We hypothesize that it is likely because they contain fine-grained, orientation-sensitive, or highly color-dependent details.

We also look at geographical disparities of these changes in orientation in Figure 14 and Figure 15. We observe the disparity in model robustness also emerges clearly. For example, models such as Aya Vision 32B, GPT4o-mini and Gemini 3 12B show very different sensitivity across both a) perturbations and b) regions which were affected. We hypothesise that architectural and training differences might be influencing how models process image orientation and color. While gray-scaling may reduce performance due to the loss of visual detail or color-dependent cues, rotation disrupts spatial reasoning and object orientation, which are critical for geographic or cultural recognition.

These findings highlight the importance of evaluating model performance under realistic image distortions, especially for applications where images may not be clean or consistently formatted as image characteristics can vary widely.

5.3 Effect of Input Variations on Confidence

Despite the drop in Overall accuracy by all of the tested models due to either of the image perturbations, the confidence of the open-weight models didn't have a significant change while the proprietary models displayed a visible drop in confidence compared to the original images. Compared to rotation of images, Gray-scaling had a larger impact on the response accuracies. The average confi-

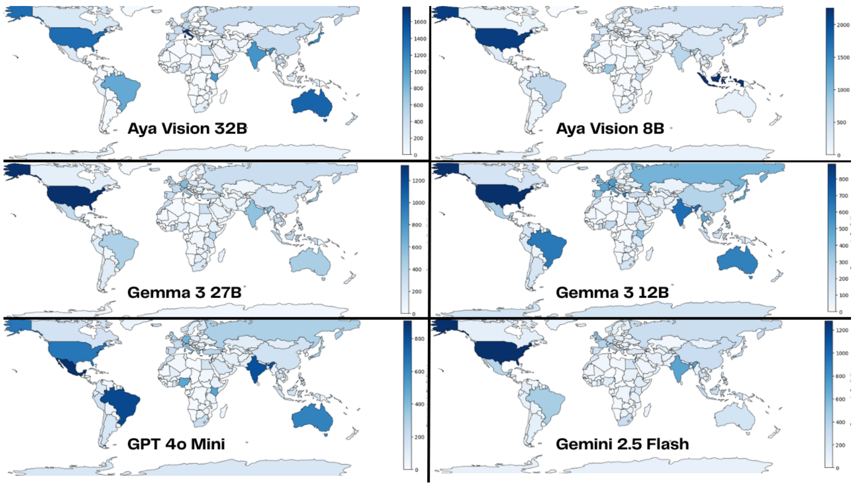

Figure 6: Country-wise response distribution in the open-ended prompt format. There is a consistent trend of models predicting USA, but otherwise, no clear bias towards predicting Western countries.

dence of each VLM with each adversarial setting compared to the original can be seen in Figure 5. The closed-weights models exhibited a drop in confidence when a rotated or grayscale image was provided than the corresponding originals, but this wasn't the case with open-weight models we tested.

5.4 Image Feature categories VS accuracy

Apart from the experiments, the original 21.1 K images were also labeled multi-way based on the key features they contain using larger VLMs like Gemini-2.5-Pro, o4-mini, Grok-2-Vision. Later a majority vote of each label was considered. The quality was later manually verified over a subset by multiple people. We have used 9 sub-categories for this categorization. The descriptions of each of these categories can be seen in Table 3. A large variance was observed between each feature category and the country level accuracies obtained. Additionally there was also a large variation between how accuracy was affected for each country/feature based on model/perturbation used. This can be also be seen in Figure 13. The extent to which each category's images were recognized by VLMs can be seen in Figure 4. External architecture and native language texts' presence in the background helped the VLMs recognize the culture better compared to the other features.

5.5 Distribution of Predicted countries

The distribution of responses in an open ended approach can be seen in Figure 6. The output distributions varied largely among models, even those within the same family (i.e between Gemma-3-27B, Gemma-3-12B and Aya-vision-32B, Aya-vision-8B). The results obtained contradict the usual assumption about western biases in generative models, and was observed over a few nations with likely high training data proportion.

Notably, all models consistently overpredict certain countries—particularly the USA, India, and Brazil—regardless of actual ground truth. We hypothize that these countries are likely overrepresented in the models' pretraining data or benefit from more visually distinctive cues. Biases seem to cluster around a few highly represented or visually salient countries rather than reflecting broader geopolitical landscape.

These results show that model predictions are likely highly influenced by data availability and image characteristics rather than a generic global bias. It also underscores the need for better interpretabilty regarding the geographic composition of VLM training datasets to fully understand such biases.

5.6 Misclassification Analysis

The mapping of misclassification of samples was not limited to similar or neighboring nations. This can be observed in Figure 20 to Figure 34. These misclassifications varied by each individual feature and provide a better fine-grained insights of cultural biases. For instance, Apart from neighboring / similar countries, most images from Africa and rural regions of South America were classified as India. A specific example is shown in Figure 7 where out of the 600 images (100 * 6 models), roughly 80-120 belong to this category for most

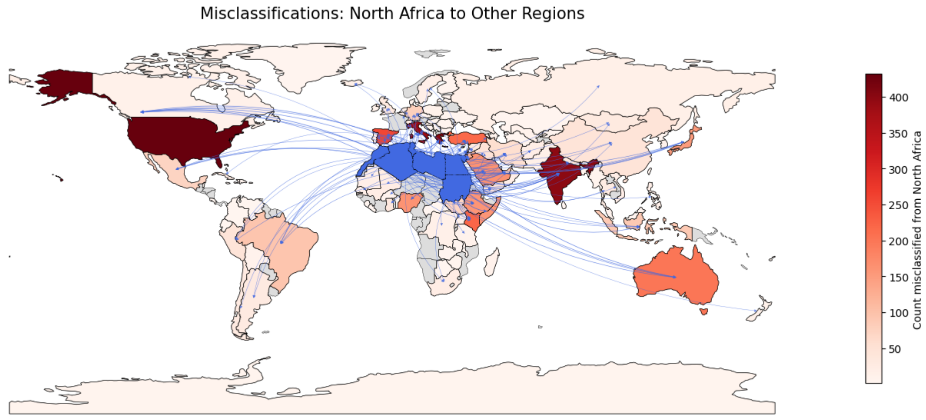

Figure 7: Mis-classification map for North African countries. There is a clear trend of models predicting USA, India, Australia, or geographically close countries in Europe and the Middle East.

countries, while many countries had most of their misclassified as originating from India.

6 Discussion

Our study presents a comprehensive analysis of cultural biases in Vision-Language Models (VLMs) using a geographically balanced dataset across 211 countries. The key difference between the samples in our datasets and those from the other works is the complexity of images provided. In most of the images, we used it is difficult to understand and predict the country of origin without a strong vision ability. For example, the Afghanistan example from Figure 2 is difficult to classify without spotting the hint of Afghan flag from one of the people from the image.

We evaluated popular models across multiple prompting strategies, e.g. open-ended, multiple-choice (random and similar distractors), and multilingual settings. Open-ended formats showed the greatest disparity in country-level accuracy, particularly in underrepresented regions such as Central Africa and parts of South America.The use of culturally similar distractors proved to be the most effective in revealing fine-grained errors, highlighting limitations in models' cultural discrimination abilities.

We further assessed the models' robustness to image perturbations like gray-scaling and rotation. While gray-scaling affected only a few specific countries, rotation led to a broad and uniform drop in performance, confirming that VLMs rely heavily on image orientation. We further observed that performance also varied by semantic image content—categories like architecture, textual cues, and attire were more predictive of cultural origin, especially in unaltered images.

Language variation in prompts had minimal impact on average accuracy, though countries closely tied to the input language (e.g., Spain with Spanish, Brazil with Portuguese) showed slight gains. However, this trend was inconsistent and did not generalize across all culturally linked regions.

Finally, our misclassification analysis shows that models frequently confuse images from low-resource or visually ambiguous countries with a few dominant nations, reinforcing the role of training data bias. These findings emphasize the need for transparent reporting of dataset composition in VLM development and call for more robust, culturally diverse benchmarks to ensure equitable global performance

7 Conclusion

We investigated cultural biases in Vision-Language Models using a diverse country-level image classification task. Our findings show that biases are not uniformly Western but instead reflect overrepresentation of certain countries in training data. Model performance varied across prompt types, languages, image features, and perturbations—highlighting limitations in robustness and cultural generalization.

These results call for greater transparency in dataset composition and the need for more culturally inclusive evaluation methods to ensure fairer and more globally representative VLMs.

Limitations

Our study has a few important limitations to keep in mind. First, the use of country-level labels as a proxy for culture, while common for large-scale analysis, inherently overlooks intra-country cultural diversity and multicultural populations, potentially obscuring sub-national or regional nuances. The country labels used don't account for political complexities like disputed territories.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2025. See it from my perspective: How language affects cultural bias in image understanding. In *The Thirteenth International Conference on Learning Representations*.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.

Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2025. How culturally aware are vision-language models? In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, volume CFP2540Z-ART, pages 1–6.

Rochelle Choenni and Ekaterina Shutova. 2024. Self-alignment: Improving alignment of cultural values in llms via in-context learning.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gallé, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2025. Aya vision: Advancing the frontier of multilingual multimodality.

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Gemini Team ... etal. 2025. Gemini: A family of highly capable multimodal models.

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

A. L. Kroeber, Wayne Untereiner, and Clyde Kluckhohn. 1985. *Culture: A critical review of concepts and definitions*. Vintage Books.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024b. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. 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, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

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.

Zheng Ma, Mianzhi Pan, Wenhan Wu, Kanzhi Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2023. Food-500 cap: A fine-grained food caption benchmark for evaluating vision-language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5674–5685.

Rohin Manvi, Samar Khanna, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34654–34669. PMLR.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Joan Nwatu, Oana Ignat, and Rada Mihalcea. 2023. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10686–10702, Singapore. Association for Computational Linguistics.

