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

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

1 Background and Problem Setting

Vision-language models (VLMs) such as CLIP combine image and text embeddings in a shared latent space, enabling tasks like image retrieval, visual question answering, and grounding. These models are typically trained on large-scale web data, much of which originates from Western sources—particularly in North America and Europe—where English dominates. The GLOBALRG benchmark illustrates the impact of this bias: while CLIP accurately identifies Western-oriented images (e.g., eggs and toast for "breakfast," white dresses for "wedding"), it struggles with culturally specific items from other regions, like the Mexican "molinillo" [2]. We hypothesize that this discrepancy arises because the latent space reflects Western-centric training, causing Western image clusters to be more tightly grouped than non-Western ones [7]. Both text and image datasets tend to favor Western norms, influencing how CLIP represents different cultures internally.

This Western emphasis leads to digital inequality by restricting fair access to AI tools worldwide. For example, in education, VLM-driven platforms may highlight Western topics—such as U.S. history or European literature—over African or Asian content, leaving non-Western learners underserved [4]. In healthcare, VLM-based systems sometimes underdiagnose patients from non-Western populations, posing serious risks [9]. In media, services that rely on VLMs could inadvertently amplify Western perspectives and diminish local cultures, reinforcing a single dominant worldview. Left unchecked, these biases can widen global inequities and undermine AI's goal of serving everyone.

Although many studies document biased outputs, few investigate whether the bias originates in the model's internal embedding space. To address this gap, our work targets two central questions: (1) How do cultural variations in images align (or fail to align) with CLIP's textual representations of universal concepts like "breakfast"? and (2) To what extent is this bias encoded in the latent space of the model? Using the diverse GLOBALRG dataset, we aim to trace output biases back to their latent roots, offering strategies to develop more culturally equitable AI.

2 Related Work

Cultural bias in vision-language models (VLMs) has been examined, but relatively few studies focus on how it forms within the model's internal representations [1]. Several evaluations of models like CLIP compare Western and Eastern images, showing higher accuracy for Western examples and linking this discrepancy to English-focused training data [7]. Other research finds that restricting data to English reduces performance for regions like Africa, though these approaches rely on precision metrics without investigating the underlying embeddings [6]. The GLOBALRG benchmark employs precision@k and diversity@k to assess retrieval results but does not analyze the model's internal encoding of cultures. Some studies address social biases in CLIP—like gender and race—by measuring classification errors across demographic datasets, noting output disparities without examining embedding structures [10]. Additional work attempts to reduce bias in text embeddings by removing stereotypical directions, but does not address the image side of VLMs [3].

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

In contrast, the field of natural language processing (NLP) has a longer history of embedding-level analysis. Researchers often apply t-SNE to large language models, revealing clusters that reflect Western-centric training [5]. Our approach differs by focusing on both image and text encoders in CLIP, rather than limiting the analysis to outputs or text embeddings. We integrate the GLOBALRG dataset (50 countries) with clustering and similarity measures to explore cultural disparities in the latent space. This approach goes beyond output-level metrics, aiming to pinpoint where bias arises inside the model.

3 Experiment Plan

3.1 Data

We will use the GLOBALRG retrieval dataset, containing 3,000 image-text pairs spanning 50 countries and 20 universal concepts (e.g., "breakfast," "wedding"). To ensure our subset captures sufficient cultural diversity, we will:

- **Selection of Universals and Countries:**
  We have chosen 10 universals (*breakfast*, *farming*, *festival*, *marriage*, *dance*, *funeral*, *instrument*, *clothing*) that vary significantly across cultures (e.g., "breakfast" can be toast in Western contexts or tamales in Mexico). We also selected 16 countries representing eight regions and covering multiple continents and socio-economic backgrounds. 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 yields 480 image-text pairs in total (10 universals $\times$ 16 countries $\times$ 3 images), allowing us to detect meaningful patterns in embedding spaces. Table 1 shows an overview of the chosen regions and countries.

