Cultural Awareness in Vision-Language Models: A Cross-Country Exploration

Avinash Madasu[♡] Vasudev Lal[♡] Phillip Howard[♢†]
[♡]Intel Labs [♢]Thoughtworks
{avinash.madasu, vasudev.lal}@intel.com phillip.howard@thoughtworks.com

Abstract

Vision-Language Models (VLMs) are increasingly deployed in diverse cultural contexts, yet their internal biases remain poorly understood. In this work, we propose a novel framework to systematically evaluate how VLMs encode cultural differences and biases related to race, gender, and physical traits across countries. We introduce three retrieval-based tasks: (1) Race to Country retrieval, which examines the association between individuals from specific racial groups (East Asian, White, Middle Eastern, Latino, South Asian, and Black) and different countries; (2) Personal Traits to Country retrieval, where images are paired with trait-based prompts (e.g., Smart, Honest, Criminal, Violent) to investigate potential stereotypical associations; and (3) Physical Characteristics to Country retrieval, focusing on visual attributes like skinny, young, obese, and old to explore how physical appearances are culturally linked to nations. Our findings reveal persistent biases in VLMs, highlighting how visual representations may inadvertently reinforce societal stereotypes.

1. Introduction

Vision-Language Models (VLMs) [4, 5, 11, 17, 19–21], pre-trained on large-scale image-text datasets, have achieved state-of-the-art results on a variety of tasks, including image retrieval, captioning, and visual question answering (VQA). Despite these advancements, recent studies [3, 6–9, 22, 23] have highlighted the presence of gender, racial, and intersectional social biases in these models. As VLMs are increasingly deployed in socially and culturally sensitive contexts across the globe, ensuring that they capture a broader and more inclusive understanding of diverse cultures has become imperative. However, prior work [12–16, 18] has demonstrated that VLMs exhibit a strong bias toward Western-centric concepts in tasks such as text-to-image generation, image captioning, and object detection. In response, recent efforts [1, 2, 14] have introduced novel evaluation benchmarks to assess the multicultural competence of VLMs. Nonetheless, these benchmarks are often limited to evaluating between western and non-western countries. However, this binary choice of evaluation is reductive in nature often ignoring multiple cultures.

In this work, we propose a comprehensive evaluation framework to assess the cultural understanding of Vision-Language Models (VLMs). Specifically, we focus on six major racial and ethnic groups[*]: White, Black, South Asian, East Asian, Middle Eastern, and Latino, selecting the top ten countries with significant populations for each group. Using the FairFace [10] and Social Counterfactuals [8] datasets, we evaluate the distribution of countries retrieved by VLMs given an image of a person from a specific racial group. This allows us to investigate whether VLMs can accurately associate visual appearances with culturally appropriate geographic contexts. Additionally, we extend this analysis by pairing images with text prompts describing personal traits (e.g., honest, lazy, criminal, etc.) and physical characteristics (e.g., skinny, obese, old, etc.) attributed to different countries, and measure the proportion of countries associated with each racial group. This enables us to examine potential stereotypes or cultural biases embedded within VLMs when associating personal and physical traits with race and nationality.

2. Methodology

2.1. Task Definition

We propose a new task, namely **Retrieval across Countries**, aimed at evaluating the geographical and cultural understanding of Vision-Language Models (VLMs). Specifically, given an input image I of a person and a set of textual prompts describing various countries $T = T_1, T_2, T_3, ..., T_n$, the goal is to identify the country that the VLM most closely associates with the input

[†]Work completed while at Intel Labs.

[*]We use the terms race and ethnicity loosely in this work

Race /Ethnicity	Countries
East asian	China, Indonesia, Japan, Philippines, Vietnam, Thailand, Myanmar, South Korea, Malaysia, North Korea
White (Caucasian)	United States, Russia, Brazil, Germany, France, Italy, United Kingdom, Spain, Ukraine, Argentina
Middle Eastern	Egypt, Iran, Turkey, Iraq, Saudi Arabia, Yemen, Syria, Jordan, United Arab Emirates, Israel
Latino	Brazil, Colombia, Argentina, Peru, Venezuela, Chile, Ecuador, Bolivia, Paraguay, Uruguay
South Asian	India, Pakistan, Bangladesh, Nepal, Sri Lanka, Afghanistan, Bhutan, Maldives
Black	Nigeria, Ethiopia, Democratic Republic of Congo, United States, Brazil, South Africa, Tanzania, Kenya, Uganda, Ghana

Table 1. List of countries covered in the retrieval task for each race or ethnicity.

Task	Prompt
Race to Country	A person from China.
Physical Characteristics to Country	A skinny person from Peru.
Physical Traits to Country retrieval	A talented person from Kenya.

Table 2. Example prompts used in different cultural tasks.

image. We then measure the proportion of total images associated with a particular country, denoted as:

$$\text{R@C} = \frac{\sum_{i=1}^{N} \mathbb{1}\left[\arg\max_{j \in \mathcal{C}} \text{Sim}(I_i, T_j) = C\right]}{N}$$

where N is the total number of images, C is the set of all country text embeddings, I_j is image embedding for the i^{th} image, $Sim(I_i, T_j)$ is the similarity score between image I_i and T_j and $\mathbb{1}[\cdot]$ is an indicator function.

2.2. Tasks

We focus on three distinct tasks to evaluate the cultural understanding and potential biases exhibited by Vision-Language Models (VLMs). The countries and prompts used for each task are are shown in Tables 1 and 2 respectively.

Race to Country retrieval: This task measures how often VLMs associate images of individuals from specific racial groups (East Asian, White, Middle Eastern, Latino, South Asian, and Black) with different countries. It uncovers potential racial biases and stereotypes in the models' learned representations.

Personal Traits to Country retrieval: This task pairs images with prompts describing traits like Smart, Honest, Criminal, and Violent, and observes which countries are retrieved. It explores whether VLMs exhibit biases by linking positive or negative personal traits to specific countries or racial groups.

Physical Characteristics to Country retrieval: In this task, images based on physical attributes like skinny, young, obese, and old are used to analyze how VLMs associate physical appearance with countries. It helps reveal culturally-ingrained visual biases linked to body types and age.

2.3. Selection of Countries

For each racial/ethnic group, we select the top-10 most populated countries where these groups are predominantly found. Table 1 lists the countries considered in this study. This ensures a globally representative evaluation, focusing on regions with significant demographic presence for each racial/ethnic group, and providing a balanced basis for analyzing cultural bias in VLMs.

2.4. Datasets and Models

Datasets: FairFace [10] and SocialCounterFactuals [8].
Models: ALIP [20], LACLIP [5], OpenCLIP [4] and BLIP-2 [11].

