Toward Equitable AI: Probing Cultural Bias in CLIP Embeddings via Latent-Space Evaluation

Annie (Zining) Wang
Department of Computer Science
University of British Columbia
zining.wang@ubc.ca

1 Introduction

Vision–language models (VLMs) such as CLIP combine image and text embeddings in a shared latent space, enabling powerful cross-modal capabilities like retrieval, captioning, and visual question answering. However, these models are trained on large-scale web data that is predominantly Western and English-centric, raising growing concerns about their cultural representativeness. For example, the GLOBALRG benchmark shows that CLIP readily recognizes Western-centric depictions—such as eggs and toast for "breakfast" or white dresses for "wedding"—but often fails on culturally specific content like the Mexican *molinillo* [2]. We hypothesize that this gap stems from biases embedded in CLIP's internal latent space, where Western imagery forms tighter clusters than representations from other regions [10].

This representational skew has real-world implications: in education, VLM-powered platforms may surface Western topics at the expense of African or Asian content [7]; in healthcare, vision models have underdiagnosed patients from non-Western populations [12]; and in media, such biases risk reinforcing a monolithic global perspective. While prior work has highlighted biased outputs, few studies directly probe whether these disparities originate in the model's learned representation space.

In this paper, we study whether and how cultural bias is encoded within CLIP's latent representations. Specifically, we examine two questions: (1) To what extent does CLIP embed culturally similar images into coherent clusters, and are Western images more tightly grouped than others? (2) How do CLIP's textual representations align with images from different regions for common, universal concepts?

Using the diverse GLOBALRG dataset spanning 19 universal concepts, we perform a two-part analysis: (a) an *image–image similarity* probe measuring how visual embeddings group by cultural region, and (b) a *text–image alignment* probe evaluating how closely caption embeddings align with images from various regions. Our findings reveal that for many concepts—particularly in daily life and rituals—CLIP embeds Western imagery into denser clusters and aligns textual prompts more closely with Western image embeddings. However, this trend is not uniform: concepts like *transport* and *headcoverings* show stronger alignment with non-Western regions.

Contributions. Our work makes the following contributions: (1) We introduce a two-part probing analysis to quantify cultural asymmetries in VLM embeddings; (2) We provide empirical evidence that cultural bias is encoded in both the visual and multimodal manifolds of CLIP, with concept-dependent variations; and (3) We release code and visualizations to support further research in culturally equitable vision–language modeling.[1]

[1] See https://github.com/aw814/EECE570-project-clip for visualizations and full results.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

2 Related Work

2.1 Cultural Bias in CLIP and Vision–Language Models

Recent studies reveal that CLIP exhibits cultural bias in both its outputs and internal representations. For example, Bhatia et al. [2] introduce the GLOBALRG benchmark, which shows that CLIP retrieves Western-centric images (e.g., eggs and toast for "breakfast") more reliably than culturally diverse alternatives. Similarly, Pouget et al. [10] and Nayak et al. [8] find that CLIP performs worse on non-Western content, particularly when trained on English-only data. Nwatu et al. [9] use the Dollar Street dataset to show that CLIP's similarity scores increase with regional wealth—images from low-income African households consistently rank lower than Western equivalents for the same object. These findings suggest that CLIP's latent space encodes geographic and socio-economic bias, favoring high-income, Western visual features.

2.2 Probing Bias in Latent Embeddings

Several recent works focus on uncovering how cultural and social biases are embedded in the internal representations of VLMs. Hamidieh et al. [6] introduce the So-B-IT benchmark and use association tests to reveal stereotypical alignments between concepts like "terrorist" and images of Middle Eastern men. Their study also proposes the Conditional Association Score (C–ASC) to quantify group-level bias in embedding space. Chuang et al. [4] and others explore methods to identify and remove biased directions in text embeddings, though they often neglect the image side. Meanwhile, Burda-Lassen et al. [3] introduce the Cultural Awareness Score (CAS) to assess whether model-generated captions reflect culturally relevant context. These probes go beyond output accuracy, offering insight into how models encode and prioritize different cultural features internally.

