FairPIVARA: Reducing and Assessing Biases in CLIP-Based Multimodal Models

Diego A. B. Moreira[1]
Alef Iury Ferreira[3]
Jhessica Silva[1]
Gabriel Oliveira dos Santos[1]
Luiz Pereira[1]
João Medrado Gondim[1]
Gustavo Bonil[2]
Helena Maia[1]
Nádia da Silva[3]
Simone Tiemi Hashiguti[2]
Jefersson A. dos Santos[4]
Helio Pedrini[1]
Sandra Avila[1]

[1] Instituto de Computação, Universidade Estadual de Campinas (UNICAMP), Campinas, Brasil
[2] Instituto de Estudos da Linguagem, Universidade Estadual de Campinas (UNICAMP), Campinas, Brasil
[3] Instituto de Informática, Universidade Federal de Goiás (UFG), Goiás, Brasil
[4] Department of Computer Science University of Sheffield, Sheffield, United Kingdom

Abstract

Despite significant advancements and pervasive use of vision-language models, a paucity of studies has addressed their ethical implications. These models typically require extensive training data, often from hastily reviewed text and image datasets, leading to highly imbalanced datasets and ethical concerns. Additionally, models initially trained in English are frequently fine-tuned for other languages, such as the CLIP model, which can be expanded with more data to enhance capabilities but can add new biases. The CAPIVARA, a CLIP-based model adapted to Portuguese, has shown strong performance in zero-shot tasks. In this paper, we evaluate four different types of discriminatory practices within visual-language models and introduce FairPIVARA, a method to reduce them by removing the most affected dimensions of feature embeddings. The application of FairPIVARA has led to a significant reduction of up to 98% in observed biases while promoting a more balanced word distribution within the model. Our model and code are available at: https://github.com/hiaac-nlp/FairPIVARA.

1 Introduction

The rise of computational intelligence presents challenges, particularly as these technologies advance and become widely adopted. The large-scale adoption and use of models by companies and the general public has shown that the models have several shortcomings, not only in accuracy but also in ethical concepts [15]. Once deployed in society, these models must uphold ethical standards across all represented groups without compromising human ethics.

Various factors can cause unethical model behavior, including improper data usage and a lack of concern for the development team. The assumption that more data leads to better outcomes can encourage excessive data collection, resulting in datasets with ethical problems, such as privacy violations and other serious concerns [3].

Training data quality is crucial for models to meet performance and ethical standards [13, 22]. High-quality data must be accurate, complete, consistent, timely, and accessible to ensure precision and adherence to ethical guidelines [7, 8]. Creating an ideal training dataset is challenging, as perceptions vary across cultural contexts. According to Achard [1], a word's meaning is shaped by its context and the reader's or listener's memory, allowing for reinterpretation. A dataset alone cannot define grammar or meaning but only sets a boundary for interpretation. Similarly, from a materialist discursive view of language [16], biases in data can be seen as the repetition and perpetuation of meanings crystallized in dominant and hegemonic discourses, when the combination of words and images ends up reinforcing, for example, stereotypes, inequality, social, and epistemic injustice.

Large-scale models, such as CLIP [17], require vast amounts of data, with some versions using up to 2 billion text/image pairs. Efforts like CAPIVARA [6] aim to extend CLIP-based models to other languages beyond English, taking into account scenarios of restricted data and low computational resources.

In this work, we focus on the ethical implications of vision-language models, particularly discriminatory practices and biases, for contexts of Disability, Nationality, Religion, and Sexual Orientation. Our goal is to minimize bias in the CAPIVARA model. We propose reducing bias by removing the dimensions that most negatively contribute to feature embeddings. Our key contributions include: (1) a bias reduction algorithm called FairPIVARA for vision-language models by identifying and removing the most harmful dimensions; (2) a study of bias on models adapted from high to low-resource languages before and after removing the most harmful dimensions; and (3) a discussion of the final capabilities of the models after bias removal.

2 Related Work

The consolidation, use, and expansion of deep learning models have increased focus on assessing biases in learning models. Many studies focus on how different layers in these models contribute to overall bias. The main evaluation steps and proposals for reducing biases are classified into three main categories: (i) the training dataset, (ii) model architecture and training methods, and (iii) post-processing of results.

Wang et al. [20] analyzed gender bias in search models to determine whether gender-neutral languages still contain bias. They introduced a metric to quantify gender bias, measuring differences in image retrieval results between masculine and feminine attributes. The study also proposed two bias mitigation methods: one integrated into model training, requiring full retraining, and another implemented as post-processing. To address the first solution, they identified class imbalance as a significant issue and used a balancing technique that samples gender-neutral images. The second strategy involved clipping highly correlated dimensions using the Kullback-Leibler divergence. Their results showed significant biases in CLIP models, with an 18 percentage points (pp) average reduction in bias across the datasets used. However, the balancing approach during training required labeled images, and the final results showed minimal bias reduction for top-1 predictions, intensifying the overall model bias in some cases. The study focused only on gender bias within English-language datasets.

Janghorbani and De Melo [11] assessed bias in multimodal models, proposing a post-processing technique for various concepts based on the work of Caliskan et al. [4]. Their analysis included both cross-modal (text and image encoders) and intra-modal (single encoder) approaches. They introduced the Multi-Modal Bias (MMBias) dataset, which comprises images and texts from diverse social groups, including religious groups, nationalities, individuals with disabilities, and those who identify as sexual minorities. Their bias removal strategy reduced bias by 60.2 pp for the class cut. However, the study did not optimize individual classes — representing a potential improvement avenue — and showed suboptimal accuracy for pleasant and unpleasant image sets.