Nick Obradovich, Ömer Özak, Ignacio Martín, Ignacio Ortuño-Ortín, Edmond Awad, Manuel Cebrián, Rubén Cuevas, Klaus Desmet, Iyad Rahwan, and Ángel Cuevas. 2022. Expanding the measurement of culture with a sample of two billion humans. *Journal of the Royal Society Interface*, 19(190):20220085.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz

Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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 *Advances in Neural Information Processing Systems*, volume 37, pages 106474–106496. Curran Associates, Inc.

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

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D' Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Naome Etori, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, 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 Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 workshop: Machine Learning for the Developing World*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Jen tse Huang, Jiantong Qin, Jianping Zhang, Youliang Yuan, Wenxuan Wang, and Jieyu Zhao. 2025. Visbias: Measuring explicit and implicit social biases in vision language models.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 121475–121499. Curran Associates, Inc.

Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. 2024. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15669–15680, Miami, Florida, USA. Association for Computational Linguistics.

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 Wilie, 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 Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. 2025. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines.

Srishti Yadav, Zhi Zhang, Daniel Hershcovich, and Ekaterina Shutova. 2025. Beyond words: Exploring cultural value sensitivity in multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7592–7608.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VL-StereoSet: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538, Online only. Association for Computational Linguistics.

A Overall Accuracies Before and After Image Perturbations

Figure 8, Figure 9, Figure 10 display the accuracy obtained for each image perturbation used compared to the original through each of the VLMs tested.

B Overall Accuracy VS Models used : In each MCQ setting

Figure 11, Figure 12 display the accuracy obtained through each model in each MCQ experiment.

C Reproducibility

Inference was done through Cohere's API for Aya-Vision-8B and Aya-Vision-32B through the default hyperparameters with a seed value of 1024. The rest of the models were used through OpenRouter's API through the default hyper-parameters with a seed value of 1024. The experiments were repeated thrice and the overall accuracy varied between 1-1.2%, with some countries' accuracy varying up to 1.5%. The costs associated with all experiments combined were 850$ through OpenRouter and 250$ Cohere API credits. The experiments were run on TPUs costing 0.35$/hr with the costs reaching 60$.

D Prompt Usage

The prompts used for each experiment setting can be seen below.

```
## OPEN ENDED
Respond in the following format :
{
  "Country": {
  "type": "string",
  "description": "Just Country Name, Nothing
      else."
  },
  "Reasoning": {
  "type": "string",
  "description": "Why you feel the image is
      from that particular country. Explain
      what features/objects/items/visuals
      of the image made you think so.
      Should be fine-grained and clear, but
      in just 2-3 sentences."
  },
  "Confidence": {
  "type": "integer",
  "description": "A integer score on a scale
      of 0 to 100 on how confident you are
      about the classification."
  },
  "Features": {
  "type": "string",
  "description": "Which of the features of
      the image were primarily used for
      determining the country labels,
      separated by commas, ordered with
      primary feature in the beginning."
  }
}
```

```
## MCQ
prompt_header="Respond with a label from the
    choices : What country is this image
    likely from ?\n"
for l,c in
    zip(labels,choices):prompt_header+=f"{l})
    : {c}\n"
fixed_prompt=prompt_header+"""Respond in the
    following format :
{
  "Label": "Uppercase Alphabet",
  "Country": "The chosen label's country
      name exactly as it was",
  "Confidence": "Integer between 0 to 100 in
      numeric format"
}"""
```

E Other Plots

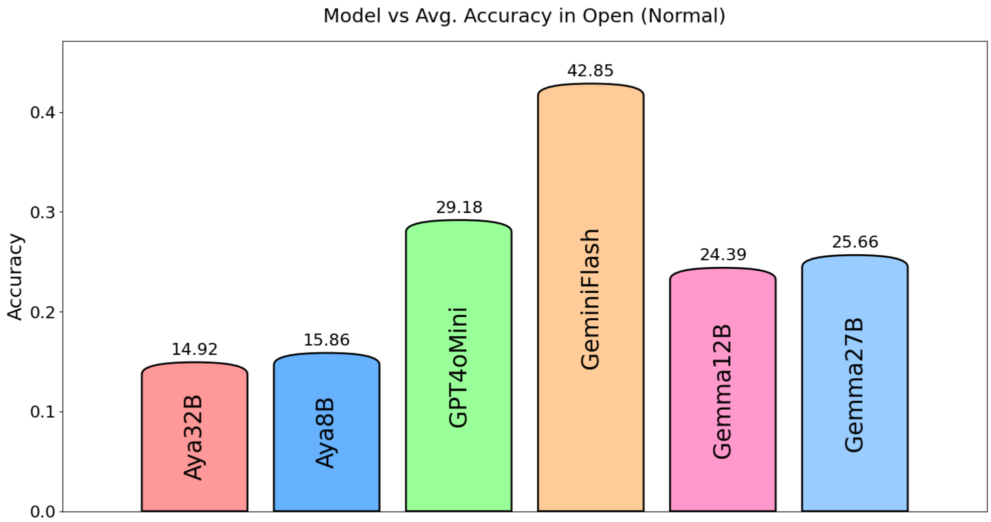

Figure 8: Overall Accuracy : Open Ended (Normal)

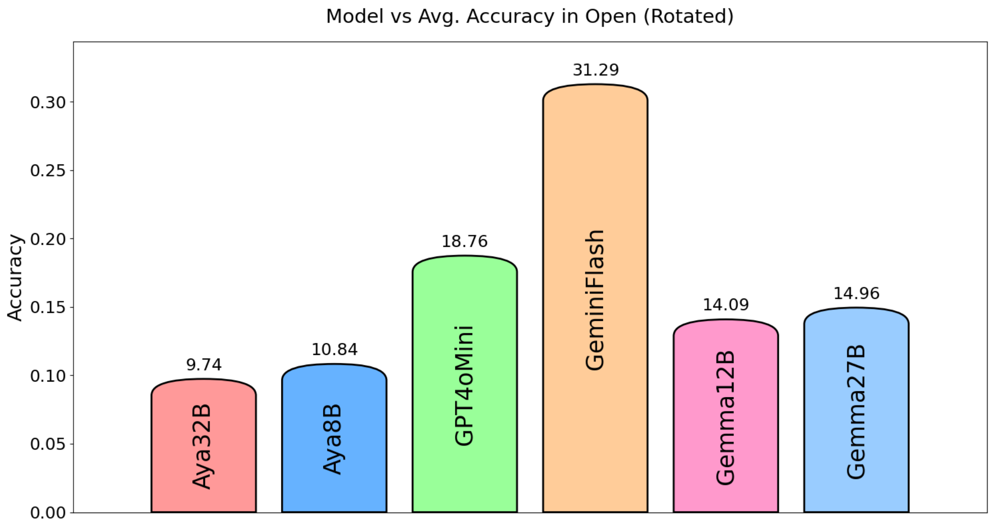

Figure 9: Overall Accuracy : Open Ended (Rotated)

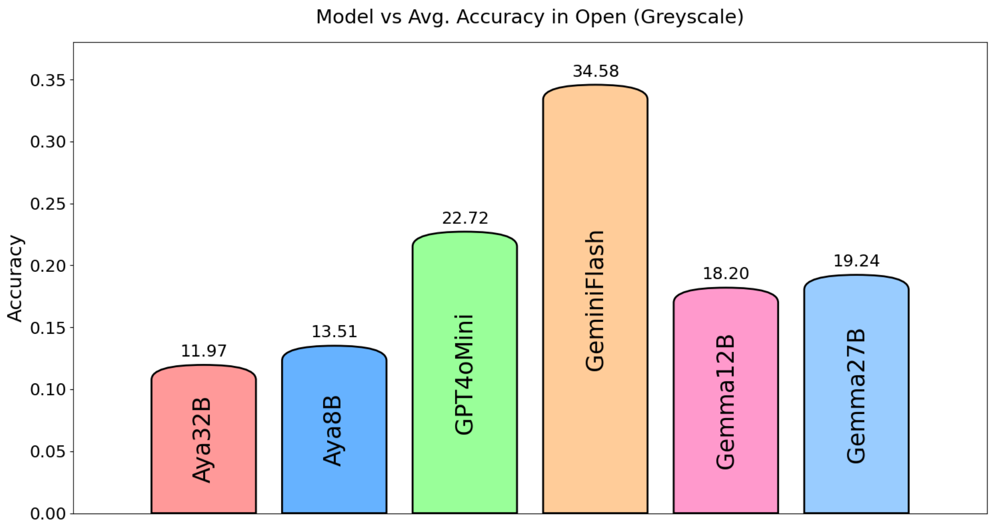

Figure 10: Overall Accuracy : Open Ended (Grayscale)

Category	Description
Appearance (Attire)	Attires of some people from the image, clothes being hanged in the background, etc.
Appearance (People)	Appearance / visual perception of people's ethnicity, presence of any celebrities, etc.
Architecture (Exterior)	Building facades, monuments, bridges, outdoor structures, and any external architectural elements visible in the scene.
Architecture (Interior)	Indoor environments e.g. rooms, corridors, staircases, furniture, and interior design details.
Landscape (Water)	Bodies of water such as oceans, rivers, lakes, waterfalls, ponds, and any aquatic scenery.
Landscape (Air)	Aerial / bird's-eye views, landscapes captured from above, clouds, sky scenes, and horizon vistas.
Landscape (Vegetation)	Forests, grasslands, gardens, crops, shrubs, foliage patterns, plant life, or visible greenery.
Texts/Scripts/Posters	Signs, banners, billboards, labels, handwritten or printed text, posters, and any other written or graphic messaging.
Patterns/Designs	Decorative motifs, surface textures, fabric prints, wallpaper or tile patterns, abstract designs, and repetitive graphical elements.

Table 3: Overview of the image categories used to analyse model performance as a function of the type of image.