2.3 Multicultural Benchmarks and Datasets

To evaluate cultural bias more systematically, several multicultural benchmarks have emerged. The GLOBALRG dataset [2] evaluates retrieval and grounding across 50 countries and highlights disparities in model performance on non-Western content. CVQA [11] and CulturalVQA [8] test VLMs on culturally contextual visual question answering across dozens of countries and languages, revealing that even state-of-the-art models underperform on content from underrepresented regions. The Dollar Street dataset [9] provides cross-income visual variation for basic concepts, allowing researchers to test model robustness across socio-economic diversity. MOSAIC-1.5k [3] focuses on culturally rich imagery (e.g., rituals, dress, symbols), enabling fine-grained probing of captioning and retrieval models. Together, these benchmarks form a growing ecosystem for evaluating cultural fairness in VLMs beyond Western-centric metrics.

While prior work has documented cultural bias in model outputs, few studies directly probe the structure of these biases in latent space across both visual and textual modalities. Our work builds on these foundations by systematically analyzing CLIP's image–image clustering and text–image alignment across 19 universal concepts using the GLOBALRG dataset, surfacing how and where cultural disparities emerge inside the model.

3 Dataset

3.1 GLOBALRG-Retrieval Dataset

We use the recently published GLOBALRG-RETRIEVAL dataset on Hugging Face, which provides cross-cultural visual content across multiple universal concepts [2]. This dataset is designed for evaluating vision-language retrieval performance in culturally diverse contexts, offering a valuable testbed for examining representational bias. It contains 3,000 image–text pairs spanning 50 countries and 20 universal concepts (e.g., "breakfast," "wedding").

3.1.1 Selection of Universals and Countries:

We included all 20 universal concepts but excluded one—*farming*—due to two corrupted image entries, leaving a total of 19 universals. These include diverse themes such as *breakfast*, *festival*,

marriage, *dance*, *funeral*, *instrument*, and *clothing*. The complete list of universal concepts can be found in Table 1.

Table 1: List of Universal Concepts

Universal Concepts		
marriage	dessert	breakfast
clothing	sports	transport
ritual	drinks	instrument
religion	headcoverings	eating habits
music	funeral	lunch
festival	greeting	dance
dinner		

We selected 16 countries that span six continents and reflect a balance of Western and non-Western cultural contexts. Countries were chosen to ensure regional and cultural representation across North America, Europe, Asia, Africa, Oceania, and South America. Among those 7 of which are Western cultural, and 9 of them are non-Western. Definition of western and nonwestern. In this study, we classify countries as Western or Non-Western based on cultural and historical frameworks, including the Inglehart–Welzel cultural map [1]. Western countries (e.g., in Europe, North America, Oceania) share common values rooted in European traditions, while Non-Western countries represent culturally distinct regions across Asia, Africa, and Latin America. This distinction helps us analyze potential cultural bias in vision-language models.

Table 2: Selected countries across continents and cultural groups. Western countries are shaded in blue; Non-Western countries are shaded in orange.

Country	**Continent**	**Cultural Group**
Japan	Asia	Non-Western
China	Asia	Non-Western
India	Asia	Non-Western
Germany	Europe	Western
Spain	Europe	Western
United Kingdom	Europe	Western
South Africa	Africa	Non-Western
Nigeria	Africa	Non-Western
Egypt	Africa	Non-Western
United States of America	North America	Western
Canada	North America	Western
Argentina	South America	Non-Western
Brazil	South America	Non-Western
Australia	Oceania	Western
New Zealand	Oceania	Western
Fiji	Oceania	Non-Western

Each universal has three distinct images per country, as curated in GLOBALRG, ensuring consistent cultural representation (e.g., three unique "breakfast" images for Japan). This subset design guarantees equal representation across all selected countries and concepts, as well as a equal representation on the Western and non-Western perspectives, enabling fair comparisons. This results in a total of 912 image–text pairs (19 universals $\times$ 16 countries $\times$ 3 images). Table 2 shows an overview of the chosen countries grouped by region.