- **Data Verification:**
  We will manually inspect a random 10% sample (48 pairs) to check image clarity and label accuracy. This ensures that mislabeled or low-quality images are minimized, improving the reliability of our cultural representations.

Region	**Countries**
East Asia	China, Japan
South East Asia	Vietnam, Thailand
South Asia	India, Pakistan
Europe	Italy, Netherlands
Africa	Tanzania, Kenya
Latin America	Brazil, Argentina
Oceania	Australia, Fiji
North America	USA, Canada

Table 1: Regions and corresponding countries used in our subset.

3.2 Candidate Models or Pipelines

We will analyze CLIP [8], specifically the ViT-B/32 variant, given its open-source availability on `Hugging Face` and its established performance in aligning image and text embeddings. The GLOBALRG benchmark [2] has identified cultural bias in CLIP (e.g., 72.5 precision@5 and low diversity@5), making it a suitable baseline for our investigation.

Our pipeline begins by extracting embeddings. Images are resized to 224×224 pixels, normalized, and processed by the ViT-B/32 vision encoder, while text queries are tokenized with CLIP's standard tokenizer and fed to the text encoder. Each modality produces 512-dimensional embeddings. Consequently, each image-text pair yields one image embedding (unique to that image) and one text embedding (common to the concept).

3.2.1 Visualizations

We will use t-SNE (t-distributed Stochastic Neighbor Embedding) to visualize embeddings in a 2D space, tuning hyperparameters like perplexity. We then apply k-means clustering with $k = 8$, matching our eight regions. We will also calculate silhouette scores (ranging from -1 to 1) to measure cluster quality.

3.2.2 Bias Quantification

Bias is quantified using pairwise cosine similarities,

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|},$$

computed across and within regions. We conduct two main comparisons:

1. Image-to-Image Similarity for the Same Concept: We compare image embeddings from different countries for the same concept (e.g., "breakfast" in Japan vs. "breakfast" in Nigeria) to examine how similarly CLIP encodes these images across cultures.

2. Text-to-Image Similarity: We compare a fixed text embedding (e.g., "breakfast") to each country's image embeddings. This reveals how well a single textual concept aligns with diverse visual representations.

4 Expected Outcomes

Visualizations We expect to see distinct clustering patterns when we map the embeddings to two dimensions with t-SNE. Western embeddings (e.g., U.S. or Canadian "breakfast") may cluster more tightly, suggesting higher internal coherence, whereas non-Western embeddings (e.g., Nigerian "breakfast") may be more dispersed. When k-means clustering is performed with $k = 8$, silhouette scores might be above 0.5 for Western clusters and below 0.3 for non-Western ones, reinforcing the hypothesis that cultural bias is reflected in the latent space.

Bias Quantification We will measure pairwise cosine similarities in two ways. First, by comparing image embeddings for the same concept across different countries (e.g., Japan vs. Nigeria for "breakfast"), we can assess whether CLIP finds Western images more similar to each other (e.g., a similarity score of around 0.8 for U.S. vs. Canada) than Western–non-Western pairs (around 0.6 or lower for U.S. vs. India). Statistically significant differences (e.g., $p < 0.05$) would confirm cultural bias in the visual domain. Second, by comparing a text embedding (e.g., "breakfast") with each country's image embeddings, we can see whether the text encoder favors Western images. If non-Western images score similarly to Western ones for some concepts (e.g., festivals), that would indicate more balanced representations.

If image embeddings prove more biased, broadening the visual training data could help. If text embeddings show greater disparities, techniques like multilingual captions or translations might be needed to address language-based bias. Unexpected results, such as stronger clustering or higher similarity scores for certain non-Western concepts, would suggest better generalization than anticipated, prompting further exploration of data distribution factors. Ultimately, these findings will provide a clear diagnostic of how CLIP's latent space encodes cultural variations, informing future strategies for debiasing and advancing more equitable multimodal AI.

References

[1] Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint arXiv:2406.11665*, 2024.

[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] 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.

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

[5] Zhaoming Liu. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, (0), 2024.

[6] 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.

[7] 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.

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

[9] 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.

[10] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14830–14840, 2021.