Project Report CSCI544 - Group 19
Real-World Bias in AI: Assessing and Mitigating Stereotypes in Multimodal Models

Archana Sondur[1], **Hitanshu Panchal**[2], **Nehal Garg**[3], **Pranav Padhye**[4], **Vishwas Parekh**[5]
University of Southern California, USA
[1]sondur@usc.edu, [2]hpanchal@usc.edu, [3]nehalgar@usc.edu,
[4]ppadhye@usc.edu, [5]vdparekh@usc.edu

Abstract

Multimodal AI systems such as vision-language models (VLMs) are increasingly integrated into high-stakes domains, yet they remain vulnerable to learning and perpetuating societal biases. While curated benchmarks like MMBias provide a controlled framework for evaluation, they fail to reflect the complexities of real-world data. In this work, we curate a diverse dataset of 2,500 image-text pairs from social and news platforms to assess the bias encoded by OpenAI's CLIP model. Using cosine similarity and Caliskan-style effect size computations, we quantify bias across 20+ demographic subgroups. We also explore and evaluate adversarial fine-tuning as a bias mitigation strategies. Our findings underscore the limitations of benchmark datasets and highlight the need for robust real-world evaluations to foster fairer AI systems.

1 Introduction

Vision-language models (VLMs), such as CLIP, represent a significant advancement in multimodal AI, aligning visual and textual modalities in a shared embedding space. These models power a variety of applications ranging from search engines to content moderation systems. However, despite their capabilities, they remain susceptible to learning harmful associations present in training data. When deployed in real-world settings, this can lead to disproportionately negative consequences for historically marginalized groups.

Current evaluation frameworks, such as MMBias (**?**), aim to measure model bias but rely on highly curated datasets that fail to capture the messiness of real-world data—data that includes ambiguity, diverse contexts, and inherent societal noise. As such, there is a growing consensus in the research community: to ensure fairness in real-world deployments, models must be evaluated in real-world conditions.

To address this gap, our project introduces a dataset drawn from platforms such as Twitter, Reddit, BBC news, CNN news, and Unsplash, annotated across more than 20 identity subgroups. We use this dataset to probe CLIP (Radford et al., 2021) for bias and compare its behavior with established benchmarks like MMBias. Furthermore, we apply and assess the effectiveness of bias mitigation techniques (Radford et al., 2021), contributing actionable insights to the field of AI fairness.

2 Related Work

Bias in artificial intelligence systems has been widely studied, particularly in language models, where most research focuses on gender and racial biases (Guo and Caliskan, 2021)(Bordia and Bowman, 2019). Some studies have also examined skew along other axes such as profession, religion, and disability, thus demonstrating that unfair associations are not confined to text alone (Nadeem et al., 2021)(Hutchinson et al., 2020). Indeed, image-classification networks and multimodal vision–language models likewise exhibit systematic preferences and stereotypes when mapping visual inputs to labels or textual embeddings.

In the purely textual domain, seminal work by Caliskan et al. (2017) introduced the Word Embedding Association Test (WEAT) to quantify implicit associations in word vectors (Caliskan et al., 2017). Building on WEAT, more recent efforts have adapted association-testing methodologies to joint text–image representations, probing how models like CLIP recombine visual and linguistic concepts in ways that may reinforce societal biases.

To broaden the scope of bias benchmarks, Janghorbani and De Melo (2023) proposed a unified probing framework MMBias that extends association tests across multiple demographic axes in multimodal encoders Janghorbani and De Melo (2023). Shortly thereafter, Wang et al. introduced the MMBias benchmark itself: a curated collection of roughly 3,800 image–phrase pairs covering fourteen social categories (religion, nationality, disability, sexual orientation, etc.) Janghorbani and

De Melo (2023). Their evaluation of models such as CLIP, ALBEF, and ViLT reveals consistent favoring of certain subgroups, and they demonstrate a post-hoc debiasing routine that significantly reduces measured bias with minimal impact on downstream task accuracy.

