The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models

ABEBA BIRHANE*, Mozilla Foundation & School of Computer Science and Statistics, Trinity College Dublin, Ireland
SEPEHR DEHDASHTIAN†, Michigan State University, Department of Computer Science and Engineering, USA
VINAY UDAY PRABHU, HAL51 Inc, USA
VISHNU BODDETI, Michigan State University, Department of Computer Science and Engineering, USA

'Scale the model, scale the data, scale the GPU farms' is the reigning sentiment in the world of generative AI today. While model scaling has been extensively studied, data scaling and its downstream impacts on model performance remain under-explored. This is particularly important in the context of multimodal datasets whose main source is the World Wide Web, condensed and packaged as the Common Crawl dump, which is known to exhibit numerous drawbacks. In this paper, we evaluate the downstream impact of dataset scaling on 14 visio-linguistic models (VLMs) trained on the LAION400-M and LAION-2B datasets by measuring racial and gender bias using the Chicago Face Dataset (CFD) as the probe. Our results show that as the training data increased, the probability of a pre-trained CLIP model misclassifying human images as offensive non-human classes such as chimpanzee, gorilla, and orangutan decreased, but misclassifying the same images as human offensive classes such as criminal increased. Furthermore, of the 14 Vision Transformer-based VLMs we evaluated, the probability of predicting an image of a Black man and a Latino man as *criminal* increases by *65%* and *69%*, respectively, when the dataset is scaled from 400M to 2B samples for the larger ViT-L models. Conversely, for the *smaller* base ViT-B models, the probability of predicting an image of a Black man and a Latino man as *criminal* decreases by *20%* and *47%*, respectively, when the dataset is scaled from 400M to 2B samples. We ground the model audit results in a qualitative and historical analysis, reflect on our findings and their implications for dataset curation practice, and close with a summary of mitigation mechanisms and ways forward. All the meta-datasets curated in this endeavor and the code used are shared at: https://github.com/SepehrDehdashtian/the-dark-side-of-dataset-scaling.
Content warning: This article contains racially dehumanising and offensive descriptions.

CCS Concepts: • **General and reference** → **Evaluation**.

Additional Key Words and Phrases: Audits, Evaluations, Scale, Visio-Linguistic Models, Multimodal Datasets, CLIP, Racism, Bias

ACM Reference Format:
Abeba Birhane, Sepehr Dehdashtian, Vinay Uday Prabhu, and Vishnu Boddeti. 2024. The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3630106.3658968

1 INTRODUCTION

Over the past few years, transformer-based models have come to revolutionize the field of deep learning. These models leverage a mechanism known as attention or self-attention [91], allowing them to weigh the influence of different input elements and capture long-range dependencies. Transformer models have been instrumental in tasks such as text

*Equal contribution
†Equal contribution

Manuscript submitted to ACM

and speech translation [42, 49], genomic research [22, 41, 50, 98], anomaly detection in time series [21, 89], and fraud detection [20, 93, 95]. Furthermore, the principles of transformer models have been extended to other domains, such as computer vision, with the introduction of Vision Transformers (ViT). The parallel processing capability of these models has significantly improved the efficiency of both training and inference [43]. The landscape of transformer models is dynamic, with continuous advancements contributing to the state-of-the-art where notable models include RoBERTa [52], XLNet [96], and GPT-4 [3].

However, the ubiquitous development and adoption of artificially intelligent (AI) technologies – including transformer models – into numerous societal domains has also ushered in a multitude of actual and potential risks and harms. Some of the most notable crises in current AI include functional failures [72]; disparate performance and treatment based on gender, race, and other dimensions [23, 39, 67]; exacerbation of discriminatory, stereotypical and otherwise marginalising values [8, 54, 76]; legal incompatibility, plagiarism, and copyright violations [4, 47]; privacy violations [16, 63, 92]; production and spread of misinformation [19, 48]; massive consumption of energy and resources [24, 55]; opaque and inscrutable datasets, models and practices [62, 69]; power centralization [1, 12]; the normalization of surveillance [14, 40]; and labour exploitation [31, 88]. Continual and comprehensive documentation, checks, critical scrutiny, evaluations, and testing of these systems have become pressing. Subsequently, audits of algorithmic systems – including of datasets – have emerged as one of the most effective mechanisms for diagnosing, documenting, and mitigating numerous AI risks and harms.

Mislabeling and misclassification of people's images, particularly of those from minoritized groups has been one of the major problems in computer vision systems. In 2015, Google's Photo app classified photos of Jacky Alciné and his friend (both of whom are Black) as "gorillas" [45, 80]. Eight years later in 2023, the problem remains unsolved [30]. There has since been growing awareness of racial and gender bias in computer vision and multimodal models, which has ushered in significant improvements in the accuracy of image classification. Yet, misclassification of images, particularly of minoritized races and genders remains one of the most persistent problems. Over the past years, a robust body of work has demonstrated the tendency of machine learning systems, tools, and applications to encode and exacerbate societal stereotypes and historical biases [9, 13, 14, 57, 60, 66].

In 2018, Buolamwini and Gebru [15] evaluated three commercial classification algorithms along the dimensions of gender and skin tone. In what has now become one of the canonical studies that paved the way for algorithmic auditing as a field of study, they found statistically significant disparities in performance showing up to 34.7% error rate for dark-skinned females, compared to an error rate of under 0.8% for lighter-skinned males. Numerous subsequent studies have demonstrated that computer vision models often fail to detect and/or accurately classify images of genders, races, and demographics outside the status quo. For example, computer vision models failed to detect images from non-Western demographic groups [25], image detection – in the context of pedestrian detection – showed lower rate of pedestrian detection on darker-skin tones while exhibiting a higher rate of precision for lighter-skin tones [94], and unsupervised image representation models encode implicit racial, gender, and intersectional bias [85].

Racial and gender bias in language models [2, 6, 17] and vision models [15, 34, 46, 76] is a well documented phenomena. Bias in multimodal models (models with any combination of text, image, audio, and video modalities as inputs/outputs), on the other hand, are sparsely studied. Still, a rapidly growing body of work indicates that multimodal models also encode and exacerbate societal and historical stereotypes and biases, in some cases at a much worse scale than that of models with a single modality. For example Mannering [58], found gender bias in text-to-image models using object detection, Luccioni et al. [54] found that outputs from diffusion models encode societal biases, Hendricks et al. [35] found that race and gender bias are exacerbated in downstream outputs in image captioning, Mandal et al. [57]

found that DALL-E 2 and Stable Diffusion reflect gender bias, and Bianchi et al. [11] found that image generation models encode and exacerbate societal and historical stereotypes along with complex biases in generated images. More specifically, Hundt et al. [37] audited the multimodal model CLIPort [79], which runs on a robot and is backed by CLIP, for performance on terms such as "criminal", "homemaker", and "doctor" on the eight variants of race and gender in the CFD and found significant bias and negative stereotypes. Similarly, Liu et al. [53] trained text to image generators, diffusion models in particular, on a small dataset of images and descriptions collected by residents of different countries, then generated images on multiple approaches. Humans rated the models on offensiveness, stereotypes, image to description match, and cultural representativeness. They found that CLIP cosine similarity scores get worse as the models improve on each of the aforementioned human rated metrics.

The drive towards Artificial General Intelligence (AGI) being pursued by various actors entails a triadic interplay between the verticals of computing power, model architecture and data. Thus far, constructing a defensible *moat* with computing power and model architecture advancements alone has proven elusive. Most of the key players in Big Tech and in the startup ecosystem alike source their computing resources from the same set of two or three key players that supply the silicon and orchestrate the cloud computing software. On the model architecture front, the marquee architectures have rarely been disrupted, with almost all the major players training on some variant of the U-Net-like, ResNet-like or transformer-like architectures. It is in the data vertical that players try to establish their idiosyncratic moats, often scraping and munging data from the unsuspecting corners of the internet that are not guarded by authorization checks [59], pirated book-dumps [78] and even fandom wikis, casino wholesaling websites, and even random internet comments [65] which is the considered their secret sauce and guarded fiercely. In the specific context of visio-linguistic models (VLMs), one canonical source for datasets has been the Common Crawl (CC) repository, a collection of periodically web crawled data maintained by a San Francisco based 501(c)(3) non–profit organization. This primary source has been distilled to generated datasets such as LAION-400M and LAION-5B. The recipe entails using a pre-trained black-box VLM (typically a variant of the CLIP [71] model published by OpenAI) purportedly to filter images whose alt-text description closely resemble their semantic content. For example, the *plain-vanilla* CLIP model and its ViT B/32 CLIP variant were used to distill the CC dataset into LAION-400M with 0.3 and 0.28 cosine similarity thresholds, respectively. In this study, we investigate the questions of what happens to the quality of such distilled datasets whose scale is increased by expanding coverage of the CC data-dump and manipulating the ad hoc hand-set cosine similarity thresholds as well as the subsequent downstream effects of the models' predictions trained on these datasets.

We audit pre-trained Contrastive Language-Image Pretraining (CLIP) [71] models for gender and ethnicity bias. Specifically, we evaluate 14 Vision Transformer-based VLMs from OpenCLIP [38] on a classification task using the Chicago Face Dataset (CFD) [56] as a probe dataset. CLIP model architectures comprise two encoders: a vision transformer for image inputs and a transformer-based language model for text inputs. These encoders project the input data into a shared embedding space, enabling the model to compare and relate visual and textual information. One of the key features of CLIP is its ability to perform zero-shot learning. This means that the model can "generalize" to tasks it has not been explicitly trained on, such as prediction of a novel class. This is achieved by leveraging the flexibility of natural language as a prediction space. However, CLIP models suffer numerous problems. The CLIP paper [71] itself (in Section 7.1) outlined that images belonging to the "Black" racial designation had an approximately 14% chance of being miscategorized as `['animal', 'gorilla', 'chimpanzee', 'orangutan', 'thief', 'criminal' and 'suspicious person']` in their *FairFace* dataset experiment. We replicate the Zero-Shot CLIP experiment using the CFD (see Appendix 8 for a sample of hand blurred images) as a probe dataset and study the effect of scaling the

pre-training dataset. Our analysis finds that the effect of scaling the datasets is dependent on the scale of the trained model. Larger models exhibit a greater proclivity towards predicting specific racial groups like Black and Latin faces as criminals as the scale of the pre-training datasets increases. On the other hand, smaller models exhibit a lower proclivity towards predicting specific racial groups like Black and Latin faces as criminals as pre-training datasets' scale increases.