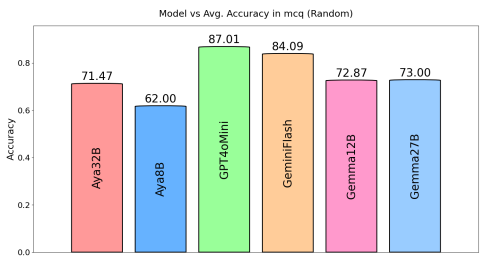

Figure 11: Overall Accuracy : MCQ-Random : Model wise

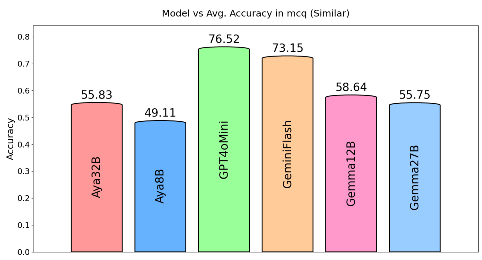

Figure 12: Overall Accuracy : MCQ-Similar : Model wise

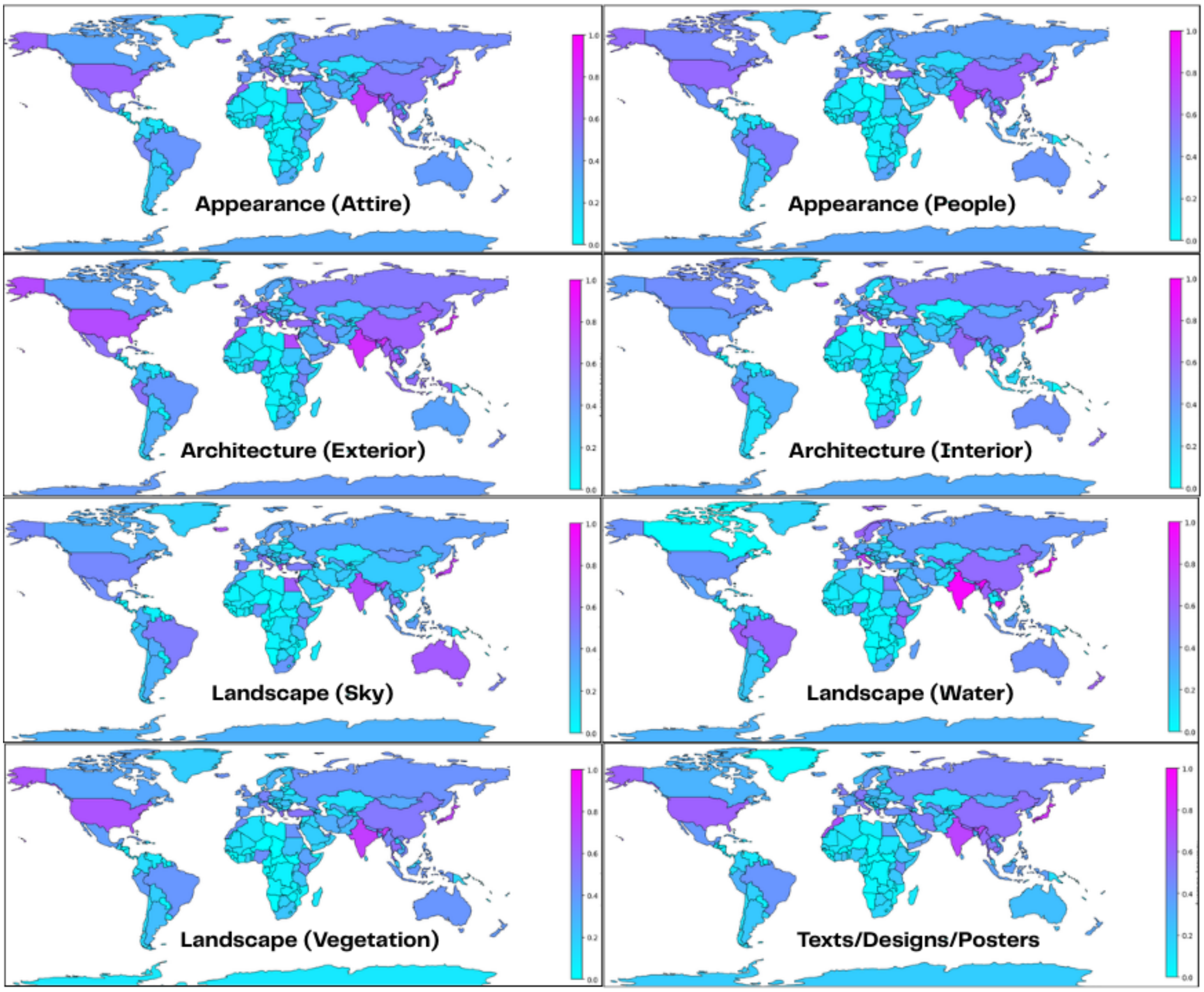

Figure 13: Image Feature categories VS Country wise Accuracy

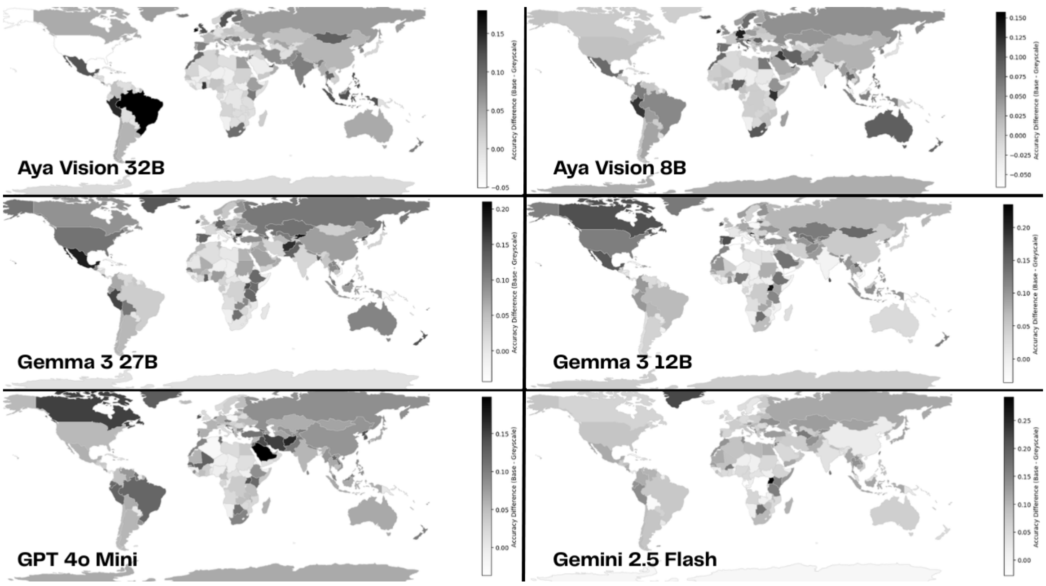

Figure 14: Effect of Gray-scaling VS change in country wise accuracies

Higher Contrast = Larger Drop in accuracy

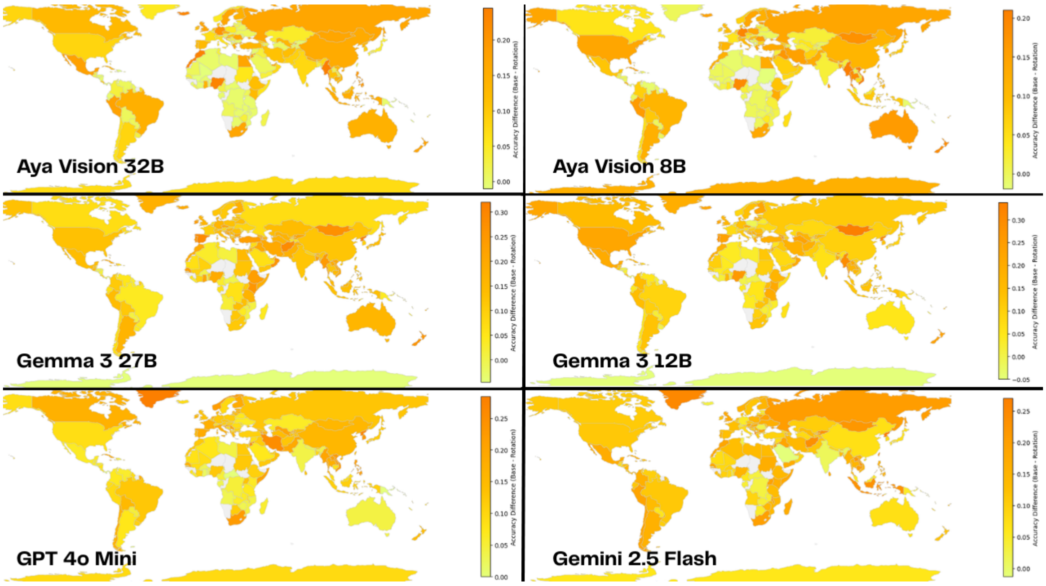

Figure 15: Effect of Rotation VS change in country wise accuracies

Higher Contrast = Larger Drop in accuracy

F Mis-Classification Map : Region-wise

The mis-classifications from one region to countries outside the region can be seen fro each region in Figure 20 to Figure 34 respectively.

G Country wise accuracies in each experimental setting

The accuracies obtained over samples of each country through each experimental setup can be seen in Table 4 to Table 8.