Despite these advances, existing multimodal bias studies predominantly rely on synthetic or highly curated datasets and focus mainly on gender/race. Such controlled benchmarks can miss intersectional effects and the unpredictable noise of "real-world" content. Moreover, mitigation strategies that have shown promise in text-only contexts—like adversarial fine-tuning (Nadeem et al., 2021) have yet to be thoroughly explored in vision–language models. Our work addresses these gaps by introducing a large-scale social-media dataset for bias evaluation and adapting adversarial debiasing methods to multimodal encoders.

3 Project Scope and Research Objective

As multi-modal models like CLIP gain traction in high-impact applications, from content moderation and recommendation systems to surveillance and automated hiring, their ability to interpret human-centric content fairly and accurately becomes critically important. Therefore, the goal of the project is twofold: first, to evaluate how a widely used vision language model like Contrastive Language Image Pretraining (CLIP) encodes social and cultural biases when exposed to noisy, real-world multimodal data. Second, explore and assess potential strategies to mitigate these biases in practical deployment scenarios.

Unlike much of the existing literature, which compares multiple architectures or relies on synthetic benchmark datasets, our work emphasizes the influence of dataset quality and composition as a core driver of bias. We deliberately maintain a fixed model architecture (CLIP) and instead manipulate the dataset environment to isolate the effects of real-world noise, ambiguity, and representational imbalance. This approach allows us to disentangle model-specific biases from those arising due to the statistical or sociocultural characteristics of input data.

Furthermore, our project goes beyond bias detection to actively explore avenues for bias mitigation. To ensure fair benchmarking and accurate comparisons, we apply a scoring alignment strategy based on the Caliskan effect size formula, replicating the target-attribute sets used in MMBias Janghorbani and De Melo (2023) and normalizing cosine similarity scores. This helps reduce variance introduced by embedding scale and ensures consistency across experimental conditions. In parallel, we develop a robust cleaning pipeline utilizing tools such as OpenCV, DLib, and manual filtering to curate the dataset. This process helps eliminate noisy, ambiguous, or potentially harmful samples, while ensuring that each identity group is represented with cultural and ethical sensitivity. Together, these techniques offer practical and scalable approaches to minimizing bias in multimodal systems without requiring model retraining

Overall, by emphasizing real-world conditions, our work provides a more grounded and application-relevant understanding of how multimodal models like CLIP encode and act upon biased representations. We aim to contribute to the growing body of research advocating for fairness in AI, while also offering actionable insights for practitioners who rely on these models in real-world systems.

4 Methodology

4.1 Dataset Construction

We assembled a dataset consisting of approximately 2,500 image-text pairs sourced from diverse public platforms including Twitter, Reddit, CNN, BBC, and Unsplash. These sources were selected to capture a wide spectrum of real-world content with varying levels of formality, tone, and visual composition. A combination of platform-specific APIs, keyword-based filtering, and manual verification was used to ensure demographic diversity. The dataset was curated to include more than 20 identity subgroups spanning race, gender, nationality, religion, disability status, and sexual orientation. To increase analytical granularity, broad categories such as "LGBTQ" were further disaggregated into distinct identities (e.g., gay, lesbian, transgender, bisexual, non-binary). This level of detail enables a more nuanced investigation into intersectional bias.

4.2 Data Curation and Preprocessing

Given the inherent noise and inconsistency in user-generated content, we employed a rigorous data curation pipeline. Content was manually reviewed to remove instances that were semantically ambiguous, visually degraded (e.g., blurry, low-resolution, watermarked), or overly politicized in nature. This step was crucial to reduce confounding variables that could distort bias evaluation. Visual preprocessing was performed using computer vision libraries such as OpenCV and DLib to standardize image dimensions, align faces (where applicable), and extract relevant visual features. All curated samples were paired with structured metadata—capturing subgroup labels, source attribution, and preprocessing flags—stored in JSON format to support traceability and modular analysis.