2 AUDIT METHODOLOGY

To quantitatively evaluate the downstream consequences of scaling the pre-training datasets, we first explored model variants where the architecture was held constant and two or more model checkpoints were provided: some trained with LAION-400M and some trained with LAION-2B-en. The emergence of OpenCLIP [38] facilitated this endeavor as (to the best of our knowledge) it remains the only resource that hosts VLM variants with *fixed model architecture* but varying dataset sizes (LAION-400M and LAION-2B-en respectively). Among the 120 models present in OpenCLIP (at the time of our experimentation), we selected the following 14 CLIP-model pairs presented in Table 1 that met our criteria. The architecture of the models that we evaluated, their pre-training dataset, the number of parameters of each architecture, and the number of **FL**oating point **OP**eration**s** (FLOPs) are listed in Table 1. The OpenCLIP project currently uses an idiosyncratic naming convention for the model checkpoints presented in the second column of Table 1. To evaluate the effect of scaling the pre-training dataset on these model variants, we used the CFD [56], as a probe. We replicated the *Zero-Shot CLIP experiment* that appeared in *Section 7.1-Bias* of the original CLIP paper [71] by OpenAI, the details of which are in Subsection 2.1.

Table 1. Architecture-Dataset variants in the OpenCLIP ecosystem we evaluated in this study.

Architecture	Dataset/Checkpoint	Number of Parameters (M)	FLOPs (B)

ViT-B-16	`laion400m_e31`
	`laion400m_e32`
	`laion2b_s34b_b88k`
		149.62	41.09

ViT-B-16-plus-240	`laion400m_e31` `laion400m_e32`	208.38	64.03

ViT-B-32	`laion400m_e31`
	`laion400m_e32`
	`laion2b_e16`
	`laion2b_s34b_b79k`
		151.28	14.78

ViT-B-32-quickgelu	`laion400m_e31` `laion400m_e32`	151.28	14.78

ViT-L-14	`laion400m_e31`
	`laion400m_e32`
	`laion2b_s32b_b82k`
		427.62	175.33

The CFD is a highly controlled dataset that consists of high resolution images of 597 unique individuals along with their *self-classified* race and gender labels belonging to Asian (109), Black (197), Latin (108), and White (183) categories. A (blurred) sample of images and additional information on the CFD dataset is shown in Appendix 8. The dataset has been meticulously standardized to control for potentially confounding causal covariates such as facial expressions, resolution, image-pixel saturation, lighting conditions, clothing, and eye gaze. The 597 images have each individual

wearing the same heather grey t-shirt. While much smaller in volume, unlike the the majority of openly available datasets, the individuals in CFD had their consent obtained, were financially compensated and were given the option to self-classify from a set of pre-defined options: from Black, White, Asian or Latin and Female or Male.

2.1 Experiment Design

The sub-phases involved in the bias analysis experiments were as follows:

1: Image pre-processing: All the 597 images with neutral expressions extracted from CFD were pre-processed using the respective OpenCLIP model's built-in `preprocess` function that entails resizing (to size 224×224), center-cropping and pixel intensity normalization sub-processes. The output of this sub-phase is a CFD-image-tensor, $\mathbf{I}_{cfd} \in \mathbb{R}^{597\times224\times224\times3}$.

2: Class-generation and tokenization: We first created an 8-class vector with the following classes [`'human being'`, `'animal'`, `'gorilla'`, `'chimpanzee'`, `'orangutan'`, `'thief'`, `'criminal'`, and `'suspicious person'`]. Except for the `'human being'` class, which was added by us, the remaining seven classes were verbatim extracted from *Section 7.1 Bias* of the OpenAI CLIP paper [71]. Next, we created the class-sentences using the "`A photo of a/an <class>`" template[1]. The output of this sub-phase is a sparse zero-padded text-token matrix, $\mathbf{T}_{8-class} \in \mathcal{I}^{8\times77}$ where $\mathcal{I} = [0, ..., N_{tokens} - 1]$ is the tokenizer-index set.

3: Forward pass, feature extraction and normalization: The pre-processed image tensors and the text-tokens generated in the previous sub-phase were now fed into the encoder of the chosen OpenCLIP model, and the output image and text features were then normalized. For all the evaluated models, the features are 768-dimensional thus rendering the text and image feature matrices over the 597 neutral-expression CFD images to be 597×768. That is, the image-feature matrix is $\mathbf{F}_I = \left[\mathbf{f}_0^I, ..., \mathbf{f}_{596}^I\right]^T \in \mathbb{R}^{597\times768}$ and the text-feature matrix would is $\mathbf{F}_\tau = \left[\mathbf{f}_0^t, ..., \mathbf{f}_7^t\right]^T \in \mathbb{R}^{768\times8}$.

To highlight how self-similar the 8×8 textual features are, we present the annotated heatmap of the $\mathbf{F}_\tau \times \mathbf{F}_\tau^T$ matrix (see Figure 7(a) in the Appendix). Similarly, we also present the heatmap of the 597×597 sized $\mathbf{F}_I \times \mathbf{F}_I^T$ matrix (Appendix Figure 7(b)). Given the fact that the 597 images were sorted and grouped by Race-Gender categories, the block-like structures visible (in the Appendix Figure 7(b)) indicate the fact that the model's output features are certainly influenced by these categorical indicators.

4: Computing softmax-matrices: Firstly, we obtain the image-text cosine similarity matrix, $\mathbf{C} \in \mathbb{R}^{597\times8}$ as:

$$\mathbf{C} = \mathbf{F}_I \mathbf{F}_\tau^T. \tag{1}$$

Then, the softmax-matrix $\mathbf{S} \in \mathcal{P}^{597\times8}$ ($\mathcal{P} = \{p|0 < p < 1\}$) is computed as:

$$\mathbf{S} = \text{softmax}\left(100 \times \mathbf{C}\right). \tag{2}$$

Here $softmax()$ is the softmax function applied row-wise. That is, if $\mathbf{C}_{i,j}$ is the i^{th} row j^{th} column element in the cosine-matrix, then the corresponding $(i, j)^{th}$ element in the softmax-matrix, $\mathbf{S}_{i,j}$ would be $\mathbf{S_{i,j}} = \frac{\exp\left(100\times\mathbf{C}_{i,j}\right)}{\sum_{k=0}^{7}\exp\left(100\times\mathbf{C}_{i,k}\right)}$.

3 RESULTS

In Figure 1 we first present the three 597 × 8 sized output softmax-matrices obtained from ViT-B-16, ViT-B-32, and ViT-L-14 with two different pre-training datasets: LAION-400M and LAION-2B-en. The $(i, j)^{th}$ element of each of these matrices captures the softmax score value of the j^{th} class ($j \in \{0, ..., 7\}$) obtained from that specific OpenCLIP

[1]As advocated in the `Interacting with CLIP` Jupyter notebook shared at https://github.com/mlfoundations/open_clip/blob/main/docs/Interacting_with_open_clip.ipynb in the context of *Zero-Shot Image Classification for CIFAR-100 dataset.* These eight sentences were then *tokenized* using OpenCLIP's tokenizer module (the `Vocab size` is 49408 for all the models considered in this paper), thus yielding an 8×77 sized token-matrix.