3.1.2 Data Verification:

To ensure high data quality, we manually inspected a random 10% sample (approximately 90 pairs) to check image clarity and label accuracy. All reviewed entries were found to be valid, requiring no further edits. This verification step ensured the dataset's reliability for subsequent analysis.

4 Methodology

In this section we describe in detail what we're asking the model to do (the task definition), how we extract and preprocess the data, and the core algorithms we apply to probe cultural bias in CLIP's latent space.

4.1 Embedding Extraction

To analyze cultural representations within vision-language models, we utilized the CLIP ViT-B/32 architecture, accessed via the Hugging Face Transformers library. Each image in our dataset was processed using the `CLIPProcessor`, which internally applies a series of preprocessing steps to prepare the images for the CLIP model. These steps include resizing the image so that its shortest edge is 224 pixels while maintaining the aspect ratio, followed by a center crop to obtain a 224×224 pixel image. The images are then rescaled by a factor of 1/255 and normalized using the mean and standard deviation values employed during CLIP's training [5]. These preprocessing steps were automatically applied by the default CLIPProcessor.

For the textual modality, the original dataset only provides short vocabulary-level labels. Since CLIP is trained primarily on full-sentence captions, we convert each universal into a sentence-level caption prompt of the form *"A photo of a <universal>"*, to better match CLIP's learned multimodal associations. The caption sentence were tokenized using CLIP's tokenizer, and embedding was extracted.

Finally, all 512-dimensional image and text embeddings were checkpointed and saved in JSONL format to support efficient reuse in downstream analyses.

4.2 Task Definition

To evaluate cultural bias within the CLIP latent space, we designed two complementary probing tasks, each centered around a single universal concept. These tasks aim to address our research questions by analyzing (1) image–image proximity and (2) text–image alignment. The first task measures how visually similar CLIP considers images of the same concept across cultures, while the second task evaluates how closely textual descriptions align with images from different regions.

4.2.1 Image–Image Similarity Task

The objective of the image–image similarity task is to determine whether images representing the same concept from Western countries cluster more tightly compared to pairs comprising Western and Non-Western images.

1. **Quantitative Analysis:** We extract pairwise cosine similarities for Western–Western (WW) and Western–Non-Western (WN) image pairs. We compute the mean similarity for each group and apply Welch's t-test, incorporating Bonferroni correction, to evaluate whether the difference between WW and WN similarities is statistically significant. A higher mean similarity among WW pairs would suggest a cultural bias in the model's representation of images.
2. **Visualization of similarity of same-concept images across countries:** For each concept, we compute a confusion matrix of mean cosine similarities between all pairs of regions. This matrix is visualized as a heatmap to facilitate intuitive interpretation of clustering patterns.

4.2.2 Text–Image Alignment Task ($\mathbf{T_{T\text{–}I}}$)

This task evaluates whether caption prompts align more strongly with images from Western regions.

1. **Quantifying Alignment with C–ASC**: To measure bias, we employ the Caption-Association Score (C-ASC) [6], adapted from SC–WEAT:

$$C - ASC(c, G) = \frac{\frac{1}{|G|}\sum_{g \in G} d(c, g) \; - \; \frac{1}{|\bar{G}|}\sum_{g' \in \bar{G}} d(c, g')}{\mathrm{sd}_{u \in D} d(c, u)} \; .$$

Here, c is the caption embedding, G is the target group (e.g., images from Western countries), and $\bar{G}$ is the complement set (e.g., images from non-Western countries). The function $d(c, x)$ computes cosine similarity between the caption and image embeddings. The numerator captures the mean similarity difference between G and $\bar{G}$, while the denominator normalizes by the standard deviation of similarities across all images D, yielding an effect-size measure analogous to Cohen's d.