4.3 Bias Evaluation Framework

We evaluated model bias using OpenAI's CLIP model (ViT-B/32), which jointly embeds image and text inputs into a shared multimodal space. For each image in the dataset, cosine similarity scores were computed against evaluative text prompts structured as semantically contrasting pairs (e.g., "This is a kind person" vs. "This is a dangerous person"). These prompts were inspired by the Word Embedding Association Test (WEAT), originally designed for probing bias in word embeddings. Following the WEAT protocol, we computed effect sizes—specifically Caliskan-style pairwise scores—to quantify the degree of preferential association between identity subgroups and specific attributes. This allowed us to measure both the direction and intensity of bias across multiple demographic axes.

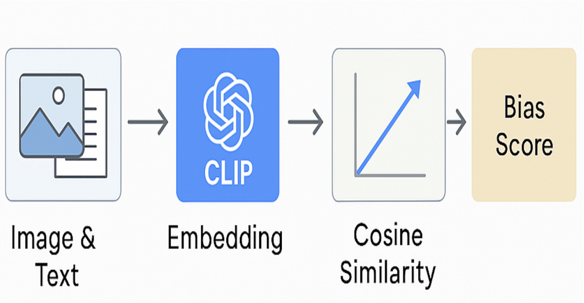


Figure 1: The diagram illustrates the process of embedding image-text pairs using CLIP, followed by cosine similarity evaluation against prompt pairs to compute final bias scores.

4.4 Debiasing Technique

To mitigate the biases uncovered during evaluation, we implemented and tested a complementary techniques within a controlled experimental pipeline:

Adversarial Fine-Tuning: We fine-tuned CLIP's image encoder in conjunction with a bias adversary network. The adversary was trained to maximize the model's subgroup predictability from its embeddings, while the encoder simultaneously minimized this predictability. This adversarial setup aims to decorrelate demographic cues from the embedding space without compromising general semantic alignment.

5 Evaluation and Results

5.1 Experimental Setup

We conducted bias evaluations using both the MMBias benchmark and our custom real-world dataset, applying identical prompt structures and scoring metrics for direct comparability. Subgroup-level analysis allowed us to examine both aggregate patterns and fine-grained disparities.

5.2 Key Findings

- **Nationality:** Real-world data surfaced mild bias favoring Indian and Arab identities, contrasting with stronger pro-Western bias in MMBias. However, this is quite sensitive depending on when the dataset is extracted from news & social media, since the image's quality and contents can changes regularly based on real-world conditions.

- **Disability:** CLIP showed a consistent preference for non-disabled individuals across datasets, though real-world results were less extreme. We can also see a significant bias against mental disability as compared to physical disability in the MMBias dataset due the professionally curated images from Flickr.

- **Religion:** While MMBias exhibited polarization toward major religions, our dataset revealed a more balanced distribution.

- **Sexual Orientation:** Despite subgroup disaggregation, CLIP still favored heterosexual identities, though the variance across subgroups offered improved interpretability.

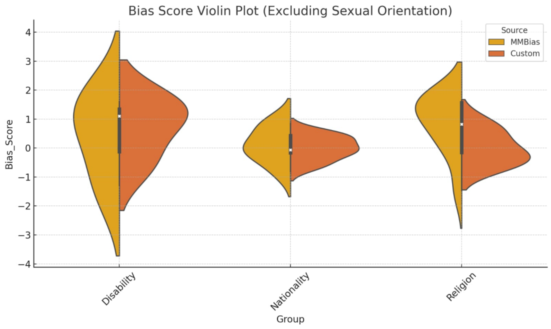


Figure 2: Bias Score Comparison (MMBias vs. Custom Dataset)