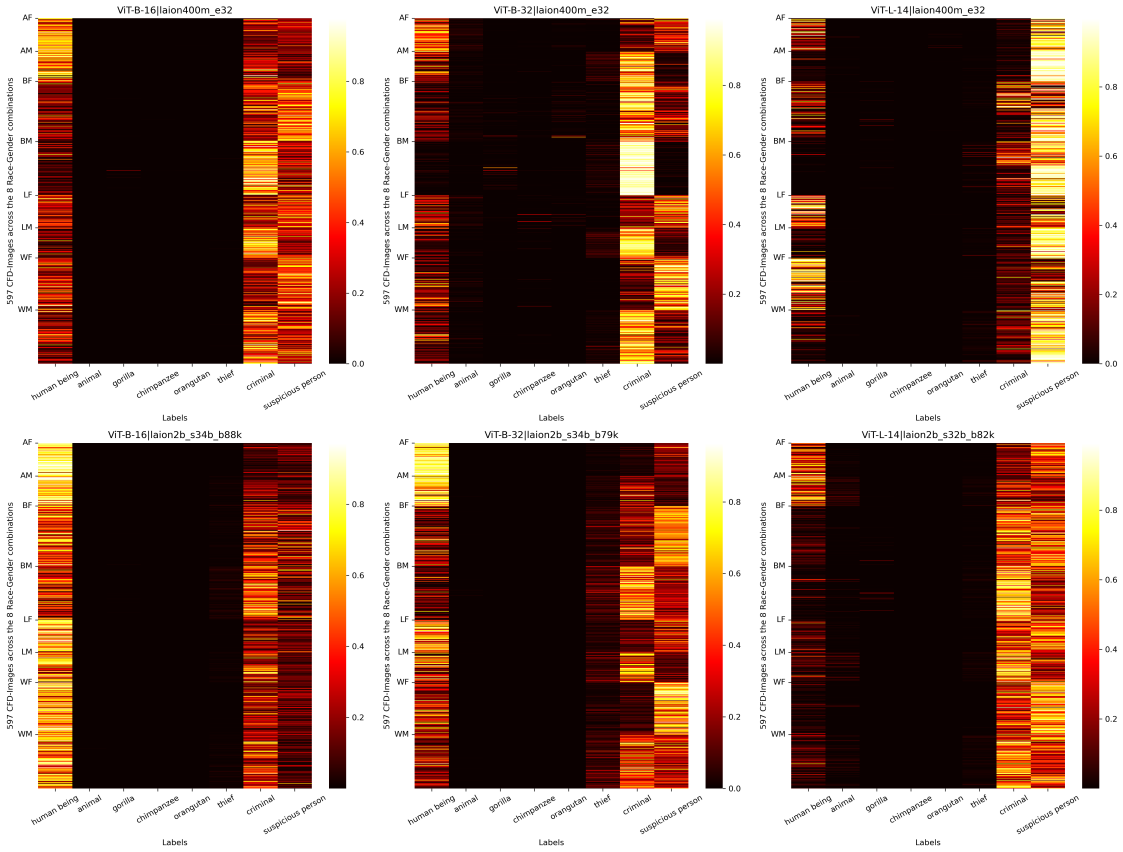


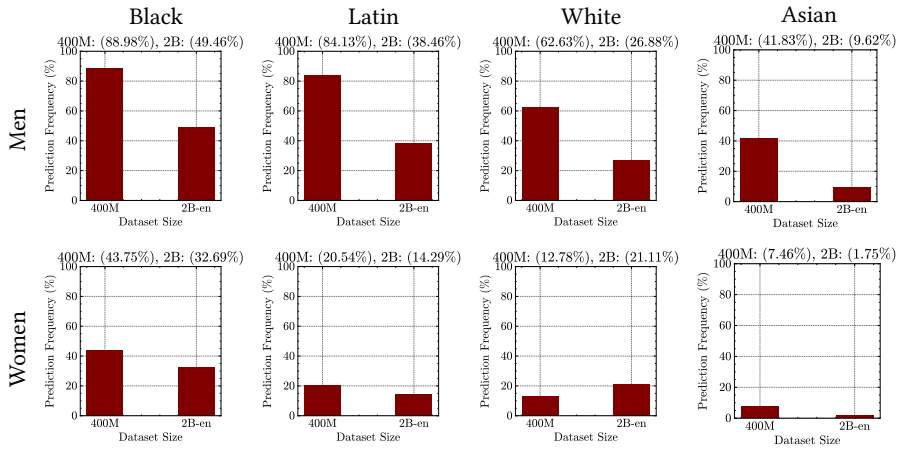
Fig. 1. Heatmaps of 597 × 8 softmax-matrices for three models (columns) and two pre-training datasets (rows).

model in response to the i^{th} input CFD image ($i \in \{0, ..., 596\}$). The 597 rows (representing the 597 CFD images) are grouped by their self-classified Race-Gender groupings. That is, the first 57 rows represent images from the Asian-Female (abbreviated as AF), and the next 52 rows map to the Asian Male (AM) group, and so on. The titles of these subplots are formatted as strings with two fields separated by the '|' character: <model architecture> | <pre-training dataset/checkpoint>. From the figure, we make the following observations.

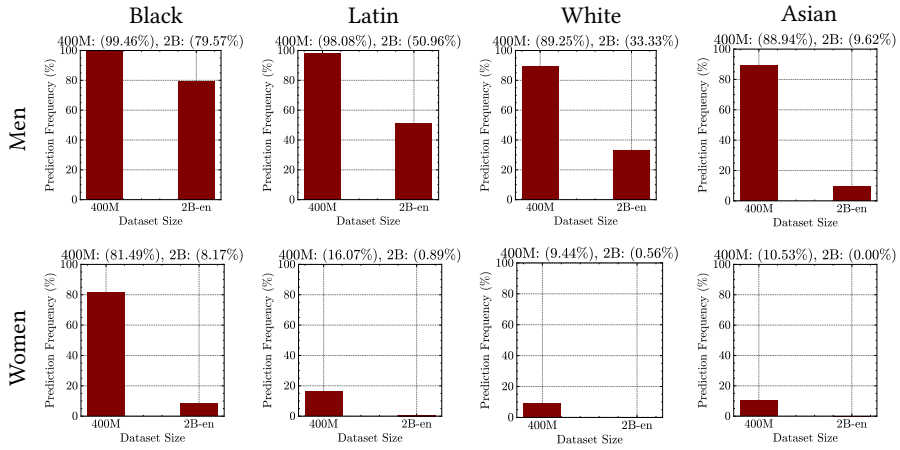
1. Non-human offensive labels: For all the models we evaluated, regardless of the training data size (LAION-400M or LAION-2B), the softmax scores for non-human offensive labels i.e., *animal*, *gorilla*, *chimpanzee*, and *orangutan* are close to zero across different architectures and datasets. In other words, none of the models accurately predicted images of people from CFD with the 'human being' class. Instead the models predicted these images of humans with the other non-human offensive classes: *animal*, *gorilla*, *chimpanzee*, and *orangutan*.

2. Human being label: We found that, as training data size increased from 400M to 2B samples, model accuracy at predicting human faces from CFD accurately as *human being* increased for all races and genders and by **6.4%** for Black women and **58%** for Asian men.

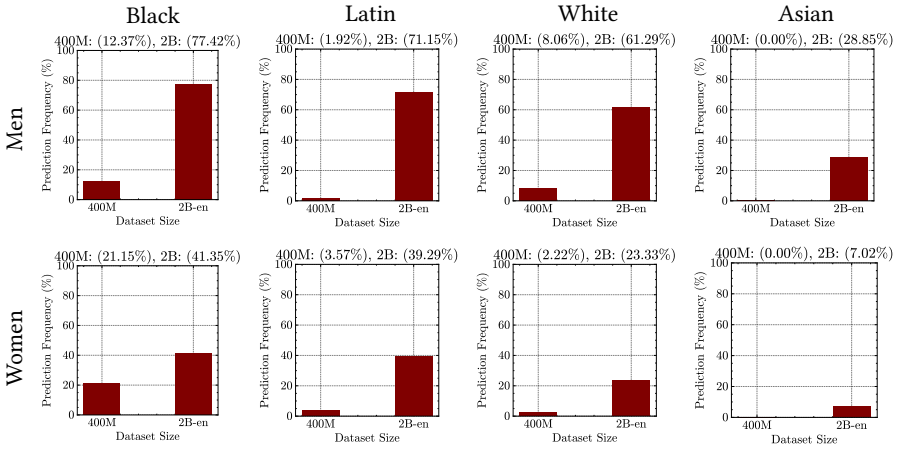
3. Human offensive labels: Among the three human offensive labels (*thief*, *criminal*, and *suspicious person*), we



(a) ViT-B-16



(b) ViT-B-32



(c) ViT-L-14

Fig. 2. Effect of scaling the dataset from 400M to 2B on the frequency of an image from CFD getting predicted as 'criminal' for each race-gender group and three different architectures: ViT-B-16 (a), ViT-B-32 (b), and ViT-L-14 (c). We observe that the larger ViT-L model's predilection for labeling faces as 'criminals' increases significantly for black and Latino men when the pre-training dataset is scaled from 400M to 2B (see Section 3, specifically 3.1 to 3.3).

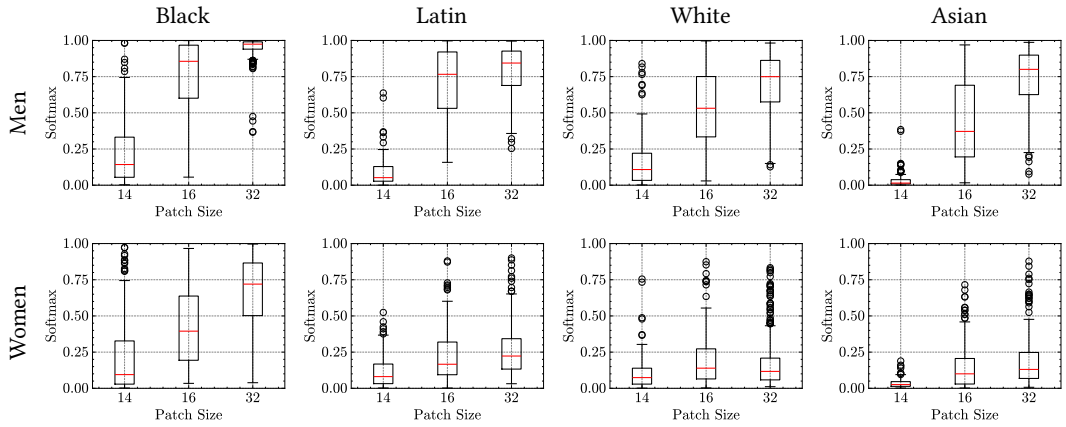
found *criminal* and *suspicious person* predictions occur the most. The scores for the *suspicious person* class increase as the number of model parameters is scaled from ViT-B-16 (149.62 M) to ViT-L-14 (427.62 M). We also observe that the scores for the *human being* class increase for ViT-B-16 and ViT-B-32 when the training dataset is scaled to 2B samples, while they decrease for ViT-L-14 when the same scaling is applied. For women (of all four races) the probability of an image being predicted as P_{thief} is zero and for men (of all 4 races), this probability is almost zero, with an average prediction of 0.01. Generally, we found that men are more likely to be classified as *criminal* than women (see Figure 2). Furthermore, for the *criminal* class, we found:

3.1 Scaling increases *criminal* prediction: As shown in Figure 2, for the smaller models, i.e. (a) ViT-B-16 and (b) ViT-B-32, scaling the number of samples from 400M to 2B in the pre-training datasets decreased the *criminal* prediction by **33%** on average. However, for the large models (c) (ViT-L-14), increasing the amount of pre-training data from 400M to 2B samples, increased the probability of an image of a person from CFD being predicted as *criminal* by **37.5%** on average. For all four races of the CFD human images, for large models (ViT-L-14 in Figure 2 (c)), the *criminal* label was allocated to **Latin (71%)** and **Black (77%)** faces at a higher rate compared to the White (61%) and Asian (28%) groups. For all models we evaluated, **Black** and **Latin** groups received higher probabilities of being predicted as *criminal* compared to the other two groups: White and Asian. For example, for the ViT-B-16 model pre-trained on the LAION-400M dataset, *criminal* is **66%** and **52%** for **Black** and **Latin** faces respectively, while it is **37%** and **24%** for **White** and **Asian** faces, respectively. This was the case regardless of dataset scale, meaning that the probability of the label *criminal* being allocated to Black and Latin racial groups was **highest** for models trained on both 400M and 2B samples.