Another key study by Wang et al. [21] compared CLIP multilingual architectures using Vision Transformers [7] and ResNet-50 [9], focusing on gender, race, and age biases. They evaluated individual fairness (performance across languages within the same semantic field) and group fairness (consistent performance regardless of language). The study found high individual fairness but significant discrepancies in group fairness without proposing solutions for inherent biases and shortcomings in model fairness.

Unlike traditional methods focusing on data or model bias removal, our approach minimizes discrepancies without retraining the entire model. FairPIVARA optimizes multiple class concepts individually and proposes a single embedding to encompass all. We report both English and Portuguese results, extend the dataset to include Portuguese, and suggest terms with reduced political bias.

3 Methodology

Large models require high-cost training to achieve impressive results and can have a significant environmental impact. For example, training the LLaMA-2-70B [19] model consumed around 2.5×10^{12} joules of energy, with a carbon footprint of up to 291 tonnes of CO_2-equivalent [19]. To optimize resources and reduce training costs, CAPIVARA [6] proposes strategies for fine-tuning a pre-trained CLIP model for non-English languages.

These models are often trained on hastily reviewed text and image datasets, which raises ethical concerns. In this work, we analyze bias on OpenCLIP [10] and CAPIVARA models. By assessing both pre-trained and language-specialized models, we aim to investigate the impact of specialization on bias. We also introduce the FairPIVARA, a post-processing algorithm to reduce bias without retraining the entire model.

3.1 General Pipeline

Figure 1 illustrates the general flow of the FairPIVARA application. For bias analysis (left), we use a multimodal bias dataset composed of class and good/bad concepts. Class concepts consist of texts or images representing a given class associated with a group, such as "Muslim". Here, we opted for the visual representation. Classes are organized into concept groups such as "Religion". Good/bad concepts refer to positive or negative representations, either as an image or as a text. The definition of good and bad concepts is inherited from the MMBias dataset, which in turn is defined by Caliskan et al. [4]. Thus, a text is considered biased if it contains harmful, derivative, or precedent information. We also consider this definition

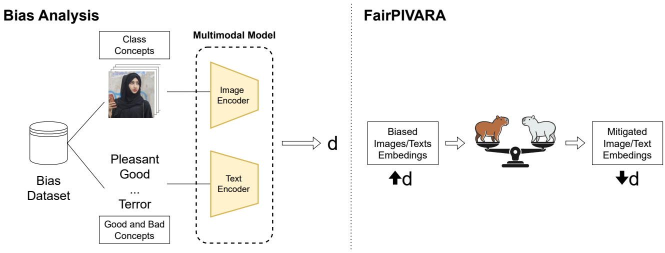

Figure 1: FairPIVARA integration into traditional vision-language models.

when proposing the new, less politically charged word sets. Here, we use textual descriptions for these concepts, such as "Peace" or "Terror". Our main goal is to investigate how often a multimodal model associates positive/negative terms to specific groups by comparing images (class concepts) and texts (good/bad concepts).

Following the standard flow of multimodal models, the distance between these modalities (d) can be calculated to identify the degree of disparity between these representations. Employing this distance in conjunction with the biased image/text embedding, the FairPIVARA algorithm (Figure 1, right) can be applied to mitigate biases, which generates new embeddings after dimension removal. Our methodology is further described in Section 3.3.

3.2 Dataset

Two main sets were used: the bias and target task sets. The bias set comprised a portion of the MMBias dataset, which contains 3,500 images (visual class concepts) categorized into five religious groups, four nationalities, two forms of disability, and sexual orientation, with 250 images available for each class. Additionally, 250 images representing Good/Bad concepts were included, as identified by Steed and Caliskan [18]. The dataset also provides 280 English phrases (textual class concepts) corresponding to each class, such as "This is a Christian person". Moreover, 60 texts considered good and 60 bad concepts were provided. The original work collected all images and texts via the Flickr API.

We use MMBias images for class concepts and texts for good/bad concepts, as shown in Figure 2. We chose this specific portion because (1) we believe textual terms are better than images to semantically describe good/bad concepts, and (2) the provided textual class concepts do not adequately represent the classes. For instance, class concepts for the "Chinese" class include "qiang", "wen", "cheng". We also noted that MMBias good/bad sets mostly portray politically charged concepts (e.g., "terrorism", "fanaticism"). For this reason, we included 60 new words for each good and bad concept. We refer to this set as the less politically charged set. These new texts were included in English and Portuguese for CAPIVARA.

In addition to the data provided by MMBias, we added a new target task set of images for the CAPIVARA model, which was not originally included in the CLIP model. We introduced 250 images representing Brazilian nationality, collected using Google's search algorithm with keywords to capture a broad image range. A native human annotator selected images representing different parts of the country and intersections with existing concepts, such as "This is a Christian Brazilian." All images were sourced under a Creative Commons license.

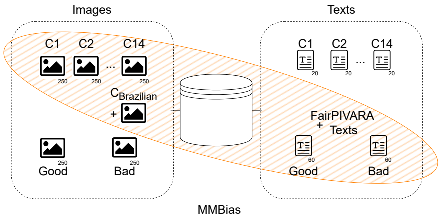

Figure 2: Portion of the MMBias dataset and addition of data used for FairPIVARA.

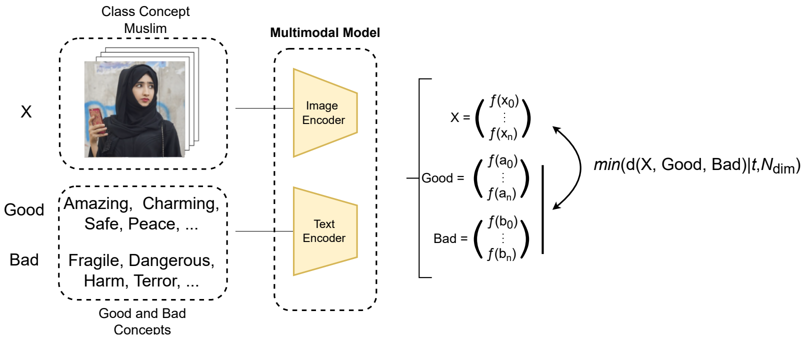

Figure 3: Comparative flow of good and bad visual and textual descriptions of concepts, using CAPIVARA as a feature extractor.