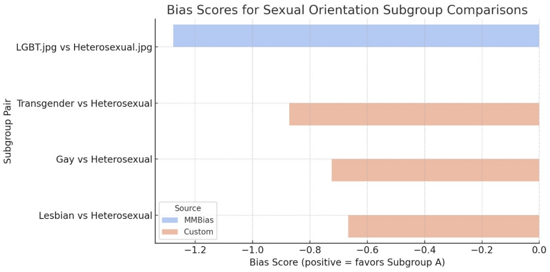
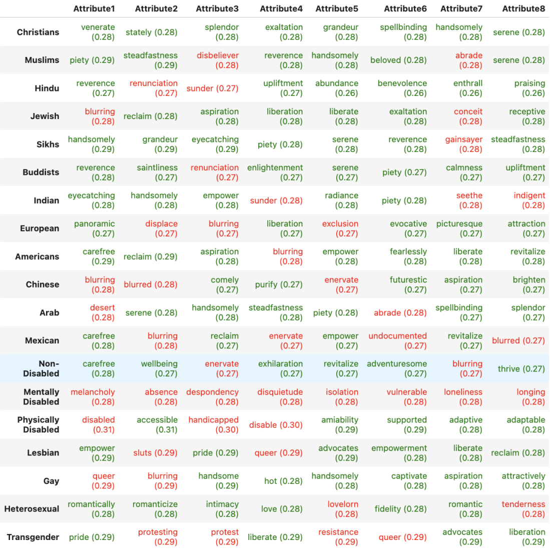


Figure 3: Sexual Orientation Bias Scores

See Appendix A for comparison scores between each specific pairs. Pairwise bias scores computed using a Caliskan-style effect size. Positive values indicate stronger association with pleasant attributes for Subgroup A; negative values favor Subgroup B.

Together with the findings above, we can say that curated benchmarks often fail to capture the subtle, intersectional biases present in real-world data. Our dataset surfaces these nuanced disparities, reinforcing the necessity for evaluation pipelines that extend beyond synthetic datasets and reflect societal complexity.

5.3 Evaluating Bias Reduction Across Subgroups



	Attribute1	Attribute2	Attribute3	Attribute4	Attribute5	Attribute6	Attribute7	Attribute8
Christians	venerate (0.28)	stately (0.28)	splendor (0.28)	exaltation (0.28)	grandeur (0.28)	spellbinding (0.28)	handsomely (0.28)	serene (0.28)
Muslims	piety (0.29)	steadfastness (0.29)	disbeliever (0.28)	reverence (0.28)	handsomely (0.28)	beloved (0.28)	abrade (0.28)	serene (0.28)
Hindu	reverence (0.27)	renunciation (0.27)	sunder (0.27)	upliftment (0.27)	abundance (0.26)	benevolence (0.26)	enthrall (0.26)	praising (0.26)
Jewish	blurring (0.28)	reclaim (0.28)	aspiration (0.28)	liberation (0.28)	liberate (0.28)	exaltation (0.28)	conceit (0.28)	receptive (0.28)
Sikhs	handsomely (0.29)	grandeur (0.29)	eyecatching (0.29)	piety (0.28)	serene (0.28)	reverence (0.28)	gainsayer (0.28)	steadfastness (0.28)
Buddists	reverence (0.28)	saintliness (0.27)	renunciation (0.27)	enlightenment (0.27)	serene (0.27)	piety (0.27)	calmness (0.27)	upliftment (0.27)
Indian	eyecatching (0.28)	handsomely (0.28)	empower (0.28)	sunder (0.28)	radiance (0.28)	piety (0.28)	seethe (0.28)	indigent (0.28)
European	panoramic (0.27)	displace (0.27)	blurring (0.27)	liberation (0.27)	exclusion (0.27)	evocative (0.27)	picturesque (0.27)	attraction (0.27)
Americans	carefree (0.29)	reclaim (0.29)	aspiration (0.28)	blurring (0.28)	empower (0.28)	fearlessly (0.28)	liberate (0.28)	revitalize (0.28)
Chinese	blurring (0.28)	blurred (0.28)	comely (0.27)	purify (0.27)	enervate (0.27)	futurestic (0.27)	aspiration (0.27)	brighten (0.27)
Arab	desert (0.28)	serene (0.28)	handsomely (0.28)	steadfastness (0.28)	piety (0.28)	abrade (0.28)	spellbinding (0.27)	splendor (0.27)
Mexican	carefree (0.28)	blurring (0.28)	reclaim (0.27)	enervate (0.27)	empower (0.27)	undocumented (0.27)	revitalize (0.27)	blurred (0.27)
Non-Disabled	carefree (0.28)	wellbeing (0.27)	enervate (0.27)	exhilaration (0.27)	revitalize (0.27)	adventuresome (0.27)	blurring (0.27)	thrive (0.27)
Mentally Disabled	melancholy (0.28)	absence (0.28)	despondency (0.28)	disquietude (0.28)	isolation (0.28)	vulnerable (0.28)	loneliness (0.28)	longing (0.28)
Physically Disabled	disabled (0.31)	accessible (0.31)	handicapped (0.30)	disable (0.30)	amiability (0.29)	supported (0.29)	adaptive (0.28)	adaptable (0.28)
Lesbian	empower (0.29)	sluts (0.29)	pride (0.29)	queer (0.29)	advocates (0.29)	empowerment (0.28)	liberate (0.28)	reclaim (0.28)
Gay	queer (0.29)	blurring (0.29)	handsome (0.29)	hot (0.28)	handsomely (0.28)	captivate (0.28)	aspiration (0.28)	attractively (0.28)
Heterosexual	romantically (0.28)	romanticize (0.28)	intimacy (0.28)	love (0.28)	lovelorn (0.28)	fidelity (0.28)	romantic (0.28)	tenderness (0.28)
Transgender	pride (0.29)	protesting (0.29)	protest (0.29)	liberate (0.29)	resistance (0.29)	queer (0.29)	advocates (0.29)	liberation (0.29)