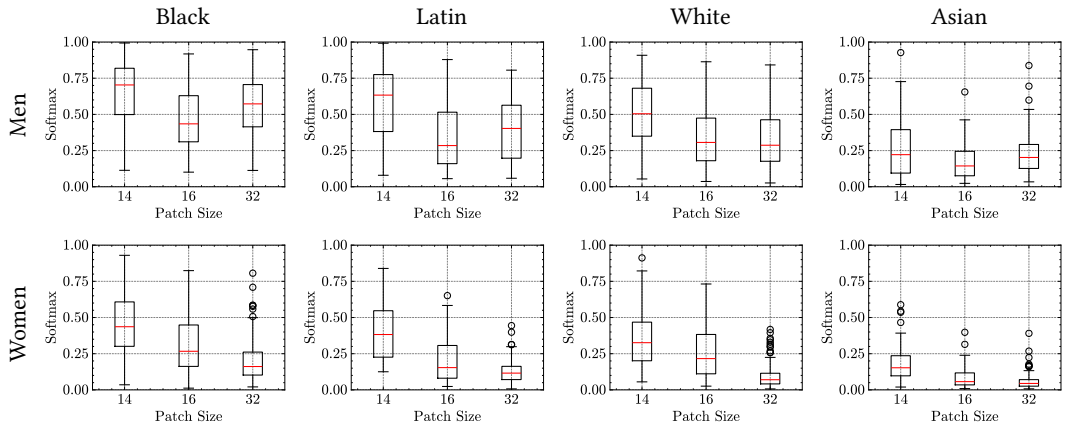
3.2 Patch size increased *criminal* prediction: We found that, within the same racial group, men are generally labeled as *criminal* at a higher rate (**45% higher** on average) than women. As shown in Figure 3 (a), for models trained on the smaller dataset, increased model patch size increases the probability of an image of a face being predicted as *criminal* for all racial and gender groups **except for White women**. However, for models trained on the larger 2B sample, the probability of *criminal* decreased for women, as the patch size of the model increased.

3.3 Frequency of *criminal* prediction versus patch size: As shown in Figure 4, the frequency of an image from CFD being labeled as *criminal* increased for all groups except White and Latina women for all models trained on 400M samples (Figure 4 (a)). In other words, the models showed bias against Black and Asian women, and all men from the four racial groups (Black, Asian, Latin and White). In other words, we see a close to 100% prediction frequency for the model with patch size 32 for Black men, which means that the *model predicts all images of Black men* from the CFD as *criminal*. Conversely, the frequency of *criminal* prediction decreased for women (of all racial groups) for models trained on the 2B sample (Figure 4 (b)), as the patch size of the model increased. This means that fewer women were classified as *criminal* when the patch size of the model increased.

4. Summary of the effect of dataset scaling on models' predictions: We summarize the effect on the models' predictions as we scale the pre-training dataset, shown in Figure 5. For all racial groups, the probability of an image of a human from the CFD being predicted as *human being* was higher in smaller models (ViT-B-16 and ViT-B-32). On the other hand, the probability of an image of a human from the CFD being predicted as *human being* decreased for Latino women, Black women, White women, and White men in the larger models (ViT-L-14). For the larger ViT-L-14 models, the heatmaps demonstrate the disparate increase in the probability of labeling human faces as *criminal* across different racial groups. Similarly, for the smaller ViT-B models, the heatmaps also demonstrate the disparate decrease in the probability of labeling faces as *criminal* across different racial groups. We also found that, in general, Latino/Latina



(a) LAION-400M



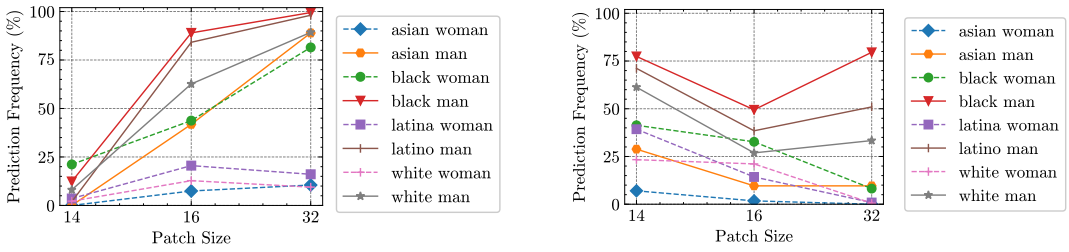
(b) LAION-2B

Fig. 3. Plots showing the effect of **patch size** on the distribution of "criminal" predictions for (a) LAION-400M and (b) LAION-2B as the pre-training datasets.

individuals were misclassified with high confidence as one of the 'Asian' classes and this *misclassification increased with dataset scaling* (see Appendix 9 for details). All of our results, as well as the meta-dataset created as a result of our audit, are available on our repository.

4 QUALITATIVE ANALYSIS: DEHUMANIZATION AND CRIMINALIZATION OF BLACK BODIES

A rich body of work within Science and Technology Studies (STS), Black studies, and critical data and algorithm studies has emphasized the tendency of ML research, tools, and applications to encode and exacerbate societal stereotypes and historical injustice [9, 14, 60, 66]. As presented in Section 3, our findings extend this rich body of work by demonstrating that not only do large VLMs encode such historical trend that dehumanizes Black bodies but also, as these models and datasets increase in scale, such dehumanization is further exacerbated.



(a) LAION-400M (b) LAION-2B

Fig. 4. Frequency of 'criminal' prediction versus patch size for the LAION-400M (a) and for the LAION-2B (b) datasets.

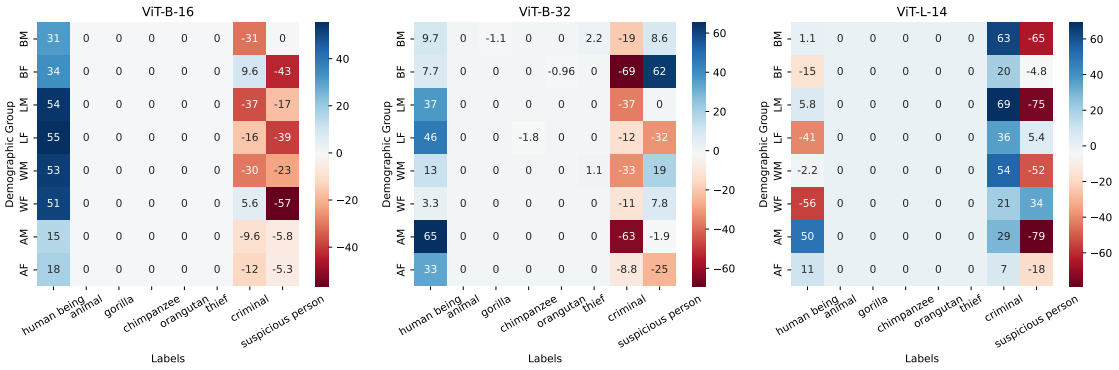
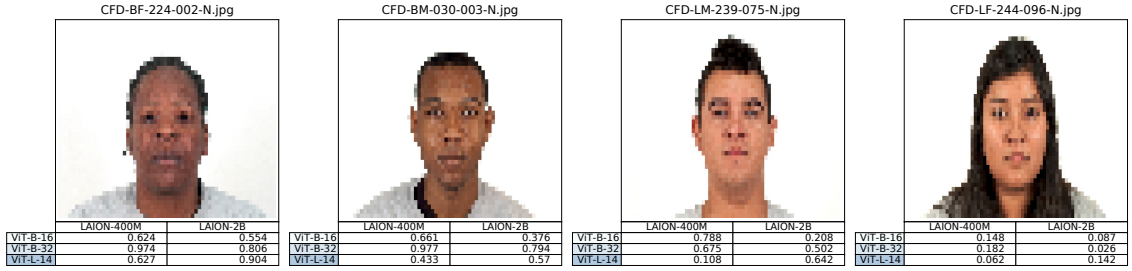


Fig. 5. The effect of dataset scaling on the predictions of the models. The numbers show the change in probabilities when the pre-trained dataset is scaled from 400M to 2B for ViT-B-16, ViT-B-32, and ViT-l-14. Positive values mean an increase in the probability when the number of pre-training samples is increased. All values are in percentage.

During the institution of slavery, Black people, particularly Black men, were depicted as "brute" and "docile" creating and reinforcing the idea that the most fitting position for them is slavery [81]. Scientific racism enabled racial classification in 18th and 19th century that justified slavery, legal segregation and discrimination [74]. Arbitrary racial classifications emerged that portrayed Caucasians (white Europeans) as the epitome of humanity at the top of the hierarchy, while these arbitrary systems placed African and African-Americans at the bottom of the racial hierarchy. Although practices such as chattel slavery and legal segregation were eventually abolished, systemic racism – which is rooted in these vacuous underlying conceptions – remain embedded in institutions, social structures and processes and continue to be pervasive and ingrained in societal systems [26, 27].