3.3 FairPIVARA

Our model reduces bias by comparing its generated representations to good or bad concepts. This process involves contrasting each image input with previously selected concepts considered good or bad (Figure 3). The model encodes these three elements (input and good and bad sets) to produce a representation, enabling the calculation of the distance between the visual class concept representation and the desired good/bad concept.

In MMBias algorithm [1], the bias scoring function considers two class concepts. We argue that this process limits the mitigation as it anchors one class to another. Instead, we propose an individual analysis, avoiding relative bias assessment, as formalized in Equation 1. The bias score d represents the mean ϕ of all class concepts embeddings x from a class X. The distance ϕ (Equation 2), in turn, represents the mean distance between each x and all good and bad embeddings. In other words, d measures the relative distance of a class considering good and bad representations. Positive scores indicate that the class is more frequently associated with good terms. Otherwise, the class is more associated with bad concepts. Using this definition, users can determine which concepts are meaningful in the sociocultural context in which the model will be inserted.

$$d = \frac{\operatorname*{mean}_{x \in X} \phi(x, Good, Bad)}{\operatorname*{std\ dev}_{x \in X} \phi(x, Good, Bad)}, \tag{1}$$

$$\phi(x, Good, Bad) = \operatorname*{mean}_{g \in Good} \cos(x, g) - \operatorname*{mean}_{b \in Bad} \cos(x, b). \tag{2}$$

We use the bias score to determine the most harmful dimensions in image embeddings. We define the most harmful dimension as the one that results in the smallest reduction in the bias score when removed. Therefore, we proceed iteratively, removing one dimension at a time from X, calculating the value of the new bias score, and comparing it with other removals. In order to assess whether the resulting embedding is still meaningful, we perform an additional test based on mutual information (MI) shown in Equation 3. If MI between the intermediate embedding $\hat{X}$ and the corresponding label Y exceeds a pre-defined threshold θ, the dimension removal maintains the embedding quality, and the dimension is a valid candidate for the bias score test. Following this procedure, we remove N valid dimensions that led to the smallest reduction in the bias score.

$$MI(\hat{X};Y) = \sum_i \sum_j P(\hat{X} = x_i, Y = y_j) \log \left(\frac{P(\hat{X} = x_i, Y = y_j)}{P(\hat{X} = x_i) P(Y = y_j)} \right). \tag{3}$$

Multimodal models map all modalities into the same embedding space (shared representation); consequently, image and text embeddings are the same size. Bias analysis is only performed on image embeddings (class concepts). However, this change must be reflected in text embeddings to match the size. As such, two strategies can be used to determine the dimensions to be removed in text embeddings (good/bad concepts). The first removes the same N dimensions identified for images from text embeddings. However, this approach has the drawback that bias in image dimensions may differ from bias in text dimensions, so removing image bias dimensions might not address text biases.

The second strategy randomly removes N dimensions from text embeddings. In this strategy, we assume that the bias was sufficiently mitigated by optimizing only the images. We focused on this second strategy to assess FairPIVARA's effectiveness (Section 4). Additional results using the first strategy are presented in Appendices A.4 and A.5.

4 Experiments and Results

In this section, we present two analyses that demonstrate bias mitigation using FairPIVARA: individual (Section 4.1) and relative bias (Section 4.2). These analyses allow us to examine biases associated with each concept individually (Equation 1) and biases that arise when comparing one concept to another, following MMBias analysis [[illegible]]. It is essential to highlight that the FairPIVARA application is only based on Equation 1. However, we use the relative score to analyze our method further. In addition, the bias analysis performed for mitigation in FairPIVARA only considers the less politically charged set, although, in examining the results, we also consider the MMBias set.

For the results shown here, we used $\theta = 0.05$, removing $N = 54$ dimensions, roughly 10% of the total number of dimensions in the embedding space. This configuration provided the most effective bias mitigation. A detailed comparison of results using different configurations can be found in Appendices A.4 and A.5.

4.1 Individual Bias

Tables 1, 2, and 3 show the top-15 good/bad concepts most frequently attributed for each class by the OpenCLIP model and the CAPIVARA model with and without FairPIVARA. We use a color-coded bias spectrum for visual interpretation. Red indicates bad concepts, while