Country name	**Open-Ended**	**MCQs with Similar choices**	**MCQs with Random choices**
Afghanistan	41.33	68.90	81.56
Albania	20.00	42.80	67.64
Algeria	10.50	29.73	65.71
Andorra	12.00	59.63	72.41
Angola	4.67	48.07	58.83
Anguilla	2.00	15.27	58.51
Antarctica	34.83	84.80	83.57
Antigua and Barbuda	7.67	31.67	70.64
Argentina	30.67	84.17	71.39
Armenia	42.33	66.23	80.07
Aruba	17.67	55.67	78.96
Australia	44.50	87.90	69.58
Austria	18.83	42.13	80.69
Azerbaijan	20.00	46.83	66.45
Bahamas	24.83	69.47	78.13
Bahrain	21.00	63.00	73.94
Bangladesh	42.50	59.30	87.48
Barbados	17.67	39.50	72.07
Belarus	13.33	45.60	72.98
Belgium	21.00	44.93	72.21
Belize	11.67	59.13	68.49
Benin	7.50	51.47	78.75
Bermuda	20.67	62.63	67.61
Bhutan	59.17	66.03	90.70
Plurinational State of Bolivia	26.33	76.13	78.26
Bonaire, Sint Eustatius and Saba	3.50	36.47	69.24
Bosnia and Herzegovina	23.33	44.43	73.23
Botswana	22.83	82.13	80.00
Brazil	47.67	83.37	74.70
Brunei Darussalam	8.67	21.73	48.78
Bulgaria	25.33	46.47	77.12
Burkina Faso	7.50	60.83	74.72
Cabo Verde	10.17	67.23	55.22
Cambodia	62.83	81.02	92.15
Cameroon	4.67	67.20	70.02
Canada	41.50	69.43	81.16
Cayman Islands	6.67	28.07	68.78
Central African Republic	0.83	16.67	50.21
Chile	20.83	65.90	67.78
China	58.83	78.73	81.48
Colombia	23.83	75.73	69.25
Democratic Republic of Congo	6.83	40.70	56.60
Cook Islands	3.83	22.23	68.28

Table 4: Country wise accuracies through various experimental settings : Part 1/5

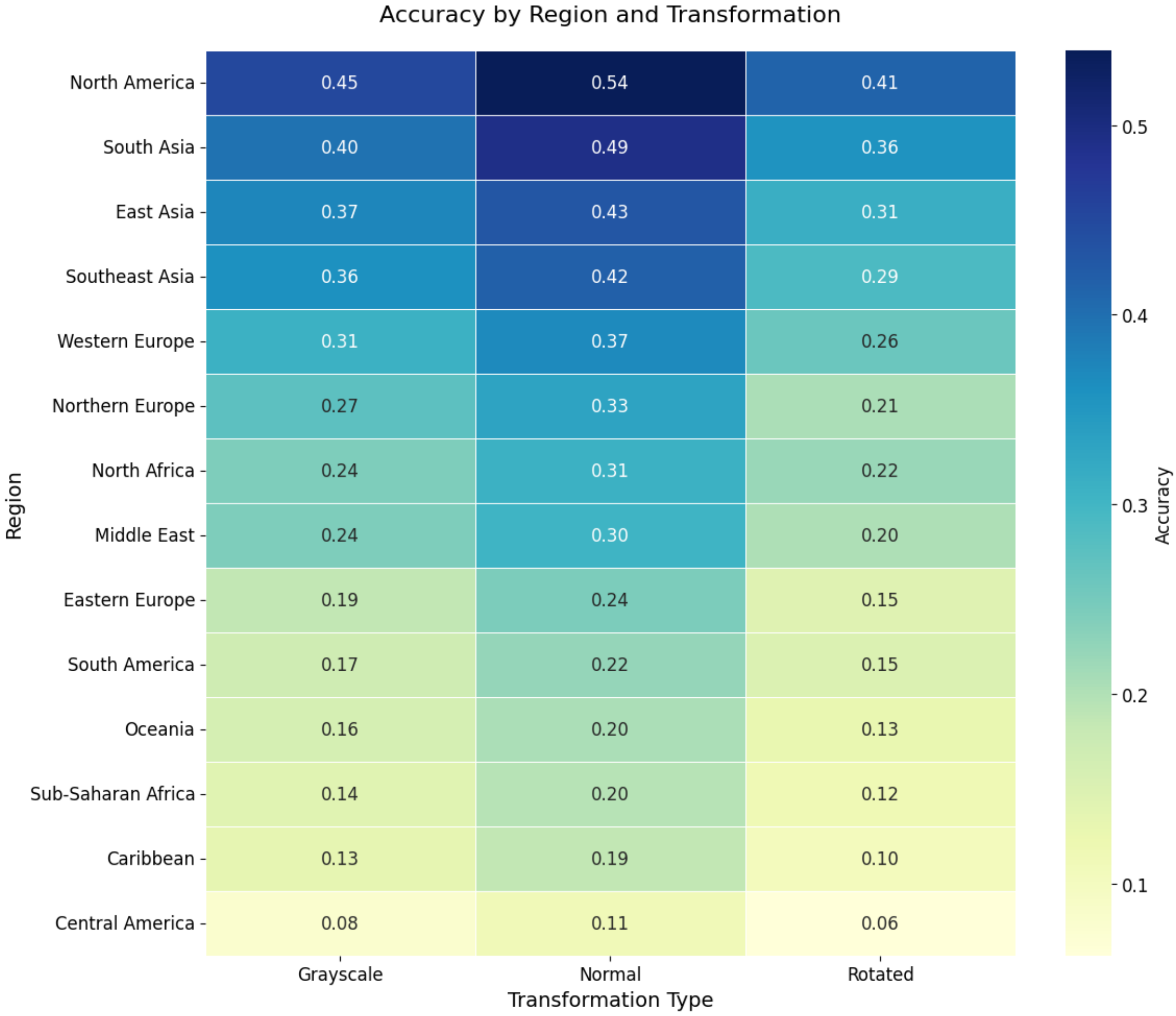

Figure 16: Region wise effect of perturbations

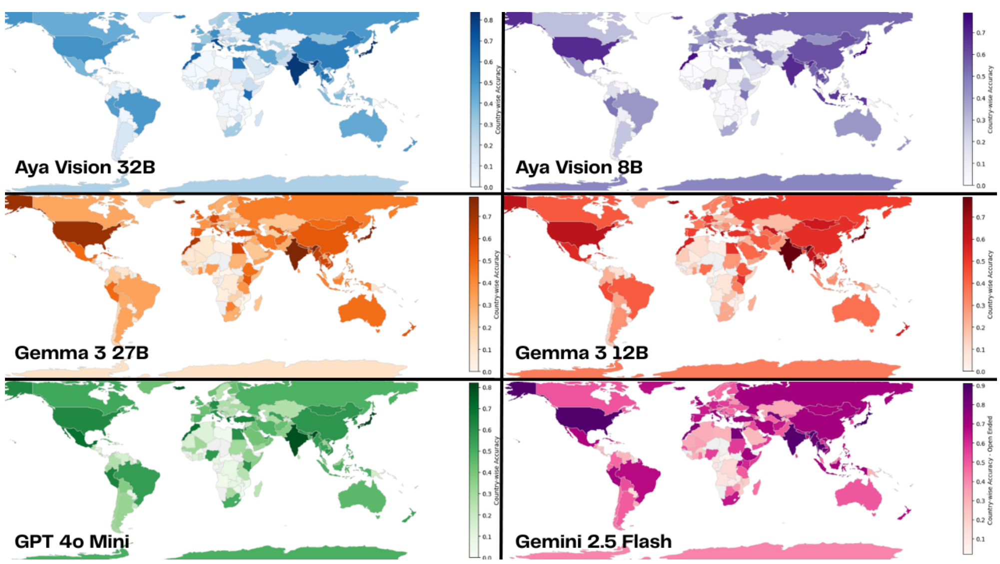

Figure 17: Accuracy over each country's images through open-ended Experiments

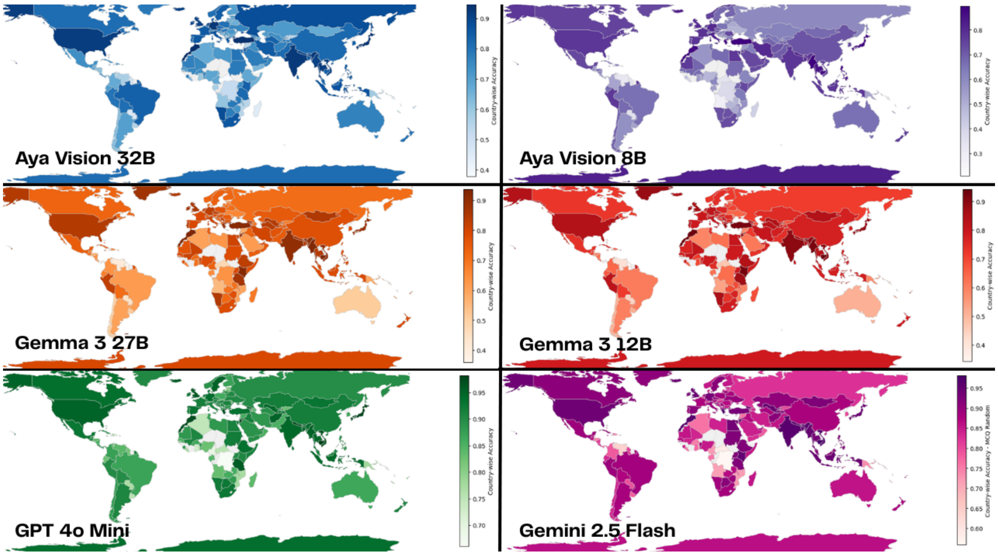

Figure 18: Accuracy over each country's images through MCQ Experiments with random distractors