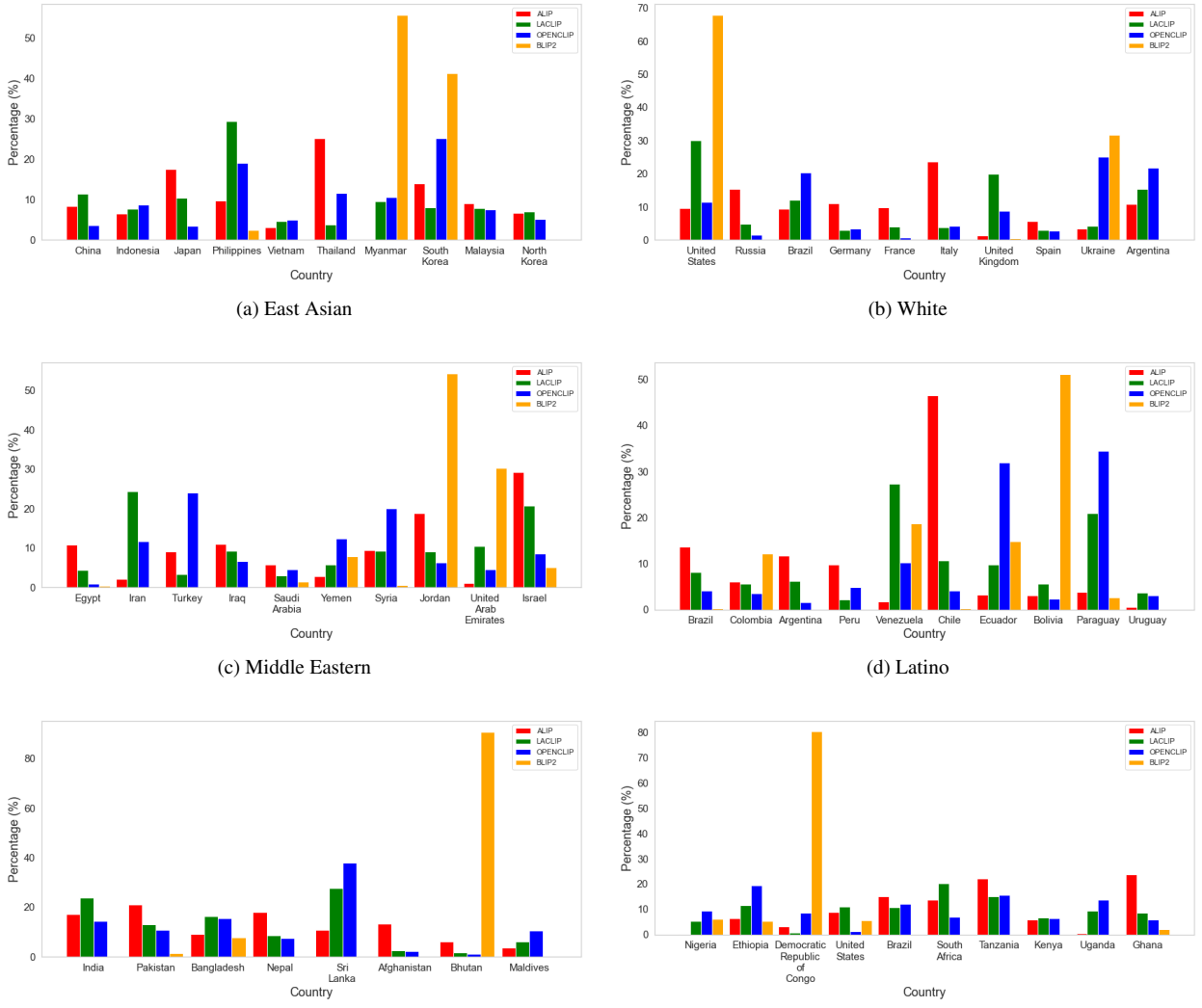
3. Results

3.1. Race to Country retrieval.

To investigate implicit regional biases in vision-language models (VLMs), we conducted a pair of country retrieval analyses conditioned on racial appearance, using two complementary datasets: FairFace and SocialCounterfactuals. Figure 1 shows the results on FairFace dataset[†]. In both cases, images of individuals from six racial categories—East Asian, White, Middle Eastern, Latino, South Asian, and Black—were paired with neutral country prompts, and the proportion of times each country was retrieved was measured. We evaluated four recent VLMs: ALIP, BLIP-2, LACLIP, and OpenCLIP.

In the FairFace setting, the results revealed distinct patterns of regional association that reflected both expected stereotypes and surprising biases. For example, BLIP-2 strongly associated White individuals with the United States (67%) and Black individuals with the Democratic Republic of Congo (80%), while OpenCLIP favored Ethiopia and Brazil for Black faces. LACLIP demonstrated more balanced retrieval across countries but still exhibited localized preferences, such as Iran for Middle Eastern and Venezuela for Latino identities. These findings highlighted the presence of learned socio-cultural and geographic biases within the latent representations

[†]Due to space limitation, results for socialcounterfactuals is in appendix



(a) East Asian

(b) White

(c) Middle Eastern

(d) Latino

(e) South Asian

(f) Black

Figure 1. **Performance comparison of different VLMs on various races and retrieved countries as a percentage of total images on FairFace dataset.**