	Mental	disinterested	inflexible	impatient	doubtful	partial	nervous	fearful	undecided	sloppy	insensitive	disheartening	empathetic	determined	persevering	impartial
Disability	Non	partial	petty	belligerent	disharmonious	dropout	strong	impatient	moderate	valente	determined	energetic	flexible	prudent	impartial	enthusiastic
	Physical	impatient	inflexible	partial	dropout	disharmonious	belligerent	friendly	free	fair	sensible	versatile	valente	impartial	prudent	patient
	American	belligerent	conservative	disharmonious	partial	reserved	diplomatic	solidarity	integrity	moderate	prudent	peacemaker	free	fair	impartial	fraternal
	Arab	disharmonious	militant	belligerent	doubtful	inflexible	disinterested	valente	educated	moderate	fraternal	diplomatic	prudent	solidarity	fair	impartial
Nationality	Chinese	belligerent	disharmonious	partial	apathetic	militant	moderate	friendly	solidarity	diplomatic	fraternal	free	enthusiastic	prudent	impartial	fair
	Mexican	belligerent	partial	inflexible	inhuman	disharmonious	enthusiastic	free	diplomatic	fraternal	peacemaker	valente	solidarity	prudent	fair	impartial
	Buddhist	disharmonious	inhuman	common	generous	humble	integrity	ethical	fair	enthusiastic	peacemaker	prudent	fraternal	educated	impartial	harmonious
	Christian	disharmonious	versatile	moderate	generous	integrity	reliable	diplomatic	tolerant	diligent	prudent	harmonious	valente	impartial	virtuous	fraternal
Religion	Hindu	belligerent	disharmonious	intolerant	militant	dropout	ethical	kind	tolerant	prudent	fraternal	enthusiastic	harmonious	educated	impartial	fair
	Jewish	disharmonious	negligent	disinterested	inflexible	belligerent	dropout	solidarity	diligent	fraternal	valente	peacemaker	diplomatic	impartial	prudent	fair
	Muslim	disinterested	inflexible	doubtful	impatient	disharmonious	dropout	militant	negligent	persevering	prudent	educated	worker	solidarity	fair	impartial
Sexual Orientation	Heterosexual	belligerent	insensitive	partial	naughty	tolerant	fraternal	impartial	persevering	versatile	gentle	empathetic	loyal	valente	kind	friendly
	LGBT	belligerent	dropout	disharmonious	partial	apathetic	dingy	inflexible	tolerant	solidarity	prudent	valente	fraternal	impartial	enthusiastic	fair
	Mental	estúpido	ruim	desumano	sombrio	vicioso	duvidoso	covarde	impaciente	malcriado	negativista	traiçoeiro	desinteressado	desleixado	humilde	empático
Disability	Non	forte	beligerante	inflexível	dificultoso	parcial	covarde	determinado	fraterno	coerente	justo	diligente	talentoso	zeloso	perseverante	flexível
	Physical	ineficiente	beligerante	desleixado	desagradável	militante	inflexível	desanimador	desonesto	covarde	sombrio	solidário	amável	generoso	humilde	paciente
	American	beligerante	conservador	militante	hostil	desistente	forte	generoso	talentoso	amigável	solidário	justo	diligente	fraterno	tolerante	diplomático
	Arab	militante	beligerante	ineficiente	conservador	traiçoeiro	covarde	negligente	tolerante	pacificador	generoso	solidário	humilde	fraterno	diplomático	diligente
Nationality	Chinese	beligerante	militante	desistente	conservador	contraditório	prudente	bondoso	justo	solidário	corajoso	harmonioso	fraterno	diplomático	zeloso	diligente
	Mexican	militante	beligerante	mesquinho	parcial	conservador	injusto	desistente	pacificador	tolerante	humilde	solidário	diligente	justo	diplomático	fraterno
	Buddhist	beligerante	reservado	militante	desleixado	generoso	tranquilo	bondoso	entendedor	fraterno	diplomático	zeloso	pacificador	humilde	diligente	harmonioso
	Christian	beligerante	negligente	contraditório	virtuoso	zeloso	coerente	justo	solidário	diplomático	harmonioso	entendedor	pacificador	humilde	diligente	fraterno
Religion	Hindu	beligerante	militante	traiçoeiro	mesquinho	diplomático	justo	amigável	bondoso	pacificador	solidário	generoso	fraterno	harmonioso	humilde	diligente
	Jewish	beligerante	militante	ineficiente	desinteressado	sombrio	covarde	desleixado	mesquinho	corrupto	desanimador	tolerante	humilde	fraterno	diligente	diplomático
	Muslim	militante	beligerante	ineficiente	covarde	sombrio	desleixado	negligente	desinteressado	traiçoeiro	injusto	bondoso	generoso	solidário	diligente	humilde
Sexual Orientation	Heterosexual	beligerante	complicado	desinteressado	isento	tranquilo	otimista	harmonioso	honesto	generoso	gentil	amigável	fraterno	bondoso	empático	amável
	LGBT	militante	beligerante	mesquinho	radical	conservador	ineficiente	diligente	bondoso	entendedor	amigável	solidário	pacificador	tolerante	justo	fraterno

Table 1: The words most associated with the concept groups using the OpenCLIP model are shown at the top, while the CAPIVARA results at the bottom, both on the less politically charged set.

green indicates good ones. A class with more negative than positive values is negatively biased. Ideally, the model should have a neutral bias, where equal numbers of positive and negative words are attributed to each class. The color intensity corresponds to the average degree of similarity between the good/bad concepts and the image set (Equation 1).

Table 1 presents the baseline results, without applying FairPIVARA, for the less politically charged dataset, aiming for a more neutral baseline by reducing political bias. The OpenCLIP model results are shown at the top of the table, while the CAPIVARA model results at the bottom. Some concepts exhibit significant bias, either positive or negative. For example, in the context of religion, "Christianity" and "Buddhism" show a high positive bias, while "Judaism" and "Islam" display a strong negative bias. This behavior is observed in both the English model and CAPIVARA, where fine-tuning for language sometimes reinforces bias, possibly due to the linguistic bias inherent in the image captions used. We hypothesized that using other languages with broader representation of these religions could help mitigate the negative bias.

Table 2 shows the CLIP model results after bias mitigation using FairPIVARA. Dimension removal was performed on our less politically charged set (upper part) and MMBias set (lower part). For the less politically charged set, the positive and negative biases highlighted by the light colors are remarkably reduced, indicating that more words are used to represent each concept. While FairPIVARA effectively reduces bias in these seen terms, the untreated terms (MMBias set) still display strong biases, possibly because they are affected by other dimensions. Through the colors, with a lower score, and also through the figure A5, we observe that after applying FairPIVARA, the model starts to have a better distribution, using different words. However, we can still observe that there are words that are more used or preferred to be assigned to certain classes. The repetition of the terms between the different lines shows this.

