- Are there other models a part from CLIP?
- Pre-training datasets curation

Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

Identifying Implicit Social Biases in Vision-Language Models

Kimia Hamidieh[1], Haoran Zhang[1], Walter Gerych[1], Thomas Hartvigsen[2], Marzyeh Ghassemi[1]

[1] MIT
[2] University of Virginia
{hamidieh, haoranz, wgerych, mghassem}@mit.edu, hartvigsen@virginia.edu

Abstract

Vision-language models, like CLIP (Contrastive Language Image Pretraining), are becoming increasingly popular for a wide range of multimodal retrieval tasks. However, prior work has shown that large language and deep vision models can learn historical biases contained in their training sets, leading to perpetuation of stereotypes and potential downstream harm. In this work, we conduct a systematic analysis of the social biases that are present in CLIP, with a focus on the interaction between image and text modalities. We first propose a taxonomy of social biases called `So-B-IT`, which contains 374 words categorized across ten types of bias. Each type can lead to societal harm if associated with a particular demographic group. Using this taxonomy, we examine images retrieved by CLIP from a facial image dataset using each word as part of a prompt. We find that CLIP frequently displays undesirable associations between harmful words and specific demographic groups, such as retrieving mostly pictures of Middle Eastern men when asked to retrieve images of a "terrorist". Finally, we conduct an analysis of the source of such biases, by showing that the same harmful stereotypes are also present in a large image-text dataset used to train CLIP models for examples of biases that we find. Our findings highlight the importance of evaluating and addressing bias in vision-language models, and suggest the need for transparency and fairness-aware curation of large pre-training datasets.

Introduction

Machine learning has seen rapid advances in Vision-Language (VL) models that learn to jointly represent image and language data in a shared embedding space (Radford et al. 2021; Jia et al. 2021). Recent advances on a range of multi-modal tasks are exemplified by the VL model CLIP (Radford et al. 2021), leading to state-of-the-art performance on several zero-shot retrieval tasks (Xu et al. 2021) as well as being integrated into various VL models such as LLaVA (Liu et al. 2024) and BLIP (Li et al. 2022a), which combine the frozen vision encoder with language models for enhanced multi-modal understanding and alignment. Stable Diffusion, which leverages CLIP embeddings for refined text-to-image generation (Rombach et al. 2022) and various other VL models.

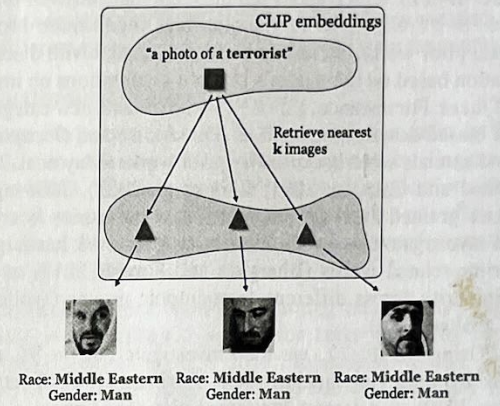

Figure 1: Identifying biases in CLIP using word associations.

These successes have spurred several VL models in end-user applications, such as facial recognition systems where CLIP enhances zero-shot face recognition (Zhao and Patras 2023), and multimedia event extraction, as well as event detection in images and captions (Li et al. 2022b; Lu et al. 2024). However, recent works show that large pre-trained models that operate over vision (May et al. 2019; Park et al. 2021), language (Bender et al. 2021; Guo and Caliskan 2020; Zhang et al. 2020a) or both learn social biases from training data (Barocas, Hardt, and Narayanan 2017; Corbett-Davies and Goel 2018), which risks perpetuating bias into downstream retrieval and generation tasks (Silva, Tambwekar, and Gombolay 2021; Luccioni et al. 2023; Weidinger et al. 2021). In VL models specifically, terms related to race have been found to be associated more with people of color (Agarwal et al. 2021), and women are generally underrepresented in image retrieval tasks (Wang, Liu, and Wang 2021). However, these existing works focus only on very specific forms of bias while probing disparities using a small set of curated words (Bhargava and Forsyth 2019).

Given that more extensive and *intersectional* forms of bias may exist in VL models, there is a need to expand these

* Background workings of CLIP.
* Models vs. datasets
* SoBIT ↔ relationship with Sonia's work?

experiments to a richer taxonomy of potential biases. Further, the size of current datasets used to train such models makes it more difficult for humans to effectively identify low-quality, toxic, or harmful samples (Hanna and Park 2020; Kreutzer et al. 2022). Methods to find and describe biases in datasets are crucial to ensure safe adoption of VL models, yet few methods exist. Recent findings of child sexual abuse images (Thiel 2023) in LAION-5B (Schuhmann et al. 2022), a popular training VL training dataset (Ilharco et al. 2021), further highlights the need to audit the relationships learned by VL models in particular.

In this work, we target identifying and describing bias in pre-trained VL models at scale. We first propose a large new taxonomy, called **Social Bias Implications Taxonomy (So-B-IT)**, which spans ten different categories of biases. So-B-IT allows us to examine bias much more broadly than prior works, including biases associated with discrimination based on the model's implicit assumptions on images of *faces*. For instance, So-B-IT implements new categories of biased description, such as *Appearance* and *Occupation*, and extends word lists used by prior works (May et al. 2019; Steed and Caliskan 2021; Berg et al. 2022), allowing for finer grained analysis. Including new categories is crucial to investigate bias in VL models, as past work has targeted crime-related words (Bhargava and Forsyth 2019) or self-similarity across different demographic groups (Wolfe and Caliskan 2022).

Using So-B-IT, we then investigate bias in VL models by retrieving images from FairFace (Kärkkäinen and Joo 2019) — a dataset containing pictures of peoples' faces along with their age, gender, and race — that the model associates with the words in our taxonomy. For each category in So-B-IT, we quantify the demographic distributions of these retrieved images. As each image contains *only* a persons face, the association that a VL model makes between these images and the words in our taxonomy should be exclusively explained by the biases inherent to the model itself (Figure 1). Our analysis, based on studying four CLIP-based models (OAICLIP (Radford et al. 2021), OpenCLIP (Ilharco et al. 2021), FaceCLIP (Zheng et al. 2022), and DebiasCLIP (Berg et al. 2022), confirms that these systems encode significant racial and gender biases.