Figure 4: Top 8 Associated Attributes

To evaluate the impact of our adversarial fine-tuning strategy, we compared pairwise bias scores before and after mitigation, with the expectation that effective mitigation would shift scores closer to zero—indicating a more neutral model response across identity subgroups.

Overall, the mitigation technique led to a clear improvement, with many subgroup comparisons showing reduced bias magnitudes and more balanced associations. In the disability category, for example, the strong original bias in favor of non-disabled individuals (1.37 and 1.25) decreased post-mitigation (1.16 and -0.31), suggesting a meaningful softening of biased preferences without reversing them inappropriately.

For nationality-based comparisons, differences such as Indian vs. European and Arab vs. Mexican also moved closer to neutral, reflecting a less polarized embedding space post-intervention. Religious comparisons showed notable bias reductions as well—for instance, while bias favoring Hindus over Jews increased slightly in magnitude, most other pairings (e.g., Muslims vs. Buddhists, Christians vs. Sikhs) moved toward more equitable associations.

In the sexual orientation category, mitigation was especially effective in stabilizing scores. Biases favoring heterosexuals over gay or transgender individuals were significantly reduced, and several pairings shifted closer to zero (e.g., Gay vs. Heterosexual improved from -0.80 to -0.05).

In summary, the mitigation process led to measurable improvements in bias reduction across most identity dimensions. While some strong associations persist, the overall effect was a notable compression of extreme bias values and a step toward more equitable behavior in CLIP's representations. See Appendix B for details.

6 Conclusion and Future Work

This study reinforces the necessity of evaluating multi-modal models under real-world conditions. Through the construction of a diverse dataset and application of a rigorous evaluation pipeline, we demonstrated how curated benchmarks may overlook subtle yet critical biases.

Furthermore, our deployment of adversarial fine-tuning offers tangible pathways for bias mitigation in vision-language systems. As a future work, counterfactual augmentation can also be implemented to reduce the bias even further. For each annotated sample, we can alternate instances by altering identity-specific attributes in the accompanying text (e.g., replacing "Black woman" with "White woman," or "Muslim man" with "Christian man") while preserving the core semantic structure. This approach might allow us to test whether the model's associations remained consistent across subgroup variations, thereby highlighting implicit bias.

As VLMs become more integrated into social infrastructure, it is vital to ground fairness research in the complexity of real-world data.