Trait	ALIP						LACLIP					
	East Asian	South Asian	White	Middle Eastern	Latino	Black	East Asian	South Asian	White	Middle Eastern	Latino	Black
Smart	Philippines (45.05)	Nepal (29.22)	Russia (44.94)	Egypt (33.21)	Chile (67.68)	Ethiopia (35.97)	Philippines (29.49)	SriLanka (26.73)	Ukraine (22.77)	Iran (21.18)	Venezuela (30.93)	South Africa (21.92)
Honest	Philippines (40.4)	SriLanka (31.81)	Argentina (33.88)	Jordan (34.16)	Argentina (33.01)	South Africa (46.79)	Philippines (21.05)	SriLanka (27.46)	Ukraine (26.74)	Iran (25.52)	Paraguay (22.73)	South Africa (19.52)
Successful	North Korea (19.34)	Nepal (31.37)	Spain (27.71)	Egypt (43.73)	Chile (46.34)	South Africa (27.65)	Philippines (29.88)	Bangladesh (23.52)	Ukraine (21.38)	Iran (20.12)	Venezuela (23.32)	Ethiopia (19.73)
Talented	Philippines (35.49)	SriLanka (43.87)	Germany (25.87)	Jordan (22.56)	Brazil (39.08)	Brazil (28.87)	Philippines (30.48)	SriLanka (24.14)	USA (19.49)	Israel (19.08)	Ecuador (22.65)	Tanzania (20.51)
Creative	Japan (31.24)	SriLanka (40.06)	France (29.75)	Israel (30.99)	Chile (45.79)	Ethiopia (26.27)	Philippines (34.48)	SriLanka (29.29)	Ukraine (28.75)	Syria (23.24)	Venezuela (35.29)	Ethiopia (25.15)
Hardworking	China (16.9)	Bhutan (23.53)	Spain (24.92)	SaudiArabia (40.92)	Uruguay (27.65)	Tanzania (28.64)	Philippines (32.58)	SriLanka (26.85)	Ukraine (28.02)	Iran (30.73)	Venezuela (27.69)	South Africa (24.26)
Criminal	China (24.99)	SriLanka (50.05)	Germany (34.0)	Israel (23.22)	Chile (50.94)	SouthAfrica (32.24)	Philippines (21.09)	India (29.15)	Ukraine (17.6)	Iran (23.14)	Ecuador (16.63)	Ethiopia (18.86)
Lazy	Japan (36.66)	India (34.81)	Germany (53.56)	Israel (51.26)	Chile (37.23)	Ethiopia (49.49)	Philippines (25.09)	SriLanka (29.09)	Ukraine (47.73)	Iran (20.67)	Venezuela (27.23)	Ethiopia (23.53)
Dangerous	Japan (30.79)	SriLanka (39.07)	France (26.65)	Jordan (33.5)	Chile (37.4)	SouthAfrica (37.98)	Philippines (23.6)	Bangladesh (23.04)	Ukraine (22.13)	Israel (23.6)	Venezuela (31.08)	SouthAfrica (27.49)
Poor	Philippines (29.72)	Nepal (30.03)	France (45.57)	Jordan (36.28)	Peru (30.3)	SouthAfrica (58.94)	Indonesia (14.92)	SriLanka (23.35)	USA (20.36)	Iran (28.34)	Venezuela (40.09)	SouthAfrica (32.8)
Violent	China (25.9)	SriLanka (40.97)	Argentina (61.52)	Israel (37.41)	Argentina (39.5)	Kenya (23.73)	Philippines (18.14)	India (25.38)	USA (17.9)	Iran (24.86)	Venezuela (15.94)	Ethiopia (20.74)
Illiterate	Japan (26.15)	Pakistan (21.21)	Ukraine (34.66)	Israel (23.38)	Paraguay (28.44)	SouthAfrica (24.26)	Philippines (21.77)	India (22.7)	Ukraine (30.2)	Iran (22.25)	Venezuela (32.24)	SouthAfrica (22.82)

Table 3. **Comparison of personal traits across demographics for ALIP and LACLIP models on FairFace dataset.**

of VLMs. We observe similar set of results on social-counterfactuals dataset. Taken together, these analyses reveal that vision-language models encode stereotypical geographic associations.

Trait	OpenCLIP						BLIP-2					
	East Asian	South Asian	White	Middle Eastern	Latino	Black	East Asian	South Asian	White	Middle Eastern	Latino	Black
Smart	Myanmar (20.96)	SriLanka (25.36)	Brazil (25.19)	Syria (28.08)	Ecuador (31.45)	Ethiopia (29.42)	SouthKorea (61.75)	Bhutan (80.83)	Ukraine (54.03)	Jordan (85.63)	Colombia (63.17)	Ethiopia (49.11)
Honest	Myanmar (27.53)	Bangladesh (30.77)	Argentina (30.91)	Syria (30.77)	Ecuador (57.65)	Ethiopia (31.17)	Myanmar (31.88)	Bhutan (82.58)	Ukraine (39.81)	Jordan (98.59)	Colombia (76.05)	Ethiopia (71.45)
Successful	Myanmar (21.53)	Bangladesh (25.25)	Argentina (31.82)	Syria (35.76)	Ecuador (29.38)	Ethiopia (24.05)	Japan (48.71)	Bhutan (84.9)	Italy (30.59)	Jordan (82.78)	Bolivia (59.27)	Ethiopia (61.65)
Talented	Philippines (18.3)	SriLanka (26.07)	Argentina (25.28)	Syria (24.74)	Paraguay (29.99)	Ethiopia (23.0)	Japan (76.79)	SriLanka (46.65)	UK (52.33)	Jordan (97.71)	Colombia (85.52)	Ethiopia (74.85)
Creative	Philippines (21.15)	SriLanka (31.46)	Argentina (28.95)	Iraq (23.1)	Ecuador (35.36)	Ethiopia (26.79)	Myanmar (52.96)	Bhutan (51.21)	Germany (84.26)	Jordan (98.4)	Colombia (85.97)	DRC (62.67)
Hardworking	Philippines (22.64)	SriLanka (30.3)	Ukraine (20.72)	Syria (28.58)	Ecuador (23.86)	Brazil (24.12)	Philippines (72.86)	Bhutan (83.26)	Ukraine (65.05)	UAE (40.51)	Colombia (53.41)	DRC (91.03)
Criminal	Philippines (16.41)	SriLanka (31.21)	Brazil (25.06)	Turkey (23.78)	Ecuador (24.0)	Brazil (21.74)	Philippines (94.79)	Bhutan (85.59)	USA (54.04)	UAE (81.35)	Ecuador (67.34)	DRC (86.18)
Lazy	Philippines (17.82)	India (36.23)	Ukraine (21.27)	Syria (25.95)	Paraguay (26.83)	Ethiopia (26.35)	SouthKorea (57.65)	Bhutan (98.86)	Ukraine (36.21)	Jordan (97.93)	Colombia (55.99)	Ethiopia (71.66)
Dangerous	SouthKorea (25.48)	India (25.64)	Argentina (23.38)	Syria (22.27)	Ecuador (32.25)	Ethiopia (24.12)	SouthKorea (53.54)	Bhutan (90.72)	Ukraine (59.24)	Jordan (98.37)	Colombia (98.43)	Ethiopia (56.81)
Poor	Thailand (17.6)	SriLanka (30.01)	Argentina (20.86)	Israel (22.79)	Ecuador (30.05)	Brazil (24.66)	SouthKorea (50.91)	Bhutan (85.33)	Ukraine (51.3)	Jordan (94.9)	Colombia (64.49)	DRC (88.44)
Violent	Philippines (15.86)	India (30.04)	Argentina (24.38)	Iran (18.92)	Ecuador (30.38)	Ethiopia (20.72)	Philippines (88.2)	Bhutan (62.38)	Ukraine (56.99)	UAE (95.8)	Colombia (95.8)	DRC (59.36)
Illiterate	SouthKorea (18.36)	SriLanka (32.67)	Ukraine (25.29)	Israel (29.16)	Venezuela (21.05)	Brazil (29.89)	SouthKorea (85.42)	Bhutan (71.69)	Ukraine (83.46)	SaudiArabia (49.15)	Colombia (58.8)	DRC (80.27)

Table 4. **Comparison of personal traits across demographics for OpenCLIP and BLIP-2 models on FairFace dataset.**

3.2. Personal Traits to Country Retrieval.

Next, we assess the cultural understanding of personal traits across racial groups using FairFace dataset. In an image-text matching setup, we prompted models with *"A [trait] person from [country]"*, measuring the maximum retrieval probability across different races (East Asian, South Asian, White, Middle Eastern, Latino, Black) for both positive and negative personal traits. Tables 3 and 4 show the results of this setup.