A high positive C-ASC score indicates that the model associates the caption more strongly with images from the target group. We compute this score per concept and compare alignment levels across Western and Non-Western groups. We interpret $|C - ASC|$ using conventional cutoffs: *small* (<0.2), *medium* (≈0.5), and *large* (≥0.8).

2. **Visualization of Image–Text Alignment**:
   - *Latent Space Projection*: We project both image and caption embeddings for each concept into a 2D space using UMAP. We apply K-means clustering and visualize the decision boundary to explore the separation between Western and Non-Western representations.
   - *Global Alignment Map*: To evaluate cultural bias in text–image associations, we visualize the average alignment between caption embeddings and image embeddings on a country-level basis. For each universal concept, we compute the cosine similarity between the caption embedding (e.g., *"A photo of a breakfast"*) and all corresponding image embeddings. We then calculate the average similarity for each country and rank all countries based on these scores, from highest to lowest. By aggregating these ranks across all concepts, we derive an overall alignment score for each country, reflecting how consistently its images are closely aligned with textual descriptions. These scores are visualized on a world map, where darker regions indicate stronger average alignment—implying that the model tends to associate caption prompts more strongly with those countries' visual representations. In contrast, lighter regions suggest weaker alignment and possible under-representation in the model's learned associations.

5 Results

We present empirical findings from two probing tasks designed to uncover cultural bias in CLIP's embedding space, combining visualizations with statistical analysis.

5.1 Image–Image Similarity

5.1.1 Quantitative Analysis of Image–Image Similarity

To determine whether CLIP embeds concept images from Western countries more closely than those from culturally diverse regions, we computed pairwise cosine similarities between all image embeddings within each universal concept. Specifically, we compared similarity distributions for two types of region pairs: Western–Western (WW) and Western–Non-Western (WN). For each concept, we calculated the mean cosine similarity for both groups and conducted Welch's t-tests to test for statistical significance.

The results indicate a consistent trend: for several concepts, images from Western countries were significantly more similar to each other than to those from non-Western countries. As shown in Table 3, 8 out of 19 concepts demonstrated statistically significant WW > WN similarity differences even after Bonferroni correction ($p < 0.05$), including culturally salient concepts like *funeral* ($t = 8.23$, $p < 10^{-10}$), *marriage* ($t = 7.06$, $p < 10^{-8}$), and *dessert* ($t = 5.68$, $p < 10^{-5}$). Other examples include *breakfast*, *transport*, and *religion*. This suggests that, for these concepts, CLIP embeds Western imagery into denser latent clusters than cross-cultural counterparts, pointing to potential bias in visual representation that could influence downstream applications such as image retrieval or content moderation.

By contrast, the remaining 11 concepts did not exhibit statistically significant differences. Some, like *music*, *ritual*, and *festival*, showed near-zero or slightly negative t-values, indicating relatively balanced or even more dispersed Western representations. These concepts may reflect more globally shared or abstract visual attributes, resisting cultural clustering effects.

Table 3: Image–Image Similarity Analysis Across Concepts: Welch's t-test comparing cosine similarities between Western–Western (WW) and Western–Non-Western (WN) image pairs.