Because So-B-IT is more fine-grained than prior work, we uncover previously-unknown, intersectional biases in CLIP models. For example, OpenCLIP not only strongly associates *Homemaker* with *Women* significantly more than it does with *Men* (Stanovsky, Smith, and Zettlemoyer 2019; De-Arteaga et al. 2019), but overwhelmingly associates *Homemaker* with *Indian Women* more than it does for women of other races, which is previously uncharacterized in VL models. Our analysis also uncovers that debiasing VL models for *Gender* can significantly *increase* the racial bias of the model. This extends prior work showing the propensity of vision models to lean more strongly on remaining shortcuts after debiasing (Li et al. 2023) to VL models. We also extend our experiments to seek the *sources* of bias in VL training data. Our investigation into training data associated with biased terms confirms the non-representative demographic distributions we identify experimentally. While our experiments are based on CLIP due to its ubiquity (Rombach et al. 2022; Gao et al. 2023; Zhou et al. 2023), our analysis and the So-B-IT taxonomy is directly applicable to any VL model with a joint image and text encoding.

Our contributions can be summarized as follows:

- We propose a taxonomy, So-B-IT, that covers more categories of bias than prior work and at a finer grain. So-B-IT allows us to categorize a VL model's capacity to perpetuate societal bias in more representative tasks, and can be used broadly for vision and language auditing.
- Using So-B-IT, we audit four different versions of CLIP, finding that these models encode various forms of societal bias and stereotyping across gender and racial groups.
- Our findings indicate that debiasing with respect to one sensitive attribute, such as gender, does not necessarily eliminate other forms of bias, particularly racial bias.
- We investigate the source of such biases using CLIP's pre-training data, finding that disproportionate demographic representation may be a root cause of identified biases.

Related Work

Vision-Language Models. Recently, Vision-Language (VL) models have shown great potential for learning general visual representations and enabling prompting for zero-shot transfer to a range of downstream classification tasks (Radford et al. 2021; Jia et al. 2021; Zhang et al. 2020b). In this work, we focus our experiments on CLIP-based models (Radford et al. 2021). CLIP models utilizes an image encoder and a text encoder to match vector representations for images and text in a multi-modal embedding space. The training objective for CLIP is to maximize the cosine similarity between an image and its corresponding natural language caption, while minimizing the similarity between the image and all other captions in the batch, a training technique known as contrastive learning (Chen et al. 2020; Mnih and Kavukcuoglu 2013). By learning a meaningful joint representation between text and images, CLIP achieves strong zero-shot performances on vision benchmarks while also benefiting downstream VL tasks (Shen et al. 2021). The original CLIP model, OACLIP (Radford et al. 2021), was trained on a large scale image-text pair dataset. According to its creators, this dataset consists of approximately 300 million images and their associated text descriptions, but the source of this data was not specified.

Bias in Vision-Language Models. A number of prior works have focused on harmful biases of CLIP. Agarwal et al. (2021) conducted a preliminary study on racial and gender bias in the CLIP model showing that CLIP associates a "white" text label with the white racial label less than associating in the individuals belonging to the other racial groups with their group. Dehouche (2021) show that CLIP has a gender bias when prompted with gender neutral text. Wolfe, Banaji, and Caliskan (2022) show that multiracial people are more likely to be assigned a racial or

ethnic label corresponding to a minority or disadvantaged racial group. Wolfe and Caliskan (2022) show that biases related to the marking of age, gender and race in CLIP, reflect the biases of language and society which produced the training data. For instance, the default representation of a "person", is close to representations of white middle-aged men. In contrast to prior works, which only touch on harmful associations to gender and racial groups using a smaller list of captions only containing crime-related words (Agarwal et al. 2021), or consider self-similarity and markedness across different demographic groups (Wolfe and Caliskan 2022), we focus on identifiers of biases related to face images while providing a wider taxonomy of biases that could be attributed to human faces by a VL model.

One of the sources of bias in VL models is the lack of diverse and representative data. When the data used to train models is biased, the resulting models may also exhibit bias. This has been observed in a number of studies (Bhargava and Forsyth 2019; Birhane, Prabhu, and Kahembwe 2021; Tang et al. 2021b). More importantly, offensive and biased content can be found in open-source training corpora (e.g. LAION (Schuhmann et al. 2021)) that are used to train open-source versions of CLIP (Birhane, Prabhu, and Kahembwe 2021). Prior work has found that such datasets contain pornographic, misogynistic, and stereotypical images and accompanying text captions. Different types of representational harms has also been studied in the context of image captioning (Wang et al. 2022).

Debiasing Vision-Language models To address lack of diversity in the training data, Bhargava and Forsyth (2019) have proposed methods such as data augmentation and balancing as a means of reducing bias in the training data. Another approach to addressing bias in vision models is through model-level adjustments. Srinivasan and Bisk (2021) proposed the use of bias mitigation techniques such as debiasing the input representation and adversarial training to reduce bias in pre-trained vision-and-language models. Zhang et al. (2020b) suggested the use of environment re-splitting and feature replacement to diagnose environmental bias in vision-and-language navigation. Cho, Zala, and Bansal (2022) proposed the evaluation of visual reasoning skills and social biases as a means of identifying and addressing biases in text-to-image generation. More recently, Berg et al. (2022) have proposed prepending learned vision embeddings to text queries that are trained with adversarial can help debias the representation space.

Creating a Taxonomy of Social Biases in Vision-Language Models

We propose a taxonomy of VL model biases called So-B-IT (Social Bias Implications Taxonomy), which categorizes 374 words into nine types of bias as show in Table ?? in the Appendix. We define bias as *a disproportionate association between a word or concept and a specific demographic group in comparison to others* (Operario and Fiske 2001; Levinson and Young 2009), and especially focus on gender and racial identities.

In the first step, we consider different categories of biases. Our first step in creating the taxonomy involved a review of existing literature in biases. There are many papers that propose different types of biases in AI models, but we selected those that are either actionable or have higher sentiment associated with them, or they exhibit allocative harms (Nadeem, Bethke, and Reddy 2020; Steed and Caliskan 2021). Allocative harms involve making assumptions about people that can lead to unfair resource distribution, whereas representational harms involve the misrepresentation of people that can perpetuate stereotypes (Barocas et al. 2017). Our taxonomy focuses on representational harms that could turn into allocative harms. By identifying these biases, we aim to mitigate potential negative impacts in real-world applications. While our taxonomy is not exhaustive, it is designed to be easily extendable. Below, we discuss the different types of categories included in our taxonomy.

Algorithmic Governance Areas

To examine potential biases in VL models, we refer to the top AI use cases by policy areas proposed by Engstrom et al. (2020), that are specifically related applications that can harm marginalized groups by using images or videos of faces. So-B-IT contains potentially-biased words from the following categories.