3.2.1. Positive Traits

Regional Consistency: ALIP and LACLIP consistently retrieved *Philippines*, *Sri Lanka*, and *Nepal* for positive traits associated with East Asians and South Asians, respectively. Similarly, positive associations for Black individuals centered around *Ethiopia*, *South Africa*, and *Brazil*.

Model-Specific Trends: BLIP-2 exhibited a striking over-association, retrieving *Bhutan* for nearly all positive traits among South Asians, often with extremely high percentages (80–98%). This suggests a model collapse toward a singular national representation for a racial group.

White Race Associations: Unlike traditional Western dominance, *Ukraine* surfaced as the most frequent country associated with positive traits for Whites across LACLIP, OpenCLIP, and BLIP-2, indicating a potential data or alignment shift toward Eastern Europe.

Latino and Middle Eastern Groups: Positive traits for Latino individuals were commonly linked to *Chile* and *Ecuador*, while Middle Eastern individuals were most associated with *Egypt*, *Jordan*, and *Iran*.

3.2.2. Negative Traits

Stereotypical Reinforcement: Countries such as *Philippines* (East Asian) and *India* (South Asian) were frequently retrieved for traits like *criminal* and *lazy* across all models, exposing latent stereotypes embedded in VLMs.

Black Race Negative Bias: In ALIP, LACLIP, and OpenCLIP, *South Africa* dominated the negative trait retrievals. However, BLIP-2 exhibited an even stronger and problematic bias, overwhelmingly associating all negative traits with the *Democratic Republic of Congo (DRC)*, further highlighting a collapse into a singular representation.

Middle Eastern and Latino Trends: Middle Eastern countries such as *Israel*, *Iran*, and *Jordan* were frequently retrieved for *criminal* and *dangerous*, pointing to the persistence of geopolitical stereotypes. Among Latinos, *Paraguay*, *Argentina*, and *Ecuador* were common retrievals for negative traits.

3.2.3. Model Comparison

ALIP and LACLIP demonstrated relatively greater diversity and cultural nuance in their associations compared to OpenCLIP and BLIP-2. Although biases were still present, they were less extreme. OpenCLIP tended toward flattening associations but still retrieved stereotypical mappings. BLIP-2 showed the most severe overgeneralization for both positive and negative traits within racial categories, indicating critical fairness and robustness issues.

4. Conclusion

In this work we proposed a comprehensive framework for cultural evaluation of VLMs. Specifically we evaluated four models on three tasks such as race to country retrieval, physical traits to country retrieval and physical characteristic to country retrieval. Our results show that VLMs exhibhit stereotypical positive and negative biases towards just a few set of countries. These results reveal the lack of cultural understanding of VLMs across different races and nationalities.

References

[1] Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. See it from my perspective: How language affects cultural bias in image understanding. In *The Thirteenth International Conference on Learning Representations*. 1

[2] Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eun-Jeong Hwang, and Vered Shwartz. 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, 2024. 1

[3] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023. 1, 2

[5] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023. 1, 2

[6] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6957–6966, 2023. 1

[7] Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *arXiv preprint arXiv:2301.11100*, 2023.

[8] Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11985, 2024. 1, 2

[9] Sepehr Janghorbani and Gerard De Melo. Multi-modal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, 2023. 1

[10] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 1, 2

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[12] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, 2021. 1

[13] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351, 2023.

[14] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Steenkiste, Lisa Hendricks, Karolina Stanczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, 2024. 1

[15] Joan Nwatu, Oana Ignat, and Rada Mihalcea. Bridging the digital divide: Performance variation across socioeconomic factors in vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10686–10702, 2023.

[16] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37: 106474–106496, 2024. 1

[17] 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. 1

[18] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. 1

[19] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022. 1

[20] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023. 2

[21] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

[22] 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. 1

[23] Kankan Zhou, Yibin LAI, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. Association for Computational Linguistics, 2022. 1

Cultural Awareness in Vision-Language Models: A Cross-Country Exploration

Supplementary Material

5. Additional Results

5.1. VLMs encode persistent racial-regional biases across datasets

To further test the robustness of these associations, we performed analysis on socialcounterfactuals dataset while retaining the same racial categories and neutral prompts. Figure 2 shows the results on this dataset. The results reaffirmed the presence of racialized regional biases, with BLIP-2 again showing sharp, isolated associations—favoring the United States for White individuals (57.7%) and the Democratic Republic of Congo for Black individuals (45.6%). ALIP distributed retrievals more evenly, though stereotypes persisted, linking Asian faces predominantly to the Philippines (39.8%) and South Asian faces to India (19.3%) and Sri Lanka (24.5%). OpenCLIP demonstrated persistent preferences, notably associating Black faces with Nigeria (47.9%) and Indian faces with India (45.4%) and Sri Lanka (43.6%). LACLIP maintained a comparatively balanced retrieval pattern, though regional clustering was still apparent—such as Asian faces around China and the Philippines, and Middle Eastern faces toward Iran and Turkey.

5.2. Personal Traits to Country Retrieval.

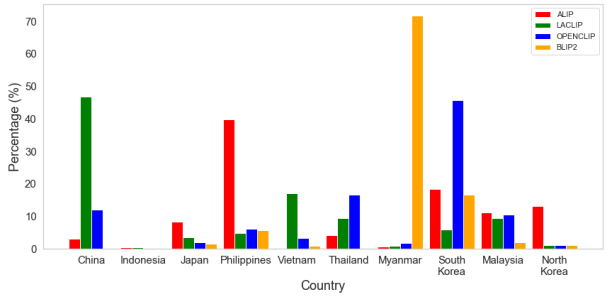
We also experiment personal traits to country retieval on SocialCounterFactuals dataset. Tables 5 and 6 show the results for ALIP, LACLIP, OpenCLIP and BLIP-2 models. The findings reveal notable patterns of cultural attribution and bias across models. In ALIP, positive traits like *smart*, *honest*, *successful*, and *talented* for East Asians were most associated with the Philippines, suggesting a strong stereotypical mapping in the model's representation. South Asians were similarly associated with positive traits but through a different set of countries such as Pakistan, Afghanistan, and India, though with slightly lower confidence percentages compared to East Asians. White individuals were predominantly linked with European countries like Russia, Germany, and France for positive traits, reflecting a Western-centric bias in competence-related attributes. For Middle Eastern and Latino groups, associations were more mixed, but traits like *successful* and *hardworking* were often linked to countries like UAE, Peru, and Uruguay, indicating a modest positive bias.