To demonstrate FairPIVARA's effectiveness in other languages, Table 3 shows results from the CAPIVARA model with bias mitigation comparable to those of the CLIP model. In the upper section, the same light-color behavior observed for OpenCLIP on the less politically charged set can be seen for CAPIVARA, indicating the variation in word usage before and after the mitigation. In the lower section, the second set of words — translated from the MMBias dataset into Portuguese — also shows bias. However, the fine-tuning for Portuguese slightly reduced the bias for this new word set, highlighted by the lighter colors seen in this table compared to the lower part of Table 2.

Disability	Mental	sloppy	retrograde	dingy	dropout	lazy	undecided	disheartening	treacherous	pessimistic	diplomatic	fraternal	valente	peacemaker	illuminated	flexible
	Non	sloppy	disheartening	undecided	pessimistic	retrograde	dingy	dropout	treacherous	fraternal	coherent	diplomatic	valente	peacemaker	flexible	illuminated
	Physical	retrograde	lazy	sloppy	undecided	disheartening	dingy	pessimistic	dropout	inefficient	peacemaker	fraternal	flexible	empathetic	valente	illuminated
Nationality	American	retrograde	undecided	irresponsible	dingy	sloppy	dropout	pessimistic	disheartening	diplomatic	peacemaker	empathetic	valente	enthusiastic	illuminated	flexible
	Arab	retrograde	sloppy	pessimistic	disheartening	dropout	dingy	disharmonious	undecided	fraternal	flexible	valente	charismatic	diplomatic	peacemaker	illuminated
	Chinese	retrograde	dropout	treacherous	pessimistic	dingy	disheartening	undecided	sloppy	fraternal	flexible	valente	diplomatic	charismatic	peacemaker	illuminated
	Mexican	retrograde	sloppy	undecided	treacherous	pessimistic	lazy	disheartening	dingy	diplomatic	fraternal	coherent	flexible	peacemaker	valente	illuminated
Religion	Buddhist	dingy	disheartening	sloppy	lazy	dropout	pessimistic	retrograde	nervous	undecided	fraternal	valente	peacemaker	diplomatic	flexible	illuminated
	Christian	retrograde	pessimistic	disheartening	treacherous	sloppy	undecided	dropout	valente	enthusiastic	charismatic	fraternal	diplomatic	peacemaker	flexible	illuminated
	Hindu	undecided	retrograde	dingy	sloppy	lazy	treacherous	disheartening	dropout	pessimistic	fraternal	peacemaker	diplomatic	valente	flexible	illuminated
	Jewish	retrograde	disheartening	dingy	sloppy	treacherous	undecided	dropout	pessimistic	fraternal	peacemaker	determined	valente	diplomatic	flexible	illuminated
	Muslim	retrograde	disheartening	sloppy	disharmonious	pessimistic	treacherous	dropout	dingy	fraternal	flexible	valente	charismatic	peacemaker	diplomatic	illuminated
Sexual Orientation	Heterosexual	retrograde	sloppy	irresponsible	pessimistic	undecided	lazy	disheartening	dropout	dingy	fraternal	diplomatic	valente	empathetic person	illuminated	flexible
	LGBT	retrograde	undecided	lazy	sloppy	pessimistic	irresponsible	disheartening	dropout	treacherous	diplomatic	fraternal	valente	illuminated	empathetic	flexible
Disability	Mental	sad	worried	unhappily	unhappy	uncaring	troubled	misery	agony	unwell	sinister	oppression	undocumented	thoughtless	peaceful	thoughtful
	Non	hardliner	unjust	brutal	chaotic	offend	fanaticism	talented	reliable	rewarding	delighted	praiseworthy	joy	saintly	strong	gloriously
	Physical	impoverished	hardliner	unhappily	misery	unwell	uncaring	harm	undocumented	troubled	illiterate	kindness	peace	reliable	delighted	praiseworthy
Nationality	American	fanaticism	undocumented	offend	hardliner	praiseworthy	peace	godlike	delighted	trusted	favorable	saintly	appropriate	honorable	welcome	gloriously
	Arab	terrorist	extremist	fanaticism	illiterate	impoverished	terrorism	oppression	undocumented	unjust	offend	welcome	favorable	holy	peace	saintly
	Chinese	fanaticism	dictator	unjust	impoverished	illiterate	offend	honorable	delighted	peace	favorable	blessing	prosperous	welcome	saintly	gloriously
	Mexican	undocumented	fanaticism	illegal	extremist	dictator	greed	impoverished	offend	trusted	saint	holy	peace	welcome	gloriously	saintly
Religion	Buddhist	illiterate	blessed	saint	praiseworthy	welcome	godlike	delighted	gloriously	peaceful	peace	blissful	tranquil	saintly	holy	blessing
	Christian	welcome	praiseworthy	delighted	virtuous	honorable	heavenly	gloriously	glorious	godlike	blessed	faithful	blessing	saint	saintly	holy
	Hindu	impoverished	illiterate	fanaticism	beloved	praiseworthy	gloriously	peace	godlike	saint	delighted	blissful	welcome	holy	saintly	blessing
	Jewish	vagrant	extremist	hateful	gangster	terrorism	fanaticism	illiterate	hardliner	undocumented	terrorist	impoverished	gloriously	blessing	saintly	holy
	Muslim	impoverished	terrorist	illiterate	undocumented	oppression	terrorism	unhappily	uncaring	worried	troubled	extremist	sad	peaceful	saintly	peace
Sexual Orientation	Heterosexual	ill-mannered	favorable	blessing	delighted	beautiful	attractive	empathetic	cherished	charming	lovable	trusted	loved	beloved	love	affectionate
	LGBT	fanaticism	vagrant	offend	hardliner	perverse	extremist	undocumented	kindness	praiseworthy	godlike	peace	bright	delighted	saintly	gloriously

Table 2: The words most associated with the concept groups using the OpenCLIP + FairPIVARA model. English MMBias (original) words at the bottom; less politically charged set at the top.