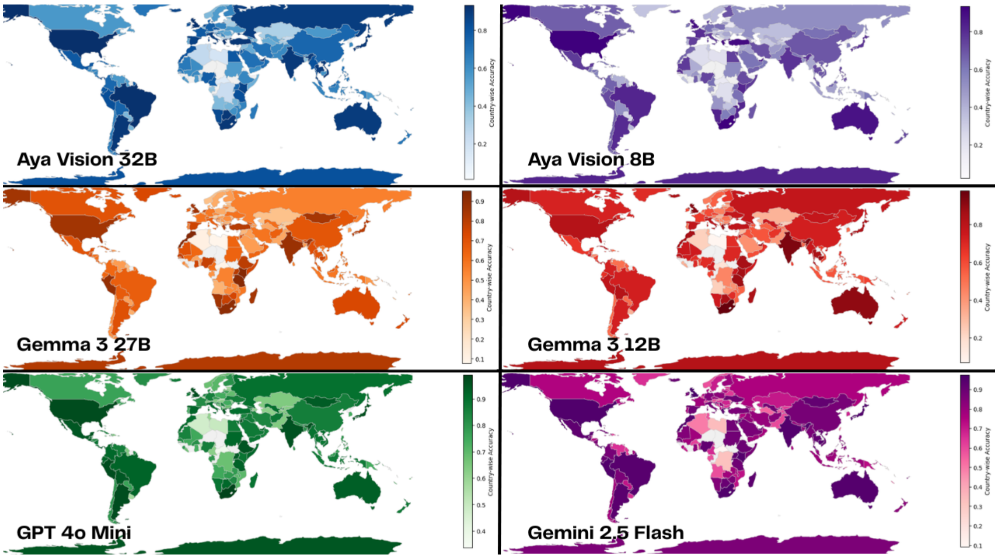

Figure 19: Accuracy over each country's images through MCQ Experiments with similar distractors