Negative traits, however, showed clearer evidence of problematic stereotyping. In ALIP, *criminal*, *lazy*, and *dangerous* traits were more frequently associated with South Asian and Black images. For instance, Black individuals were heavily associated with the Democratic Republic of Congo (DRC) across multiple negative traits such as *dangerous* and *criminal*. Latino groups showed a tendency to be matched with lower economic status attributes such as *poor* or *illiterate*, with countries like Uruguay and Venezuela frequently appearing.

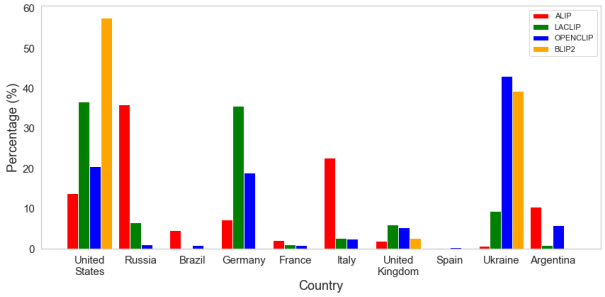
LACLIP displayed both continuities and deviations from ALIP. Although it generally reduced the extent of extreme associations, it still presented patterns of cultural bias. Notably, LACLIP showed higher positive associations for Middle Eastern individuals compared to ALIP, linking them with traits like *hardworking* and *successful* with countries such as UAE at very high confidence scores ($> 80\%$). Nonetheless, Black individuals continued to be strongly associated with Nigeria and Ethiopia for negative traits, suggesting persistent racial biases that are resistant across architectures.

OpenCLIP presented a somewhat different picture. The model was relatively more consistent in associating positive traits with South Asians, predominantly India, across almost all positive attributes. Confidence scores were notably high, often exceeding 70%, for positive traits such as *smart*, *successful*, and *hardworking*. However, for negative traits, Black and Middle Eastern groups were still more heavily penalized, with countries like Nigeria, South Africa, and UAE frequently appearing for *criminal* and *dangerous* traits. OpenCLIP's results suggest that while it is capable of strong positive cultural attributions, it remains vulnerable to negative racial profiling.

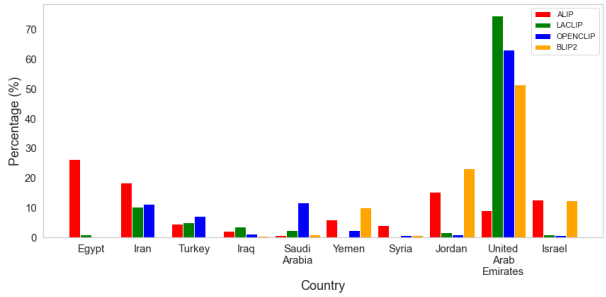
BLIP-2, the most recent and arguably the most powerful model among the four, exhibited the most polarized results. On positive traits, Latino individuals were overwhelmingly associated with Colombia across traits like *smart*, *successful*, *creative*, and *hardworking*, with extremely high matching percentages often above 75%. This points towards a more homogenized cultural mapping in BLIP-2, where a single country dominates the representation of positive traits for a racial group. For Black individuals, the Democratic Republic of Congo continued to be linked with both positive (*hardworking*, *creative*) and negative (*dangerous*, *criminal*) traits. Interestingly, in contrast to other models, BLIP-2 associated Black individuals with positive traits at a higher rate than earlier models but still maintained stereotypes for negative traits, indicating a partial but incomplete mitigation of bias.



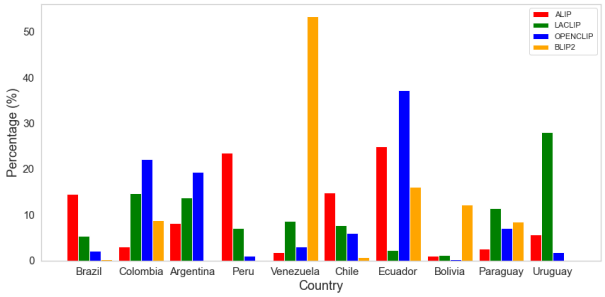
(a) East Asian



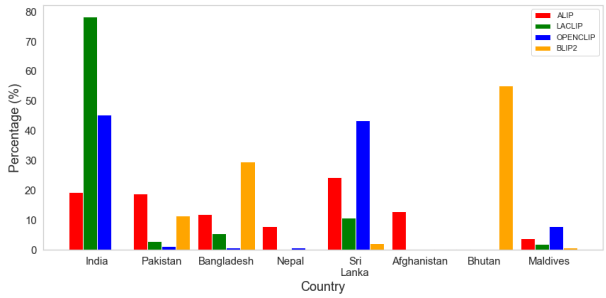
(b) White



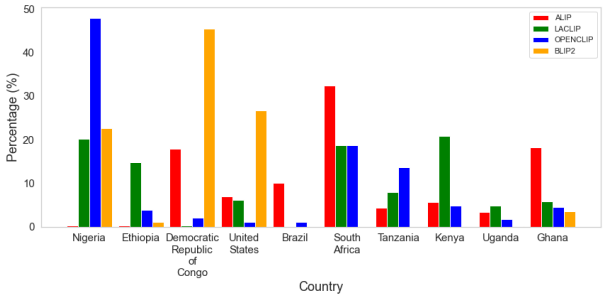
(c) Middle Eastern



(d) Latino



(e) South Asian



(f) Black

Figure 2. **Performance comparison of different VLMs on various races and retrieved countries as a percentage of total images on Socialcounterfactuals dataset**