Concept	n_{WW}	n_{WN}	**Mean WW**	**Mean WN**	t**-stat**	p**-val**	p**-bonf**	**Sig**
Funeral	21	63	0.614	0.547	8.23	$7.04e{-11}$	$1.34e{-9}$	✓
Marriage	21	63	0.700	0.590	7.06	$1.98e{-9}$	$3.76e{-8}$	✓
Dessert	21	63	0.715	0.670	5.68	$1.32e{-6}$	$2.51e{-5}$	✓
Breakfast	21	63	0.694	0.637	5.45	$2.60e{-6}$	$4.94e{-5}$	✓
Transport	21	63	0.635	0.574	4.90	$5.51e{-6}$	$1.05e{-4}$	✓
Headcoverings	21	63	0.493	0.454	4.45	$2.28e{-5}$	$4.33e{-4}$	✓
Eating Habits	21	63	0.658	0.605	4.58	$4.16e{-5}$	$7.91e{-4}$	✓
Religion	21	63	0.601	0.549	4.28	$6.18e{-5}$	$1.17e{-3}$	✓
Dance	21	63	0.638	0.614	2.64	$5.17e{-3}$	$9.82e{-2}$	
Greeting	21	63	0.549	0.498	2.46	$9.46e{-3}$	$1.80e{-1}$	
Lunch	21	63	0.654	0.638	2.21	$1.53e{-2}$	$2.91e{-1}$	
dessert	21	63	0.664	0.625	2.23	$1.64e{-2}$	$3.11e{-1}$	
Sports	21	63	0.531	0.511	1.94	$2.95e{-2}$	$5.61e{-1}$	
Dinner	21	63	0.643	0.624	1.89	$3.22e{-2}$	$6.12e{-1}$	
Instrument	21	63	0.593	0.570	1.39	$8.69e{-2}$	1.00	
Clothing	21	63	0.496	0.482	0.88	$1.93e{-1}$	1.00	
Ritual	21	63	0.526	0.530	-0.24	$5.92e{-1}$	1.00	
Music	21	63	0.534	0.537	-0.29	$6.14e{-1}$	1.00	
Festival	21	63	0.583	0.587	-0.39	$6.50e{-1}$	1.00	

Overall, these results confirm that latent-space bias is not uniform across concepts. It appears most strongly in domains where visual aesthetics and cultural practices differ widely. The presence of such clustering effects highlights how CLIP may encode and amplify Western-centric norms in its vision-based representations.

5.1.2 Heatmap Visualizations of Image–Image Similarity

While statistics reveal overall trends, heatmaps allow us to visually locate where cultural convergence and divergence appear. We generated heatmap visualizations showing pairwise cosine similarities between countries for each universal concept. In these heatmaps, rows and columns represent countries, and the color intensity reflects average cosine similarity between their image embeddings.

Figure 1a shows a sharp cultural boundary for *marriage*, where Western countries exhibit high intra-group similarity (brighter yellow), and Non-Western regions show low similarity (darker tones). This aligns with our statistical findings and highlights how the model captures culturally-specific aesthetics. Beyond the Western/Non-Western divide, finer-grained patterns emerge. For example, Japan and China form a distinct cluster, while Argentina, Brazil, and India appear closer to Western imagery—possibly due to shared media influences, colonial histories, or hybrid cultural traits. We include representative image samples in Figure **??** to contextualize these patterns.

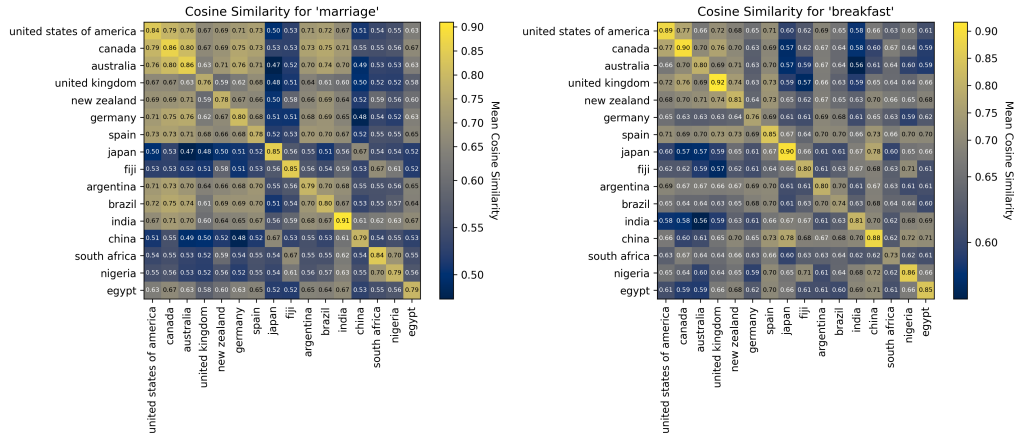
The *breakfast* heatmap (Figure 1) reveals a more blended but still observable divide. Western countries like the United States, Canada, Germany, and the UK cluster brightly in the top-left, while Non-Western countries such as China, Nigeria, and Egypt appear with darker tones, indicating lower similarity.

