A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning

Hugo Berg*, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain

Oxford Artificial Intelligence Society, University of Oxford

Abstract. Vision-language models can encode societal biases and stereotypes, but there are challenges to measuring and mitigating these harms. Prior proposed bias measurements lack robustness and feature degradation occurs when mitigating bias without access to pretraining data. We address both of these challenges in this paper: First, we evaluate different bias measures and propose the use of retrieval metrics to image-text representations via a bias measuring framework. Second, we investigate debiasing methods and show that optimizing for adversarial loss via learnable token embeddings minimizes various bias measures without substantially degrading feature representations.

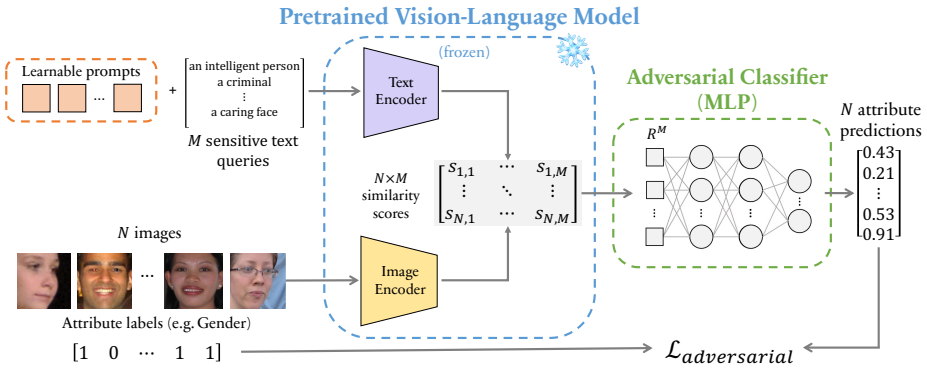

Fig. 1: Our proposed debiasing method for pretrained vision-language models. Sensitive text queries and images (with labeled attributes, e.g. Gender) are fed to their respective frozen text and image encoders. We employ an adversarial classifier which aims to predict the image attribute labels from similarity scores between the outputs of the two encoders. Learnable "debiasing" prompt tokens are prepended to the sensitive text queries and optimized to maximize the error of the adversary. In this way, biased correlations between text-image similarity scores and attribute labels are reduced whilst preventing significant feature degradation.

* Corresponding author: hugo@hugob.se

1 Introduction

Large-scale, pretrained vision-language models are growing in popularity due to their impressive performance on downstream tasks with minimal additional training data. Their success can be attributed to three main advances: the success of transformers in NLP [16, 14, 8], cross-modal contrastive learning [68, 46] and the increasing availability of large-scale multimodal web datasets [12, 5, 13]. These models, including OpenAI's CLIP [50], are now readily available through APIs [1, 2], allowing non-technical users to capitalize on their high performance 'out of the box' on zero-shot tasks [40]. Despite these benefits, an expansion in scope for downstream applications comes with greater risk of perpetuating damaging biases which the models have learned during pretraining. The training data, consisting of web crawled image-text pairs, is not audited for quality nor toxicity [6]. Cultural and temporal specificity is also of concern, given models are trained on a snapshot in space and time [62, 15, 51]. These snapshots, or alternatively, "frozen moments" [29] can reinforce negative stereotypes that may otherwise naturally alter through societal pressures and change of norms.

The risk and type of societal harm intimately interacts with the downstream task at hand. Clearly, using vision-language models for dog-species classification poses very different dangers to projecting the similarity of human faces onto axes of criminality [20, 64] or homosexuality [61]. Even in more benign applications such as image search, there may be harmful consequences arising from representational and/or allocational harms. Representational harms come from the technological entrenchment of stereotypical perceptions, where associations of certain genders or ethnicities are made through a biased lens. For example, the over-representation of one gender when querying for a profession (e.g. "nurse" versus "doctor") or one ethnicity when querying for explicit and NSFW content [6]. Allocational harms arise when an individual's or a social group's opportunities or access to resources and opportunity are differentially impacted [62]. For example, if images returned in searches for specific occupations contain specific demographics, then inadequately represented peoples might self-select themselves out of applying for such jobs. The ordering of images in search results can shift people's perceptions about what these distributions might appear like in the 'real-world' [38].

While frameworks to measure bias have been established for NLP and image classification, none currently exist for vision-language [3]. However, tools to measure bias alone do not provide actionable progress unless less-biased models are also available. Thus, investigating methods to mitigate model-level bias is equally as important. Debiasing large-scale pretrained models is not without its challenges, namely a lack of access to the original training data and the infeasible amount of compute required for retraining. For the successful and safe adoption of vision-language models by downstream users, we need both effective measures of bias as well as efficient methods of debiasing. To this end, this work makes the following three contributions: (i) we investigate and evaluate different measures of bias for vision-language models, showing that some prior measures, such as *WEAT*, are inappropriate for vision-language models (ii) we evaluate gender and racial bias in state-of-the-art models on two face datasets: FairFace

[37] and UTKFace [70]; and finally (iii) we provide a framework for debiasing vision-language models requiring only image-attribute label supervision and show that optimizing for unbiasedness via an array of learnable prompt tokens is the best strategy for mitigating bias without substantially degrading the quality of feature representations.

2 Related Works

Recent advances have seen the release of powerful, open-source vision-language models [50, 47, 5]; but research into bias measurement and mitigation has not kept pace, with only a few papers to date tackling these topics for vision-language [3, 71, 31]. However, inspiration can be drawn from research in computer vision and natural language processing fields, which investigates the measurement and mitigation of dataset- and model-level biases [45].

Bias in NLP. Large-scale language models are optimized to accurately reflect language by learning underlying statistical patterns, which can be problematic if training datasets contain harmful or misrepresentative language [62]. In NLP, prior work has documented gender bias [7, 72], racial bias [43, 21] and their intersections [28, 40]. One commonly-deployed bias metric is the Word Embedding Association Test (*WEAT*) [11], derived from the Implicit Association Test (IAT) [26] which measures the time-delay that human subjects take in associating a given demographic group with positive or negative descriptors. *WEAT* uses cosine differences to measure the similarity in vector space between a demographic group and a given contextual prompt, e.g. the man/woman vector and a science/humanities profession. Applying *WEAT* to sentences embeddings [44] has been shown to lack robustness and consistency. In this work, we adapt *WEAT* to the vision-language setting and evaluate its appropriateness as a measure.