In the U.S., the rise of the for-profit prison industrial complex is a primary example that embodies systemic racism maintaining the cycle of systemic racism through unjust incarceration of Black people and unrealistic depictions of Black men as "thug", "criminal", and "suspicious" [27, 81]. Many prison companies mandate that municipalities have a 90-95% prison occupancy rate increasing targeted association of Black people and crime [5, 7]. Such stereotypes and racist ideologies have fueled racial violence, criminalization, and mass incarnation of Black men, especially in the US. Black bodies, according to [10], are often perceived as a threat and typecast as "gangster," "rapist", and "ghetto". The "Black-as-criminal" stereotype, subsequently, can result in non-violent acts of Black men being perceived as violent and aggressive while violent acts performed by white men are perceived as unintentional or get attributed to external factors and uncontrollable causes such as mental health [18].



	LAION-400M	LAION-2B
ViT-B-16	0.624	0.554
ViT-B-32	0.974	0.806
ViT-L-14	0.627	0.904

	LAION-400M	LAION-2B
ViT-B-16	0.661	0.376
ViT-B-32	0.977	0.794
ViT-L-14	0.433	0.57

	LAION-400M	LAION-2B
ViT-B-16	0.788	0.208
ViT-B-32	0.675	0.502
ViT-L-14	0.108	0.642

	LAION-400M	LAION-2B
ViT-B-16	0.148	0.087
ViT-B-32	0.182	0.026
ViT-L-14	0.062	0.142

Fig. 6. Example images of Black individuals from CFD and the tendency of the OpenCLIP models studied to associate them with the "A photo of a criminal" sentence. The title(s) indicates the file name, and the table under each image shows P_{criminal} for three different architectures and two pre-training datasets. The images are hand-blurred to preserve the privacy of the data subjects.

Contrary to these racial stereotypes, a robust body of work, especially in the context of the U.S., documents that Black men commit crimes at a far lower rate than whites, while Black people constitute the group that are victims of violent crimes at far higher rates than whites [29, 32]. Innocent Black people, according to Gross et al. [32], are seven-and-a-half times more likely to be convicted of murder than whites, and convicted Black people are 80% more likely to be innocent than other convicted murderers. In 2002, Black people were six times more likely to be murdered than whites, and this number was much higher during previous decades, where 47% of victims were African Americans during the 1976-2002 period [73]. Conversely, a 2018 United Nations Report on racial disparities [70] shows that "African-American adults are 5.9 times as likely to be incarcerated than whites" and more likely than whites to be arrested; once arrested, more likely to be convicted; and once convicted, more likely to be incarnated than whites. Studies on drug use across demographers in the US reveal a similar trend. Although African Americans and whites use illegal drugs at similar rates, Black people are 19 times more likely to be convicted of drug crimes than whites [32, 73].

Erroneous stereotypes have historically (and currently) served to explicitly, implicitly, and systematically place Black people, particularly Black men, as "suspects", "criminals", or "persons of interest" [81]. Along with past work that has highlighted the risk of models amplifying racial stereotypes [11, 77, 90], our findings confirm this trend. As outlined in Section 3, we observe that current SoTa models encode and exacerbate racial stereotypes. Furthermore, as outlined in 4, the likelihood of a Black man being classified as "criminal" *increases as training datasets get bigger*. As illustrated in Figure 4 (a), the prediction frequency for Black men as "criminal" for model patch size 32 was close to 100%, where all the CFD samples for Black men were predicted as "criminal". (See Figure 6, for a randomly selected example images of four Black individuals from the CFD dataset each showing criminality prediction for three different model architectures). As reported in the tables, the association of Black and Latin faces with `'A photo of a criminal'` increases for the large model (ViT-L-14) while it decreases for smaller models such as ViT-B-16 and ViT-B-32 by scaling the dataset from 400M to 2B. To summarise, the findings from our audits align with the rich body of work within Black studies, critical data studies, and critical race scholarship that have examined, underscored and challenged systemic racism. To that end, as training datasets get larger, they further exacerbate deeply ingrained negative societal and historical stereotypes and racial dehumanisation, particularly against Black people.

5 DISCUSSION AND RECOMMENDATIONS

In this paper, we have systematically examined the impact of dataset scaling and model architecture by evaluating 14 Vision Transformer-based VLMs pre-trained on two datasets: LAION 400M and LAION-2B-en. Our results show

that with a larger ViT-L model the predilection for labeling faces as 'criminals' increased statistically significantly for black and Latino men when the pre-training dataset was scaled from 400M to 2B samples. Datasets are fundamental backbones to models and party determine whether a model is equitable, just, robust, and well-performing. Subsequently, a transparent and just sourcing and rigorous evaluation, audit, curation, and management of datasets is critical for advancing the field towards a healthy and sustainable direction.

We strongly highlight the need to avoid interpreting the empirical results from a reductionist lens where the emphasis is erroneously laid on the specific details of the metrics introduced (such as P_{human} and $P_{bf/bm \rightarrow \text{criminal}}$) and model variants used. It is evident that the *brittleness* of these models certainly allows for trivially flipping the results to favor another narrative by smartly changing either the choice of labels, the choice of default-class (replacing `human being` with a synonym for example), the class sentence construction template or the model architecture variants (Using Vit-B/16/32 for example). Furthermore, parameters beyond our control (such as batch size used during pre-training, choice of tokenizer, and number of training epochs used) also likely played an important role in influencing these results. Instead, what we are conveying through these results is simply this: despite making the most templated design choices on all the aspects of the pipeline, and despite verbatim replication of the empirical orchestration straight from the example code notebooks in the official Github repositories, and despite using an extremely controlled *easy* probe dataset and class-design, *it was verifiably hard to avoid the glaring negative impact on the biases measured that could be directly attributed to dataset scaling.*

Below we present a set of observations that we hope the ML community, dataset curators, as well as other stakeholders might find helpful towards advancing not only data curation but also the field as a whole in a manner that is transparent, rigorous, responsible, and accountable.

Avoid ad-hoc decision-making for dataset curation hyperparameters. In the *CLIP inference at the post-processing stage* section of the LAION-5B dataset announcement, we encounter the fact that the dataset curators estimated the cosine similarity between an image and its alt-text description using the `ViT B/32` CLIP model and discarded all images with cosine similarity score of less than the manually set threshold of 0.28. This is a marked departure from the procedure published during the LAION-400M release where the curators stated that *"We use OpenAI's CLIP model (the 'ViT-B-32' version) to compute the image and alt text embeddings. Then we calculate the cosine similarity of both embedding vectors and drop all samples with a similarity below 0.3. We chose this threshold after trying different values and using human evaluations of how well the texts fit the images. Lower values like 0.28 or 0.29 also seemed okay in many cases, but after further inspections, we decided to choose the conservative value of 0.3"*. The reasoning behind this decision is not clear. However, such a decision might have been taken to boost the dataset size past the 5 billion mark, a pre-mandated milestone perhaps. Given these decisions have a significant consequence for dataset quality (subsequently model performance and potentially concrete lives through deployment), we recommend such processes be rigorously justified, well documented, and made transparent a la scientific practices.

Beware of CFD physiognomy. Scholars have warned about the the rebirth of phrenology and physiognomy via the by-lanes of Computer Vision [82, 83]. Similarly, some of our preliminary investigations that emerged when we dug into the *whyness* of criminality-association of some CFD faces by the models under consideration show high correlations with metrics such as Facial Width-to-Height Ratio (fWHR) and Cheekbone Prominence that are recorded as metadata in the CFD dataset. Well-informed and in-depth awareness of this pernicious development as well as mitigation mechanisms against phrenology is crucial. To this end, we encourage future research in line with those such as Hundt et al. [37] to build upon this finding through a statistical experiment mapping the objective face-measurement-metrics found in 'Study-1 and Table-1' of [56] to the model outputs to further investigate the rebirth of phrenology and develop

regulatory and mitigation mechanisms.

Dataset sub-sampling: Only for ethics checks? There is an emergent trend within the broad culture of internal audits (self-audits within big corporations and institutes) focusing *subsample-only-for-ethics-auditing* when it comes to handling large datasets, despite the abundant resources at their disposal. As far as training a monetizable model is concerned, scale is deemed a virtue and not a hindrance as exemplified by frequent aggressive crawl-scrape-scoop strategies. On the contrary, scale is deemed as an impediment when it comes to auditing, evaluating, and stress-testing datasets and models for critical concerns including checking for quality of data, encoded racial stereotypes, and bias. For example, we observed that the CLIP model was trained on a black-box Web-Image-Text (WIT) dataset spanning 400 million image text pairs. However, when it came to measuring the racial biases baked into the model, sub-sampling was resorted to a comparatively *small* dataset, the FairFace dataset [44], which only contains 0.027% (108,501 images) of the training dataset. Moreover, the bias-measurement exercise is minimal, limited only to running inference (read forward pass) through the model that is an order of magnitude less computationally intensive compared to training the model (backward pass). As stated in *Section 7.1: Bias* in the CLIP paper [71], only 10000 images (0.0025% of the training dataset size) were used from this FairFace dataset for the bias-check-inference task (that we have used in our experiments (see Section 2)). We recommend audit, evaluation, and general critical and ethics work is carried out to the highest possible standards and scientific rigour. Otherwise, it risks ethics and audit washing.

Legal and policy implications. The LAION datasets we audited serves as a critical backbone for numerous popular, influential and impactful models including Google's Imagen and Stable Diffusion variants. Increased integration of these models into numerous societal domains and practices means that these models are not purely intellectual exercises but result in direct or indirect impact on actual people, particularly marginalised groups. Yet, neither datasets nor information around dataset creation, curation, documentation, filtering and detoxifying mechanisms used are made available for most of these popular and influential models. Restricting access as well as active obfuscation of information around these datasets present a major obstacle to carrying out independent audits and developing appropriate regulatory guidelines and guardrails. Open access is a prerequisite to independent audits, particularly those aiming to examine, diagnose and challenge societal and historical injustices that datasets and AI models encode and amplify. We hope this work serves for legal and policy experts and authorities as a reminder for the urgent need to both encourage and develop legally enforceable mechanisms to allow access for independent audit and evaluation of training datasets. Our work also illustrates the importance of dataset curation, filtering and management. We highly recommend such practices become part and parcel of model development.