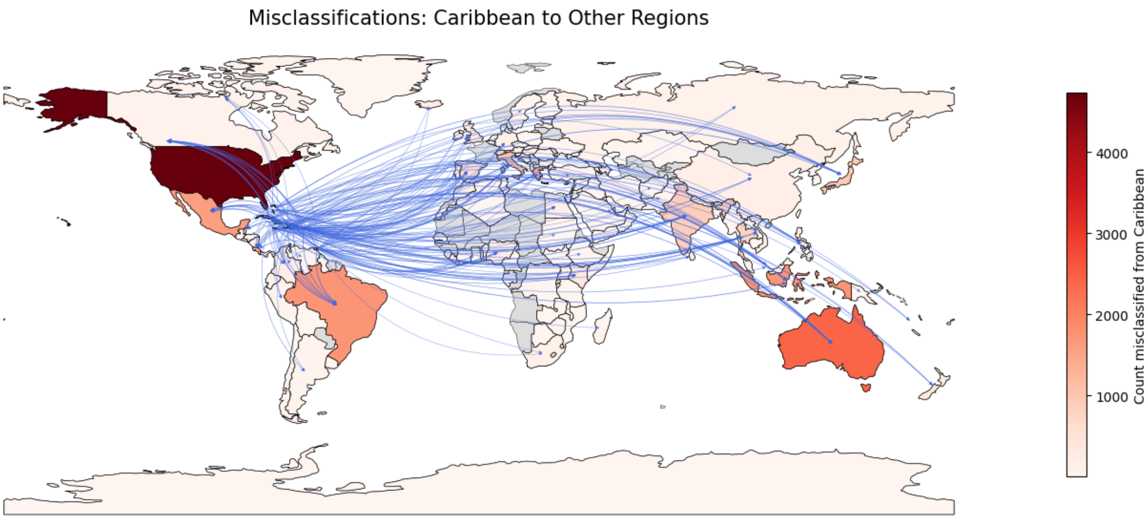

Figure 20: Mis-classification map : Caribbean

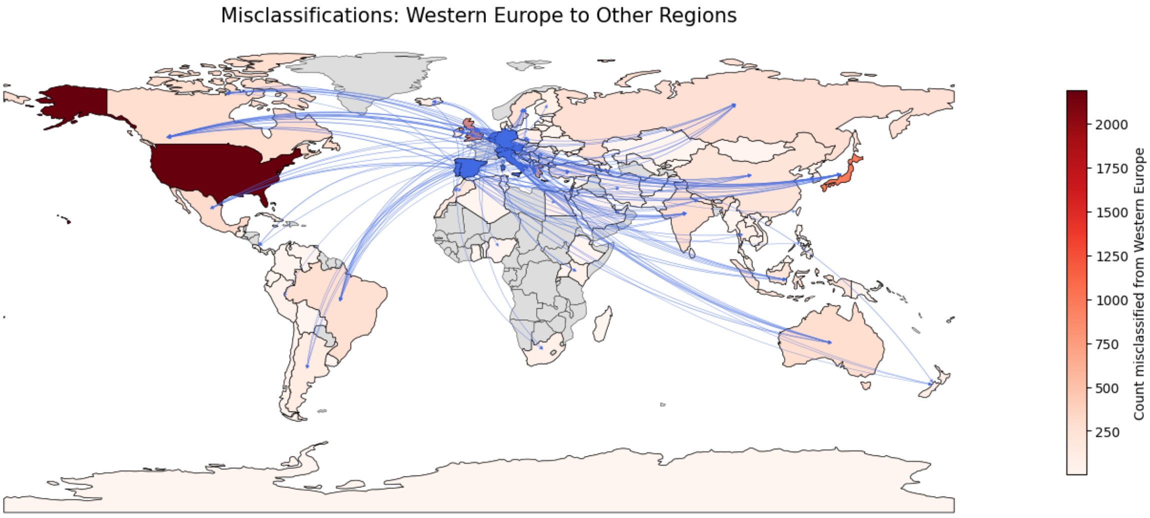

Figure 21: Mis-classification map : Western Europe

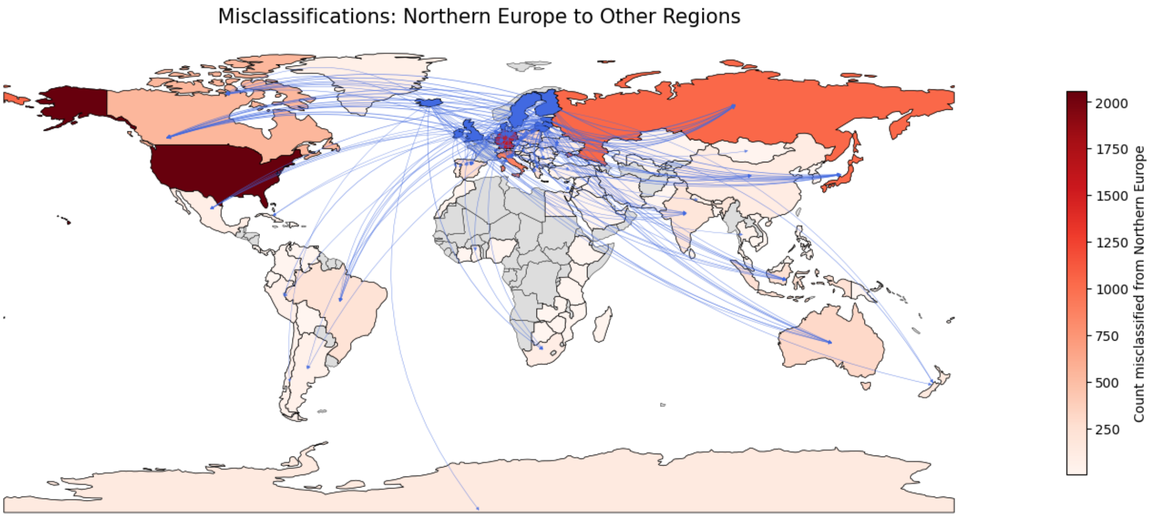

Figure 22: Mis-classification map : North Europe

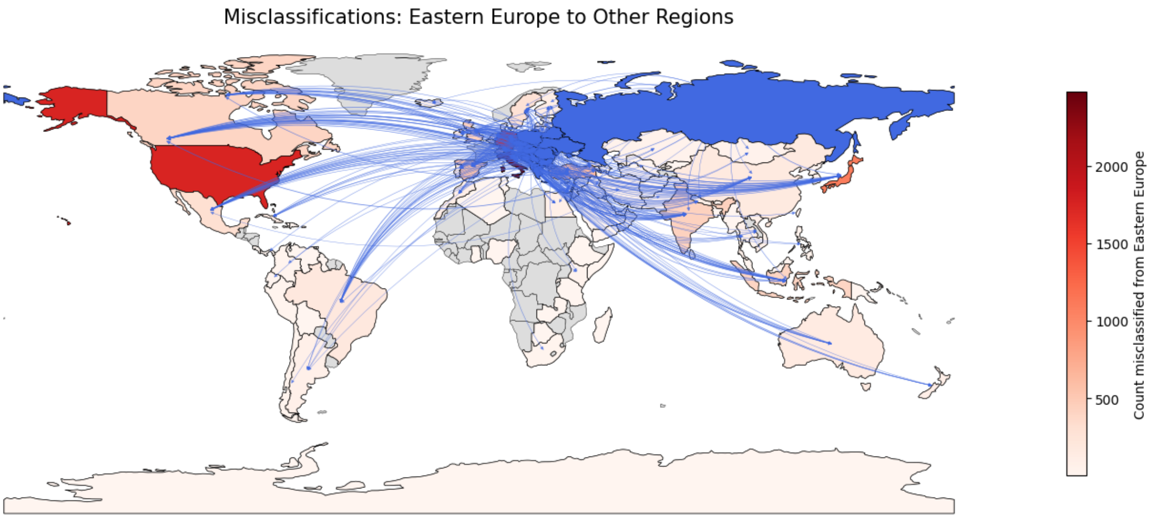

Figure 23: Mis-classification map : Eastern Europe

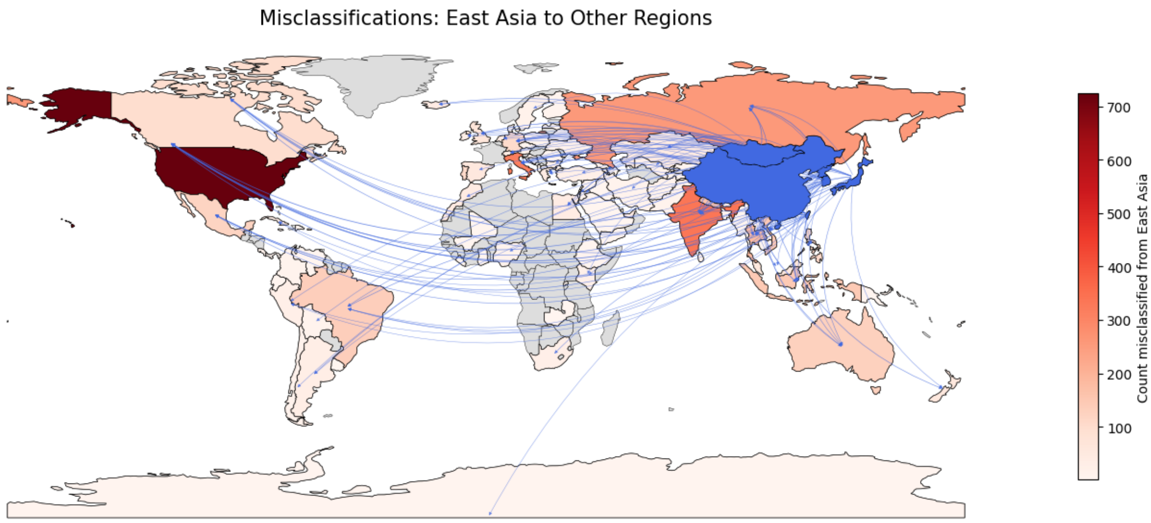

Figure 24: Mis-classification map : East Asia

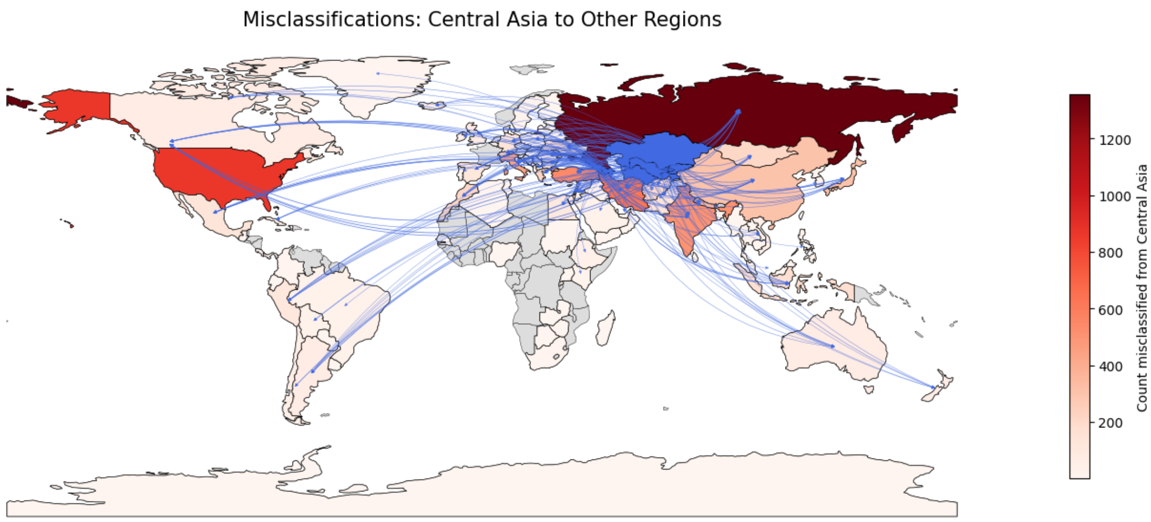

Figure 25: Mis-classification map : Central Asia

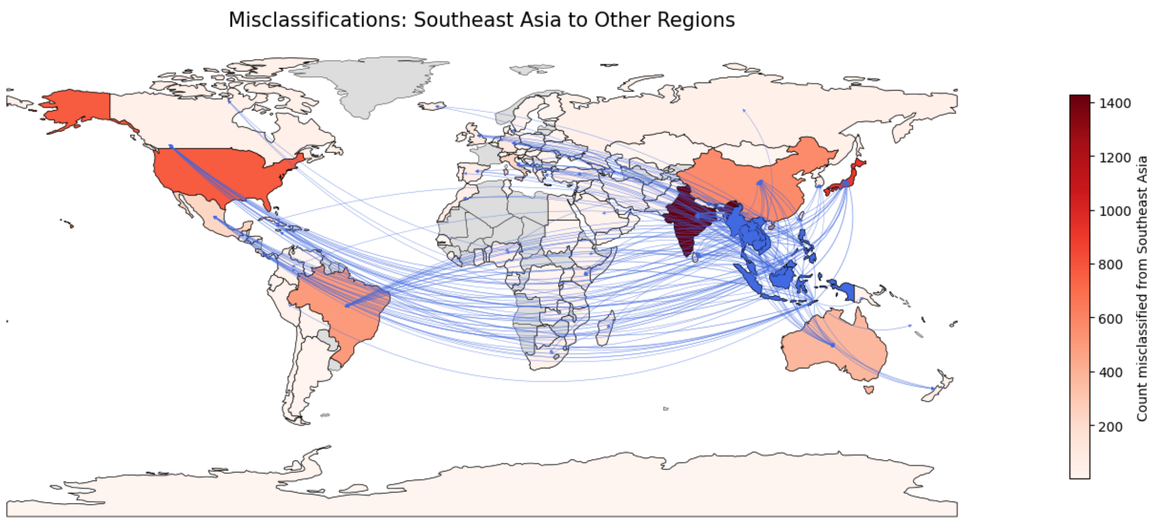

Figure 26: Mis-classification map : South East Asia

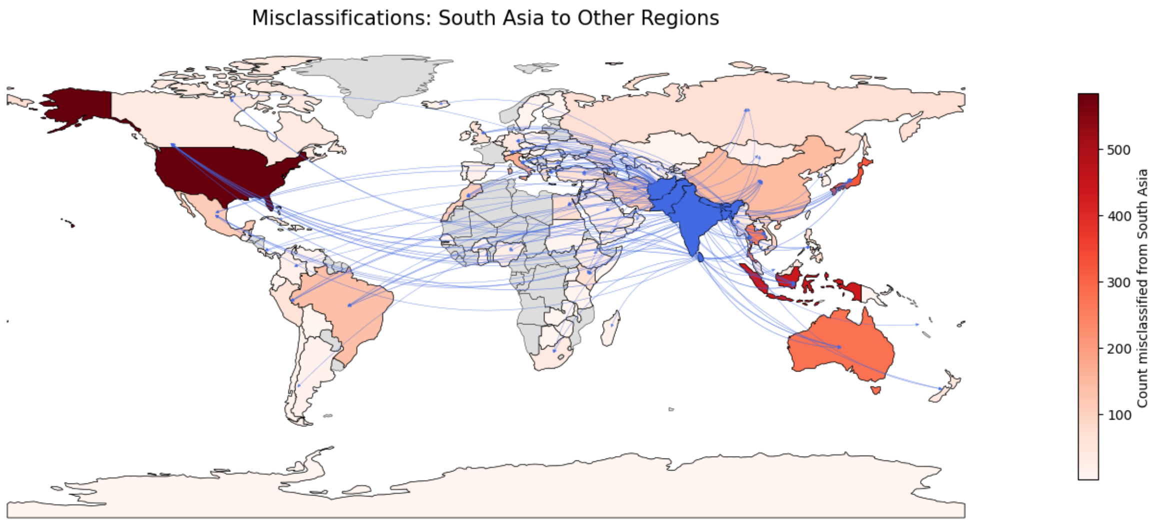

Figure 27: Mis-classification map : South Asia

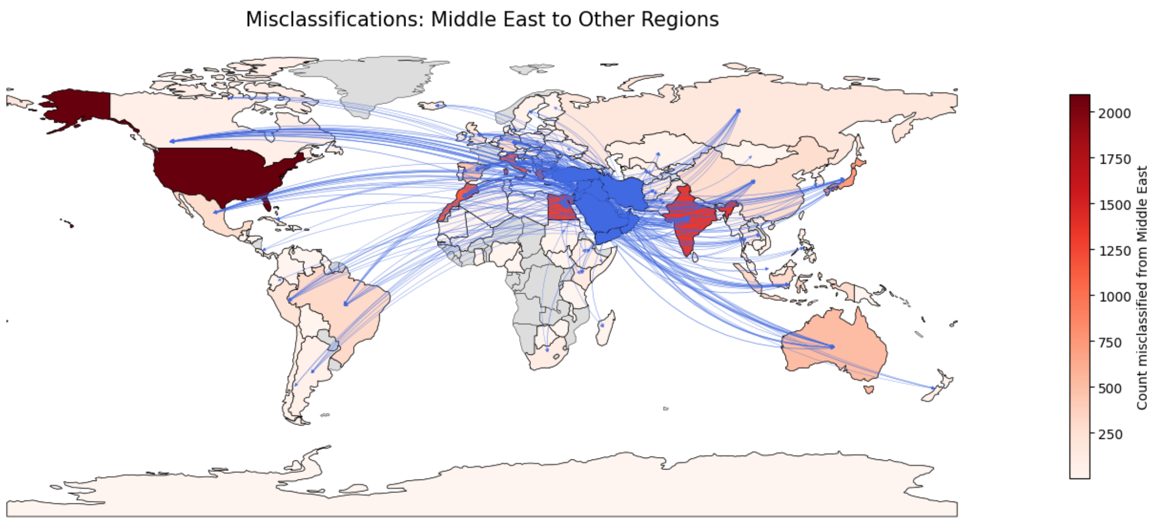

Figure 28: Mis-classification map : Middle East

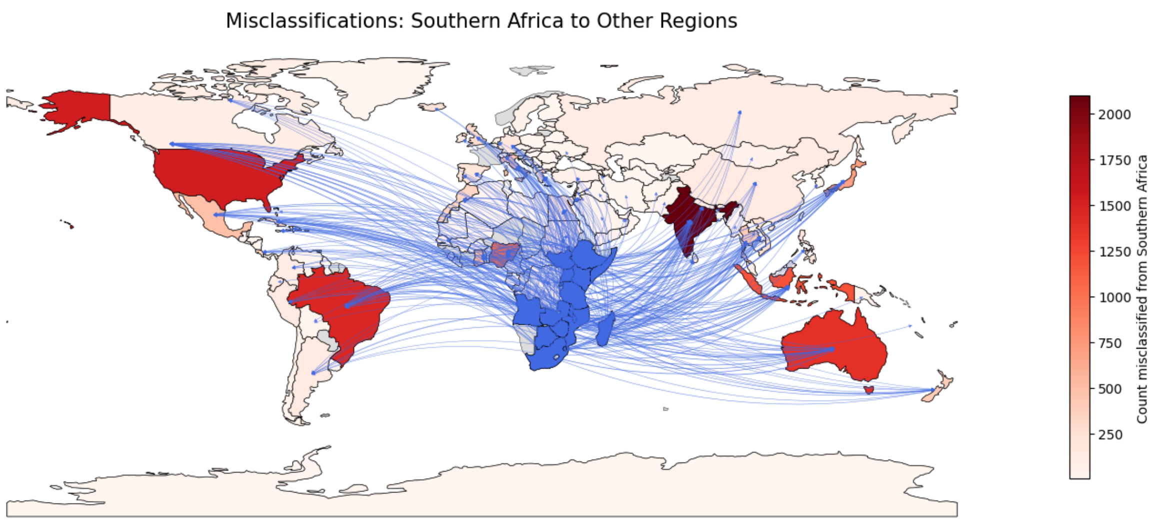

Figure 29: Mis-classification map : Southern Africa

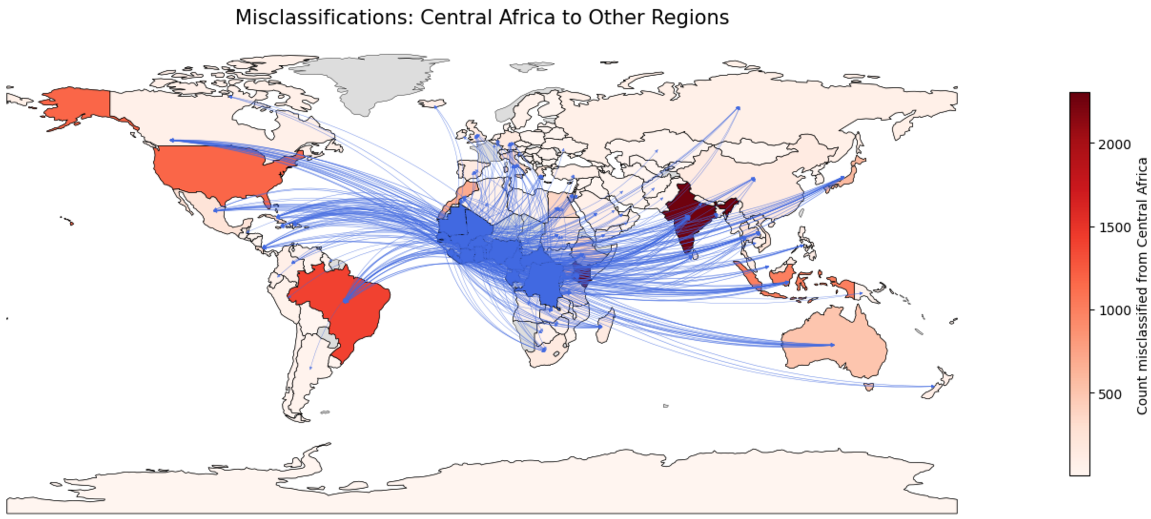

Figure 30: Mis-classification map : Central Africa

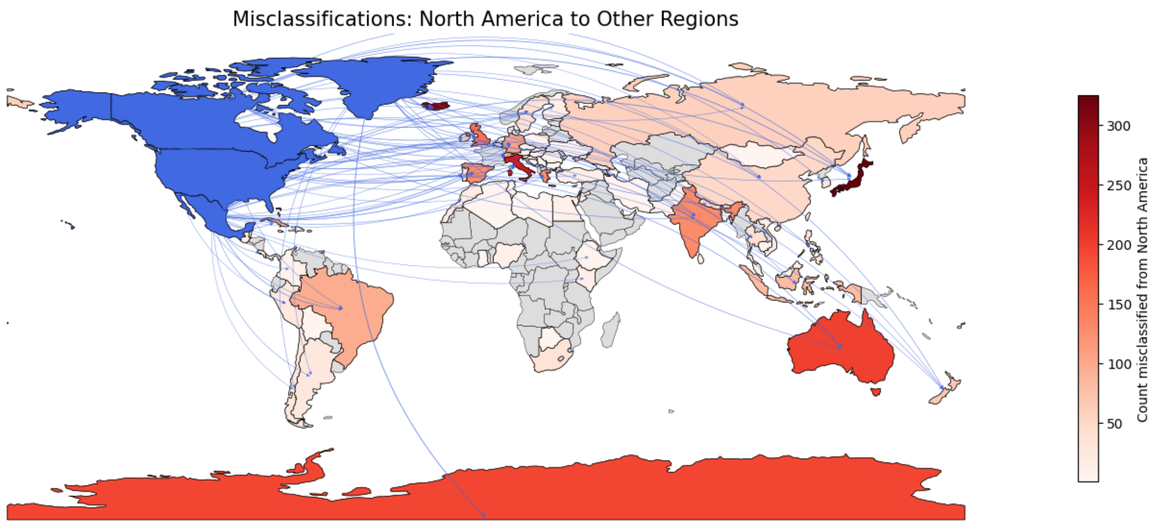

Figure 31: Mis-classification map : North America

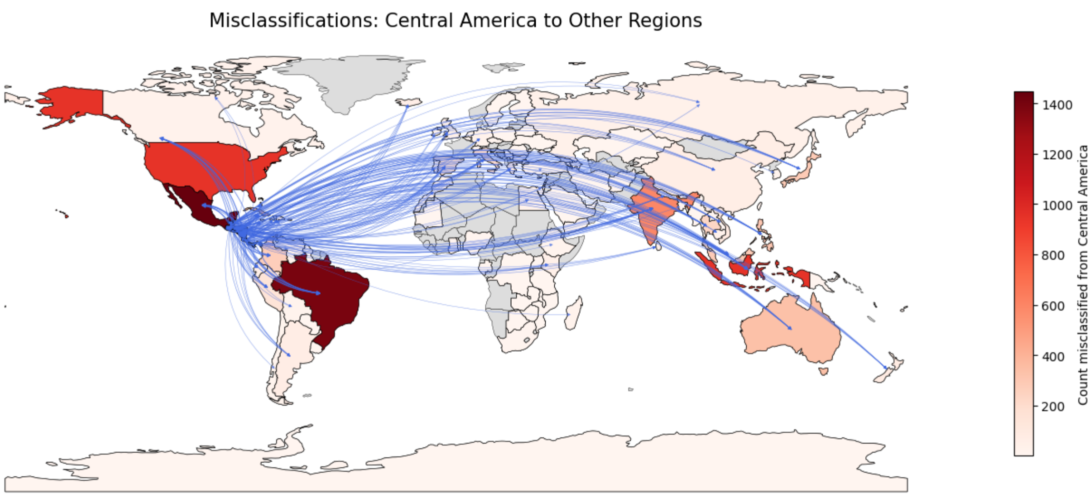

Figure 32: Mis-classification map : Central America

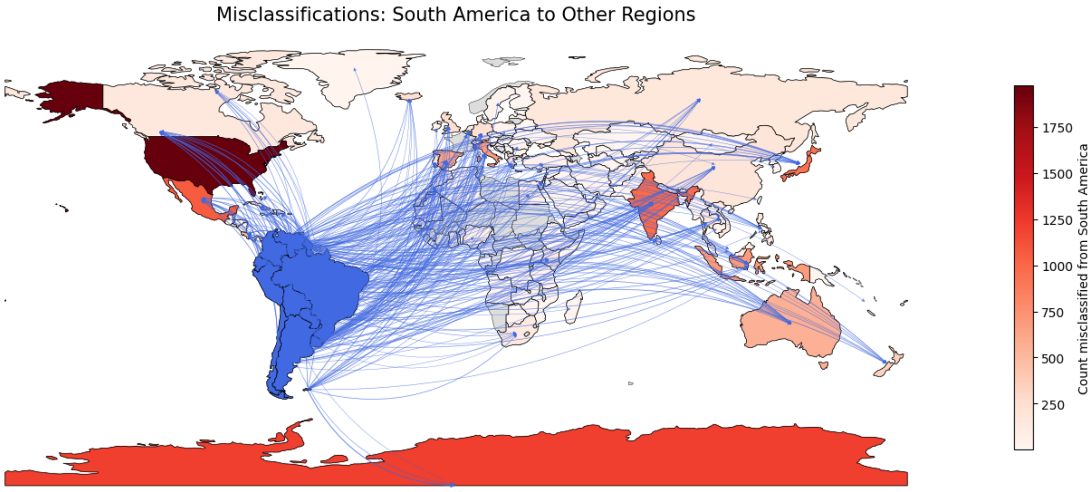

Figure 33: Mis-classification map : South America

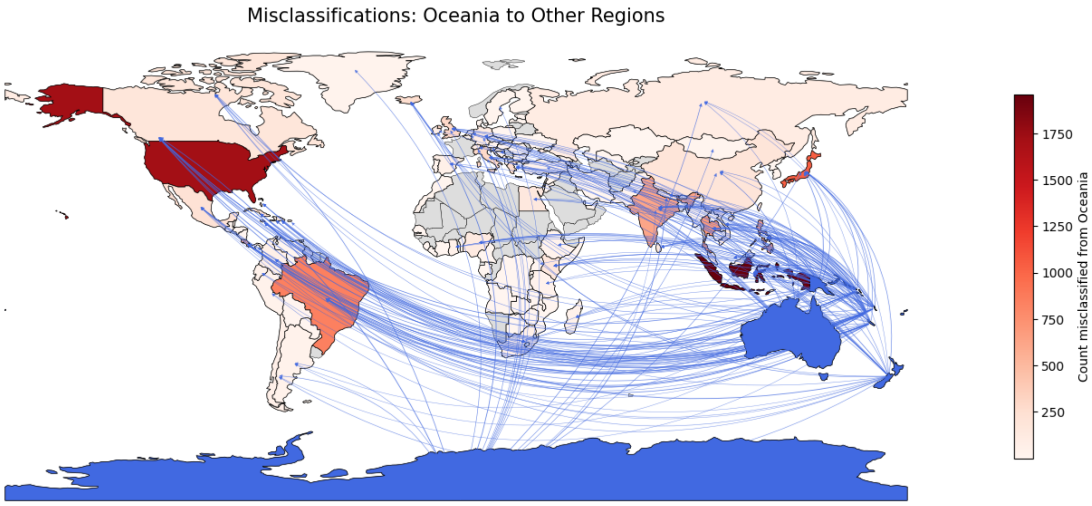

Figure 34: Mis-classification map : Oceania

Country name	Open-Ended	MCQs with Similar choices	MCQs with Random choices
Costa Rica	26.00	73.23	72.16
Croatia	47.83	72.83	83.92
Cuba	47.50	76.83	77.92
Curaçao	20.83	61.07	80.96
Cyprus	13.67	59.33	69.19
Czechia	40.50	66.07	83.90
Côte d'Ivoire	13.33	60.00	71.47
Denmark	32.50	66.93	78.54
Dominica	15.17	61.17	67.04
Dominican Republic	15.00	56.37	70.43
Ecuador	21.50	76.10	73.05
Egypt	60.50	77.07	83.84
El Salvador	4.83	65.93	63.40
Estonia	21.83	43.90	70.30
Eswatini	0.50	28.70	53.07
Ethiopia	41.00	80.93	80.24
Falkland Islands	8.83	92.13	90.35
Faroe Islands	30.33	71.30	90.66
Fiji	22.83	64.10	76.93
Finland	32.33	67.80	76.31
France	40.83	73.77	83.70
French Guiana	3.00	64.73	53.93
French Polynesia	24.67	81.90	83.97
Gabon	5.33	56.00	66.67
Gambia	3.33	41.40	53.16
Georgia	32.00	71.40	83.37
Germany	54.83	71.30	87.54
Ghana	26.33	70.53	67.80
Gibraltar	19.00	62.27	79.48
Greece	66.67	91.00	91.06
Greenland	27.00	65.43	84.90
Grenada	3.50	37.77	63.75
Guadeloupe	1.50	48.80	71.09
Guam	11.33	70.43	55.57
Guatemala	19.67	75.50	74.38
Guernsey	1.83	66.50	81.14