Trait	ALIP						LACLIP					
	East Asian	South Asian	White	Middle Eastern	Latino	Black	East Asian	South Asian	White	Middle Eastern	Latino	Black
Smart	Philippines (49.46)	Pakistan (44.3)	Russia (58.96)	Egypt (23.48)	Chile (26.77)	DRC (49.79)	China (32.67)	India (69.4)	Ukraine (36.31)	UAE (72.13)	Peru (33.23)	Nigeria (24.43)
Honest	Philippines (65.39)	Afghanistan (35.19)	Argentina (37.35)	Jordan (38.6)	Peru (41.05)	South Africa (24.22)	China (35.96)	India (58.42)	USA (34.44)	UAE (83.93)	Peru (28.78)	Nigeria (23.77)
Successful	Philippines (31.68)	India (22.35)	Germany (43.79)	Egypt (51.85)	Peru (32.21)	DRC (37.28)	China (34.39)	India (44.85)	Ukraine (44.66)	UAE (64.18)	Peru (37.27)	Nigeria (27.92)
Talented	Philippines (77.8)	SriLanka (33.0)	Italy (33.22)	UAE (31.43)	Brazil (39.44)	DRC (30.73)	China (53.61)	India (44.63)	USA (31.33)	UAE (68.53)	Peru (46.08)	Nigeria (26.15)
Creative	Philippines (30.73)	SriLanka (27.51)	Russia (37.04)	Israel (29.79)	Chile (35.81)	DRC (45.76)	China (38.43)	India (39.35)	Ukraine (42.65)	UAE (75.96)	Peru (26.93)	Uganda (26.25)
Hardworking	Philippines (56.22)	Bangladesh (46.68)	Russia (29.85)	UAE (37.36)	Uruguay (50.77)	DRC (63.31)	China (32.62)	India (60.29)	Ukraine (40.29)	UAE (81.82)	Uruguay (28.45)	Nigeria (24.55)
Criminal	Philippines (29.51)	SriLanka (40.22)	France (35.81)	Turkey (17.54)	Bolivia (26.18)	SouthAfrica (34.27)	China (54.76)	India (82.5)	Germany (42.91)	UAE (76.84)	Peru (42.16)	Ethiopia (27.99)
Lazy	Japan (38.93)	India (27.62)	Germany (53.47)	Turkey (33.39)	Uruguay (39.35)	SouthAfrica (63.41)	China (32.76)	India (60.23)	Ukraine (47.29)	UAE (83.8)	Paraguay (21.02)	Ethiopia (29.56)
Dangerous	Philippines (29.93)	SriLanka (40.55)	France (32.47)	Jordan (39.91)	Brazil (57.53)	Brazil (32.03)	China (41.69)	India (63.33)	Germany (31.63)	UAE (79.33)	Uruguay (33.53)	Ethiopia (25.68)
Poor	China (28.76)	India (29.5)	Germany (45.61)	Jordan (52.71)	Uruguay (32.02)	SouthAfrica (25.47)	China (44.13)	India (68.46)	Germany (43.17)	UAE (79.4)	Colombia (29.98)	SouthAfrica (35.84)
Violent	Philippines (28.21)	Maldives (34.29)	Argentina (54.56)	Egypt (25.55)	Argentina (39.03)	Brazil (44.42)	China (53.13)	India (79.21)	Germany (35.5)	UAE (59.12)	Peru (42.7)	Uganda (32.21)
Illiterate	Philippines (31.97)	SriLanka (44.11)	USA (51.66)	UAE (37.82)	Venezuela (40.95)	DRC (36.53)	China (48.25)	India (69.21)	USA (30.93)	UAE (83.18)	Uruguay (38.52)	Ethiopia (30.0)

Table 5. **Comparison of personal traits across demographics for ALIP and LACLIP models on SocialCounterFactuals dataset.**

Across all models, White individuals were frequently linked with Western nations like Germany, Ukraine, and the USA for both positive and, to a lesser extent, neutral or mildly negative traits. There were comparatively fewer

Trait	OpenCLIP						BLIP-2					
	East Asian	South Asian	White	Middle Eastern	Latino	Black	East Asian	South Asian	White	Middle Eastern	Latino	Black
Smart	Vietnam (50.14)	India (74.22)	Germany (28.99)	UAE (61.31)	Ecuador (38.94)	Nigeria (28.63)	SouthKorea (63.8)	Bangladesh (45.06)	USA (43.57)	Jordan (54.66)	Colombia (75.48)	DRC (54.67)
Honest	China (25.38)	India (61.08)	Ukraine (70.04)	UAE (58.07)	Ecuador (56.28)	Nigeria (61.44)	Philippines (31.27)	SriLanka (56.5)	Ukraine (49.03)	Jordan (84.39)	Colombia (82.25)	Ghana (38.25)
Successful	Vietnam (52.53)	India (64.79)	Ukraine (37.12)	UAE (53.7)	Colombia (39.96)	South Africa (35.01)	Philippines (50.35)	SriLanka (60.99)	Ukraine (31.92)	Jordan (78.52)	Colombia (62.15)	DRC (51.66)
Talented	Vietnam (36.44)	India (74.93)	USA (35.23)	UAE (71.02)	Venezuela (38.48)	Nigeria (56.01)	Japan (62.01)	SriLanka (78.9)	USA (42.48)	Jordan (92.33)	Colombia (77.79)	Ghana (46.29)
Creative	Vietnam (40.4)	India (69.2)	Ukraine (50.37)	UAE (79.23)	Ecuador (28.86)	Nigeria (50.39)	Myanmar (27.34)	SriLanka (76.51)	Germany (58.39)	Jordan (86.45)	Colombia (70.5)	DRC (61.45)
Hardworking	SouthKorea (39.74)	India (78.11)	Germany (58.52)	UAE (41.11)	Colombia (28.6)	South Africa (55.61)	Philippines (66.12)	SriLanka (42.85)	Ukraine (48.15)	UAE (50.67)	Colombia (65.45)	DRC (76.48)
Criminal	China (41.41)	India (77.84)	Germany (39.28)	UAE (55.25)	Ecuador (33.22)	SouthAfrica (62.44)	Philippines (75.96)	Bhutan (39.16)	USA (63.67)	UAE (90.01)	Ecuador (42.65)	DRC (58.6)
Lazy	North Korea (30.15)	India (82.77)	Ukraine (61.2)	UAE (36.04)	Ecuador (44.3)	Nigeria (63.24)	Myanmar (36.49)	Bhutan (87.9)	Ukraine (38.21)	Jordan (82.48)	Colombia (63.41)	Ghana (37.45)
Dangerous	SouthKorea (41.82)	India (70.89)	USA (53.89)	UAE (72.34)	Venezuela (44.43)	Nigeria (60.37)	SouthKorea (42.36)	Bhutan (50.87)	Ukraine (53.13)	Jordan (93.57)	Colombia (84.51)	Ghana (69.75)
Poor	China (29.75)	India (49.78)	Germany (50.84)	UAE (59.56)	Ecuador (28.52)	SouthAfrica (40.53)	SouthKorea (40.32)	Bhutan (40.61)	United Kingdom (34.25)	Jordan (87.54)	Colombia (66.44)	DRC (80.49)
Violent	China (42.44)	India (73.48)	Germany (68.78)	UAE (43.76)	Peru (39.24)	Ethiopia (36.09)	Philippines (69.75)	Bhutan (46.66)	Ukraine (45.02)	UAE (85.15)	Colombia (94.97)	DRC (48.42)
Illiterate	SouthKorea (24.74)	SriLanka (37.97)	USA (50.88)	UAE (68.05)	Chile (28.79)	SouthAfrica (47.41)	SouthKorea (75.48)	SriLanka (46.76)	Ukraine (74.19)	SaudiArabia (46.74)	Colombia (40.69)	DRC (60.24)

Table 6. **Comparison of personal traits across demographics for OpenCLIP and BLIP-2 models on SocialCounterFactuals dataset.**

