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

Hugo Berg[*], Siobhan Mackenzie Hall, Yash Bhalgat,
Hannah Rose Kirk, Aleksandar Shtedritski, 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 multimodal harms due to lacking measurement robustness and feature degradation. To address these challenges, we investigate bias measures and apply ranking metrics for image-text representations. We then investigate debiasing methods and show that prepending learned embeddings to text queries that are jointly trained with adversarial debiasing and a contrastive loss reduces various bias measures with minimal degradation to the image-text representation.

1 Introduction

Large-scale, pretrained vision-language (VL) models are growing in popularity due to their impressive performance on downstream tasks with minimal finetuning. Their success can be attributed to three main advances: the rise of transformers in natural language processing (NLP) (Devlin et al., 2018), cross-modal contrastive learning (Zhai and Wu, 2018) and the availability of large multimodal web datasets (Changpinyo et al., 2021). These models, including CLIP (Radford et al., 2021), are readily available through APIs (Evertrove; HuggingFace), allowing non-technical users to capitalize on their high performance 'out of the box' on zero-shot tasks (Kirk et al., 2021).

Despite these benefits, an expansion in scope for downstream applications comes with greater risk of perpetuating damaging biases that the models learn during pretraining on web-scraped datasets which are too large to be manually audited for quality (Birhane et al., 2021). Cultural and temporal specificity is also of concern given models are trained on a snapshot in space and time (Haraway, 2004), thus reinforcing negative stereotypes that may otherwise naturally alter through societal pressures and norm change.

The risk and type of societal harm intimately interacts with the downstream task at hand. Clearly, using VL models for dog-species classification poses very different dangers to projecting the similarity of human faces onto axes of criminality (Wu and Zhang, 2016; Fussell, 2020) or homosexuality (Wang and Kosinski, 2018). Applications of this kind are extremely hard to ethically motivate and there may be no appropriate use case that justifies their associated risks. 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; for instance, the over-representation of one gender when querying for a profession (e.g., "nurse" versus "doctor") or one ethnicity in explicit and NSFW content (Birhane et al., 2021). Allocational harms arise when an individual's or group's access to resources and opportunity are differentially impacted (Weidinger et al., 2021); for instance, if the ordering of images in search results shifts recruiters' perceptions about the real-world suitability of different peoples for different jobs (Kay et al., 2015).

In this paper, we focus on the risk of representational harms when large-scale VL models are used to map sensitive text queries, such as "a photo of a criminal" onto face datasets. While frameworks to measure bias have been established for NLP and computer vision (CV) separately, there is considerably less work on VL (Agarwal et al., 2021). Appropriate debiasing techniques for large-scale VL models are also sparse and face challenges from 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 VL models, we need both effective measures of bias as well as efficient methods of debiasing. To this end, we make three contributions: (i) we investigate and evaluate different measures of bias for VL models,

[*] Corresponding author: hugo@hugob.se

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 806–822
November 20–23, 2022. ©2022 Association for Computational Linguistics

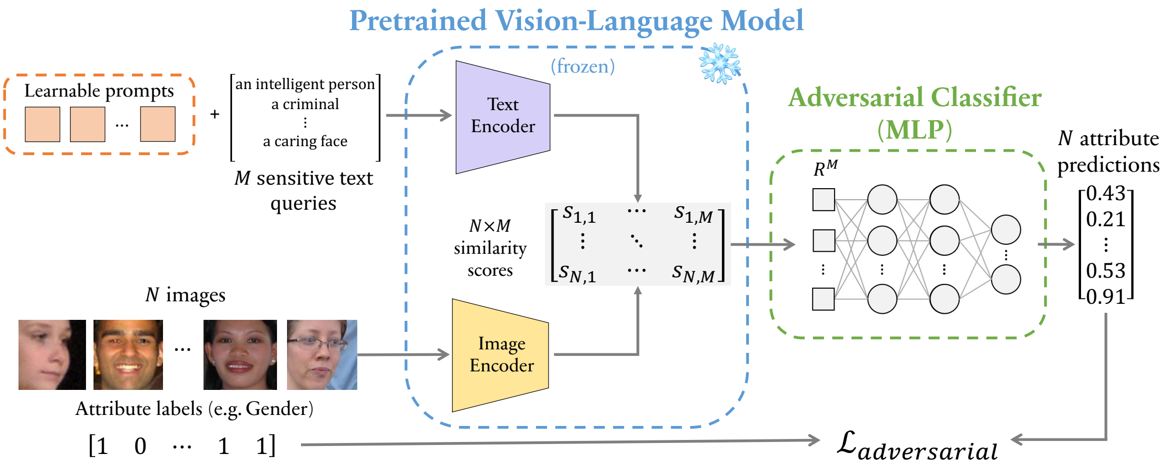
Pretrained Vision-Language Model
(frozen)
Learnable prompts
···
+
an intelligent person
a criminal
⋮
a caring face
M sensitive text queries
Text Encoder
$N \times M$ similarity scores
$s_{1,1}$ ··· $s_{1,M}$
$s_{N,1}$ ··· $s_{N,M}$
N images
···
Image Encoder
Attribute labels (e.g. Gender)
[1 0 ··· 1 1]
Adversarial Classifier (MLP)
R^M
N attribute predictions
0.43
0.21
⋮
0.53
0.91
$\mathcal{L}_{adversarial}$