In terms of debiasing, [7, 43] focus on altering the embedding space but these methods have been criticized by [24] who show that gender bias is still reflected in distances between "gender neutralised" words; thus we do not pursue embedding-level debiasing as a viable method in our work. [72, 9] propose dataset-level debiasing techniques through data augmentation and perturbation, and [48] implement supervised fine-tuning on data checked by humans. While these techniques have shown promise, they either require access to the original training data, or compute to retrain large models; so, are not feasible with the large-scale, pretrained vision-language models under investigation in our work.

Bias in Computer Vision. Investigations into computer vision models have also uncovered evidence of gender bias [73, 55], racial bias [63, 55], and their intersectionality [10, 55]. Whilst not the focus of this paper, works such as [6, 32, 33, 49, 22, 34, 58] focus on bias stemming from dataset curation and annotation, and outline more appropriate practices to ensure fair representation. Model-based debiasing methods are more similar to our work, these include optimizing confusion [4], domain adversarial training [18], or training a network

to *unlearn* bias information [27]. While the latter of these approaches is out-of-scope because it requires retraining the entire network from scratch with the original training data, we adopt the idea of adversarial fine-tuning in our work.

Bias in Vision-Language. Several approaches make initial steps towards measuring bias in vision-language representations. The authors of the original CLIP paper investigated manifestations of bias within their own model [3] by assessing the misclassification of faces by age or race with non-human and criminal categories. Given the lack of robustness in measures such as *WEAT*, other metrics such as ranking metrics for search and recommendation tasks [23] may be better suited. These works can be applied to vision-language by measuring the attribute distribution of returned images for a given text prompt. The sparse literature on debiasing vision-language models falls into two categories: (1) dataset-level debiasing [71] and (2) model-level debiasing [31]. On the dataset side, imbalanced training data has been proposed as a point of blame [71] but balancing the data is not a panacea, with [60] finding exaggerated gender stereotypes in tasks unrelated to gender recognition, even after balancing the number of labels by gender. In large-scale, multi-modal datasets, like those used to train vision-language models, the disproportionate representation of certain genders and ethnicities in various roles can lead to misclassifications, such as overly strong associations of black people playing basketball [31] or certain genders and ethnicities with explicit content [6]. However, correcting bias at the dataset-level requires access to the original training data and retraining the model on alternative data. Even if an end-user had these resources, it may still be infeasible to capture all proxies for demographic bias [31], so it is possible that the data necessary to combat bias has not been curated yet [62]. On model-level adjustments, [31] train an image captioning model to confidently predict gender when there is gender evidence, and be cautious in its absence. Specifically, they improve the captioning model by attending to the right visual evidence, such as the person in the image as opposed to the background. However, retraining the model with costly masked supervision is beyond the scope of this work.

Domain adaptation of Pretrained Models. Adapting pretrained generalist models to target downstream tasks is similar to our work in that the goal is to adapt the model to be less biased without significant degradation of the feature representation. Prompting has become the de-facto domain adaptation technique for vision-language models [74, 35, 59, 75], as well as large language models [53, 42]. Constraining the optimization space to input tokens leads to training less than 0.1% of the total model size but is surprisingly effective at model adaptation with minimal training data [76] while avoiding overfitting. Similarly, [69] show that optimizing over only the text encoder and freezing the image encoder is superior to full finetuning and improves generalization.

In our work, we build on prior bias measurements, namely an adaptation of *WEAT*, harmful image misclassifications, and ranking metrics. Our aim is to demonstrate a debiasing protocol which does not assume access to the original training data, nor the resources to retrain from scratch. Thus, we focus our efforts on fine-tuning, prompting and debiasing protocols with low computational cost.

3 Measuring Bias

We consider the problem of learning unbiased joint text-image representations, but we must first establish a framework for quantitatively measuring the amount of bias in these representations. In this section, we outline such a framework and how to apply bias measures proposed in prior works.

Given a dataset of image-attribute pairs (I, A) where I is an image and A is its corresponding attribute from a set of disjoint protected attribute labels $\mathcal{A} = \{A_1, ..., A_l\}$, for example photos of faces with gender labels. Suppose there is a set of sensitive text queries, $\mathcal{T} = \{T_1, ..., T_m\}$ with corresponding concepts $\mathcal{C} = \{C_1, ..., C_m\}$, such as the sentences "a photo of a good person", "a photo of a bad person" and their corresponding concepts "good" and "bad".

Our goal is to learn a joint vision-language model Ψ that: (i) outputs a similarity score for image-text pairs, $s = \Psi(I, T)$, where semantically similar image-text pairs are scored highly; and (ii) is unbiased, defined as outputting similar distributions of scores across attributes for a given text query which *should* be unrelated to demographic affiliation. Specifically, we consider the case where Ψ is initialized as a pretrained model which already achieves (i) but not (ii) – as is the case with current large, pretrained vision-language models, which are often used for zero-shot classification, as well as image and video retrieval. We evaluate the bias of a model when applied to these two scenarios.

3.1 WEAT

The Word Embedding Association Test (*WEAT*) [11] is used to measure the bias of word and sentence embeddings [11, 44], and more recently has been adapted to evaluate the the bias of vision encoders [55]. *WEAT* measures the differential association between a set of two target concepts $\mathcal{C} = \{C_1, C_2\}$ (e.g. 'career' and 'family') and a set of attributes $\mathcal{A} = \{A_1, ..., A_l\}$ (e.g. 'male' and 'female'). Here each concept C_i and attribute A_i contain embeddings in a common space for stimuli associated with them (e.g. 'office', and 'business' for the concept 'career', and 'boy', 'father' and 'man' for the attribute 'male'). Now the differential association between concepts C_1 and C_2 and attributes A_1 and A_2 is defined as

$$s(C_1, C_2, A_1, A_2) = \sum_{c_1 \in C_1} s(c_1, A_1, A_2) - \sum_{c_2 \in C_2} s(c_2, A_1, A_2), \tag{1}$$

where

$$s(w, A_1, A_2) = \text{mean}_{a_1 \in A_1} \cos(w, a_1) - \text{mean}_{a_2 \in A_2} \cos(w, a_2) \tag{2}$$

measures the differential association of w with the attributes using cosine similarity. The significance of this association is computed using a permutation test. Denoting all the equal-size partitions of $C_1 \cup C_2$ by $\{(C_1^i, C_2^i)\}^i$, we generate a null-hypothesis of no bias and compute the p-value

$$P_{r_i}[s(C_1^i, C_2^i, A_1, A_2) > s(C_1, C_2, A_1, A_2)]. \tag{3}$$

Finally, the effect size, i.e., the normalized measure of the separation between the associations of the targets and attributes, [11] is defined as

$$\frac{\text{mean}_{c_1 \in C_1} s(c_1, A_1, A_2) - \text{mean}_{c_2 \in C_2} s(c_2, A_1, A_2)}{\text{std}_{c \in C_1 \cup C_2} s(c, A_1, A_2)}. \tag{4}$$