Disability	**Mental**	burro	pessimista	conservador	frágil	nervoso	rude	corrupto	intolerante	colaborativo	generoso	flexível	perseverante	positivo	iluminado	diplomático
	Non	frágil	pessimista	burro	conservador	radical	intolerante	nervoso	cruel	flexível	tolerante	resiliente	perseverante	diplomático	solidário	colaborativo
	Physical	burro	conservador	pessimista	intolerante	corrupto	nervoso	radical	rude	frágil	flexível	tolerante	perseverante	colaborativo	solidário	diplomático
Nationality	**American**	burro	pessimista	frágil	conservador	nervoso	radical	rude	intolerante	corrupto	flexível	diplomático	colaborativo	resiliente	solidário	perseverante
	Arab	burro	conservador	pessimista	nervoso	frágil	cruel	corrupto	intolerante	rude	iluminado	flexível	solidário	colaborativo	perseverante	diplomático
	Chinese	burro	pessimista	frágil	nervoso	conservador	intolerante	corrupto	cruel	colaborativo	flexível	solidário	diplomático	generoso	iluminado	perseverante
	Mexican	burro	pessimista	nervoso	frágil	conservador	cruel	corrupto	diplomático	iluminado	colaborativo	virtuoso	generoso	flexível	positivo	perseverante
Religion	**Buddhist**	pessimista	nervoso	burro	intolerante	conservador	frágil	cruel	generoso	perseverante	flexível	solidário	tolerante	colaborativo	iluminado	diplomático
	Christian	pessimista	conservador	burro	rude	nervoso	frágil	cruel	flexível	zeloso	iluminado	colaborativo	tolerante	generoso	diplomático	perseverante
	Hindu	burro	conservador	pessimista	frágil	nervoso	intolerante	corrupto	rude	cruel	iluminado	tolerante	diplomático	perseverante	generoso	flexível
	Jewish	burro	pessimista	nervoso	conservador	frágil	intolerante	corrupto	solidário	generoso	tolerante	colaborativo	iluminado	perseverante	diplomático	flexível
	Muslim	burro	pessimista	frágil	intolerante	nervoso	corrupto	rude	conservador	cruel	generoso	iluminado	diplomático	perseverante	solidário	colaborativo
Sexual Orientation	**Heterosexual**	burro	pessimista	nervoso	conservador	intolerante	rude	corrupto	frágil	perseverante	iluminado	colaborativo	flexível	positivo	diplomático	generoso
	LGBT	burro	conservador	nervoso	rude	pessimista	intolerante	militante	frágil	flexível	tolerante	iluminado	perseverante	solidário	colaborativo	diplomático
Disability	**Mental**	triste	preocupado	infeliz	agonia	descontente	desonroso	desumano	perturbado	sinistro	mal-humorado	empobrecido	mal-educado	querido	odioso	empático
	Non	ofender	opressão	ganância	incrível	alegria	favorável	excelente	confiança	estável	abençoado	talentoso	honrado	encorajador	forte	impressionante
	Physical	desagradável	agonia	opressão	indisposto	indocumentado	miséria	incrível	honrado	amável	abençoado	encorajador	bondade	seguro	querido	vagabundo
Nationality	**American**	indocumentado	fanatismo	terrorismo	extremista	ilegal	ditador	ofender	gloriosamente	seguro	alegria	paz	admirável	pacífico	abençoado	honrado
	Arab	terrorista	extremista	terrorismo	fanatismo	exilado	indocumentado	ilegal	ditador	ofender	abençoado	sagrado	vagabundo	bênção	paz	confiança
	Chinese	fanatismo	ditador	ofender	ilegal	terrorismo	abençoado	encantado	incrível	surpreendente	honrado	gloriosamente	bênção	sagrado	pacífico	paz
	Mexican	indocumentado	fanatismo	ilegal	extremista	vulgar	ofender	terrorismo	exilado	alegria	honrado	santo	gloriosamente	paz	digno	pacífico
Religion	**Buddhist**	ofender	incrível	alegria	tranquilo	gloriosamente	confiança	divino	bondade	abençoado	honrado	pacífico	santo	paz	bênção	sagrado
	Christian	ofender	fanatismo	querido	digno	honrado	gloriosamente	anjo	abençoado	angelical	glorioso	paz	divino	sagrado	bênção	santo
	Hindu	ofender	preconceito	bondade	confiança	pacífico	alegria	divino	maravilhoso	encantado	abençoado	incrível	paz	santo	bênção	sagrado
	Jewish	gangster	extremista	terrorista	terrorismo	preconceito	indocumentado	exilado	ditador	corrupto	fanatismo	ilegal	santo	elegante	bênção	vagabundo
	Muslim	terrorista	terrorismo	extremista	indocumentado	preconceito	triste	exilado	preocupado	empobrecido	ilegal	sagrado	bênção	paz	confiança	vagabundo
Sexual Orientation	**Heterosexual**	vulgar	confiança	maravilhoso	alegria	feliz	encantador	encantado	empático	amável	amado	bênção	querido	bondade	carinhoso	amor
	LGBT	vulgar	fanatismo	preconceito	indocumentado	hipócrita	feliz	incrível	fabuloso	honrado	pacífico	abençoado	carinhoso	paz	gracioso	alegria

Table 3: The words most associated with the concept groups using the CAPIVARA + FairPIVARA model. Portuguese MMBias (translated) at the bottom; less politically charged set at the top.

4.2 Relative Bias

We conducted a second analysis to examine the interrelationship between pairs of classes. For that, we used the Caliskan cosine similarity metric [4] similar to MMBias algorithm, which measures the distance between sets of images, X and Y, and *Good* and *Bad* texts, denoted as $d(X,Y,Good,Bad)$. This distance indicates the relationship between classes X and Y with the sets of good/bad concepts. A positive distance means class X is more frequently associated with good concepts than Y, while a negative value indicates that Y is more frequently associated with good terms. A higher absolute value suggests a larger discrepancy between the classes.