Figure 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 image-text similarity scores and attribute labels are reduced whilst preventing significant degradation of the joint image-text representation. Additionally, we jointly train with a contrastive loss on generic image-text pairs to further avoid degradation of the joint representation (not shown for clarity).

showing that some measures, such as *WEAT*, are inappropriate; (ii) we evaluate gender and racial bias in state-of-the-art VL models on two face datasets: FairFace (Kärkkäinen and Joo, 2021) and UTKFace (Zhang et al., 2017); and (iii) we provide a framework for debiasing VL models (see Fig. 1), requiring only sensitive attribute labels of images as supervision, and show that jointly optimizing for unbiasedness and image-text contrastive (ITC) losses via an array of learnable tokens prepended to text embeddings is the best strategy for mitigating bias without substantially degrading the quality of the image-text representation.

2 Defining and Measuring Bias

2.1 Problem Statement

We consider the problem of learning unbiased joint text-image representations. We first establish a framework for measuring the degree of bias in these representations. Consider 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 (see Sec. 2.2). Specifically, we consider the case where Ψ is initialized as a pretrained model that already achieves (i) but not (ii) – as is the case with current pretrained VL 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 this scenario.

2.2 Sensitive Attributes and Relevancy

Some statistical associations between demographic groups and text queries are required for accurate and relevant text-image pairing in VL models. This is especially true with historical or contextual associations; for instance, the expected over-representation of men in the query '19th century dockworker' or various minoritized groups in '1960s civil rights marches'. However, our framework assumes there is a reasonably concrete normative view that there exists a set of 'neutral' text queries like "a good/bad person" which hypothetically should be independent of demographic categories. This aligns with a notion of statistical parity (Dwork et al., 2012), where maintaining high-quality feature representations alongside debiasing specifically relates to *conditional* statistical parity (Corbett-Davies et al., 2017). Under this treatment of fairness, some associations with a sensitive attribute are legitimate and explainable, while others are illegitimate and unjust (Makhlouf et al., 2021). While this assumption underpins existing bias evaluations such as the Implicit Association Test (Greenwald et al., 1998), it is necessarily a simplification and does not resolve deep tensions in ontology and normative ethics, including questions over what sensitive attributes are relevant, what a 'legitimate' association is or what a fair society should look like. These issues require ongoing, multi-disciplinary and multi-stakeholder discussions. We demonstrate a method for measuring and debiasing associations between a set of text prompts and demographic attribute labels but the specification of the prompts and sensitive attributes can and should be adapted to the context and culture under which the VL model is applied and how the downstream task is defined.

2.3 Bias Metrics

WEAT. We first investigate the suitability of the Word Embedding Association Test (*WEAT*) (Caliskan et al., 2017) for measuring bias in VL models. *WEAT* is derived from the Implicit Association Test (IAT) (Greenwald et al., 1998) which measures the time-delay that human subjects take in associating a given demographic group with positive or negative descriptors. *WEAT* is used to measure the bias of word and sentence embeddings (Caliskan et al., 2017; May et al., 2019), and more recently has been adapted to evaluate the the bias of vision encoders (Steed and Caliskan, 2021). The mathematical implementation of *WEAT* for the VL setting is described in App. A.

ranking metrics. We also apply bias measures from the information retrieval literature (Geyik et al., 2019; Yang and Stoyanovich, 2017) to the setting of text-image retrieval. This is a natural application given that VL models are increasingly used for semantic image search, introducing biases from the attributes which get ranked higher than others in the top k results. We describe the mathematical implementation of these metrics, namely *Skew*, *MaxSkew* and Normalized Discounted Cumulative KL-Divergence (*NDKL*) in App. B.

harmful zero-shot image misclassification. Agarwal et al. (2021) propose using the zero-shot misclassification rates of people into derogatory criminal and non-human categories. Implementation details for zero-shot image classification experiments are described in App. G.