In the case of *WEAT*, all attributes and categories are word embeddings. In our experiments, we have cross-modal interactions where the target concepts $\mathcal{C}$ are inferred from the text queries $\mathcal{T}$ and are the corresponding embeddings from the text encoder of the vision-language model, and attributes $\mathcal{A}$ are the image embeddings from the vision encoder.

3.2 Ranking Metrics

We also apply bias measures from the information retrieval literature [23, 66] to the setting of text-to-image retrieval. This is a natural application given that vision-language models are increasingly being used for semantic image search [1], introducing biases from the attributes which get ranked higher than others in the top k results. Let τ_y be a ranked list of images $\mathcal{I}$ according to their similarity to a text query T, and τ_T^k be the top k images of the list.

Skew@k [23] measures the difference between the desired proportion of image attributes in τ_T^k and the actual proportion. For example, given the text query "this person has a degree in mathematics", a desired distribution of the image attribute gender could be 50% to ensure equal representation.

Let the desired proportion of images with attribute label A in the ranked list be $p_{d,T,A} \in [0, 1]$, and the actual proportion be $p_{\tau_T,T,A} \in [0, 1]$. The resulting $Skew$ of τ_T for an attribute label $A \in \mathcal{A}$ is

$$Skew_A@k(\tau_T) = \ln \frac{p_{\tau_T,T,A}}{p_{d,T,A}} \tag{5}$$

This measurement gives an indication of possible representational bias [62], with certain attributes being under-represented in the top k search results (i.e. a negative $Skew_{A_i}@k$). However, $Skew_{A_i}@k$ has a couple of disadvantages: (i) it only measures bias with respect to a single attribute at a time, and so must be aggregated to give a holistic view of the bias over all attributes A; (ii) different chosen values of k gives different results, so more than a single $Skew$ value would need to be computed for each attribute. These disadvantages form the basis of the next two measures, proposed by [23], which address each of these limitations.

MaxSkew@k is the maximum $Skew@k$ among all attribute labels A of the images for a given text query T

$$MaxSkew@k(\tau_T) = \max_{A_i \in \mathcal{A}} Skew_{A_i}@k(\tau_T). \tag{6}$$

This signifies the *"largest unfair advantage"* [23] belonging to images within a given attribute. The desired outcome is 0, implying that the real distribution is equal to the desired distribution (e.g. all genders are equally represented in the ranked images, when the desired distribution is uniform).

Normalized Discounted Cumulative KL-Divergence (*NDKL*) employs a ranking bias measure based on the Kullback-Leibler divergence, measuring how much one distribution differs from another. This measure is non-negative, with larger values indicating a greater divergence between the desired and actual distributions of attribute labels for a given T.

Let $D_{\tau_T^i}$ and D_T denote the discrete distribution of image attributes in τ_T^i and the desired distribution, respectively. *NDKL* is defined by

$$NDKL(\tau_T) = \frac{1}{Z}\sum_{i=1}^{|\tau_y|}\frac{1}{\log_2(i+1)}d_{KL}(D_{\tau_T^i}||D_T) \tag{7}$$

where $d_{KL}(D_1||D_2) = \sum_j D_1 \ln\frac{D_1(j)}{D_2(j)}$ is the KL-divergence of distribution D_1 with respect to distribution D_2, and $Z = \sum_{i=1}^{|\tau_r|}\frac{1}{\log_2(i+1)}$ is a normalization factor.

The KL-divergence of the top-k distribution and the desired distribution is a weighted average of $Skew_A@k$ measurements (averaging over $A \in \mathcal{A}$). So, this aggregation overcomes the first disadvantage of *Skew*. However, *NDKL* is non-negative, and so it cannot distinguish between two "opposite-biased" search procedures.

Aggregation over text queries In practice, there are many sensitive text queries that we would like to measure and mitigate bias against and multiple ways to construct such text queries. In order to produce single bias measures, we aggregate across the list of text queries (see Table 1) using arithmetic mean over all templates.

3.3 Harmful Zero-shot Image Misclassification

[3] proposed an alternative measure of the bias in a vision-language model using the zero-shot misclassification rates of people into derogatory criminal and non-human categories. They found that certain demographics might be disproportionately affected by using CLIP as a zero-shot classifier, where black people were more likely to be classified as non-human, and women were more likely to be classified into lower status occupations. We follow their protocol by using CLIP to classify images from the FairFace validation set into different categories, the $7 \cdot 2 = 14$ FairFace ethnicity-gender class pairs, non-human categories (animal, gorilla, chimpanzee, and orangutan) and crime-related words (thief, criminal and suspicious person). We then look at the percentage of images that are misclassified into the non-human and crime classes. The original implementation [3] is lacking

in details, and it is unclear if they use a template approach. We use the template "a photo of a {}", since it's the standard for all other CLIP measurements. We also perform the test without using a query template but classification accuracy was significantly reduced for all images.

4 Debiasing

The proposed debiasing method has two components: (i) the objective function to minimize in order to reduce bias; and (ii) the choice of parameters to optimize over in the vision-language model Ψ in order to minimize (i).