These visualizations not only reinforce statistical patterns but also highlight specific inter-country relationships, offering insight beyond binary cultural groupings. Full heatmaps for all 19 concepts are available in our GitHub repository: `https://github.com/aw814/EECE570-project-clip`.

5.2 Text–Image Alignment

5.2.1 Quantitative Analysis with C–ASC

We assessed text–image alignment bias by computing the Caption–Association Score (C–ASC) for each concept. C–ASC quantifies how closely a caption embedding aligns with Western versus



(a) *marriage* (b) *breakfast*

Figure 1: Cosine similarity heatmaps for two representative concepts, *marriage* and *breakfast*. Countries are ordered with Western countries (e.g., United States – Spain) listed first, followed by Non-Western countries (e.g., Japan – Egypt). Brighter (more yellow) cells indicate higher similarity between country pairs for the same concept, while darker cells indicate lower similarity. These visualizations help illustrate how the model may embed Western imagery more tightly than cross-cultural representations.

Figure 2: Images related to universal 'marriage' for different countries: Fiji, Canada, Japan, Argentina, South Africa and China. We can identify clear differences between the clothings and traditions associated with this universal activity.

Non-Western image embeddings. Positive scores indicate stronger alignment with Western images; negative scores suggest stronger association with Non-Western representations.

Out of 19 concepts, 7 showed moderate to strong Western alignment, including *breakfast* (C–ASC = 1.29, μ_W = 0.271, μ_N = 0.242), *dessert* (0.87), *dessert* (0.65), and *eating habits* (0.61). These findings suggest CLIP's text embeddings often gravitate toward Western imagery, especially for food-related and daily life concepts. Concepts like *religion*, *funeral*, and *dance* also show noticeable Western preference (> 0.4).

On the other hand, several concepts align more closely with Non-Western representations. *Transport* recorded the lowest score (-0.83), followed by *headcoverings* (-0.56) and *lunch* (-0.55). These trends suggest stronger proximity between captions and Non-Western visual content for these domains. A few concepts, such as *music*, *ritual*, and *dinner*, showed near-zero C–ASC values, indicating relatively balanced semantic alignment. Full results are available in Appendix Table 4.

In summary, the C–ASC distribution reveals that CLIP's multimodal representations encode uneven cultural associations, with a tendency to favor Western alignment in some domains while amplifying Non-Western alignment in others.

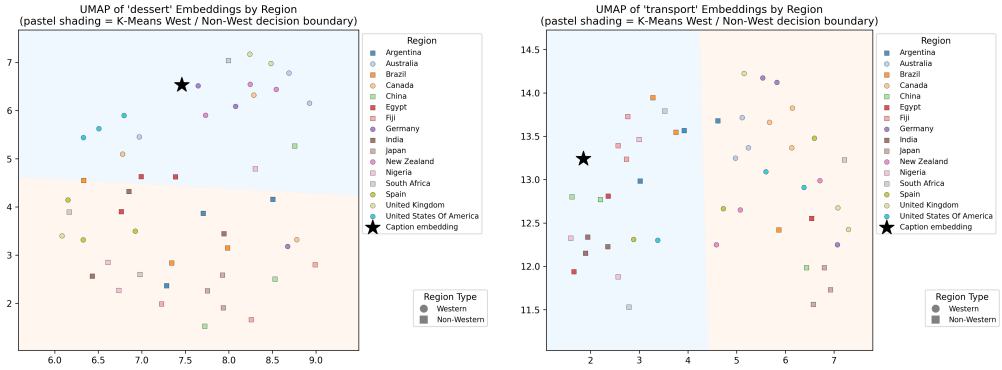
5.2.2 Visualizing Text–Image Alignment

To complement the C–ASC analysis, we present two visualizations that highlight how caption embeddings align with image representations across countries.