- **Criminal Justice**. Machine learning models have been deployed in criminal justice for tasks including recividism prediction (Berk 2017; Tolan et al. 2019), predictive policing (Shapiro 2017), and criminal risk assessment (Berk, Berk, and Drougas 2019). These models have also been shown to have disparate perform across demographic groups. For example, models used to predict recividism risk have been shown to exhibit a higher false positive rate for Black inmates (Wadsworth, Vera, and Piech 2018). We probe the relations learned by CLIP and concepts associated with historical biases (Alexander 2020) such as "criminal", "delinquent", and "terrorist".
- **Education and wealth**. Discrimination based on education level is common, and automated inference can lead to real harm (Brown and Tannock 2009). For instance, in education-based hiring (Tannock 2008), candidacy can be overlooked for those with less education (Van Noord et al. 2019). Moreover, given recent use of ML to predict student dropout in university admission decisions, detecting educational bias in VL models is increasingly important (Liu et al. 2022).
- **Health**. There is a long history of bias and discrimination in healthcare (Govender and Penn-Kekana 2008). Such bias can worsen outcomes for people struggling with mental health (Thornicroft, Rose, and Kassam 2007) and for the aging population (Kydd and Fleming 2015), especially for racial minorities (Peek et al. 2011). We check for health-based biases using words like "disabled," "mentally ill," and "addicted".
- **Occupation**. Different occupations are unfairly associated with different groups of people. For example, many recent works have studied associations between gender and occupation (Singh et al. 2020). A well-established

example is people subconsciously stereotyping doctors as men and nurses as women (Banaji and Hardin 1996). Then, well-known biases can slip into trained models (De-Arteaga et al. 2019; Bolukbasi et al. 2016). We thus define a long list of occupations. We include some with known biases like "nurse" and "doctor," but also also include new occupations like "painter" and "geologist" to investigate new biases.

Stereotypical Markers

In addition to algorithmic governance areas, we also consider categories that are not directly related to known applications. However, these categories may be used spuriously as a proxy for a particular gender or racial demographic group. Probing VL model biases in these categories can help prevent the misrepresentation or under-representation of certain groups, which can have serious consequences for individuals' lives and opportunities. For instance, a biased model that associates specific physical traits or behaviors with a particular gender or racial group may result in unfair or discriminatory hiring practices in the employment sector. Similarly, a model that perpetuates harmful stereotypes about a specific group may contribute to the over-criminalization of that group in the criminal justice system (Alexander 2020). For instance, in recommendation-based models such as TikTok's algorithm, Karizat et al. (2021) have shown how participants alter their behavior and thus their algorithmic profile to resist the suppression of marginalized social identities via individual and collective action. Therefore, we include words from the following categories in `So-B-IT`.

- **Appearance**. Our self-worth is often tied to our perceived physical appearance (Patrick, Neighbors, and Knee 2004). For example, comparing oneself to cultural beauty standards can be detrimental, especially for members of minority groups (Mahajan 2007). To investigate appearance-related biases in CLIP, we look for disproportionate associations between racial and gender identities and the set of Appearance words in Table **??**. We focus on subjective descriptors of cultural attractiveness like "beautiful" and "chubby" but also include words that may correlate with appearance like "old" and "tall".
- **Behavior**. Bias can stem from assumptions about others' behavior. For example, incorrect assumptions about behavior can occur in interracial interactions, often to the detriment of minority populations (Dovidio et al. 2002). We study such behavior bias using mostly adjectives like "aggressive" or "calm," which describe interactions with the world.
- **Portrayal in media**. How people are depicted can reinforce historical biases. For example, recent media coverage of Russia's war against Ukraine compares Ukraine to the Middle East, perpetuating harmful "war-torn" connotations (Al Lawati and Ebrahim 2022). Similar portrayal biases are common in social media (Singh et al. 2020; Hartvigsen et al. 2022). To investigate such biases, we use words like "third-world" and "savage" along with other stereotypes associated with different regions like "hypersexual" and "exotic".
- **Politics**. The US Congress has a lengthy history of lacking gender and racial diversity among its members (Reny 2017). As such, large machine learning models trained on historical data may learn to associate positions of power with specific groups (Andrich and Domahidi 2022), which may further propagate such disparities. In addition, associating specific political beliefs with certain demographic groups can lead to further polarization and discrimination (Gordon, Babaeianjelodar, and Matthews 2020), especially when such models are used in online advertising and recommender systems. Here, we evaluate association of demographic groups with political affiliations such as "liberal", "conservative", and "liberterian" in VL models.
- **Religion**. Religious discrimination is common around the world (Fox 2007). For example, there is a long history of religious persecution in the workplace (Ghumman et al. 2013) and in justice systems (Al-Qattan 1999). While the persecuted groups are different, there is a disturbing and consistent trend around the world for religious majorities to persecute local minorities. We investigate CLIP's religious bias using both religions like "christian" and "muslim" but also stereotypes like "intolerant" and "superstitious".

For word selection, we started by examining word lists from prior work on bias in language models and image captioning (Nadeem, Bethke, and Reddy 2020; Steed and Caliskan 2021). We manually selected words representative of potential biases in VL models and focused on those that we believed could lead to harmful associations with racial or gender groups. To expand the taxonomy in each category beyond existing works, we also used GPT-3.5 assistance to generate a larger candidate list which we filtered manually, utilizing around 40% of the language model's suggestions. This taxonomy is non-exhaustive, and we acknowledge that our choices are skewed towards a Western focus. Our aim was to provide a concrete starting point for auditing many biases, grounded in real-world applications and societal stereotypes.

Our proposed taxonomy covers many potential applications of CLIP and other VL models. For instance, a VL model may be used for affect detection in airport security based on people's appearance, ultimately determining who should be screened. Disproportionately attributing a word like "anxious" to one demographic group may then target them. As another example, consider the task of object detection with co-occurring human faces. A biased CLIP model with ingrained stereotypes about certain demographic groups may perform disparately between such demographic groups on the object detection task (Hall et al. 2023).

Vision-Language Model Bias Identification Pipeline: Exploring Different Types of Biases Across Demographic Groups

We propose a simple framework for evaluating potential biases of VL models in facial recognition tasks. We focus on harmful associations present in the model, specifically based on retrieved images of people from different demographic groups as defined by the intersection of race and gender.

Setup for Vision-Language Bias Identification Pipeline

To identify biases in VL models, we employ a word-association approach that focuses on identifying biases based on a given adjective or word's association with individuals from a certain demographic group. Specifically, we measure the similarity between the VL model's encoding of the word and its encoding of images of human faces from the FairFace dataset belonging to each demographic group.