6 FUTURE WORK AND CONCLUSION

We have carried out an extensive audit investigating the impact of dataset scale and model architecture on VLMs trained on the LAION datasets. In this regard, the emergence of projects such as `openclip` [38] have been instrumental in allowing for easy orchestration of the type of investigations executed and presented here. This section presents a list of natural extensions of our work.

BLIP and other CLIP models: In the associated GitHub repository, we have shared image-class cross-tabulated softmax matrices akin to the ones presented in Figure 1 for the other non-SoTA CLIP models presented in Table 1 for which we could run the *fix-architecture-vary-training-datasets* experiments presented in the Results Section 3. We highly encourage for these experiments to be replicated across the other models including BLIP [51] and the new variants emerging on the scene. We hope that this will help the ML community to intimately understand (and mitigate) the role that model architectures play in encoding harmful biases as the dataset scales.

Choice of prompt template and class design: In this paper, we converted the categorical class labels into sentences using the format *"A photo of <class>"* to maintain consistency with the CLIP [71] paper results. We posit that varying this prompt template with its rephrased variants such as *"This a **picture** of <class>"* would result in variations of the results shown in Section 3. Similarly, we also expect that replacing the word `person` with the self-declared race-gender identifier (such as `asian-man`) will also result in variations to the cosine similarity value output by the models under consideration. Accordingly, future research might unearth the *fairness-optimal* prompt template by both paraphrasings as well as choosing alternative-identifiers for the word `human being`.

Extension across other expressions and other face datasets: In this paper, we have restricted our experimentation to the neutral expression images of the CFD dataset for the sake of brevity. One avenue for future work might be to investigate if holding the individuals' faces constant and varying the facial expressions makes a marked difference in the results. Also, inspired by the CFD project, we have seen the emergence of other similar datasets such as MR2 [86], Bogazici face database [75], the Delaware dataset [61] and the ISIEA dataset [99]. Replicating these experiments using these datasets might yield a more granular view of how these models – supposedly trained on internet sourced data – function and what biases might be baked into them.

The Race-Gender experiment: Some initial results: There also emerges the natural question with regards to the extent to which stereotypes about facial appearances are cross-related with racial identities by these VLMs. Given that the CFD has self-classified race-gender labels, we also performed a small-scale race-gender classification experiment (similar to the FairFace experiment in the CLIP paper [71]), using the subjects' self-classified race-gender labels. That is, we replaced the 8 classes of [`human being`,...,`suspicious person`] in the *human-being experiment* above with the 8 self-classified race-gender category labels [`asian man`,...,`white woman`]. The initial results are discussed in Appendix B and it appears as if faces with visible epicanthic folds (that occur across a broad spectrum of racial identities) are solely associated with the 'Asian' race identifier. This observation merits a deeper analysis especially given the wide availability of meta-data that is associated with the images in CFD that can be a rich source of confounding factors.

6.1 Conclusion

We have carried out a dataset audit of two visio-linguistic multimodal datasets, LAION-400M and LAION 2B-en, and 14 Vision Transformer-based VLMs trained on them. We found evidence of misclassification in the models, particularly towards Black men and Latino men as 'criminal', which exacerbates with training dataset size. We cannot stress the importance of open-source and in audit endeavors such as ours, since any kind of quantitative and qualitative dataset exploration hinges upon access to the artifacts themselves. We are saddened to see an increasing number of ML organizations fail to provide access to their datasets and models since we believe that this is an essential element to scientific advancement and a healthy, equitable, and innovative research community.

Today's state-of-the-art visio-linguistic multimodal models are trained with massive carbon footprints, massive data infrastructure, and massive funding. These models are currently being deployed in the real-world including in recommendation systems, information-retrieval systems, semantic search systems, and image captioning systems, although as we have illustrated in this paper, they can predict photographs of those with Black and Latin racial backgrounds as 'criminal'. Given that such failures can result in dire consequences on real people, often those at the margins of society, we implore the research community as well as those developing and deploying these systems to carry out due diligence with rigorous audits of these models and training datasets and take necessary actions, including refraining from use in high-stake scenarios.

ACKNOWLEDGMENTS

Abeba Birhane is supported by Science Foundation Ireland via the ADAPT Centre of Digital Content Technology funded under the European Regional Development Fund (ERDF) through Grant No #13/RC/2106_P2. Sepehr Dehdashtian and Vishnu Naresh Boddeti are partly supported by the National Science Foundation under Grant No. #2147116. We would like to thank Ellen Rushe, Thomas Laurent, Andrew Hundt and the anonymous FAccT reviewers for their extensive and helpful feedback.

ETHICAL CONSIDERATIONS

In this work, we have audited existing openly available datasets and VLM variants trained on them. We recognize these datasets pose numerous ethical concerns including the sourcing of these content that forms these datasets without consent, awareness or financial compensation for people in these datasets. The audit results are also disturbing and distressing, particularly to Black men who were predicted as "criminal" with close to 100% frequency for the model with patch size 32. We hope by bringing these to light, existing structures and systems of oppression can be challenged. In using the CFD as a probe dataset, we have been fully transparent about its limitations (see below) to help contextualise our findings. We believe our audit work poses no harm or risk to individuals or groups. To further minimise any potential privacy risk to individuals behind the CFD, we have hand-blurred all instances of the CFD throughout the paper.

LIMITATION

The racial and gender construction and limited categories of the CFD adheres to gender and race essentialism. The binary gender (female or male) and the seemingly clean racial (Black, White, Asian, or Latin) categories used in the CFD fail to capture genders and races the real world presents. Far from these binary categories, genders and races are fluid, complex, multivalent, and multidimensional in actuality. Furthermore, as [37] point out, the individuals of CFD were provided with limited pre-defined categories (as opposed to given the agency to self-identify) to select their identities from. Yet, despite this limitation, we believe the dataset presents a useful proxy in the context of our experiments.

Additionally, we also details four sources of confounding factors that scholars investigating these biases need to consider that are beyond the scope of the work published here.

- Shortcomings of the cosine similarity metric during dataset curation process
- CLIP-like models suffering from Concept Association Bias (CAB)
- CLIP-like models exhibiting Bags-Of-Words like behavior
- CLIP-like models being vulnerable to Identity Inference Attack (IDIA)

6.2 Effect of cosine similarity metric during dataset curation

During the dataset curation stage of LAION datasets, cosine similarity between the text and image embeddings has been used to filter images that had *reasonable* textual explanation associated with them in the alt-text field. It was hand-set to 0.3 during the LAION 400M curation process and reduced to 0.28 during the LAION-5B curation process. It has recently come to light by the work presented in Steck et al. [84] that this metric of cosine similarity can potentially yield *"arbitrary and therefore meaningless similarities"* for the learned embeddings in deep models such as CLIP and this constitutes another source of caution for future dataset curation practices.

6.3 CLIP-like models suffering from Concept Association Bias (CAB)

CLIP-like VLMs have been shown to suffer from Concept Association Bias (CAB) [87] on account of being trained on contrastive losses (in lieu of autoregressive losses). Recently, Tang et al. [87] have uncovered an interesting behavior that they termed as Concept Association Bias (CAB) that resulted in VLMs treating inputs as a bag of concepts and attempting to fill in the other missing concept crossmodally often resulting in unexpected zero-shot prediction. We belive that it will be an intriguing downstream study to critique our work and disentangle the contribution of this CAB that resulted in racially biased results.

6.4 Bags-Of-Words like behavior

CLIP-like VLMs fail to encode the compositional relationships between objects and attributes in the images thus displaying Bags-Of-Words like behavior [97]. Similar to Tang et al. [87], Yuksekgonul et al. [97] also discovered the settings in which CLIP-like VLMs treated the constituent objects in an input image as bags-of-words thus displaying limited relational understanding and order insensitivity. They also advocated for composition-aware hard negative mining (CAHNM) as a potential solution and juxtaposing the performance of plain-vanilla VLMs and CAHNM-improved VLMs will be an interesting vector of research exploration.

6.5 Data-leakage and Identity Inference Attacks (IDIA)

CLIP-like large scale VLMs are presented with millions of human images during training and thus are vulnerable to data-leakage and Identity Inference Attacks (IDIA). When we run benchmarking tests on VLM models that have been trained on internet-scale datasets, there exists a real possibility that the individuals who appear in the probe-test datasets might well have appeared in the VLM's training dataset. In a recent paper titled *"Does CLIP Know My Face?"*, Hintersdorf et al. [36] empirically demonstrated how VLMs trained on the LAION-400M dataset trapped and leaked information about individuals appearing less than 25 times in the dataset and how one could preemptively check for this before finalizing on the probe images. We have not done any data-leakage checks in this work and it is an interetsing topic to explore in a future dissemination.

POSITIONALITY STATEMENT

We acknowledge any research process and subsequent knowledge produced cannot be entirely separable from various structural, institutional and personal factors. Seemingly invisible influencing factors include current trends in the field, interests of funding bodies, availability of resources, as well as the interests, motivations, goals, perspectives, and backgrounds of the researchers themselves. Thus, acknowledgement of these factors and transparency (and not hiding behind the veil of objectivity) is instrumental for research excellence. Our team is multi-racial and multi-gender and includes graduate and post-graduate researchers, senior researcher and an independent researcher engaged with AI, machine learning, computer vision, cognitive science, critical race theories, and algorithm and data audits. Having said that, we may have gaps in representing what might be most important to communities at the margins of society. Furthermore, we are all housed within Western universities, a privilege which enabled us to carry out and publish this research with relative ease compared to our peers who may not have the resources or compute available to carry out similar work, for example, those in non-Western universities.

REFERENCES

[1] Mohamed Abdalla and Moustafa Abdalla. 2021. The Grey Hoodie Project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 287–297.