Guyana	8.83	52.33	52.66
Haiti	27.83	71.23	65.98
Vatican City State	8.67	43.77	74.31
Honduras	4.67	64.13	66.98
Hong Kong	22.67	65.23	86.70
Hungary	24.83	49.00	78.44
Iceland	69.00	82.27	89.04

Table 5: Region wise accuracies through various experimental settings : Part 2/5

Country name	Open-Ended	MCQs with Similar choices	MCQs with Random choices
India	78.33	90.03	90.10
Indonesia	48.83	67.76	84.97
Iran	50.83	70.40	83.27
Iraq	28.67	60.60	76.84
Ireland	48.33	74.57	87.63
Isle of Man	6.17	52.03	77.91
Israel	35.67	76.33	73.99
Italy	60.00	82.30	85.40
Jamaica	28.17	60.20	70.58
Japan	81.17	88.92	91.75
Jersey	3.67	50.37	71.69
Jordan	44.00	79.03	89.04
Kazakhstan	18.33	44.73	77.73
Kenya	56.00	88.57	88.15
North Korea	47.33	25.64	81.26
South Korea	47.83	67.23	79.90
Kuwait	12.83	52.30	68.70
Kyrgyzstan	20.17	37.30	69.48
Laos	26.50	38.53	80.25
Latvia	17.00	41.63	72.47
Lebanon	27.00	73.63	78.09
Liberia	9.33	50.97	65.37
Libya	6.67	22.87	73.10
Liechtenstein	6.17	34.03	72.29
Lithuania	24.00	54.43	74.40
Luxembourg	13.33	21.90	62.29
Macao	17.00	66.42	85.38
Madagascar	24.17	81.20	65.40
Malawi	8.33	54.80	66.39
Malaysia	28.33	73.28	83.65
Maldives	39.33	80.20	82.08
Mali	13.83	65.43	80.11
Malta	47.67	79.57	90.95
Martinique	4.33	53.60	72.85
Mauritania	12.00	76.77	80.28
Mauritius	38.33	92.00	79.52
Mexico	53.17	79.77	79.69
Moldova	7.67	35.23	63.57
Monaco	30.17	54.83	69.69
Mongolia	50.83	82.41	81.39
Montenegro	22.17	44.37	81.00
Morocco	67.83	85.40	93.75
Mozambique	5.17	66.57	63.78
Myanmar	61.50	76.56	92.62

Table 6: Region wise accuracies through various experimental settings : Part 3/5

Country name	Open-Ended	MCQs with Similar choices	MCQs with Random choices
Namibia	0.00	83.40	85.35
Nepal	65.00	72.53	89.72
Netherlands	46.00	74.63	86.86
New Caledonia	7.50	55.03	64.98
New Zealand	53.83	76.40	82.58
Nicaragua	6.83	69.87	69.64
Nigeria	47.33	79.13	73.78
North Macedonia	10.17	44.27	74.44
Norway	32.50	48.17	79.45
Oman	31.67	71.40	77.59
Pakistan	30.33	53.57	79.32
Palau	15.83	71.23	71.97
Palestine, State of	9.00	73.53	83.59
Panama	4.33	80.17	60.86
Papua New Guinea	13.50	61.87	63.38
Paraguay	6.17	52.23	54.29
Peru	54.83	85.73	83.61
Philippines	43.67	74.82	85.94
Poland	28.83	62.00	79.17
Portugal	43.50	58.60	84.39
Puerto Rico	16.67	68.97	72.52
Qatar	19.50	56.63	66.04
Romania	31.50	56.43	79.02
Russian Federation	52.67	73.13	77.18
Rwanda	29.50	71.73	73.72
Réunion	5.33	90.87	69.21
Saint Helena, Ascension and Tristan da Cunha	3.33	71.40	57.44