3 Debiasing

The proposed debiasing method has two components: (i) the objective function to minimize for bias reduction; and (ii) the choice of parameters to optimize over in the VL model Ψ to minimize (i).

3.1 Fairness Objective with Adversarial Debiasing

We follow a common approach in bias mitigation (Edwards and Storkey, 2015; Elazar and Goldberg, 2018; Xu et al., 2021) 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 $\mathcal{T}$

$$\hat{A} = \theta_{\text{adv}}(S) \tag{1}$$

where $S = [s_1, ..., s_M] \in \mathbb{R}^M$ and $s_m = \Psi(I, T_m)$. 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

$$\mathcal{L}_{\text{adv}} = -\sum_{A \in \mathcal{A}} A \log \theta_{\text{adv}}(S). \tag{2}$$

In this work, we define an unbiased representation as being blind to the sensitive attributes over the set of 'neutral' text queries so optimize the VL model to maximize this adversarial loss.

3.2 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. In this case, the feature representation loses all semantic information of the input, making it effectively useless for downstream tasks. We thus investigate regularization techniques (discussed below) that restrict the set of parameters in the image-text model Ψ which can be optimized over, as well as joint training of debiasing and image-text similarity objectives.

finetuning depth. Instead of optimizing all model parameters, a common regularizing adaption technique is to finetune the layers in the image-text encoders to a certain depth (Zhuang et al., 2021). We instantiate Ψ as a dual stream encoder (Radford et al., 2021; Mu et al., 2021), with text and image embeddings encoded via independent streams,

Table 1: **Templates and 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

$s = \Psi(x, y)$ where $\Psi(x, y) = \Psi_i(x)^T \Psi_t(y)$, and choose different finetuning depths for each encoder $\Psi_i(x)$, Ψ_t, noting that Zhai et al. (2021) show finetuning only the text encoder Ψ_t improves generalization and reduces catastrophic forgetting of the original pretrained 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 (Zhou et al., 2021; Wang et al., 2021b). 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 (Zhou et al., 2021), we prepend learnable text tokens to the text queries after they have been embedded by the token embedding layer (see App. F).

joint training with image-text similarity. To debias the model without losing strong image-text similarity performance, we add an auxiliary image-text contrastive (ITC) loss which is computed from batches of image-text pairs. ITC loss is used to train various VL models, including CLIP (Radford et al., 2021), however, this can be substituted with any image-text matching loss.

$$\mathcal{L} = \mathcal{L}_{\text{adv}} + \lambda \mathcal{L}_{\text{itc}} \tag{3}$$

4 Experiments

4.1 Datasets

The original IAT literature, from which this work draws inspiration, relies on the association between faces of different demographics and text attributes for measuring bias. We also use two commonly-used face datasets as a comparable baseline for the novel application of these these principles to the VL subdomain but discuss limitations in Sec. 6. **FairFace (Kärkkäinen and Joo, 2021)** consists of 108,501 images of GAN-generated faces. This dataset has emphasis on a balanced composition by age, gender and ethnicity. The ethnicities 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 (Thomee et al., 2016). **UTKFace cropped image dataset (Zhang et al., 2017)** contains 20,000 images with 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 has different characteristics to FairFace, in terms of variance in lighting conditions, color quality and angle of portraits.

4.2 Experimental Protocol

text query generation. We select pairwise adjectives from the IAT dataset.[1] We use pairs of words which are uncorrelated with facial expressions or sensitive attributes, e.g., not "happy/sad" or "beautiful/handsome" (see Tab. 1). We expand the test set with unseen templates and concepts to assess generalizability. In order to produce single bias measures, we aggregate across text queries using the arithmetic mean over all templates.

bias metrics. Of the metrics defined in Sec. 2.3, we find that the effect size of *WEAT* is overly sensitive to changes in model architecture, evaluation dataset, as well as minor syntactic changes in text queries (see App. C). *MaxSkew@k* with $k = 1000$ and *NDKL* were found to be more robust measures so are used in the following experiments. Additional results for harmful zero-shot misclassification are presented in App. G.

downstream performance metrics. We report the zero-shot (ZS) performance on (i) flickr$_{R@5}$: recall@5 text-to-image retrieval on the Flickr-1k