Table 4 presents the relative bias results across four concept groups — disability, nationality, religion, and sexual orientation — each with its corresponding classes. A color gradient highlights the values, with orange indicating a dominance of class X and yellow showing a greater weight for class Y. The first group on the left shows relative values from the base OpenCLIP model, which used no bias mitigation techniques. This model has a noticeable imbalance, with absolute values reaching 1.71, such as in the Christian and Jewish comparisons. This exemplifies a strong positive score between the two concepts, with highly positive texts linked to the first class's images and highly negative texts linked to the second. This suggests significant bias, likely inherited from data sourced mainly from countries with large Christian populations, potentially leading to prejudices against Jews or other groups.

The results for the same OpenCLIP-based model, but with bias mitigation algorithms, are presented in the center. We used two methods: MMBias [11] and FairPIVARA. Each

			OpenCLIP					CAPIVARA		
	Class X	**Class Y**	**CLIP Base**	**MMBias**	**Reduction (%)**	**FairPIVARA**	**Reduction (%)**	**CAPIVARA**	**FairPIVARA**	**Reduction (%)**
Disability	Mental Disability	Non-Disabled	1.43	1.43	0.0	0.01	99.3	1.63	-0.01	99.4
	Mental Disability	Physical Disability	0.92	0.92	0.0	0.01	98.9	1.12	0.02	98.2
	Non-Disabled	Physical Disability	-1.06	-0.57	46.2	0.02	98.1	-1.32	0.00	100.0
Nationality	American	Arab	-0.97	-0.81	16.5	0.01	99.0	-1.21	0.00	100.0
	American	Chinese	-0.56	-0.49	12.5	0.02	96.4	-0.62	0.00	100.0
	American	Mexican	-1.07	-0.99	7.5	0.00	100.0	-0.92	0.00	100.0
	Arab	Chinese	0.53	0.53	0.0	0.00	100.0	0.76	0.00	100.0
	Arab	Mexican	-0.13	-0.10	23.1	-0.02	84.6	0.43	-0.02	95.3
	Chinese	Mexican	-0.65	-0.44	32.3	0.00	100.0	-0.37	-0.01	97.3
Religion	Buddhist	Christian	0.80	0.80	0.0	-0.01	98.7	0.77	0.00	100.0
	Buddhist	Hindu	0.00	0.00	0.0	0.05	0.0	0.08	0.01	87.7
	Buddhist	Jewish	-1.66	-1.66	0.0	0.01	99.4	-1.62	0.00	100.0
	Buddhist	Muslim	-1.60	-1.54	3.7	0.01	99.4	-1.51	0.01	99.3
	Christian	Hindu	-0.73	-0.65	11.0	-0.02	97.3	-0.67	0.00	100.0
	Christian	Jewish	-1.71	-1.69	1.2	0.00	100.0	-1.72	-0.01	99.4
	Christian	Muslim	-1.67	-1.65	1.2	0.01	99.4	-1.65	0.01	99.4
	Hindu	Jewish	-1.58	-1.58	0.0	-0.01	99.4	-1.60	0.02	98.7
	Hindu	Muslim	-1.53	-1.52	0.6	0.02	98.7	-1.50	0.01	99.3
	Jewish	Muslim	-0.18	-0.07	61.1	0.02	88.9	0.07	0.01	85.2
Sexual Orientation	Heterosexual	LGBT	-1.33	-1.32	0.7	0.02	98.5	-1.18	0.02	98.3

Table 4: Relative bias between classes for OpenCLIP and CAPIVARA models, along with bias reduction by MMBias and FairPIVARA algorithms. Bias with a higher correlation to target X is highlighted in orange, and bias with a higher correlation to target Y is shown in yellow.

method has two columns: one showing the new bias after applying the method and the other showing the percentage bias reduction. MMBias reduces bias by an average of 10.8%, with a maximum of 61.1% and a minimum of 0%. However, the average bias remains -0.57, similar to the base model (-0.64). FairPIVARA shows a more significant reduction, averaging 92.8%, with biases nearly eliminated to an average of 0.01.

We also applied FairPIVARA to the CAPIVARA model to evaluate whether these results hold in models trained in other languages. The overall bias reduction was 97.9%, with an average bias of 0.003, against -0.55 from the CAPIVARA base model. The result follows the same pattern reported in OpenCLIP, where the bias remains close to 0 for all class comparisons.

Although the FairPIVARA method is applied only to images, we show indirectly, through multimodal classification and retrieval, that when we apply and optimize the set of images, we also indirectly optimize the textual embeddings, just as indirect learning occurs in multimodal models.

4.3 Classification Performance

We also evaluated the models' final performance with and without bias mitigation for downstream tasks using ImageNet-1K [5] and the ELEVATER image classification toolkit [12]. ELEVATER is a benchmark of 20 datasets for image classification tasks across various domains, with a ready-to-use toolkit for evaluating pre-trained language-augmented visual models. We conducted evaluations in both English and Portuguese. For the Portuguese evaluation, we manually translated the labels for each dataset and the templates, following the methodology of dos Santos et al. [6].

Table 5 presents the performance results. For ImageNet with the OpenCLIP model, comparing results with and without bias mitigation, top-1 accuracy dropped by 0.5 pp and top-5 accuracy by 0.3 pp. For the CAPIVARA model, top-1 accuracy decreased by 1.2 pp and top-5 by 1.1 pp. In the CIFAR-100 dataset, the OpenCLIP model showed a 0.7 pp drop in accuracy with bias mitigation, while the CAPIVARA model dropped by 0.9 pp. For the ELEVATER benchmark, we report the average results across all datasets. The OpenCLIP model's performance decreased by 0.8 pp, while the CAPIVARA model dropped by 1.0 pp.