Data and Model FairFace (Kärkkäinen and Joo 2019) is a face image dataset that is balanced in terms of race and gender. It includes 108,501 images from seven different racial groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino/Hispanic. The images were collected from the YFCC-100M Flickr dataset and labeled with information about race, gender, and age groups. In order to capture social biases in the face images, we use the taxonomy of social biases as described previously and shown in the taxonomy table in the appendix.

Caption generation To generate the captions, we design templates for four categories of words: *adjectives*, *profession or political nouns*, *object*, and *activities*. Then, for each word in our taxonomy, we use the caption `a photo of a/an [adjective] person` for adjectives, `a photo of a/an [noun]` for nouns, and `a photo of a person who is [gerund verb for activity]`. We then calculate the similarity of the CLIP model's response to all images in the training set of the FairFace dataset for each category of prompts. We obtain the similarity scores using the cosine similarity between the prompt embedding and the image embedding in CLIP's representation space.

Measuring Image-Caption Association for Demographic Groups

In the next step, we want to measure how descriptive a caption is for a certain demographic group in comparison to the rest of the groups. As both caption and image representations lie within a joint representation space, we use cosine similarity $d(c, x)$ to measure the similarity between caption c and image x. To measure the level of association between captions and demographic groups, we employ a method that is inspired by the Word Embedding Association Test (WEAT) (Caliskan, Bryson, and Narayanan 2017) measure in natural language processing. Specifically, we select a target demographic group G, such as a particular race or gender, and compute the average cosine similarity between a given caption c and the image representations of the images belonging to that group: $\sum_{g \in G} d(c, g)/|G|$ – as well as the representations of all other images in the dataset: $\sum_{g' \in \bar{G}} d(c, g')/|\bar{G}|$. The difference between the two is a measure of how closely the caption is associated with group G, as determined by the VL model's representations.

To obtain a normalized metric that accounts for the overall variance of similarity scores in the dataset $D = G \cup \bar{G}$, we divide the difference between the average similarity score of the selected demographic group and the average similarity score of all other groups combined by the standard deviation of the cosine similarity scores between captions and all images in the dataset as below:

$$\texttt{C-ASC}(c, G) = \frac{\frac{1}{|G|}\sum_{g \in G} d(c, g) - \frac{1}{|\bar{G}|}\sum_{g' \in \bar{G}} d(c, g')}{\text{std}_{u \in D}\, d(c, u)}$$

This normalized metric corresponds to Cohen's effect size in the single category WEAT measure, which quantifies the degree of separation between target group and the rest of samples in image embeddings, as well as lower standard deviation, or more concentrated similarities of the caption to images in the dataset. Note that in the case where a sensitive attribute takes multiple values, we consider one group e.g. people from a certain race as G and the rest of the samples in the dataset as $\bar{G}$.

By applying this metric to image-caption similarities in the CLIP representation space, we can evaluate the level of inductive bias that may be present towards certain demographic groups. This approach allows us to identify potential harmful associations that may exist in the model, and to develop strategies for mitigating any biases that are identified.

Identifying Bias with Caption-Association Image Retrieval

For each category of bias, given the similarities of captions corresponding to words in the taxonomy of the bias type, we retrieve the top-k samples with the highest similarity scores for each caption. We use k=100 for our experiments. For each prompt, we focus on the demographic composition of the top-k samples by computing the distribution of the race and gender of people in the retrieved images. Since the FairFace dataset has an equal number of samples across gender and racial groups, we do not need to normalize the proportions. Thus, if the distribution of the demographic group is uniform across the top-k images, we infer that the VL model exhibits no social bias for this particular word. Conversely, if the proportion of a certain demographic group in the top-k samples is significantly higher or lower than expected, we infer the presence of bias. We repeat this process for each prompt and analyze the results to identify prevalent categories of biases in the VL model.

Auditing Demographic Biases in Vision-Language Models

Our taxonomy `So-B-IT` can be applied to audit any VL model. Here, we use it to audit the following four CLIP models:

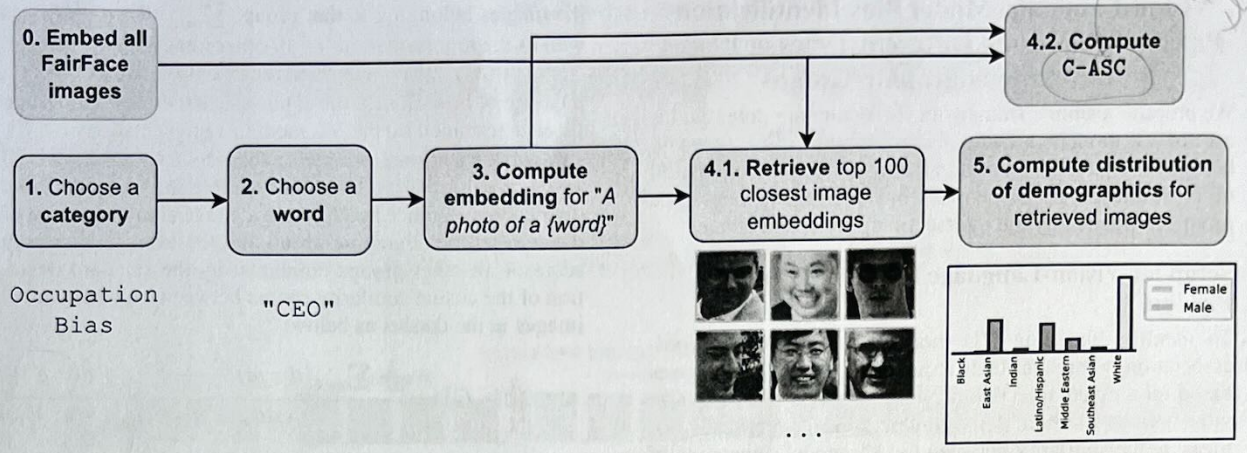

Figure 2: Flowchart demonstrating the process for image retrieval in FairFace. For each word of interest in each category, we compute its embedding with the CLIP text encoder, and retrieve the top 100 closest images by cosine similarity. We then examine the demographic distribution of retrieved images, and compute the C-ASC score.