Latent Space Projection. We project image and caption embeddings for each concept into a 2D UMAP space (Figure 3). K-means clustering and decision boundaries reveal separation patterns between Western and Non-Western representations.

In many concepts, caption embeddings are surrounded by Western image clusters, with Non-Western images more dispersed. This spatial configuration reinforces C–ASC trends. Notably, the alignment varies across concepts. For *dessert*, the caption sits within a dense Western image cluster, while for *transport*, it is embedded among Non-Western images, showing stronger alignment with those regions.

These projections validate that CLIP's latent associations differ by concept, sometimes favoring Western imagery, sometimes not. Full UMAP plots are available in our GitHub repository.



(a) *dessert* (b) *transport*

Figure 3: Latent space visualizations for two concepts. Shapes indicate Western vs. non-Western image embeddings; colors are for contrast only. The black star marks the caption embedding (e.g., "A photo of a dessert"). These projections show cultural clustering patterns in CLIP's image–text space.

Global Alignment Map. Finally, we aggregate text–image alignment ranks across all concepts and visualize them as a choropleth map. Countries are ranked per concept by cosine similarity between caption and image embeddings. These ranks are averaged to produce a global alignment score.

As shown in Figure 4, darker shades indicate stronger overall alignment. A consistent pattern emerges: countries in Southern Africa—including South Africa, Nigeria, and Egypt—are frequently

ranked near the bottom, suggesting significant under-representation in CLIP's learned multimodal associations.

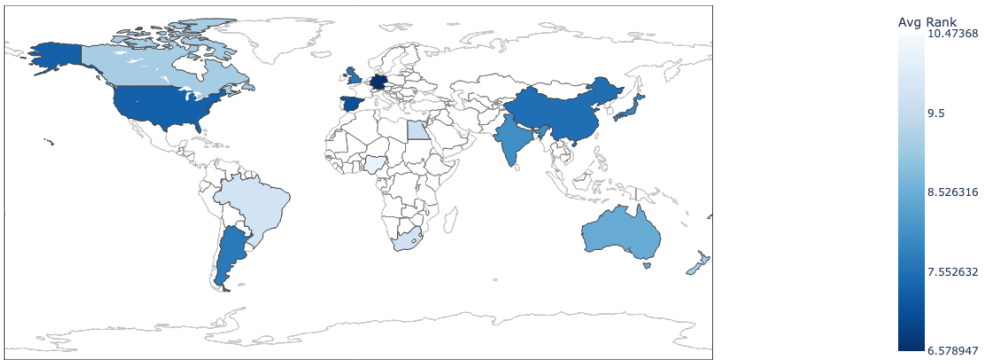


Figure 4: **Average Text–Image Alignment Rank by Country.** Darker regions reflect stronger average alignment with CLIP's textual representations, while lighter regions indicate consistently weaker alignment. Notably, several Southern African countries—including South Africa, Nigeria, and Egypt—are repeatedly ranked lower

6 Discussion

6.1 Limitatoin and Future work

While our findings provide important insights into cultural bias in CLIP's embedding space, several limitations must be acknowledged.

6.1.1 Scope and Model Coverage

This research was conducted as a course project and constrained by both time and prior experience in visual computing. As a result, our analysis focuses exclusively on CLIP, the most widely used vision–language model. However, the two-task probing framework can be readily applied to other VLMs. Future research could expand this framework to perform comparative analysis across multiple models, including newer architectures or multilingual models. Such comparisons could surface model-specific strengths and weaknesses in representing cultural diversity, resulting in more actionable insights not only for mitigating bias, but also for selecting the most culturally robust models for downstream tasks.