Saint Kitts and Nevis	14.17	41.23	64.61
Saint Lucia	16.83	61.40	79.33
Saint Martin (French)	4.00	45.43	69.48
Samoa	23.33	68.43	71.19
San Marino	10.17	35.00	54.01
Saudi Arabia	26.00	65.53	74.69
Senegal	21.83	78.73	78.20
Serbia	24.33	58.70	79.14
Seychelles	26.33	92.87	76.83
Sierra Leone	8.83	56.53	75.23
Singapore	51.33	74.91	80.15
Saint Martin (Dutch)	7.17	50.77	75.14
Slovakia	12.33	32.33	67.41
Slovenia	24.00	53.40	75.09
Solomon Islands	3.33	22.53	69.22
Somalia	24.67	75.30	78.46

Table 7: Region wise accuracies through various experimental settings : Part 4/5

Country name	Open-Ended	MCQs with Similar choices	MCQs with Random choices
South Africa	38.50	94.43	82.91
South Georgia and the South Sandwich Islands	7.17	80.70	77.99
South Sudan	25.83	65.83	82.31
Spain	51.00	83.13	84.71
Sri Lanka	37.00	61.40	82.72
Sudan	25.33	70.63	81.25
Svalbard and Jan Mayen	0.00	74.13	89.45
Sweden	35.50	54.63	81.22
Switzerland	42.17	62.53	76.40
Syrian Arab Republic	13.00	51.63	64.82
Taiwan, Province of China	23.00	51.01	80.16
Tajikistan	10.83	44.43	81.04
Tanzania, United Republic of	24.83	84.37	84.89
Thailand	64.17	84.49	89.08
Timor-Leste	7.83	41.77	69.67
Togo	2.33	31.67	65.98
Tonga	1.33	19.60	44.73
Trinidad and Tobago	8.00	56.23	53.62
Tunisia	20.33	40.00	75.53
Turkmenistan	22.67	48.73	82.83
Türkiye	56.33	86.10	92.24
Uganda	26.83	79.90	80.27
Ukraine	22.83	67.63	72.82
United Arab Emirates	53.00	85.30	85.30
United Kingdom	50.17	92.17	89.05
United States	67.17	91.03	87.76
Uruguay	14.17	46.33	61.10
Uzbekistan	47.17	68.63	83.07
Vanuatu	5.50	18.00	57.04
Venezuela, Bolivarian Republic of	11.17	57.63	53.41
Viet Nam	55.50	78.74	89.77
Virgin Islands, British	6.83	38.00	79.60
Virgin Islands, U.S.	9.67	46.73	81.72
Kosovo	6.50	28.70	65.53
Yemen	27.17	69.80	76.46
Zambia	9.50	54.80	73.29
Zimbabwe	11.67	71.03	76.05
Åland Islands	0.17	29.00	62.02
Overall	**25.14**	**61.92**	**75.06**

Table 8: Region wise accuracies through various experimental settings : Part 5/5

Country : Nepal
Confidence : 85
Features : ['prayer flags', 'architectural style', 'ornate windows']
Reasoning : The presence of colorful prayer flags is a strong indicator of Nepalese culture, commonly found in Buddhist and Hindu contexts. Additionally, the architectural style of the buildings in the background reflects traditional Nepali design, which further supports the classification.

Figure 35: A sample from our dataset and its corresponding response (GPT-4o-Mini)