- OAICLIP (Radford et al. 2021): The original CLIP model (with a ViT-B/32 transformer architecture) released by OpenAI. Note that the pretraining data is not available.
- OpenCLIP (Ilharco et al. 2021): A CLIP ViT-H/14 model trained with the LAION-2B dataset (Schuhmann et al. 2022) using the OpenCLIP library. We note that this dataset has since been removed due to concerns regarding the presence of child sexual abuse material (Thiel 2023), though the model weights are still publicly available. We further discuss the influence of pretraining data and transparency in the discussions.
- FaceCLIP (Zheng et al. 2022): A variant of CLIP ViT-B/16 which has been pretrained on the LAION-FACE dataset (Zheng et al. 2022), which is a subset of LAION-400m (Schuhmann et al. 2021) consisting of only face images.
- DebiasCLIP (Berg et al. 2022): A variant of CLIP ViT-B/16 which has been debiased with respect to *gender* using the debiasing approach proposed by Berg et al. (2022).

We start by reporting aggregate bias statistics for each category of bias across all models. Next, we examine the effect of debiasing, by comparing OAICLIP with DebiasCLIP. Then, we dive into the biases of OpenCLIP by examining select words across specific categories. Finally, we conduct an experimental analysis of the presence of occupation stereotypes across gender in the LAION-400m subset (Schuhmann et al. 2021), as a potential explanation for the biases learned by OpenCLIP.

So-B-IT Identifies That VL Models Harbor Racial and Gender Biases

In this section, we use So-B-IT to audit all four CLIP models in order to compute aggregate biases for each category.

To quantify bias for each word, we compute the *normalized entropy* of the discrete probability distribution over groups defined by the top-k retrieval procedure, using k = 100. Here, a normalized entropy of 1 corresponds to a uniform distribution of retrieved images over groups, and thus is the most fair by our definition. Conversely, a lower normalized entropy corresponds to greater bias. We then compute the bias of a category for a model as the average normalized entropy of all words in that category, with images retrieved using the model.

We plot the normalized entropies for each model for gender and race as sensitive attributes in Figure 3 and Figure 4 respectively. Our audit reveals several interesting findings. First, we find that on aggregate, bias across gender appears most prominently in the occupation category by a large margin, while biases across race appears largely in the religion, political, and education categories. Next, we find that DebiasCLIP indeed exhibits lower bias by gender compared to other models. However, debiasing by gender does not mitigate biases across race, and in fact, DebiasCLIP exhibits the most bias across racial groups out of all models. This reveals the weakness of debiasing approaches which can only target a given set of sensitive attributes (Li et al. 2023). Finally, we find that OpenCLIP is the most fair model with respect to race, but is the most biased model with respect to gender. To provide an explanation for these gender biases, we conduct an analysis of the training data of OpenCLIP, focusing on the category of greatest bias – occupation stereotypes.

Debiased Models Are Still Biased

We now perform an intersectional audit of CLIP debiasing. Recent methods have been proposed to debias VL models with respect to protected attributes such as race and gender (Berg et al. 2022; Zhu et al. 2023; Seth, Hemani, and Agarwal 2023). However, whether debiasing for one attribute im-

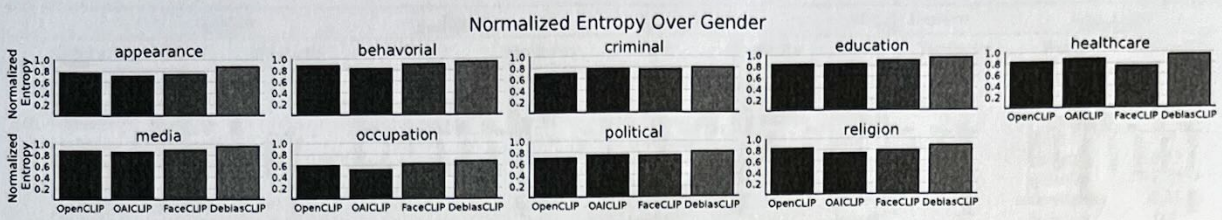

Figure 3: Normalized entropy of the top-k distribution over *gender* for each category in So-B-IT. Higher values indicate less gender bias. The gender bias of VL models is most stark for the occupation category. As expected, DebiasCLIP exhibits the least gender bias.

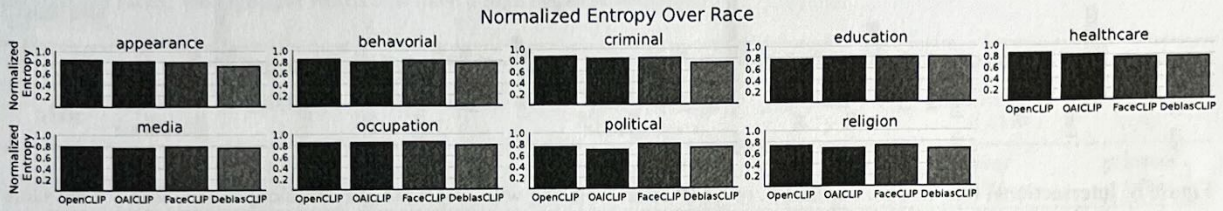

Figure 4: Normalized entropy of the top-k distribution over *race* for each category in So-B-IT. Higher values indicate less racial bias. The racial bias of VL models is most prominently seen in the religion, political, and education categories.

proves, maintains, or degrades the bias of the remaining attributes is unknown. We thus use So-B-IT to perform an intersectional evaluation of the bias in OAICLIP, and compare it with CLIP that has been debiased with respect to *gender* (DebiasCLIP).

Figure 5 shows CLIP's association of each specific race and gender for a set of words highly associated with males. Stark gender imbalance is clearly evident; each word is much more strongly associated with males regardless of race. However, a racial bias is also evident: middle eastern men are much more associated with the words "terrorist" and "barbaric", while East Asians are associated with the words "ambitious" and "rich".

We now observe CLIP's association for the same set of words *after* debiasing the model with respect to gender (Figure 6). Notably, the effect of the debiasing clearly decreases the relative differential between males and females — for white people. In particular, for white males and females the gender differential is most starkly decreased for "positive" words like "ambitious" and "rich". However, this debiasing had an unintended effect of making the model more strongly associate white people with positive words in general. Before debiasing, East Asians were significantly more associated with "ambitious, "rich" and "jock" than white people were. After debiasing *with respect to gender*, South East Asians of both genders were significantly less associated with these words, while the association with white people increased substantially. Interestingly, for a very negative word like "terrorism", this debiasing *increased* the gender disparity for middle eastern people: now middle eastern men are even more strongly associated with "terrorist", while middle eastern women are less associated with it. This phenomenon closely mirrors the Whac-A-Mole dilemma previously observed in computer vision systems (Li et al. 2023), where correcting for one source of spuriouss correlation results in models leaning more heavily on other shortcuts.