[2] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Patronus AI. 2024. Introducing CopyrightCatcher, the first Copyright Detection API for LLMs. https://www.patronus.ai/blog/introducing-copyright-catcher

[5] Michelle Alexander. 2020. *The new Jim Crow: Mass incarceration in the age of colorblindness*. The New Press.

[6] April H Bailey, Adina Williams, and Andrei Cimpian. 2022. Based on billions of words on the internet, people= men. *Science Advances* 8, 13 (2022), eabm2463.

[7] John K Bardes. 2018. Redefining Vagrancy: Policing Freedom and Disorder in Reconstruction New Orleans, 1862–1868. *Journal of Southern History* 84, 1 (2018), 69–112.

[8] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To" see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–31.

[9] Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the new jim code*. John Wiley & Sons.

[10] Marquis Bey. 2016. "Bring Out Your Dead" Understanding the Historical Persistence of the Criminalization of Black Bodies. *Cultural Studies? Critical Methodologies* 16, 3 (2016), 271–277.

[11] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759* (2022).

[12] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[13] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1536–1546.

[14] Simone Browne. 2015. *Dark matters: On the surveillance of blackness*. Duke University Press.

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[16] Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The guardian* 17, 1 (2018), 22.

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

[18] Reshawna L Chapple, George A Jacinto, Tameca N Harris-Jackson, and Michelle Vance. 2017. Do# BlackLivesMatter? Implicit bias, institutional racism and fear of the black body. *Ralph Bunche Journal of Public Affairs* 6, 1 (2017), 2.

[19] Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788* (2023).

[20] Yizhou Chen, Heng Dai, Xiao Yu, Wenhua Hu, Zhiwen Xie, and Cheng Tan. 2021. Improving Ponzi scheme contract detection using multi-channel TextCNN and transformer. *Sensors* 21, 19 (2021), 6417.

[21] Zekai Chen, Dingshuo Chen, Xiao Zhang, Zixuan Yuan, and Xiuzhen Cheng. 2021. Learning graph structures with transformer for multivariate time-series anomaly detection in IoT. *IEEE Internet of Things Journal* 9, 12 (2021), 9179–9189.

[22] Sanghyuk Roy Choi and Minhyeok Lee. 2023. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology* 12, 7 (2023), 1033.

[23] Charlene H Chu, Rune Nyrup, Kathleen Leslie, Jiamin Shi, Andria Bianchi, Alexandra Lyn, Molly McNicholl, Shehroz Khan, Samira Rahimi, and Amanda Grenier. 2022. Digital ageism: Challenges and opportunities in artificial intelligence for older adults. *The Gerontologist* 62, 7 (2022), 947–955.

[24] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

[25] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 52–59.

[26] Amanuel Elias. 2024. Brief History of Racism. In *Racism and Anti-Racism Today*. Emerald Publishing Limited, 29–56.

[27] Joe Feagin. 2013. *Systemic racism: A theory of oppression*. Routledge.

[28] Constance L Fry, Thomas C Naugle Jr, Shelley A Cole, Jonathan Gelfond, Geetha Chittoor, Angeline F Mariani, Martin W Goros, Barrett G Haik, and Venkata Saroja Voruganti. 2017. The Latino eyelid: anthropometric analysis of a spectrum of findings. *Ophthalmic plastic and reconstructive surgery* 33, 6 (2017), 440.

[29] Shytierra Gaston. 2019. Enforcing race: A neighborhood-level explanation of Black–White differences in drug arrests. *Crime & Delinquency* 65, 4 (2019), 499–526.

[30] Nico Grant and Kashmir Hill. 2023. Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's.

[31] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[32] Samuel R Gross, Maurice Possley, Ken Otterbourg, Klara Stephens, Jessica Paredes, and Barbara O'Brien. 2022. Race and Wrongful Convictions in the United States 2022. *Available at SSRN 4245863* (2022).

[33] Thomas F Gross. 2009. Own-ethnicity bias in the recognition of Black, East Asian, Hispanic, and White faces. *Basic and Applied Social Psychology* 31, 2 (2009), 128–135.

[34] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–13.

[35] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*. 771–787.

[36] Dominik Hintersdorf, Lukas Struppek, Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. 2022. Does CLIP Know My Face? *arXiv preprint arXiv:2209.07341* (2022).

[37] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. Robots enact malignant stereotypes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 743–756.

[38] 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. 2021. *OpenCLIP*. https://doi.org/10.5281/zenodo.5143773 If you use this software, please cite it as below..

[39] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the web conference 2021*. 3767–3778.

[40] Pratyusha Kalluri et al. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (2020), 169–169.

[41] Minji Kang, Sangseon Lee, Dohoon Lee, and Sun Kim. 2020. Learning cell-type-specific gene regulation mechanisms by multi-attention based deep learning with regulatory latent space. *Frontiers in Genetics* 11 (2020), 869.

[42] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2021. Transformer-based direct speech-to-speech translation with transcoder. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 958–965.

[43] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 449–456.

[44] Kimmo Kärkkäinen and Jungseock Joo. 2019. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913* (2019).

[45] Jana Kasperkevic. 2015. Google says sorry for racist auto-tag in photo app. *The Guardian* 1 (2015), 2015.

[46] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.

[47] Ido Kilovaty. 2019. Legally cognizable manipulation. *Berkeley Tech. LJ* 34 (2019), 449.

[48] Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. FABLES: Evaluating faithfulness and content selection in book-length summarization. *arXiv preprint arXiv:2404.01261* (2024).

[49] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. *arXiv preprint arXiv:2011.00747* (2020).

[50] Dohoon Lee, Jeewon Yang, and Sun Kim. 2022. Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nature Communications* 13, 1 (2022), 6678.

[51] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086* (2022).

[52] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[53] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. 2024. SCoFT: Self-Contrastive Fine-Tuning for Equitable Image Generation. *arXiv preprint arXiv:2401.08053* (2024).

[54] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. *arXiv preprint arXiv:2303.11408* (2023).

[55] Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2023. Power hungry processing: Watts driving the cost of ai deployment? *arXiv preprint arXiv:2311.16863* (2023).

[56] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47 (2015), 1122–1135.

[57] Abhishek Mandal, Susan Leavy, and Suzanne Little. 2023. Multimodal Composite Association Score: Measuring Gender Bias in Generative Multimodal Models. *arXiv preprint arXiv:2304.13855* (2023).

[58] Harvey Mannering. 2023. Analysing Gender Bias in Text-to-Image Models using Object Detection. *arXiv preprint arXiv:2307.08025* (2023).

[59] Kieran McCarthy. 2023. Web Scraping for Me, But Not for Thee. https://blog.ericgoldman.org/archives/2023/08/web-scraping-for-me-but-not-for-thee-guest-blog-post.htm. (Accessed on 04/30/2024).

[60] Dan McQuillan. 2022. *Resisting AI: an anti-fascist approach to artificial intelligence*. Policy Press.

[61] Peter Mende-Siedlecki, Jennie Qu-Lee, Jingrun Lin, Alexis Drain, and Azaadeh Goharzad. 2020. The Delaware pain database: A set of painful expressions and corresponding norming data. *Pain reports* 5, 6 (2020).

[62] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[63] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *The Twelfth International Conference on Learning Representations*.

[64] Lisa Nakamura, Shilpa Davé, LeiLani Nishime, and Tasha G Oren. 2005. 'Alllooksame'? Mediating Asian American Visual Cultures of Race on the Web. *East main street: Asian American popular culture* (2005), 262–272.

[65] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035* (2023).

[66] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of oppression*. New York University Press.

[67] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the conference on fairness, accountability, and transparency*. 89–89.

[68] George Pacheco Jr. 2008. *Rhetoric with humor: An analysis of Hispanic/Latino comedians' uses of humor*. The University of Southern Mississippi.

[69] Frank Pasquale. 2015. *The black box society: The secret algorithms that control money and information*. Harvard University Press.

[70] Sentencing Project. 2018. Report to the United Nations on racial disparities in the US criminal justice system. (2018).

[71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

[72] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[73] Katherine J Rosich. 2007. Race, ethnicity, and the criminal justice system. (2007).

[74] Angela Saini. 2019. *Superior: the return of race science*. Beacon Press.

[75] S Adil Saribay, Ali Furkan Biten, Erdem Ozan Meral, Pinar Aldan, Vít Třebický, and Karel Kleisner. 2018. The Bogazici face database: Standardized photographs of Turkish faces with supporting materials. *PloS one* 13, 2 (2018), e0192018.

[76] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[77] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW1 (2020), 1–35.

[78] Nika Schoonover. 2023. Microsoft, Meta and Bloomberg accused of using pirated books in AI development | Courthouse News Service. https://www.courthousenews.com/microsoft-meta-and-bloomberg-accused-of-using-pirated-books-in-ai-development/. (Accessed on 04/30/2024).

[79] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2021. CLIPort: What and Where Pathways for Robotic Manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*.

[80] Tom Simonite. 2018. When it comes to gorillas, google photos remains blind. *Wired, January* 13 (2018).

[81] CalvinJohn Smiley and David Fakunle. 2016. From "brute" to "thug:" The demonization and criminalization of unarmed Black male victims in America. *Journal of human behavior in the social environment* 26, 3-4 (2016), 350–366.

[82] Rory W Spanton and Olivia Guest. 2022. Measuring Trustworthiness or Automating Physiognomy? A Comment on Safra, Chevallier, Gr\ezes, and Baumard (2020). *arXiv preprint arXiv:2202.08674* (2022).

[83] Luke Stark and Jevan Hutson. 2021. Physiognomic artificial intelligence. *Fordham Intell. Prop. Media & Ent. LJ* 32 (2021), 922.

[84] Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is Cosine-Similarity of Embeddings Really About Similarity? *arXiv preprint arXiv:2403.05440* (2024).

[85] Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 701–713.