6.1.2 Interpretability of Dimensionality Reduction

Our methodology relies on UMAP to project high-dimensional embeddings (512D) into two dimensions for visualization. While this helps reveal general trends and clustering patterns, it also introduces interpretability challenges. UMAP is a nonlinear dimensionality reduction technique, and the resulting 2D plots do not preserve all information from the original space. Important semantic relationships may be distorted or lost. We address this by pairing visualization with statistical measures such as cosine similarity and the Caption–Association Score (C–ASC), which provide a more complete picture of the embedding space. Nevertheless, future work could explore techniques to better preserve semantic distances or offer explanations of how dimensionality was reduced.

6.1.3 Impact of Preprocessing on Cultural Information

The use of `CLIPProcessor` for image preprocessing—specifically, resizing and center cropping—ensures input standardization but may unintentionally remove culturally relevant visual content. Elements near the image periphery, such as traditional garments, symbolic gestures, or contextual scenery, may be cropped out. This can compromise CLIP's ability to represent cultural variation accurately, particularly in domains where subtle peripheral cues are essential. Future research should explore alternative preprocessing strategies such as adaptive cropping, padding, or bounding-box–based resizing to preserve edge information while remaining compatible with model constraints.

7 Conclusion

This paper presented a two-task probing analysis to examine cultural bias in CLIP's embedding space. Through both statistical analysis and visualization, we showed that CLIP tends to group Western images more tightly and aligns caption embeddings more closely with Western representations for several universal concepts. However, this pattern varies by concept, with some aligning more strongly with non-Western imagery. Our findings suggest that cultural bias is not just an output artifact but is encoded in the model's internal representations. The proposed methodology offers a foundation for deeper cross-model comparisons and highlights the need for more culturally aware model development and evaluation practices.

References

[1] World Values Survey Association. The inglehart–welzel world cultural map - world values survey 7 (2023), 2023. Accessed: 2025-04-21.

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

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

[4] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[5] Hugging Face. Clipprocessor documentation, 2025. Accessed: 2025-04-21.

[6] Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 547–561, 2024.

[7] Ahmed Imran. Why addressing digital inequality should be a priority. *The Electronic Journal of Information Systems in Developing Countries*, 89(3):e12255, 2023.

[8] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024.

[9] Chisom Nwatu, Omid Nouri, Gregor Pfeifer, Stefanie Zollmann, and Mehdi Mirza. Dollarstreet: Measuring socio-economic bias in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1847–1856, 2023.

[10] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *arXiv preprint arXiv:2405.13777*, 2024.

[11] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.

[12] Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J Wang, Dushyant W Sahani, and Shwetak Patel. Demographic bias of expert-level vision-language foundation models in medical imaging. *arXiv preprint arXiv:2402.14815*, 2024.

8 Appendix

Table 4: Caption–Association Scores (C–ASC) by Concept. Positive scores indicate stronger alignment with Western images.

Concept	**Mean**$_W$	**Mean**$_N$	**C–ASC**
breakfast	0.271	0.242	**1.29**
dessert	0.260	0.247	**0.87**
drinks	0.254	0.239	0.65
eating habits	0.240	0.231	0.61
funeral	0.290	0.277	0.58
religion	0.239	0.231	0.51
dance	0.261	0.254	0.45
marriage	0.270	0.264	0.38
greeting	0.244	0.241	0.14
festival	0.262	0.260	0.11
clothing	0.231	0.229	0.10
instrument	0.265	0.263	0.09
music	0.239	0.238	0.03
sports	0.247	0.247	0.00
dinner	0.246	0.247	-0.13
ritual	0.240	0.247	-0.31
lunch	0.240	0.248	-0.55
headcoverings	0.233	0.250	-0.56
transport	0.257	0.268	**-0.83**