These unexpected changes in racial biases highlight a crucial point: *debiasing for one attribute can significantly increase bias for other attributes*. It is thus imperative to perform intersectional evaluations when developing or applying debiasing strategies. Without auditing for unexpected changes in associations for a range of attributes, well-intentioned attempts to decrease bias may actually result in models that are less fair.

Diving Into The Biases In OpenCLIP

We now use So-B-IT to conduct a more fine-grained analysis to discover the specific biases encoded in OpenCLIP – a version of CLIP that that has been trained on the LAION dataset. For each category of bias in So-B-IT, we use our list of words to create captions that could lead to biased associations. We first measure the image-caption associations using C-ASC scores to find the captions that are most associated with each racial or gender group. Then, to better understand the distribution of samples that were most similar to each word in the list, we perform image retrieval using the FairFace dataset. Finally, we examine the distributions of the 100 most-similar images to each caption across race, gender, and their intersection as described in Figure 2. Due to space constraints, we base the following analysis on select words from each category. The full set of results and additional analysis is available in the appendix.

Ambitious men and bossy women We find that CLIP associates positive behaviors with men and negative behaviors with women, as shown in Table 2. Corroborating previous works (Bordia and Bowman 2019), adjectives like "ambitious" are men's most similar words and adjectives like "bossy" are women's. The associations have nothing to do

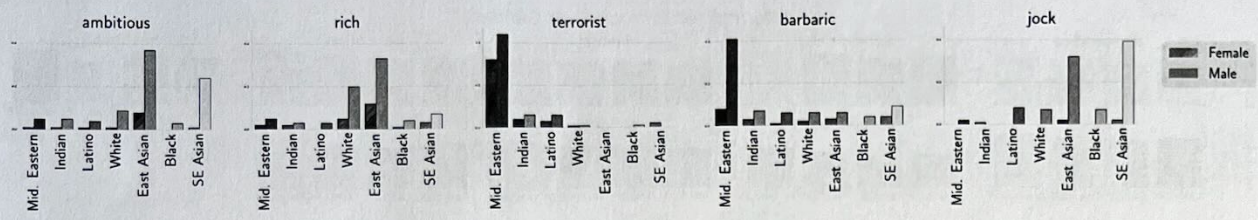

Figure 5: Intersectional bias in OAICLIP for a set of words most strongly associated with the "Male" gender.

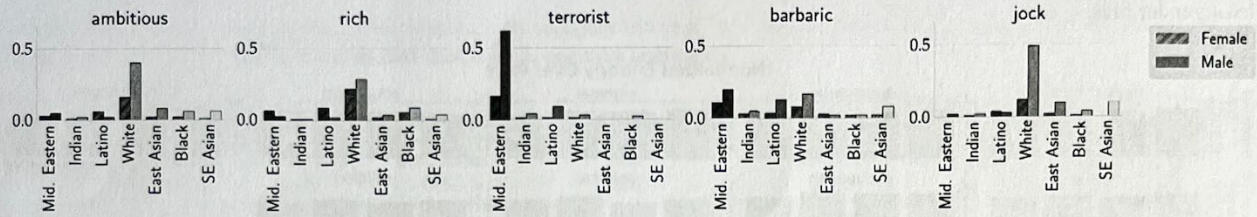

Figure 6: Intersectional bias in DebiasCLIP, or CLIP after debiasing with respect to gender (Berg et al. 2022), for the same set of words shown in Figure 5.

with gender in reality, yet pose harmful consequences, especially in high-stakes situations like hiring (Chen et al. 2014). Already, CLIP has been used for emotion detection (Bondielli and Passaro 2021), and perpetuating such associations may disadvantage women (Rhue 2018). The impact of this bias may differ by gender and is influenced by cultural norms, personality, and past experience.

Minority groups flagged as dangerous. Our analysis reveals that certain racial groups are associated with negative attributes in CLIP's representation space, with the top words associated with white, black, Latino/Hispanic, Middle Eastern, and Indian being "psychopath", "felon", "gang-related", "terrorist", and "fraud", respectively as shown in table 1.

These biases could have significant implications in the context of criminal justice. For example, as machine learning models are already starting to be used for recividism prediction (Berk 2017), such biases could lead to overestimation of recividism risk for certain demographic groups. Moreover, if the model associates Middle Eastern people with the attributes of being terrorists and militants, it could lead to biased surveillance and racial profiling.

There are real-world examples of this type of biased surveillance, such as the Screening of Passengers by Observation Techniques (SPOT) program, which has been criticized for its racial bias (III 2013). The use of biased AI models such as CLIP in such programs can exacerbate these biases and lead to the misidentification of innocent individuals as potential threats (Chamieh et al. 2018). Similarly, the use of biased AI models in risk assessment tools, such as predictive policing algorithms, can lead to overpolicing and overincarceration of certain communities.

Tribal Indians and Latino Immigrants Our experiment reveals that certain racial groups are associated with stereotypical and negative words such as "ghetto", "immigrant", "barbaric", "oriental", and "tribal", as shown in Table 1. Black, Hispanic, Middle Eastern, Asian, and Indian people are respectively strongly associated with these words.

This could particularly be harmful in applications such as content moderation, and data filtering. One potential use of CLIP is for scoring, ranking, or filtering media, as prior works have used CLIP for making an evaluation metric for image captioning (Hessel et al. 2021) and ranking video annotations (Tang et al. 2021a). However, if CLIP is used to rank media that includes these stereotypes, it could reinforce harmful stereotypes and lead to further discrimination.

Intersectional Occupations: White male CEOs and Indian woman homemakers Previous studies have reported on the presence of harmful biases in vision-language models, such as the association of certain occupations with specific genders, such as "nurse" being predominantly associated with women (Bianchi et al. 2022). Here, our experiment again highlights the presence of intersectional biases in VL models. For instance, the occupation of "maid" is more frequently associated with women of color, while "homemaker" is mostly associated with Indian women as in figure 7. These intersectional biases have not been addressed in previous studies, but it is crucial to audit vision-language models for them as they can reinforce harmful stereotypes and further marginalize underrepresented groups.

Seeking Sources of Bias in Training Data

As a proof-of-concept, we examine the set of words above for which CLIP displays significant occupation stereotyping between genders. For each word, we construct a relevant subset of the dataset by selecting all samples for which the caption contains the word of interest, as well as at least one gendered pronoun.