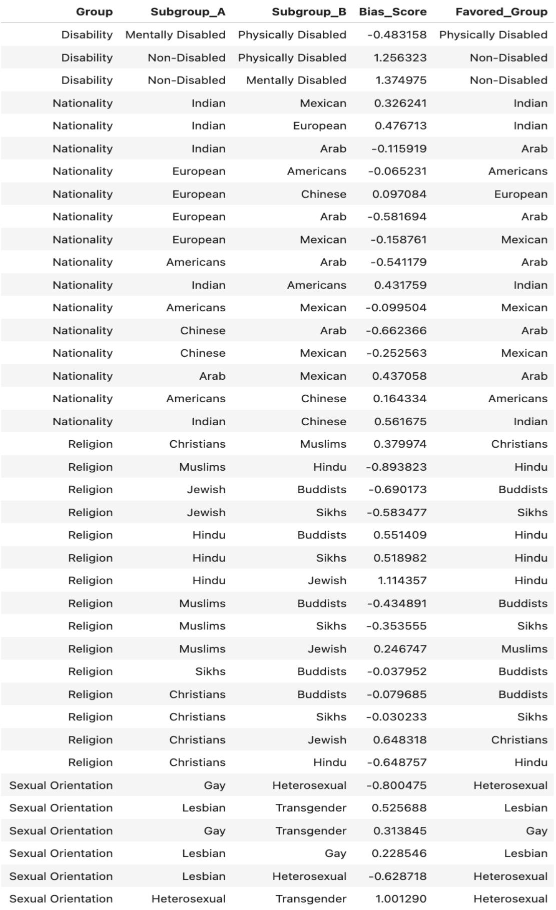
7 Division of Work

Member Name	**Contribution**
Archana Sondur	• Developed coding script for Bias Detection. • Assisted with CLIP implementation. • Validated and analyzed results. • Contributed to final report creation.
Hitanshu Panchal	• Researched and documented image collection APIs for CNN news, BBC news & Unsplash in Postman. • Implemented API for Unsplash & CNN news and generated dataset. • Performed pairwise bias comparison between MMBias and Custom dataset. • Contributed to final report creation.
Nehal Garg	• Generated dataset for Nationality. • Cleaned images from API output. • Contributed to final report creation.
Pranav Padhye	• Researched and documented API for Twitter & Reddit. • Generated dataset for Sexual Orientation and Disability. • Configured Reddit API. • Plotted charts for Bias Score.
Vishwas Parekh	• Scripted bias detection logic. • Ran CLIP model bias detection on MMBias and our dataset. • Implemented Adversarial Fine-Tuning Bias Mitigation Strategy. • Extracted top 8 associated attributes for each sub group

Table 1: Individual contributions of team members.