4.1 Fairness Objective

Adversarial Debiasing. We follow a common approach in bias mitigation [18, 19, 65, 57] and employ an adversarial classifier, θ_{adv}, whose aim is to predict the attribute label A of image I given only its similarity logits from the set of sensitive text queries $\mathcal{T}$

$$\hat{A} = \theta_{\text{adv}}(S) \quad \text{where} \quad S = [s_1, ..., s_M] \in \mathbb{R}^M \quad \text{and} \quad s_m = \Psi(I, T_m). \tag{8}$$

The adversarial classifier is trained to minimize the cross entropy loss between the predicted attribute labels $\hat{A}$ and the ground truth attribute labels A

$$\mathcal{L}_{\text{adv}} = -\sum_{A \in \mathcal{A}} A \log \theta_{\text{adv}}(S). \tag{9}$$

To encourage an unbiased representation, blind to the sensitive attributes, we optimize the vision-language model to maximize this adversarial loss.

4.2 Adaptation Methods

Naïve optimization of the above objective function without any regularization can lead to trivial solutions, such as Ψ outputting the same logits irrespective of the image or text query. This solution is undesirable since the feature representation loses all semantic information of the input, making it effectively useless for downstream tasks. To prevent this, we investigate regularization techniques which restrict the set of parameters in the image-text model Ψ which can be optimized over.

Finetuning Depth Instead of optimizing every parameter in every layer, a common regularized adaption technique is to only finetune the layers in the image-text encoders to a certain depth [77]. When Ψ is instantiated as a dual stream encoder [50, 5, 47], with the text and image embeddings encoded via independent streams, $s = \Psi(x, y)$ where $\Psi(x, y) = \Psi_i(x)^T \Psi_t(y)$, different finetuning depths can be chosen for each encoder $\Psi_i(x), \Psi_t$. [69] show that finetuning

only the text encoder Ψ_t improves generalization and reduces catastrophic forgetting of the original pre-trained representation when compared to full finetuning.

Prepending learnable text tokens Prompt learning has shown promising results for few-shot learning, when pretrained models are applied to downstream tasks with minimal additional data [74, 59, 35]. The optimization over prompt tokens of a few thousand parameters (rather than the full model which can be 100M+) enforces heavy regularization and prevents catastrophic overfitting to the few samples.

We use this method to regularize the debiasing optimization, since unconstrained training to maximize the adversary's loss can simply collapse all embeddings. Following [74], we prepend learnable text tokens to the text queries after they have been embedded by the token embedding layer. We initialize these learnable tokens as the zero-pad embeddings, minimize deviation from the original text embedding to the original text query, and optimize over the learnable tokens – the rest of the model weights are frozen. However, even with zero-pad initialized token embeddings, token embeddings of prompts are different to their non-prepended counterparts, and so the text-encoder outputs are slightly modified. This results in a degradation of model performance before any training has occurred (see Appendix).

5 Experiments

5.1 Datasets

FairFace [37], consists of 108,501 images of GAN-generated faces. This dataset has emphasis on a balanced composition by age, gender and ethnicity. The seven ethnicities included are: White, Black, Indian, East Asian, South East Asian, Middle East and Latino. The training dataset for the utilized GAN was collected from the YFCC-100M Flickr dataset [56].
UTKFace Cropped image dataset [70] contains 20,000 images and includes four distinct ethnicities: White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern). This is a notable limitation compared to FairFace which has individual classes for each of these. UTKFace is also very different to the qualitative characteristics of FairFace, in terms of large variance in lighting conditions, color quality and angle of portraits.

5.2 Experimental Protocol

Text query generation. We select pairwise adjectives from the IAT dataset[1]. These are pairs of antonyms, which do not have any correlation to facial expressions, nor direct connections to sensitive attributes, e.g. not "happy/sad" or "beautiful/handsome". The full list of templates and concepts used are shown in Table 1, where we expand the test set with unseen prompt templates and

[1] https://osf.io/y9hiq/

Table 1: Templates alongside the concepts used to populate them, for the training and testing of our debiasing protocols.

Train template (T_{train})	**Train concepts** (C_{train})	**Test templates**	**Test concepts**
A photo of a {} person	good, evil, smart, dumb, attractive, unattractive, lawful, criminal, friendly, unfriendly	T_{train} + A {} person, A {} individual, This is the face of a {} person, A photo of a {} person, A cropped photo of a {} face, This is a photo of a {} person, This person is {}, This individual is {}	C_{train} + clever, stupid, successful, unsuccessful, hardworking, lazy, kind, unkind, nasty, noncriminal, moral, immoral, rich, poor, trustworthy, caring, heroic, dangerous, dishonest, villainous, violent, nonviolent, honest

concepts to assess generalizability to varied downstream applications.

Bias metrics. Despite *WEAT* showing promising results in measuring bias in word embeddings, we find the effect size is overly sensitive to changes in model architecture, evaluation dataset, as well as minor syntactic changes in text queries (see Supplementary material). This may be explained by the fact that *WEAT* was primarily designed for single word embeddings, while we are using long prompts. [44] found *SEAT* (Sentence Embedding Association Test) to fail for analogous reasons. Accordingly, we implement *MaxSkew@1000* and *NDKL* which show consistent performance in measuring bias across different model architectures, datasets and minor syntactic changes.

Downstream performance metrics. For evaluating model performance on downstream tasks, we report the zero-shot (ZS) performance on (i) $\text{flickr}_{R@5}$: recall@5 text-to-image retrieval on the Flickr-1k test set [67]; and (ii) $\text{CIFAR}_{acc\%}$: zero-shot classification accuracy on the CIFAR100 [41] test set.

Pretrained models. CLIP [50] combines a text and image encoder, whose representations are projected to the same space. CLIP was originally trained using a contrastive loss on 400M image-text pairs from the web. During bias measurement experiments, we experiment over variants with different image encoders - ResNet50 [30], ViT [17], SLIP [47] and FiT [5].

Debiasing implementation. For debiasing, we use CLIP $\text{ViT}_{B/16}$ and prepend 2 learnable prompt embeddings to the text query. Models are trained using a NVIDIA GTX Titan X with a batch size of 256. The adversarial classifier is a multilayer perceptron (MLP) with ReLU activation, two hidden layers of size 32, input size equal to the number of training text prompts, and output size equal to the number of sensitive attributes we debias over, $\dim(A)$. We train with the *Adam* optimizer [39] and use learning rates of $2 \cdot 10^{-5}$ and $2 \cdot 10^{-4}$ for CLIP and the adversarial classifier, respectively. Following an initial two epochs of only training the adversarial model, the CLIP and adversarial model are alternately trained for 10 batches each. Minimal parameter tuning is employed due to the computational costs. Early stopping is implemented if the CLIP model

Table 2: Evaluation of gender bias on the FairFace validation set for various model architectures (arch.) and pretraining datasets. The models evaluated are as follows: CLIP[50] models trained on the WIT dataset; SLIP[47] models trained on YFCC15M with and without self-supervised learning (SSL); FiT[5] models trained on CC[52] and WebVid[5]. For each model, we additionally report the downstream zero-shot (ZS) performance of text-to-image retrieval ($\text{flickr}_{R@5}$) and image classification (CIFAR_{acc}).