First, we examine, for each word of interest, the likelihood of it associating with each gender in the caption. From Table 3, we find that gender stereotypes are clearly present in the LAION captions. For example, captions containing the

	White		Black		Latino/Hispanic		Middle Eastern		East Asian	
Category	**Word**	**Similarity**	**Word**	**Similarity**	**Word**	**Similarity**	**Word**	**Similarity**	**Word**	**Similarity**
Appearance	unattractive	0.31	young	0.19	overweight	0.19	handsome	0.26	cute	0.58
Behavioral	content	0.46	welfare-dependent	0.2	naggy	0.16	ashamed	0.34	conformist	0.84
Edu. + Employ.	dumb	0.33	underrepresented	0.39	undocumented	0.64	migrant	0.28	smart	0.44
Crime + Justice	psychopath	0.32	felon	0.37	gang-related	0.2	terrorist	1.05	abnormal	0.4
Healthcare	addicted	0.14	underprivileged	0.2	obese	0.19	addicted	0.17	lethargic	0.35
Geo. + Media	sassy	0.52	ghetto	0.42	immigrant	0.28	barbaric	0.32	oriental	1.14
Political	globalist	0.76	populist	-0.07	socialist	0.36	terrorist	1.11	authoritarian	0.31
Religion	jewish	0.59	primitive	0.13	jewish	0.16	jewish	1.08	buddhist	0.8
Occupation	attorney	0.87	porter	0.48	counselor	0.25	historian	0.44	pianist	0.5
Stereotyping	redneck	0.72	racist	0.4	cheerleader	0.2	thug	0.36	geek	0.48

Table 1: We show the most similar word from each of category based on similarity score in `OpenCLIP`'s representation space, for selected races. We highlight words that have a high negative sentiment.

		Appearance	Behavioral	Edu. + Employment	Criminal Justice	Geo. + Media	Occupation
Male	**Top Word**	handsome	ambitious	rich	terrorist	barbaric	delivery man
	Similarity	1.19	0.53	0.62	0.75	0.43	0.95
Female	**Top Word**	pretty	bossy	underrepresented	abnormal	sassy	princess
	Similarity	0.68	0.6	0.21	0.02	0.71	1.09

Table 2: We show the most similar word from each of the remaining categories based on similarity scores in `OpenCLIP`'s representation space, for female and male genders. The highlighted words have a negative sentiment. Top behavioral word for male and female groups respectively have high positive and negative sentiment.

	Male	Female	# Images
maid	27.9%	72.1%	6,917
nurse	31.0%	69.0%	18,742
housekeeper	34.3%	65.7%	787
assistant	56.4%	43.6%	12,423
porter	67.9%	32.1%	2,784
farmer	67.4%	32.6%	11,493
ceo	74.2%	25.8%	11,939

Table 3: For each word of interest, we subset LAION-400m to samples with captions containing the word and a gendered pronoun. We report the proportion of each gender associated with each word, finding that the training data for OpenCLIP contains historial biases with respect to gender and occupation.

word "nurse" are much more likely to contain a female pronoun than a male pronoun. Next, we select all images with captions containing each word and at least one gendered pronoun, and manually choose a random subset of these images which contain a human face. We visualize these images in the appendix, finding that stereotypes in the dataset also extend to the associated images. For example, captions which contain the word "nurse" are predominantly associated with images which may be conventionally identified as female-gendered.

Our analyses present a mechanism by which CLIP may have learnt the biases we observe. It also highlights the role of undesirable historical biases present in the training data, and the importance of tackling such dataset stereotypes prior to model training.

Conclusion

In conclusion, our study demonstrates that VL models such as CLIP can perpetuate harmful societal biases and stereotypes, particularly with regards to gender and racial groups. Through the use of our taxonomy, `So-B-IT`, we were able to identify biases in each category for different social groups, which highlights the importance of auditing VL models for potential societal harms.

Our findings underline the need for greater attention to be given to the potential social biases and stereotypes encoded in CLIP representations, particularly in applications that impact human lives such as criminal justice, healthcare, and employment. The harms of such biases in VL models can be far-reaching, and could potentially affect individuals' opportunities and even contribute to systemic discrimination. We believe that our work contributes to the ongoing conversation around the need for ethical and fair foundation models, particularly in the development and deployment of VL models. Our proposed taxonomy, `So-B-IT`, can be used as a tool for broader audits of vision and language models, and our analysis of CLIP's pre-training data highlights the importance of examining pre-training data for potential sources of biases.

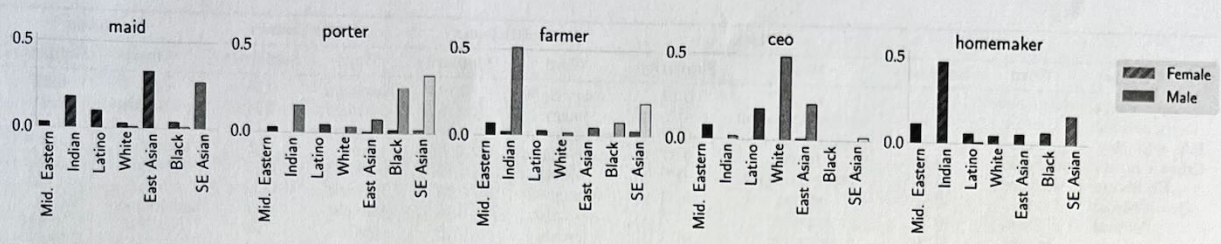

Figure 7: Intersectional biases in occupation: When the model is prompted with "a photo of a homemaker", more than 50% of the retrieved Images are Indian women, while for "farmer" and "CEO" images of Indian men and White men are retrieved respectively. For occupations other than "CEO", a few images of White people are retrieved.

Ethical Considerations

Pre-training data and transparency One of the key factors that can influence the representations learned by a VL model like CLIP is the data it is trained on. The dataset that `OAICLIP` (Radford et al. 2019) was trained on was not released, though there are some speculations about the sources of the data (Nguyen et al. 2022). This lack of transparency makes it difficult to decode a model's biases and limitations. Given the potential biases and discrimination identified in our experiments, it is important to consider the data used to train the model and how it may have influenced the representations learned by CLIP. For example, the recent publicly available LAION dataset (Schuhmann et al. 2022) has been found to contain both images and textual representations of rape, pornography (Birhane, Prabhu, and Kahembwe 2021), and child sexual abuse material (Thiel 2023). Given our finding of the correlation between gender stereotypes in LAION-400m and their presence in the CLIP model trained on such data, it is likely that other problematic correlations could have been learned by the model as well. Thus, it is critical to consider the issue of bias through a data-centric perspective (Oala et al. 2023). Manual curation of a pre-training dataset without undesirable stereotypes may be required to obtain a model truly free of such biases (Jernite et al. 2022; Gadre et al. 2023; Birhane et al. 2023).