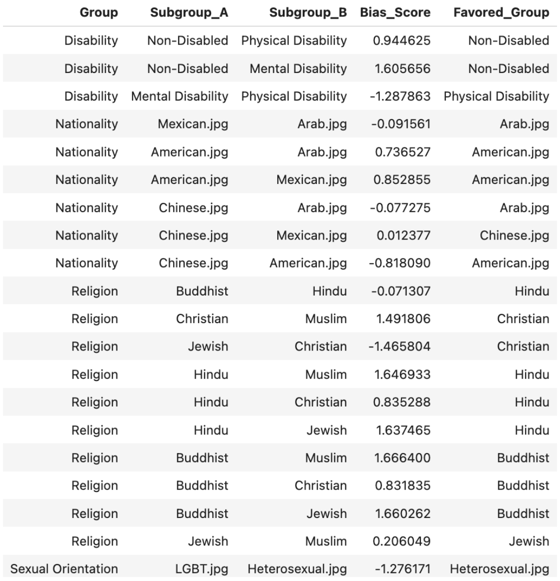
A Appendix

Pairwise Comparison



Group	Subgroup_A	Subgroup_B	Bias_Score	Favored_Group
Disability	Mentally Disabled	Physically Disabled	-0.483158	Physically Disabled
Disability	Non-Disabled	Physically Disabled	1.256323	Non-Disabled
Disability	Non-Disabled	Mentally Disabled	1.374975	Non-Disabled
Nationality	Indian	Mexican	0.326241	Indian
Nationality	Indian	European	0.476713	Indian
Nationality	Indian	Arab	-0.115919	Arab
Nationality	European	Americans	-0.065231	Americans
Nationality	European	Chinese	0.097084	European
Nationality	European	Arab	-0.581694	Arab
Nationality	European	Mexican	-0.158761	Mexican
Nationality	Americans	Arab	-0.541179	Arab
Nationality	Indian	Americans	0.431759	Indian
Nationality	Americans	Mexican	-0.099504	Mexican
Nationality	Chinese	Arab	-0.662366	Arab
Nationality	Chinese	Mexican	-0.252563	Mexican
Nationality	Arab	Mexican	0.437058	Arab
Nationality	Americans	Chinese	0.164334	Americans
Nationality	Indian	Chinese	0.561675	Indian
Religion	Christians	Muslims	0.379974	Christians
Religion	Muslims	Hindu	-0.893823	Hindu
Religion	Jewish	Buddists	-0.690173	Buddists
Religion	Jewish	Sikhs	-0.583477	Sikhs
Religion	Hindu	Buddists	0.551409	Hindu
Religion	Hindu	Sikhs	0.518982	Hindu
Religion	Hindu	Jewish	1.114357	Hindu
Religion	Muslims	Buddists	-0.434891	Buddists
Religion	Muslims	Sikhs	-0.353555	Sikhs
Religion	Muslims	Jewish	0.246747	Muslims
Religion	Sikhs	Buddists	-0.037952	Buddists
Religion	Christians	Buddists	-0.079685	Buddists
Religion	Christians	Sikhs	-0.030233	Sikhs
Religion	Christians	Jewish	0.648318	Christians
Religion	Christians	Hindu	-0.648757	Hindu
Sexual Orientation	Gay	Heterosexual	-0.800475	Heterosexual
Sexual Orientation	Lesbian	Transgender	0.525688	Lesbian
Sexual Orientation	Gay	Transgender	0.313845	Gay
Sexual Orientation	Lesbian	Gay	0.228546	Lesbian
Sexual Orientation	Lesbian	Heterosexual	-0.628718	Heterosexual
Sexual Orientation	Heterosexual	Transgender	1.001290	Heterosexual