[86] Nina Strohminger, Kurt Gray, Vladimir Chituc, Joseph Heffner, Chelsea Schein, and Titus Brooks Heagins. 2016. The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behavior research methods* 48 (2016), 1197–1204.

[87] Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. 2023. When are Lemons Purple? The Concept Association Bias of Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 14333–14348.

[88] Paola Tubaro and Antonio A Casilli. 2019. Micro-work, artificial intelligence and the automotive industry. *Journal of Industrial and Business Economics* 46 (2019), 333–345.

[89] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. 2022. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284* (2022).

[90] Emiel Van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083* (2016).

[91] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[92] Carissa Véliz. 2021. *Privacy is power*. Melville House Brooklyn.

[93] Haitao Wang, Jiale Zheng, Ivan E Carvajal-Roca, Linghui Chen, and Mengqiu Bai. 2023. Financial Fraud Detection Based on Deep Learning: Towards Large-Scale Pre-training Transformer Models. In *China Conference on Knowledge Graph and Semantic Computing*. Springer, 163–177.

[94] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* (2019).

[95] Xinze Yang, Chunkai Zhang, Yizhi Sun, Kairui Pang, Luru Jing, Shiyun Wa, and Chunli Lv. 2023. FinChain-BERT: A High-Accuracy Automatic Fraud Detection Model Based on NLP Methods for Financial Scenarios. *Information* 14, 9 (2023), 499.

[96] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[97] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it?. In *The Eleventh International Conference on Learning Representations*.

[98] Ting-He Zhang, Md Musaddaqul Hasib, Yu-Chiao Chiu, Zhi-Feng Han, Yu-Fang Jin, Mario Flores, Yidong Chen, and Yufei Huang. 2022. Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions. *Cancers* 14, 19 (2022), 4763.

[99] Zixin Zheng, Sijin Li, Licheng Mo, Weimao Chen, and Dandan Zhang. 2021. ISIEA: An image database of social inclusion and exclusion in young Asian adults. *Behavior Research Methods* (2021), 1–13.

APPENDIX

A ADDITIONAL METHODOLOGICAL DETAILS

A.1 Self-similarity matrix of CFD extracted featured

To highlight how self-similar the 8×8 textual features are, we present Figure 7(a) that has the annotated heatmap of the $\mathbf{F}_\tau \times \mathbf{F}_\tau^T$ matrix. Similarly, we also present Figure 7(b) that has the heatmap of the 597×597 sized $\mathbf{F}_I \times \mathbf{F}_I^T$ matrix. Given the fact that the 597 images were sorted and grouped by Race-Gender categories, the block-like structures visible in Figure 7(b) indicate the fact that the model's output image features are influenced by these categorical indicators.

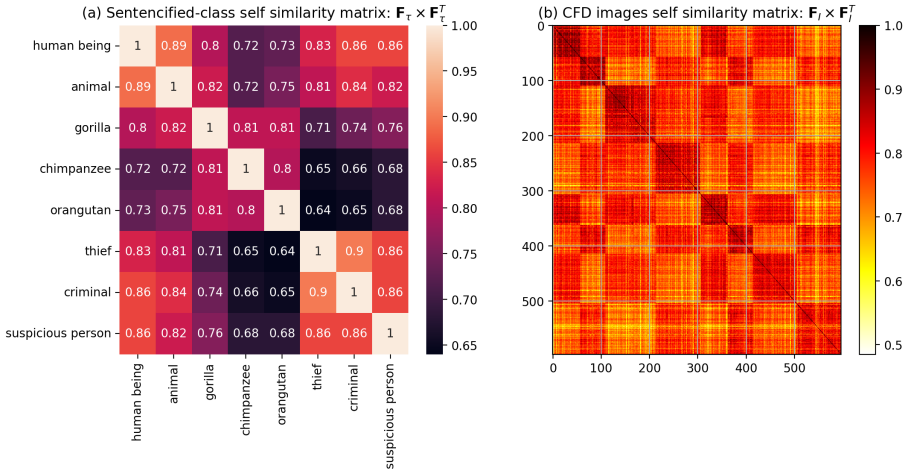


Fig. 7. Heatmap plots to help the reader visualize the (a) self-similarity matrix: $\mathbf{F}_\tau \times \mathbf{F}_\tau^T$ of the sentences corresponding to the class-labels and (b) self-similarity matrix: $\mathbf{F}_I \times \mathbf{F}_I^T$ of the features extracted from the CFD images.

A.2 Randomly selected, hand-blurred samples from the CFD

A sample of images from the Chicago Face Database (CFD) across the eight self-classified race-gender combinations. The images are sized $2444(w) \times 1718(h)$ pixels and *"equated for color temperature and placed onto a plain white background"*. Of the 597 individuals, 307 self-classified as "female" and 290 self-classified as "male". We hand-blurred these sample images for this study to preserve the anonymity of pictured individuals. The titles of each of these images here follow the exact file names given to these images in the `CFD 3.0` version that is hosted at https://www.chicagofaces.org/download/.

B ON ALLLOOKSAMEISM, NEGATIVE STEREOTYPES AND RACIAL MISCLASSIFICATION

The goal here was to understand how stereotypes about facial appearances are cross-related with racial identities. When we looked at the results (Figure 9) we saw an interesting theme emerge: the self-classified Latino/Latina individuals were misclassified with high confidence as one of the 'Asian' classes on account of the presence of *epicanthic folds* and this tendency to stereotype got worse with dataset scaling. The titles of these subplots here are formatted as strings with 4 fields separated by the '|' character: < cfd_Vit-L-14 > | < training-dataset > | < $P_{lf \to af}$ > | < $P_{lm \to am}$ >. Here, $P_{lf \to af}$ is the probability that an image belongs to the `Latina-Female` category was misclassified as `Asian-Female` (and $P_{lm \to am}$ is the probability that an image belongs to the `Latino-Male` category was misclassified as `Asian-Male`).

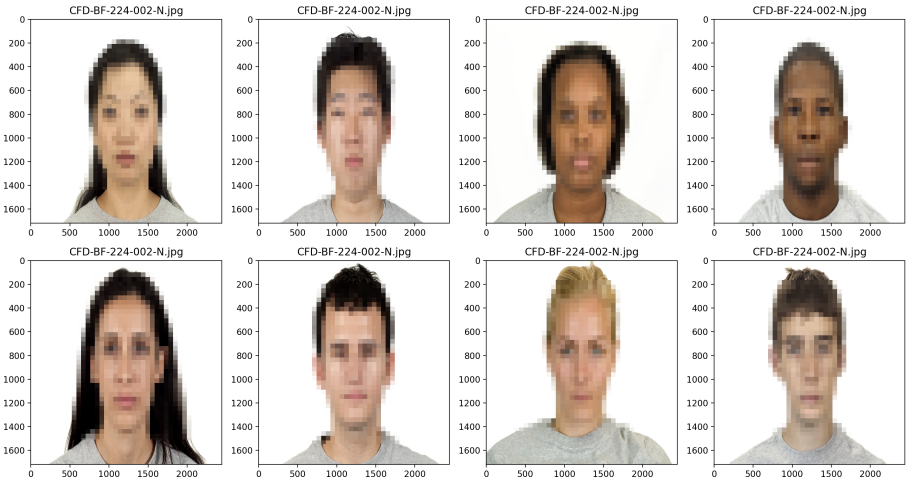


Fig. 8. A sample of images from the Chicago Face Database (CFD) across the eight self-classified race-gender combinations.

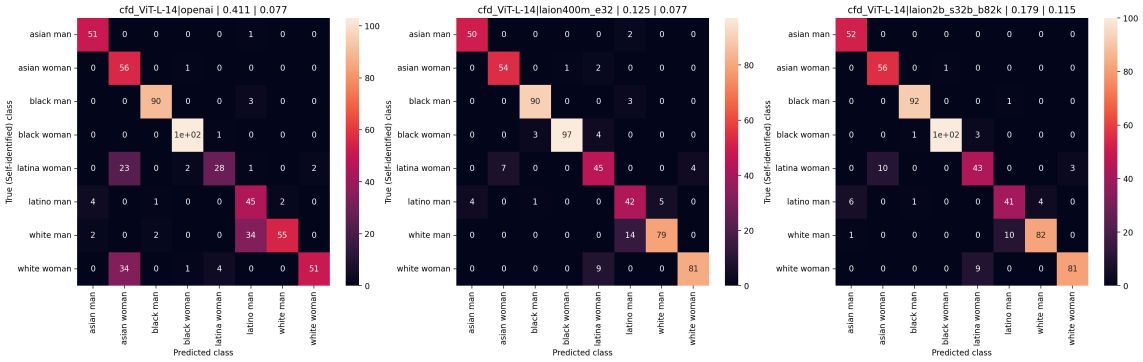


Fig. 9. Heatmap of the confusion matrix of the race-gender classification experiment showing misclassification Latino/Latina individuals as 'Asian' class. This misclassification got worse with dataset scaling.

As seen in the first of the 3 subplots (from left) that maps to the OpenAI-WIT dataset 23 of the 56 Latina women were misclassified as Asian women leading to a $P_{lf \to af} = 23/56 = 0.411$. This misclassification rate was better for the LAION-400M model (0.125) and worsened to 0.179 for the LAION-2B-En model, thereby yielding yet another example of worsening of the bias-related metrics upon scaling the dataset from 400M to 2B samples. The same trend also showed up for Latino men with the misclassification rate increasing nearly 50% from 0.077 to 0.115.

Correspondingly, there exists a substantial body of scientific literature (See [28, 33, 64, 68]) on not just the oft-ignored high levels of prevalence of the epicanthic folds in Hispanic/LatinX populations[2] but also on the sociological ramifications of this *alllooksame-ism* [64] that permeates aspects of the mainstream culture.

[2]"In Latinos, the inner canthal distance and lateral canthal angle of inclination were similar to Asians, while the lid crease spanned the range from Asians to Caucasians. Half of the Latinos had epicanthal folds" [28]