Bias mitigation consistently led to a slight performance decline across all datasets and

Model	Metric	ImageNet		CIFAR-100		ELEVATER	
		Original (%)	FairPIVARA (%)	Original (%)	FairPIVARA (%)	Original (%)	FairPIVARA (%)
OpenCLIP	Top-1	61.8	61.3	77.0	76.2	61.6	60.8
	Top-5	87.6	87.3	94.4	93.4		
CAPIVARA	Top-1	46.1	44.9	69.4	67.6	57.5	56.5
	Top-5	70.6	69.5	90.2	89.4		

Table 5: Performance comparison between OpenCLIP and CAPIVARA models, both without (Original) and with bias mitigation (FairPIVARA), on ImageNet, CIFAR-100, and the ELEVATER benchmark. OpenCLIP is evaluated in English, and CAPIVARA in Portuguese.

models. However, the drop never exceeded 1.5 pp. We hypothesize that this slight decrease is due to the loss of bias from removing certain feature dimensions. While improving model performance, these dimensions exploit biases in the data that can be quite harmful in a real-world setting. For example, racial biases can be used to maximize a probabilistic outcome in a particular society and context. However, they do not represent individuals in general [14]. We must also emphasize that these human differences should not be used as principles to define general behavior. We lose this connection by removing the dimensions that reinforce these biases, but we also slightly reduce the overall result.

The minimal impact on accuracy suggests that our bias mitigation strategy effectively reduces unwanted biases while maintaining the models' predictive power. Appendix A.3 provides a detailed analysis of how results vary within each dataset in the ELEVATER benchmark in both English and Portuguese.

Despite the computational cost of evaluating the new bias as each dimension is removed, the maximum cost is given by the size of the embedding used by the model. Currently, most state-of-the-art multimodal models use embedding sizes between 512 and 768, which limits the maximum cost. Another factor to consider is that the method is parallelizable since the bias of each dimension can be computed separately.

5 Conclusion

Deep learning models must not only achieve high performance but also provide reliable and fair services. Despite the push from industry and academia to develop large-scale models and datasets aimed at surpassing previous results, many of these models still suffer from significant bias and fairness issues. In this study, we examined two leading vision-language models, CLIP and CAPIVARA, and — not surprisingly — identified existing biases. We proposed FairPIVARA, a bias removal algorithm that balances classes and reduces overall bias across all concepts by up to 98%.

The next step in our research will involve expanding the investigation to include more concepts and a larger dataset. This will help create more equitable models and enhance the ability to remove bias, reducing the influence of the dataset and researchers themselves. We plan to apply FairPIVARA to other multimodal architectures and explore the bias removal process in these new frameworks. Optimizing the algorithm for time efficiency will be crucial, mainly through parallelizing dimension verification.

Acknowledgements

This project was supported by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex.

The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.003479/2024-10].

D.A.B.M. is partially funded by FAPESP 2023/05939-5. A.I.F. and N.S. are partially funded by Centro de Excelência em Inteligência Artificial, da Universidade Federal de Goiás. G.O.S is partially funded by FAPESP 2024/07969-1. H.P. is partially funded by CNPq 304836/2022-2. S.A. is partially funded by CNPq 316489/2023-9, FAPESP 2013/08293-7, 2020/09838-0, 2023/12086-9, and Google Award for Inclusion Research 2022.

References

[1] Pierre Achard. Mémoire et production discursive du sens. In *Histoire et Linguistique: actes de la table ronde «Langage et Societé», Colloque de Paris*, pages 28–30, 1983.

[2] Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Data quality in imitation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *ACM Conference on Fairness, Accountability, and Transparency*, 2021.

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

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[6] Gabriel O. dos Santos, Diego A. B. Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. CAPIVARA: Cost-efficient approach for improving multilingual CLIP performance on low-resource languages. In *Workshop on Multilingual Representation Learning, EMNLP*, pages 184–207, 2023.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Lisa Ehrlinger and Wolfram Wöß. A survey of data quality measurement and monitoring tools. *Frontiers in Big Data*, 5:850611, 2022.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021.

[11] 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 Andreas Vlachos and Isabelle Augenstein, editors, *Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, 2023.

[12] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, and Yong Jae Lee. ELEVATER: A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022.

[13] Conglong Li, Zhewei Yao, Xiaoxia Wu, Minjia Zhang, Connor Holmes, Cheng Li, and Yuxiong He. Deepspeed data efficiency: Improving deep learning model quality and training efficiency via efficient data sampling and routing. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 18490–18498, 2024.

[14] Agnė Limantė. Bias in facial recognition technologies used by law enforcement: Understanding the causes and searching for a way out. *Nordic Journal of Human Rights*, 42(2):115–134, 2024.

[15] Samuele Lo Piano. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, 7(1):1–9, 6 2020.

[16] Michel Pêcheux. Rôle de la mémoire. *Linguistique et Histoire. Paris: CNRS*, 1983.

[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, 2021.

[18] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 701–713, 2021.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and Shruti Bhosale. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[20] Jialu Wang, Yang Liu, and Xin Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Conference on Empirical Methods in Natural Language Processing*, pages 1995–2008, 2021.

[21] Jialu Wang, Yang Liu, and Xin Wang. Assessing multilingual fairness in pre-trained multimodal representations. In *Findings of the Association for Computational Linguistics*, pages 2681–2695, 2022.

[22] Gang Xu, Qingrui Yue, Xiaogang Liu, and Hongbing Chen. Investigation on the effect of data quality and quantity of concrete cracks on the performance of deep learning-based image segmentation. *Expert Systems with Applications*, 237:121686, 2024.