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

test set (Young et al., 2014) and (ii) IN1K$_{acc}$: image classification accuracy on the ImageNet-1k val set (Deng et al., 2009). For ablative experiments, we report CIFAR$_{acc}$: image classification accuracy on the CIFAR100 (Krizhevsky, 2009) test set.

pretrained models. CLIP (Radford et al., 2021) combines a text and image encoder whose representations are projected to the same space. CLIP was originally trained with a contrastive loss on 400M image-text pairs from the web. We experiment over variants with different image encoders: ResNet50 (He et al., 2016), ViT (Dosovitskiy et al., 2020), SLIP (Mu et al., 2021) and FiT (Bain et al., 2021).

debiasing implementation. For debiasing, we use CLIP ViT$_{B/16}$ and prepend 2 learnable prompt embeddings to the text query, as well as jointly training with an ITC loss. Further implementation details are in App. F.

debiasing baseline. We further compare our debiasing method to a simple baseline, CLIP-clip (Wang et al., 2021a), which performs feature selection on CLIP embeddings by removing the dimensions with the highest mutual information to the sensitive attribute labels of the images. The feature selection is computed on the training set and evaluated on the test set with clipping done on both the image and text embeddings.

4.3 Results

bias across model architectures and pretraining. The results in Tab. 2 indicate that higher feature quality comes from (i) models pretrained on larger datasets, and (ii) models with larger image encoders (RN50 < ViT$_{B/32}$ < ViT$_{B/16}$ < 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 higher quality datasets than YFCC15M. Increased pretraining dataset size decreases bias (both *MaxSkew* and *NDKL*). The SLIP ViT$_{B/16}$ and ViT$_{L/14}$ models trained with SSL have lower *MaxSkew* than their non-SSL counterparts, confirming the finding of Goyal et al. (2022). The best models (by feature quality) pretrained on WIT (Srinivasan et al., 2021) and YFCC100M (Thomee et al., 2016) also have low bias for their respective datasets, suggesting minimal trade-off between feature quality and model bias.

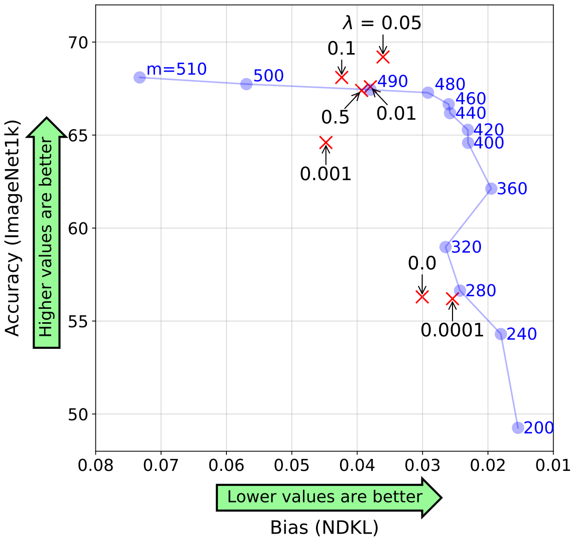

Figure 2: **The bias (*NDKL*) vs performance (*IN1K*$_{acc}$) trade-off** of our debiased models with varied ITC loss weights λ (in red) and CLIP-clip using different numbers of removed dimensions m (in blue).

effectiveness of debiasing approaches. During adversarial debiasing, we tried adding an ℓ_2 loss (Kaneko and Bollegala, 2021) between the original model embeddings and debiased model embeddings. However, finetuning in this setting did not reduce bias nor increase feature quality. To prevent the pretrained model's feature quality from degrading due to the adversarial loss, we use joint training with an ITC loss on FairFace30K (train). The results of ablation over debiasing approaches (see Tab. 3) show that while pure adversarial loss significantly reduces the bias metrics (-69% to -80%), it also reduces feature quality by up to 25%. Training only with the ITC loss shows small increase in both feature quality (0% to 5%) and bias metrics (0% to 6%). It is only when training jointly with adversarial and ITC loss that bias metrics are significantly reduced (-52% to -65%) with feature quality either improving or staying relatively unchanged (+3% to -1%) compared to the baseline. Debiasing with different ITC loss weights (λ) allows us to explore the bias-accuracy tradeoff in our framework, and we compare our results to the results of clip-clip with different numbers of cutoff dimensions (m) in Fig. 2. For $\lambda^* = 0.05$, our joint training method outperforms CLIP-clip in downstream performance for all values of m. For low values of $\lambda \leq 0.0001$, our method lies within the pareto-frontier of CLIP-clip. However, operating on this part of the curve is undesirable given that accuracy drops to 55%. There are additional benefits of our method: CLIP-clip applies heuristic feature clipping so necessarily loses more information than just gender information in debiasing