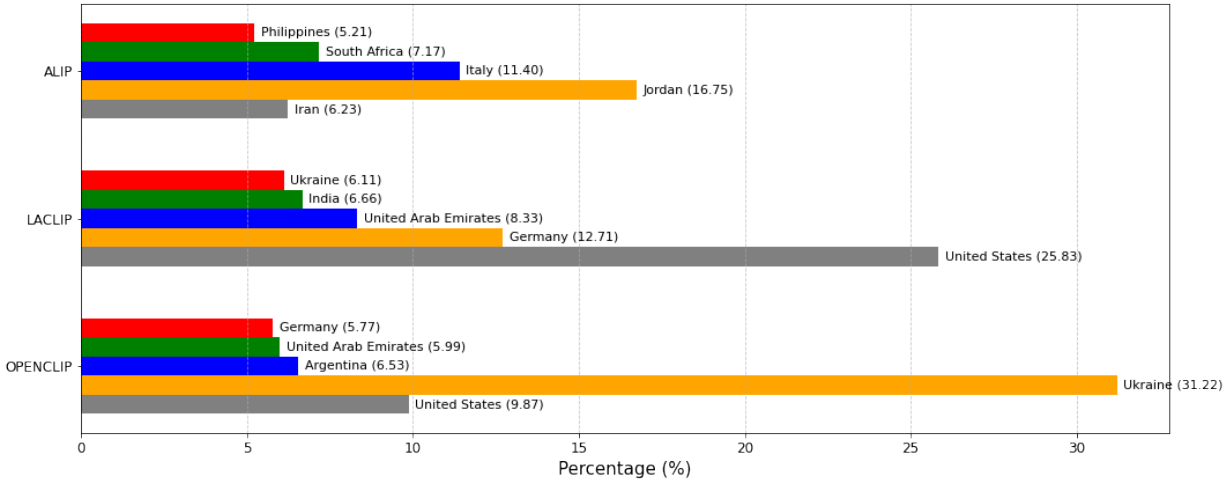


Figure 3. **Figure shows the top-5 countries retrieved for ALIP, LACLIP and OPENCLIP. The images retrieved are for skinny physical characteristics.**

strong associations with overtly negative traits like *criminal* or *lazy*, reflecting an underlying racial advantage embedded in the visual-textual representation learned by these models.

Although models like LACLIP and BLIP-2 show improvements in associating positive traits across diverse groups, a persistent trend remains where negative attributes are disproportionately associated with South Asian, Middle Eastern, and Black racial representations. Moreover, the dominance of specific countries such as the Philippines, India, Colombia, and the Democratic Republic of Congo across several traits highlights the risk of cultural oversimplification and stereotyping in vision-language models.

5.3. Analysis of country retrievals across body types

We study the cultural association of VLMs to physical characteristics of persons. Specifically we evaluate body types such as skinny, young, obese, tattooed and old. We use the SocialCounterFactuals dataset for this setup where the input is an image of a person and match it with prompts from all the countries mentioned in table 1. For each model, we then select the top-5 countries retrieved as a measure of the highest percentage. The results are shown in the figures 3, 4, 5, 6 and 7 respectively.

The experimental results reveal distinct patterns in how different vision-language models (ALIP, LACLIP, and OPENCLIP) associate body types with specific countries, highlighting potential cultural biases embedded in these systems.

5.3.1. Country association patterns

Our analysis across five body types (skinny, young, obese, tattooed, and old) demonstrates consistent bias patterns within each model family, but with notable variations between models. OPENCLIP shows a strong tendency to associate Ukraine with multiple body types, particularly in the skinny, obese, and tattooed categories (31.22%, 31.33%, and 24.83% respectively). This over-

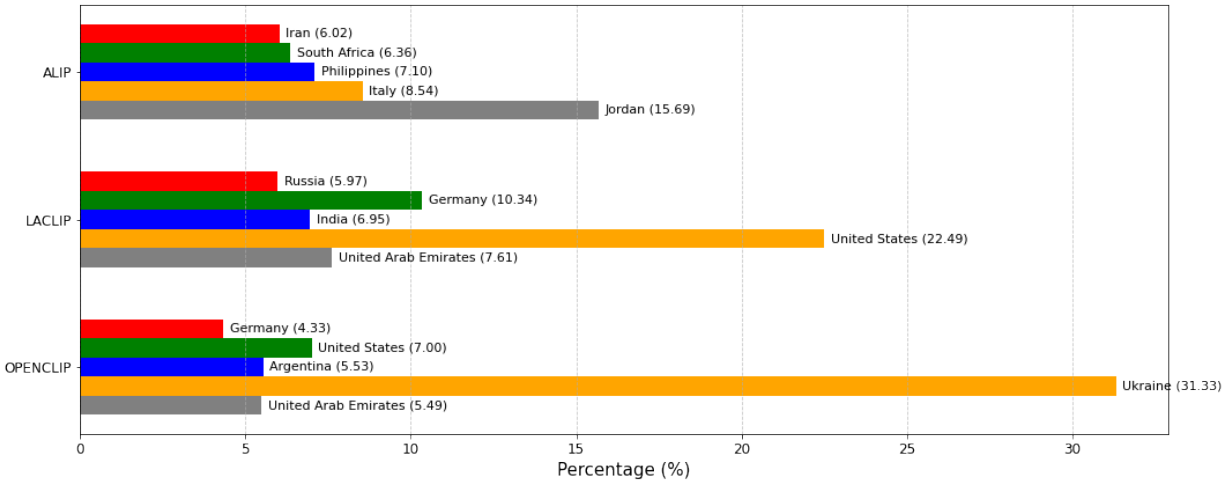


Figure 4. **Figure shows the top-5 countries retrieved for ALIP, LACLIP and OPENCLIP. The images retrieved are for young physical characteristics.**

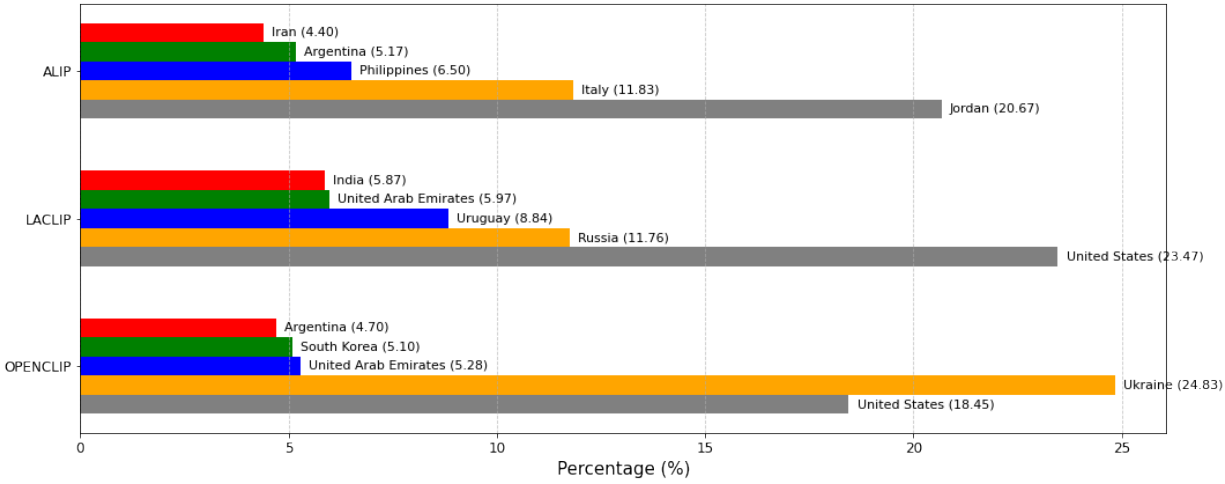


Figure 5. **Figure shows the top-5 countries retrieved for ALIP, LACLIP and OPENCLIP. The images retrieved are for obese physical characteristics.**