Pretrain Dataset	Pretrain Size	Arch.	Bias Measures↓		ZS Performance↑	
			MaxSkew@1000	*NDKL*	$\text{flickr}_{R@5}$	CIFAR_{acc}
WIT[50]	400M	RN50	0.197	0.075	83.7	40.5
		$\text{ViT}_{B/32}$	0.185	0.073	83.6	62.3
		$\text{ViT}_{B/16}$	0.233	0.103	86.1	66.5
		$\text{ViT}_{L/14}$	0.202	0.083	87.4	75.7
YFCC[47]	15M	$\text{ViT}_{B/16}$	0.259	0.115	60.1	33.7
		$\text{ViT}^{\textbf{SSL}}_{B/16}$	0.231	0.117	68.7	46.6
		$\text{ViT}_{L/14}$	0.255	0.112	61.6	40.8
		$\text{ViT}^{\textbf{SSL}}_{L/14}$	0.206	0.066	69.3	53.2
CC,WV[52, 5]	5.6M	$\text{FiT}_{B/16}$	0.292	0.174	76.3	70.4

performance as tested on CIFAR100 [41] or Flickr-1k [67] drops below 50% of the original accuracy. The small size (measured in number or size of hidden layers, or total # of parameters) of the adversarial model is motivated by the size of its input (fewer than 20 training prompts) and the size of its output (fewer than 10 sensitive attributes). We expect even the small adversarial model to remove any linear and reasonable non-linear relationships between the output logits of our vision-language models, i.e. be able to find bias if and when it exists. For finetuning, we choose to train all combinations of the last three layers of the text encoder (transformer-based with 12 layers total), the last three image encoder layers (also transformer-based with 12 layers) and the two projections from text and image feature space to the embedding space. We purposefully do not choose to train the entire model, as the expected feature quality loss is large, as well as the memory and computational requirements being significantly higher than for training only 25% of the model's parameters.

5.3 Results

Bias across different model architectures and pretraining. The results in Table 2 indicate that higher performance on feature quality metrics comes from (1) models pretrained on larger datasets , and (2) models with larger image encoders ($\text{RN50} < \text{ViT}_{B/32} < \text{ViT}_{B/16} < \text{ViT}_{L/14}$). The FiT model breaks the pattern, which may be explained by its joint training on both images (CC) and video (WV) and their cleaner quality compared to YFCC15M. Increased size of the pretraining dataset seems to result in lower bias metrics (both *MaxSkew* and *NDKL*). The SLIP $\text{ViT}_{B/16}$ and $\text{ViT}_{L/14}$ models trained with SSL have lower *MaxSkew* than their non-SSL counterparts, confirming the finding of [25] that self-supervised learning produces less biased models. The best performing models pretrained on WIT [54] and YFCC100M [56] by feature quality also have nearly the lowest bias for their respective datasets. This suggests there is no trade-off

Table 3: Comparison of adaptation techniques for debiasing gender on FairFace via adversarial learning. Bias and zero-shot downstream performance measures are displayed as absolute values with percentage change relative to the pretrained baseline, a CLIP model with $\text{ViT}_{B/16}$ architecture.

Debias Adaptation	**Bias Measures ↓**		**ZS Performance ↑**	
	MaxSkew@1000	*NDKL*	$\text{flickr}_{R@5}$	CIFAR_{acc}
PT baseline	0.233	0.103	86.1	66.5
Prompt	**0.073(-69%)**	**0.021(-80%)**	64.2(-25%)	**54.3(-18%)**
Proj. layer	0.642(+176%)	0.561(+443%)	62.3(-28%)	40.6(-39%)
Text encoder	0.691(+197%)	0.688(+566%)	**67.8(-21%)**	43.4(-35%)
Full fine-tuning	0.688(+195%)	0.664(+543%)	18.6(-78%)	6.6(-90%)

between feature quality and model bias but work is needed to establish their relationship.

Effectiveness of debiasing approaches. To preserve the pretrained model's feature quality, one could use a contrastive loss or similar objective to the original training task. However, none of the debiasing methods in Table 3 do this, since we assume no access to the training dataset, nor to large computational resources. Accordingly, the performance decrease after debiasing is to be expected.

During debiasing, we tried using squared ℓ_2 loss [36] between the original model embeddings and debiased model embeddings, as well as adversarial loss. However, finetuning and prompt learning in this setting does not reduce bias nor increase feature quality. With only the adversarial loss, prompt learning significantly reduces the bias metrics (-69% to -80%), and has the lowest decrease in feature quality (-18% to -25%). As well as the prepended token embeddings shown in Table 3, we also experimented with appended, and added learned token embeddings, as well as different initializations (e.g. zero-pad, embedding of a common initial token, and random). These showed varying feature quality at start of training, but no significant change in results. Using more learned prompt tokens results in a decrease in feature quality at start-of-training, so two tokens are chosen as a reasonable trade-off. Due to the adversarial objective, the strong regularization from having few learned embeddings is enough to keep feature quality at an acceptable level. However, finetuning larger parts of model allows it to collapse the embeddings, resulting in low performance for the adversary, but bias reduction. With more extensive hyperparameter tuning, we expect that similar reductions in bias can be achieved without degrading feature quality as much. The best bias results are achieved early on for all techniques in Table 3, and reach their optimal within 3 epochs, so this method is relatively computationally cheap ($\sim$ 3 hrs per training run on 1 GPU). We also note that for models with separate image and text encoders, training prompt embeddings allows precomputation of image embeddings, thus decreasing computational cost significantly.

Generalization across datasets and attributes. Table 4a shows the percentage change in bias measures when training the debiasing protocol (*Prompt* in Table 3) for gender attributes on FairFace then evaluating on UTKFace (and

Table 4: Debiasing generalisation results of the prompt method when training and testing on different datasets (a) and attribute types (b) for the debiasing prompt model. Bias mitigation is consistently reduced in these unseen settings.