Another potential concern is data colonialism in the training data of VL systems and other foundation models. The use of data from marginalized or colonized populations without proper consent or compensation can perpetuate existing power imbalances and contribute to the exploitation of these groups.

Regulation and Auditing Given the potential biases and discrimination identified in our experiments, regulating and auditing VL models is crucial for fairness and equality. Bias audits, impact assessments, and algorithmic accountability frameworks (Metcalf et al. 2021; Raji et al. 2020) can help evaluate performance, transparency, and fairness. Importantly, VL evaluations must be *intersectional*. Our analysis shows that considering bias for single attributes is insufficient: `OpenCLIP` associates *Homemakers* specifically with Indian women, for instance. Moreover, bias mitigation for one attribute can increase bias for others. Future debiasing approaches should use an intersectional evaluation framework like `So-B-IT` to measure effectiveness accurately.

Risks of racial erasure and dehumanization Our experiments highlight limited associations between adjectives of different categories in `So-B-IT` and racial groups such as Black and Latino/Hispanic. For instance, top-k image retrieval for `OpenCLIP` show that the model retrieves images few Black or Latino/Hispanic individuals for almost all words in behavioral category as shown in Table **??**. This limited association between adjectives and racial groups could result in a failure to recognize and tag images of people from these groups in many behavioral categories, particularly those that rely on automatic image classification, and facial recognition tasks.

The limited associations between adjectives and racial groups in the `OpenCLIP`'s representation space is linked to the ethical issue of mechanistic dehumanization. Mechanistic dehumanization (Haslam 2006; Haslam et al. 2008) refers to the denial of qualities of "human nature" to a particular group, and is a concerning issue for automated image tagging (Barlas et al. 2021). In this case, the limited associations between adjectives and racial groups could potentially lead to the denial of "human nature" qualities to certain racial groups, particularly Black and Latino/Hispanic individuals. By failing to recognize and tag these individuals with a wide range of attributes, the model could be perpetuating the view that they are interchangeable, lacking agency, and superficial, denying them the qualities of "human nature" that are afforded to other groups.

This dehumanization could have serious ethical implications, particularly in the context of machine learning models that that rely on these representations. If certain groups are denied qualities of "human nature" in these models, it could lead to biased decision-making, discriminatory practices, and perpetuation of existing power imbalances.

Limitations We recognize several limitations with our study. First, we make use of the FairFace dataset, which has several flaws. In particular, all race, gender and age attribute labels were obtained from Amazon Mechanical Turks, and so is already the product of human biases and stereotyping. In addition, the assumption of binary gender and the consideration of only seven racial groups is not representative of the full range of identities present in society, and one may also identify with a different gender or race over time. Other facial image datasets such as CelebA (Liu et al. 2015) or UTKFace (Zhang, Song, and Qi 2017) suffer from similar flaws, and conducting similar analyses on additional datasets

is an area of future work.

However, we still focus on FairFace in this work, as it has been the subject of many prior works studying bias in vision models (Cheng et al. 2021; Serna et al. 2022; Agarwal et al. 2021) and is one of the few facial image datasets which emphasized diversity and balanced race composition during data collection. Second, we only consider image retrieval tasks based on short captions on facial images. However, real-world uses of VL models may not be limited to captions of this particular format, and the images to be retrieved may not be close-up images of human faces. Future research is needed to evaluate how the biases observed here translate to a wider range of applications, as well as a larger array of VL models such as DALLE-2 (Ramesh et al. 2022). Finally, our experimental analysis is limited to the harmful associations learned by CLIP-based models, and does not account for how the images retrieved by CLIP may be interpreted by end-users. Further research and user studies are needed to understand and quantify the potential real-world consequences of these biases in deployed systems.

The protocol we propose in this work is applicable to any VL model that learns a shared embedding space between image and text representations. This includes models like BLIP (Li et al. 2022a) and LLaVa (Liu et al. 2024), both of which combine CLIP's frozen vision encoder while training the language model. We focus on CLIP in this work given its ubiquity and widespread adoption as a foundation model for multimodal representation learning. However, we emphasize that our findings have broader implications due to the core role of CLIP representations in other VL models.

References

Agarwal, S.; Krueger, G.; Clark, J.; Radford, A.; Kim, J. W.; and Brundage, M. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Al Lawati, A.; and Ebrahim, N. 2022. How the Ukraine war exposed Western media bias. *CNN*.

Al-Qattan, N. 1999. Dhimmīs in the Muslim court: legal autonomy and religious discrimination. *International Journal of Middle East Studies*, 31(3): 429–444.

Alexander, M. 2020. *The new Jim Crow: Mass incarceration in the age of colorblindness*. The New Press.

Andrich, A.; and Domahidi, E. 2022. A Leader and a Lady? A Computational Approach to Detection of Political Gender Stereotypes in Facebook User Comments. *International Journal of Communication*, 17: 20.

Banaji, M. R.; and Hardin, C. D. 1996. Automatic stereotyping. *Psychological science*, 7(3): 136–141.

Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2021. Person, Human, Neither: The Dehumanization Potential of Automated Image Tagging. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 357–367.

Barocas, S.; Crawford, K.; Shapiro, A.; and Wallach, H. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2017.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Berg, H.; Hall, S. M.; Bhalgat, Y.; Yang, W.; Kirk, H. R.; Shtedritski, A.; and Bain, M. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*.

Berk, R. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13: 193–216.

Berk, R.; Berk, D.; and Drougas, D. 2019. *Machine learning risk assessments in criminal justice settings*. Springer.

Bhargava, S.; and Forsyth, D. 2019. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*.

Birhane, A.; Prabhu, V.; Han, S.; Boddeti, V. N.; and Luccioni, A. S. 2023. Into the LAIONs Den: Investigating Hate in Multimodal Datasets. *arXiv preprint arXiv:2311.03449*.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Bondielli, A.; and Passaro, L. C. 2021. Leveraging CLIP for Image Emotion Recognition. In *NL4AI@ AI* IA*.

Bordia, S.; and Bowman, S. R. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Brown, P.; and Tannock, S. 2009. Education, meritocracy and the global war for talent. *Journal of Education Policy*, 24(4): 377–392.

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

Chamieh, J.; Al Hamar, J.; Al-Mohannadi, H.; Al Hamar, M.; Al-Mutlaq, A.; and Musa, A. 2018. Biometric of intent: a new approach identifying potential threat in highly secured facilities. In *2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, 193–197. IEEE.