Table 2: **Evaluation of gender bias on the FairFace validation set for various model architectures** (arch.) and pretraining datasets. We evaluate: CLIP (Radford et al., 2021) models trained on the *WIT* dataset; SLIP (Mu et al., 2021) models trained on *YFCC* 15M with and without self-supervised learning (SSL); FiT (Bain et al., 2021) models trained on *CC* (Sharma et al., 2018) and *WV* (WebVid) (Bain et al., 2021).

Pretrain Dataset	**Pretrain Size**	**Arch.**	**Bias**↓		**Performance**↑	
			MaxSkew@1000	*NDKL*	*flickr*$_{R@5}$	*IN1K*$_{acc}$
WIT	400M	RN50	**0.197**	**0.075**	83.7	59.1
		ViT$_{B/32}$	0.185	0.073	83.6	62.7
		ViT$_{B/16}$	0.233	0.103	86.1	68.1
		ViT$_{L/14}$	0.202	0.083	**87.4**	**74.1**
YFCC	15M	ViT$_{B/16}$	0.259	0.115	60.1	35.6
		ViT$^{\text{SSL}}_{B/16}$	0.231	0.117	68.7	40.8
		ViT$_{L/14}$	0.255	0.112	61.6	39.0
		ViT$^{\text{SSL}}_{L/14}$	**0.206**	**0.066**	**69.3**	**46.7**
CC,WV	5.6M	FiT$_{B/16}$	0.292	0.174	76.3	42.8

Table 3: **Measuring effect on gender bias and performance** of prepending prompt tokens; adversarial debiasing on FairFace; and ITC training on Flickr30k-train. Showing CLIP (Radford et al., 2021) and CLIP-clip (Wang et al., 2021a), where m denotes the remaining number of un-clipped feature dimensions, where $m = 512$ is the original dimension size of ViT-B/16.

Model	**Bias**↓		**Performance**↑	
	$MaxSkew@1K$	*NDKL*	*flickr*$_{R@5}$	*IN1K*$_{acc}$
CLIP	0.233	0.104	85.9	68.1
CLIP-clip ($m = 490$)	0.122(-48%)	0.038(-45%)	82.6(-4%)	67.4(-1%)
CLIP-clip ($m = 400$)	0.073(-69%)	0.023(-78%)	78.5(-9%)	64.6(-5%)
CLIP-clip ($m = 256$)	**0.056(-76%)**	0.023(-78%)	63.7(-26%)	55.8(-18%)
CLIP$_{+prompt}$ (debias)	0.073(-69%)	**0.021(-80%)**	64.2(-25%)	54.9(-19%)
CLIP$_{+prompt}$ (itc)	0.247(+6%)	0.104(+0%)	**90.6(+5%)**	**68.4(+0%)**
CLIP$_{+prompt}$ (debias+itc)	0.113(-52%)	0.036(-65%)	88.5(+3%)	67.6(-1%)

because no single dimension of the feature vectors is dedicated to gender information. Therefore, it is of interest to have an effective debiasing method like ours that keeps all dimensions of the image-text embeddings.

We further evaluate adversarial debiasing when training different parts of the model, as well as pure prompt learning (see App. H). The best bias results are achieved early on for all techniques in Tab. 3, and reach their optimum within 3 epochs, so our method is relatively computationally cheap ($\sim$ 3 hrs per training run on 1 GPU). We note that for models with separate image and text encoders (all VL models in this paper), 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 with adversarial loss for gender attributes on FairFace then evaluating on UTKFace (and vice-versa).[2] Training on FairFace shows larger reductions in bias metrics (-73% to -37%), than training on UTKFace (-35% to -3%). The FairFace training subset is $\sim 4\times$ larger than UTKFace which may explain the difference in reductions. When the FairFace-trained model is evaluated on UTKFace, *NDKL* is increased and *MaxSkew* is decreased, possibly due to lower diversity of facial expressions in UTKFace (Kärkkäinen and Joo, 2021). Thus, debiasing on FairFace appears to generalize better, but more work is needed to confirm this.