(a) Cross-Datasets

	Bias Measures ↓			
	MaxSkew@1000		*NDKL*	
Eval →	FairFace	UTKFace	FairFace	UTKFace
Train ↓				
PT baseline	0.233	0.034	0.103	0.014
FairFace	-68.71%	-36.82%	-72.54%	16.61%
UTKFace	-8.38%	-35.15%	4.31%	-3.23%

(b) Cross-Attribute

	Bias Measures ↓			
	MaxSkew@1000		*NDKL*	
Eval →	Gender	Race	Gender	Race
Train ↓				
PT baseline	0.233	0.549	0.103	0.209
Gender	-68.71%	-39.57%	-78.98%	-45.33%

vice-versa).[2] Training on FairFace gives large decreases in most bias measures (-73% to -37%), while training on UTKFace gives smaller reductions (-35% to -3%). The FairFace training subset is $\sim 4\times$ larger than UTKFace which may explain the larger bias reductions from training on it. The FairFace-trained model when evaluated on UTKFace shows an increase in *NDKL* and a decrease in *MaxSkew*, which may stem from the relatively lower diversity of facial expressions in UTKFace [37]. Debiasing on FairFace thus seems to generalize better, but more work is needed on additional datasets to confirm this.

Table 4b shows the percentage change in bias measures when training the same debiasing protocol with FairFace for gender attributes, then evaluating on FairFace with race attributes. The bias reductions on race (-45% to -40%) are slightly lower than the reduction on gender (-79% to -69%) but still of significant magnitude, demonstrating that debiasing on one attribute class can result in strong debiasing of other classes. Even though FairFace is well-balanced across gender, race, and their intersection, racial bias in the pretrained baseline is more than twice the gender bias (on both *MaxSkew* and *NDKL*). Given the larger prevalence of face image datasets with gender annotations, it is encouraging that debiasing on gender also reduces racial bias but further research is needed into cross-attribute debiasing generalization.

Debiasing effects on harmful image classification. We replicate [3]'s protocol for evaluating racial bias in CLIP by assessing misclassification rates of images from FairFace's 7 ethnicities with criminal and non-human categories. Table 5 shows the results directly taken from [3] alongside results from our implementation with the pretrained baseline CLIP ViT$_{B/16}$.[3] Our gender-debiased model trained on FairFace has a lower misclassification rate into crime-related

[2] Note that training and train-time evaluation on FairFace is on the training subset of FairFace, and testing is on its validation subset, while all measures for UTKFace are on the whole of UTKFace.

[3] Note that [3] do not describe their experimental protocol in detail, and no public implementation is available so the approach had to be inferred, which may explain the difference between baselines (see subsection 3.3)

Table 5: Results for harmful misclassification rate of FairFace validation images into criminal and non-human categories, by FairFace ethnicity group. We compare between the CLIP Audit paper [3], a baseline CLIP model, and a CLIP model with debiasing trained on FairFace gender attributes using learned prompt token embeddings.

Category	Model	Debiased	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related	CLIP Audit [3]	✗	16.4	24.9	24.4	10.8	19.7	4.4	1.3
	CLIP ViT$_{B/16}$	✗	3.0	26.9	2.7	4.8	8.8	0.5	0.5
	CLIP ViT$_{B/16}$	✓	1.7	14.9	0.1	1.7	4.5	0.4	0.3
Non-human	CLIP Audit [3]	✗	14.4	5.5	7.6	3.7	2.0	1.9	0
	CLIP ViT$_{B/16}$	✗	0.2	0.2	0.0	0.0	0.1	0.0	0.1
	CLIP ViT$_{B/16}$	✓	0.8	0.8	0.0	0.1	0.5	0.0	0.1

classes than the pretrained baseline. While the non-human misclassification rate was marginally higher than baseline, the absolute rates are still comparable and very low (<1%). For all ethnicities with misclassification rates greater than 1% from the pretrained baseline, our debiased model reduces the rate by half or more (-43% to -96%).

Qualitative Debiasing Results. In Figure 2 we show the top-10 ranking images for the text query: "A photo of a smart person." We see that before debiasing, CLIP produces a highly skewed top-10 distribution with only 10% of the individuals returned being female, showing a bias towards deeming male faces as semantically more similar to the concept *smart*. After debiasing with our proposed method the resulting top-10 is much more balanced with 50% of the individuals returned being female. Qualitatively, we also see a greater balance in ages with children also being represented in the top-10 too rather than just middle-aged adults.

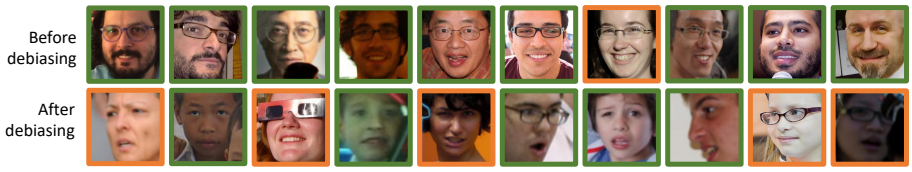

Fig. 2: Top-10 ranking images for the text query "**A photo of a smart person**", before and after debiasing CLIP ViT-B/16, from the FairFace validation set, labeled with male and female. The gender distribution of the top-10 images shifts from 90%/10% to 50%/50% (male/female) after debiasing, achieving demographic parity.

6 Conclusion

To conclude, this paper establishes a framework for measuring bias in vision-language models, demonstrating that ranking metrics (specifically *MaxSkew* and *NDKL*) are effective measures. We report these metrics on a range of pretrained vision-language models for gender and racial bias in photos of faces, finding (i) a correlation between lower bias with more pretraining data, and (ii) reductions in bias for models trained additionally with SSL. Further, we propose an adversarial debiasing method of these models via learnable "debiasing" tokens using for supervision only publicly-available face image datasets with attribute labels. The proposed method demonstrates a substantial reduction over a suite of bias metrics for gender and race attributes, with only a moderate amount of feature degradation. Future work could include (1) debiasing during the pretraining stage, with SSL showing a promising avenue in that regard, (2) utilizing a small sample of the pretraining data to help minimize reduction in downstream performance, or (3) defining a wider diversity of attributes such as removing the harmful assumption of binary gender, or considering intersectional biases. We encourage researchers in vision-language to continue to investigate bias in their models, be transparent on model cards using metrics like those proposed in this paper, and apply relatively cheap and easy debiasing protocols like ours. Our code, models and debiasing tokens are publicly available[4] for the community to use in the hope that progress can be made towards the safer and fairer use of this technology in society.