representation suggests potential dataset biases or embedding space distortions specific to this model architecture. The United States appears prominently across all models, with particularly high representation in LACLIP results (25.83% for skinny, 22.49% for young, and 25.47% for obese). This Western-centric bias likely reflects training data composition that oversamples American or Western imagery, creating an implicit association between various body types and American identity.

5.3.2. Body type specific associations

For the skinny body type (Image 1), Jordan shows significant representation in ALIP (16.75%), while LACLIP heavily favors the United States (25.83%). This divergence suggests fundamentally different conceptual associations between thinness and national identity across model architectures. Young body types (Image 2) show different top countries across models, with Jordan dominating in ALIP (15.69%), the United States in LACLIP (22.49%), and Ukraine in OPENCLIP (31.33%). This heterogeneity in results indicates unstable representations of youth across cultural contexts. For obese body types (Image 3), Jordan (20.67%) and the United States (25.47%) dominate ALIP and LACLIP respectively, while OPENCLIP presents a more balanced distribution between Ukraine (24.83%) and the United States (18.45%). These associations may reflect stereotypical Western representations of obesity prevalence. Tattooed bodies (Image 4) show strong associations with Italy in ALIP (14.66%), while LACLIP associates them primarily

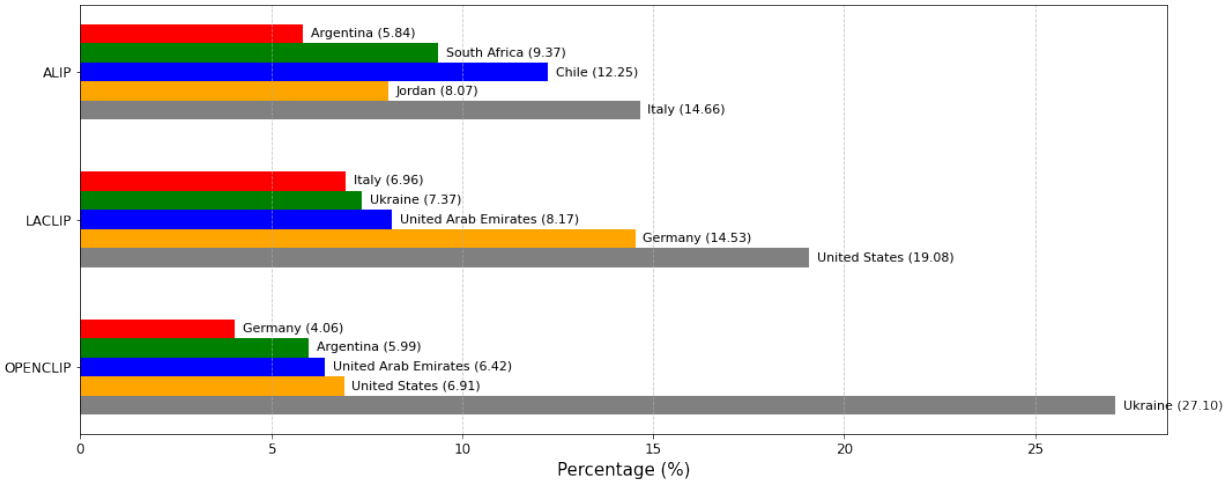


Figure 6. **Figure shows the top-5 countries retrieved for ALIP, LACLIP and OPENCLIP. The images retrieved are for skinny tattooed characteristics.**

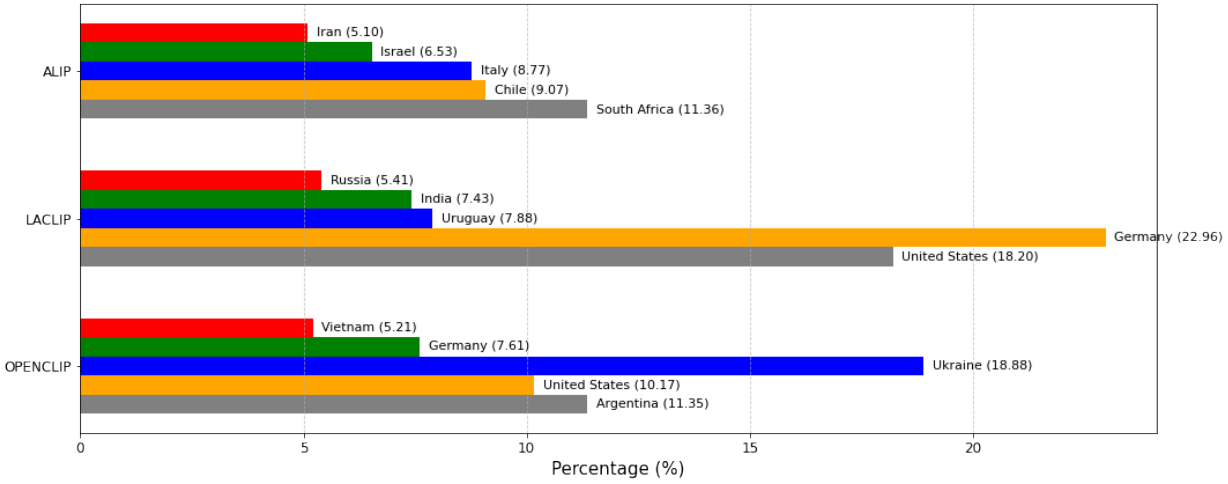


Figure 7. **Figure shows the top-5 countries retrieved for ALIP, LACLIP and OPENCLIP. The images retrieved are for old physical characteristics.**

with the United States (19.08%) and Germany (14.53%). OPENCLIP strongly associates tattoos with Ukraine (27.10%). These variations likely reflect different cultural narratives about body modification across training datasets. Elderly representation (Image 5) shows associations with South Africa (11.36%) in ALIP, while Germany dominates LACLIP results (22.96%) and Ukraine leads in OPENCLIP (18.88%). These variations suggest different cultural framings of aging across the models' training data.