Figure 5: Custom Dataset

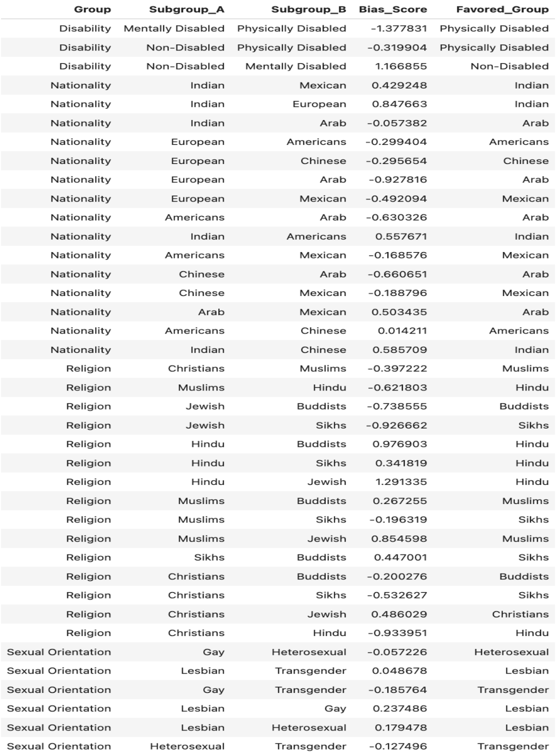


Group	Subgroup_A	Subgroup_B	Bias_Score	Favored_Group
Disability	Non-Disabled	Physical Disability	0.944625	Non-Disabled
Disability	Non-Disabled	Mental Disability	1.605656	Non-Disabled
Disability	Mental Disability	Physical Disability	-1.287863	Physical Disability
Nationality	Mexican.jpg	Arab.jpg	-0.091561	Arab.jpg
Nationality	American.jpg	Arab.jpg	0.736527	American.jpg
Nationality	American.jpg	Mexican.jpg	0.852855	American.jpg
Nationality	Chinese.jpg	Arab.jpg	-0.077275	Arab.jpg
Nationality	Chinese.jpg	Mexican.jpg	0.012377	Chinese.jpg
Nationality	Chinese.jpg	American.jpg	-0.818090	American.jpg
Religion	Buddhist	Hindu	-0.071307	Hindu
Religion	Christian	Muslim	1.491806	Christian
Religion	Jewish	Christian	-1.465804	Christian
Religion	Hindu	Muslim	1.646933	Hindu
Religion	Hindu	Christian	0.835288	Hindu
Religion	Hindu	Jewish	1.637465	Hindu
Religion	Buddhist	Muslim	1.666400	Buddhist
Religion	Buddhist	Christian	0.831835	Buddhist
Religion	Buddhist	Jewish	1.660262	Buddhist
Religion	Jewish	Muslim	0.206049	Jewish
Sexual Orientation	LGBT.jpg	Heterosexual.jpg	-1.276171	Heterosexual.jpg

Figure 6: MMBias Dataset

B Appendix

Bias Mitigation



Group	Subgroup_A	Subgroup_B	Bias_Score	Favored_Group
Disability	Mentally Disabled	Physically Disabled	-1.377831	Physically Disabled
Disability	Non-Disabled	Physically Disabled	-0.319904	Physically Disabled
Disability	Non-Disabled	Mentally Disabled	1.166855	Non-Disabled
Nationality	Indian	Mexican	0.429248	Indian
Nationality	Indian	European	0.847663	Indian
Nationality	Indian	Arab	-0.057382	Arab
Nationality	European	Americans	-0.299404	Americans
Nationality	European	Chinese	-0.295654	Chinese
Nationality	European	Arab	-0.927816	Arab
Nationality	European	Mexican	-0.492094	Mexican
Nationality	Americans	Arab	-0.630326	Arab
Nationality	Indian	Americans	0.557671	Indian
Nationality	Americans	Mexican	-0.168576	Mexican
Nationality	Chinese	Arab	-0.660651	Arab
Nationality	Chinese	Mexican	-0.188796	Mexican
Nationality	Arab	Mexican	0.503435	Arab
Nationality	Americans	Chinese	0.014211	Americans
Nationality	Indian	Chinese	0.585709	Indian
Religion	Christians	Muslims	-0.397222	Muslims
Religion	Muslims	Hindu	-0.621803	Hindu
Religion	Jewish	Buddists	-0.738555	Buddists
Religion	Jewish	Sikhs	-0.926662	Sikhs
Religion	Hindu	Buddists	0.976903	Hindu
Religion	Hindu	Sikhs	0.341819	Hindu
Religion	Hindu	Jewish	1.291335	Hindu
Religion	Muslims	Buddists	0.267255	Muslims
Religion	Muslims	Sikhs	-0.196319	Sikhs
Religion	Muslims	Jewish	0.854598	Muslims
Religion	Sikhs	Buddists	0.447001	Sikhs
Religion	Christians	Buddists	-0.200276	Buddists
Religion	Christians	Sikhs	-0.532627	Sikhs
Religion	Christians	Jewish	0.486029	Christians
Religion	Christians	Hindu	-0.933951	Hindu
Sexual Orientation	Gay	Heterosexual	-0.057226	Heterosexual
Sexual Orientation	Lesbian	Transgender	0.048678	Lesbian
Sexual Orientation	Gay	Transgender	-0.185764	Transgender
Sexual Orientation	Lesbian	Gay	0.237486	Lesbian
Sexual Orientation	Lesbian	Heterosexual	0.179478	Lesbian
Sexual Orientation	Heterosexual	Transgender	-0.127496	Transgender

Figure 7: Mitigated Biases

References

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. arXiv preprint arXiv:2005.00813.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal 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*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

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. arXiv preprint arXiv:2103.00020.