7 Acknowledgements

The authors would like to thank Mohamed Baioumy, Vit Ruzicka and Laura Weidinger for their helpful feedback. This work has been supported by the Oxford Artificial Intelligence student society, the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems [EP/S024050/1] (Y.B., A.S.), and the Economic and Social Research Council grant for Digital Social Science [ES/P000649/1](H.R.K.).

[4] github.com/oxai/debias-vision-lang

Bibliography

[1] Evertrove - The Semantic Image API, howpublished = https://evertrove.co/, note = Accessed: 2022-03-05 2, 6

[2] Hugging Face Inference API, howpublished = https://huggingface.co/inference-api, note = Accessed: 2022-03-05 2

[3] Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J.W., Brundage, M.: Evaluating clip: towards characterization of broader capabilities and downstream implications. arXiv preprint arXiv:2108.02818 (2021) 2, 3, 4, 7, 13, 14

[4] Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 3

[5] Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021) 2, 3, 8, 10, 11

[6] Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963 (2021) 2, 3, 4

[7] Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems **29** (2016) 3

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2

[9] Brunet, M.E., Alkalay-Houlihan, C., Anderson, A., Zemel, R.: Understanding the origins of bias in word embeddings. In: International conference on machine learning. pp. 803–811. PMLR (2019) 3

[10] Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018) 3

[11] Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017) 3, 5, 6, 22

[12] Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021) 2

[13] Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2020) 2

[14] Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020) 2

[15] De Vries, T., Misra, I., Wang, C., Van der Maaten, L.: Does object recognition work for everyone? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 52–59 (2019) 2

[16] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 2

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 10

[18] Edwards, H., Storkey, A.: Censoring representations with an adversary. arXiv preprint arXiv:1511.05897 (2015) 3, 8

[19] Elazar, Y., Goldberg, Y.: Adversarial removal of demographic attributes from text data. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 11–21 (2018) 8

[20] Fussell, S.: An algorithm that 'predicts' criminality based on a face sparks a furor (2020), https://www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/#:~:text=With\OT1\textquotedblleft80percentaccuracyand,deletedfromtheuniversitywebsite 2

[21] Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences **115**(16), E3635–E3644 (2018) 3

[22] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021) 3

[23] Geyik, S.C., Ambler, S., Kenthapadi, K.: Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In: Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining. pp. 2221–2231 (2019) 4, 6, 7

[24] Gonen, H., Goldberg, Y.: Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 609–614 (2019) 3

[25] Goyal, P., Duval, Q., Seessel, I., Caron, M., Singh, M., Misra, I., Sagun, L., Joulin, A., Bojanowski, P.: Vision models are more robust and fair when pretrained on uncurated images without supervision. arXiv preprint arXiv:2202.08360 (2022) 11

[26] Greenwald, A.G., McGhee, D.E., Schwartz, J.L.: Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology **74**(6), 1464 (1998) 3

[27] Grover, A., Song, J., Kapoor, A., Tran, K., Agarwal, A., Horvitz, E.J., Ermon, S.: Bias correction of learned generative models using likelihood-free importance weighting. Advances in neural information processing systems **32** (2019) 4

[28] Guo, W., Caliskan, A.: Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 122–133 (2021) 3

[29] Haraway, D.: The Haraway Reader, vol. 53. Routledge (2004) 2

[30] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 10

[31] Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 771–787 (2018) 3, 4

[32] Hu, X., Wang, H., Dube, S., Vegesana, A., Yu, K., Lu, Y.H., Yin, M.: Discovering biases in image datasets with the crowd pp. 2015–2017 (2018) 3

[33] Hu, X., Wang, H., Vegesana, A., Dube, S., Yu, K., Kao, G., Chen, S.H., Lu, Y.H., Thiruvathukal, G.K., Yin, M.: Crowdsourcing detection of sampling biases in image datasets. In: Proceedings of The Web Conference 2020. pp. 2955–2961 (2020) 3

[34] Jo, E.S., Gebru, T.: Lessons from archives: Strategies for collecting sociocultural data in machine learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 306–316 (2020) 3

[35] Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. arXiv preprint arXiv:2112.04478 (2021) 4, 9

[36] Kaneko, M., Bollegala, D.: Debiasing pre-trained contextualised embeddings. Proc. of the 16th European Chapter of the Association for Computational Linguistics (EACL) (2021) 12

[37] Kärkkäinen, K., Joo UCLA, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2021) 3, 9, 13

[38] Kay, M., Matuszek, C., Munson, S.A.: Unequal representation and gender stereotypes in image search results for occupations. Conference on Human Factors in Computing Systems (2015) 2

[39] Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2015) 10

[40] Kirk, H.R., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F.A., Shtedritski, A., Asano, Y.M.: Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. Advances in neural information processing systems (2021) 2, 3

[41] Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images (2009) 10, 11

[42] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586 (2021) 4
[43] Manzini, T., Lim, Y.C., Tsvetkov, Y., Black, A.W.: Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In: NAACL (2019) 3
[44] May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. pp. 622–628. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, Association for Computational Linguistics (ACL) (2019) 3, 5, 10
[45] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys **54** (2021) 3
[46] Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019) 2
[47] Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. arXiv preprint arXiv:2112.12750 (2021) 3, 8, 10, 11
[48] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Askell, C.Z.S.A.K.S.A.R.J.S.J.H.F.K.L.M.M.S.A., Christiano, P.W.P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. ACL (2020) 3
[49] Park, J.S., Bernstein, M.S., Brewer, R.N., Kamar, E., Morris, M.R.: Understanding the Representation and Representativeness of Age in AI Data Sets, p. 834–842. Association for Computing Machinery, New York, NY, USA (2021) 3
[50] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 2, 3, 8, 10, 11
[51] Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D.: No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536 (2017) 2
[52] Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of ACL (2018) 11
[53] Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020) 4
[54] Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning, p. 2443–2449. Association for Computing Machinery, New York, NY, USA (2021) 11

[55] Steed, R., Caliskan, A.: Image representations learned with unsupervised pre-training contain human-like biases. FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency **1**, 701–713 (2021) 3, 5

[56] Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016) 9, 11

[57] Wadsworth, C., Vera, F., Piech, C.: Achieving fairness through adversarial learning: an application to recidivism prediction. arXiv preprint arXiv:1807.00199 (2018) 8

[58] Wang, A., Narayanan, A., Russakovsky, O.: Revise: A tool for measuring and mitigating bias in visual datasets. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12348 LNCS**, 733–751 (2020) 3

[59] Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021) 4, 9

[60] Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. Tech. rep. (2018) 4

[61] Wang, Y., Kosinski, M.: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of Personality and Social Psychology **114**, 246–257 (02 2018) 2

[62] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W., Legassick, S., Irving, G., Gabriel, I.: Ethical and social risks of harm from Language Models. arXiv preprint arXiv:2112.04359 (2021) 2, 3, 4, 6

[63] Wilson, B., Hoffman, J., Morgenstern, J.: Predictive inequity in object detection. arXiv preprint arXiv:1902.11097 (2019) 3

[64] Wu, X., Zhang, X.: Automated inference on criminality using face images. ArXiv **abs/1611.04135** (2016) 2

[65] Xu, H., Liu, X., Li, Y., Jain, A., Tang, J.: To be robust or to be fair: Towards fairness in adversarial training. In: International Conference on Machine Learning. pp. 11492–11501. PMLR (2021) 8

[66] Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management. SSDBM '17, Association for Computing Machinery, New York, NY, USA (2017) 6

[67] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association of Computational Linguistics **2**, 67–78 (2014) 10, 11

[68] Zhai, A., Wu, H.Y.: Classification is a strong baseline for deep metric learning. arXiv preprint arXiv:1811.12649 (2018) 2

[69] Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning (2021) 4, 8
[70] Zhang, Zhifei, Song, Yang, Qi, Hairong: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017) 3, 9
[71] Zhao, D., Wang, A., Russakovsky, O.: Understanding and evaluating racial biases in image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14830–14840 (2021) 3, 4
[72] Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.W.: Gender bias in contextualized word embeddings. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **1**, 629–634 (2019) 3
[73] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457 (2017) 3
[74] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to Prompt for Vision-Language Models. arXiv preprint arXiv:2109.01134 (sep 2021) 4, 9
[75] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022) 4
[76] Zhu, X., Zhu, J., Li, H., Wu, X., Wang, X., Li, H., Wang, X., Dai, J.: Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. arXiv preprint arXiv:2112.01522 (2021) 4
[77] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**, 43–76 (2021) 8

Appendix

A Measuring bias across different model architectures, datasets, and syntactic changes.

In Figure 3 we report the defined bias measures (*WEAT*, *NDKL* and *MaxSkew*) across changes in vision-language model encoders, datasets and minor syntactic changes to the text queries T.

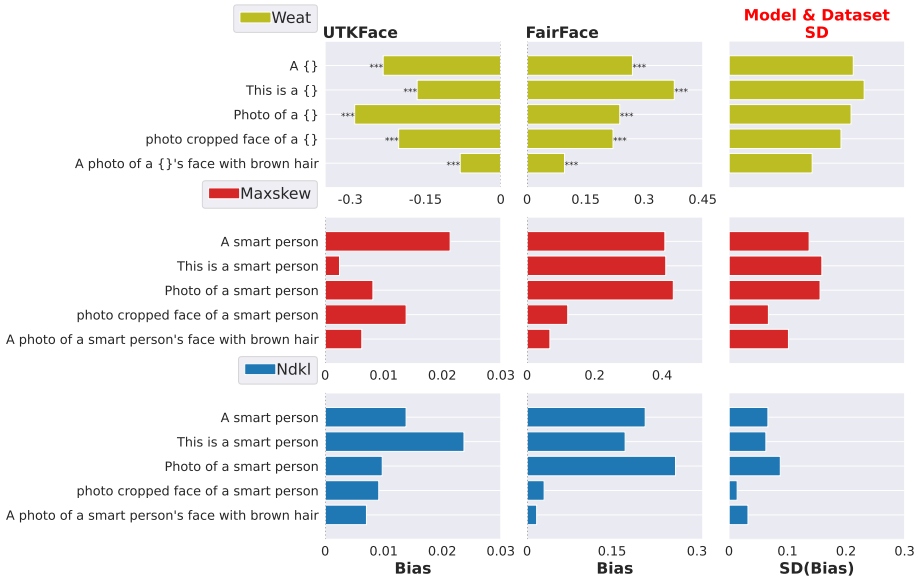

Fig. 3: Bias measures across different combinations of minor syntactic changes, models (RN50, ViT$_{B/16}$, ViT$_{B/32}$), and datasets (FairFace validation set and UTKFace). Bias is measured for gender, and we use the *WEAT* pairwise adjectives concept sets from [11]. Standard deviation of bias measurement is taken over all combinations of model architecture and datasets, for other results we use ViT$_{B/32}$

Since *WEAT* uses a template to fill in with concepts, it is not directly comparable to the text queries used in *NDKL* and *MaxSkew*. We report these results only to illustrate the high variance of bias measurement results over small changes in syntax of templates, model architecture and dataset.

We note that *WEAT* measured on UTKFace has opposing sign to *WEAT* measured on FairFace. Furthermore, with small syntactic changes in template, *WEAT* produced both positive and negative results on both FairFace and UTKFace.

B Performance effects of learnable text token initialization

In Table 6 we show the effects on zero-shot performance when adding zero-initialized text tokens to the text queries, before any debias training has occurred. We note there is a substantial drop in performance in both Flickr image retrieval and CIFAR image classification, with the drop increasing with the number of tokens added in both the prepending and appending settings. This suggests that the reduced ZS performance of the debias model is not due to the adversarial learning but rather the learnable text tokens which shift the distribution of the text query. Future work could investigate mitigating this effect.

Table 6: Results showing effect of prepending or appending with zero-pad initialized text tokens on zero-shot text-to-image retrieval and image classification.

Token Pos.	**#tokens**	flickr$_{R@5}$	CIFAR$_{acc}$
Prepend	0	85.9	66.5
	1	78.3	57.5
	2	70.1	59.4
	3	64.5	58.5
Append	0	85.9	66.5
	1	68.6	56.9
	2	68.7	58.5
	3	57.0	54.7