Next, we evaluate the change in bias measures when training the same debiasing protocol with FairFace for gender attributes, then evaluating on FairFace with race attributes (see Tab. 4b). The bias reduction on race (-45% to -40%) are lower than the reduction on gender (-79% to -69%) but still of significant magnitude, demonstrating that debiasing on one attribute class can result in 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 greater prevalence of face image datasets with gender annotations, it is encouraging that debiasing on gender also reduces

[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.

Table 4: **Generalization of debiasing results** from 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-Dataset

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

racial bias but further research is needed into cross-attribute debiasing generalization.

qualitative debiasing results. In Fig. 3, we present the top-5 ranked images for the text query: "A photo of a smart person.". Before debiasing, CLIP produces a highly skewed distribution towards male faces. After debiasing, the images are more balanced by gender and age.

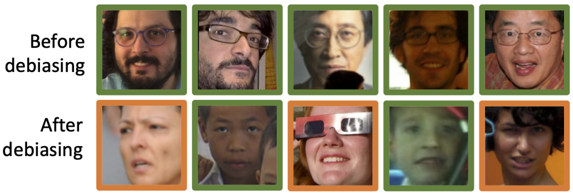

Figure 3: **Effect of debiasing CLIP ViT-B/16 by ranked images with concept of "smart"** from the FairFace validation set, labeled with male and female.

5 Related Works

There have been multiple recent releases of open-source VL models (Radford et al., 2021; Mu et al., 2021; Bain et al., 2021), but research into bias measurement and mitigation has not kept pace, with only a few papers to date tackling these topics for VL (Agarwal et al., 2021; Zhao et al., 2021; Wang et al., 2021a). In this work, we therefore drew inspiration from the literature on dataset- and model-level bias in CV and NLP (Mehrabi et al., 2021).

bias in NLP. Large-scale language models are optimized to reflect statistical patterns of human language, which can be problematic if training datasets contain harmful or misrepresentative language (Weidinger et al., 2021). Prior work has documented gender bias (Bolukbasi et al., 2016; Zhao et al., 2019; Borchers et al., 2022), racial bias (Manzini et al., 2019; Garg et al., 2018) and their intersections (Guo and Caliskan, 2021; Kirk et al., 2021). *WEAT*, as described in Sec. 2.3 is one commonly-deployed bias metric for word-embeddings (Caliskan et al., 2017; Bolukbasi et al., 2016; Manzini et al., 2019). However as Gonen and Goldberg (2019) criticize, gender bias remains in the distances between "gender neutralised" words; thus we did not pursue embedding-level debiasing as a viable method in our work. Zhao et al. (2019) and Brunet et al. (2019) propose dataset-level debiasing techniques through data augmentation and perturbation, and Ouyang et al. (2020) implement supervised finetuning on data checked by humans. While promising, these techniques were not feasible with the large-scale, pretrained VL models under investigation in our work due to the required computational resources and lack of access to the original dataset.

bias in computer vision. Similar to the body of NLP evidence, CV investigations have also shown evidence of gender bias (Zhao et al., 2017), racial bias (Wilson et al., 2019), and their intersection (Buolamwini and Gebru, 2018; Steed and Caliskan, 2021). Though not the focus of our paper, bias stemming from dataset creation practices have been widely documented (Hu et al., 2018, 2020; Park et al., 2021; Gebru et al., 2021; Wang et al., 2020; Birhane et al., 2021). Model-based debiasing methods are more similar to our work, these include optimizing confusion (Alvi et al., 2018), domain adversarial training (Edwards and Storkey, 2015), or training a network to *unlearn* bias information (Grover et al., 2019). We adopted the idea of adversarial finetuning in our work because, as well as being effective, it is computationally cheap and does not require access to the original dataset.

bias in vision-language. Some work measures bias in VL representations. The authors of the original CLIP paper investigated manifestations of bias within their own model (Agarwal et al., 2021) by assessing the misclassification of faces by age or race with non-human and criminal categories. Wang et al. (2021a) proposes a simple debiasing method via feature engineering by removing the dimensions in CLIP embeddings most associated with gender bias, however this guarantees feature degradation due to significant information loss. The sparse literature on debiasing VL models falls into two categories: (i) dataset-level debiasing (Zhao et al., 2021) and (ii) model-level debiasing (Hendricks et al., 2018). On the dataset side, simply trying to balance imbalanced data (Zhao et al., 2021) is not sufficient, with Wang et al. (2018) finding exaggerated gender stereotypes in tasks unrelated to gender recognition, despite balancing by gender. The disproportionate representation of certain genders and ethnicities in various roles can lead to misclassifications (Birhane et al., 2021). However, even if bias correction is done at the dataset-level (assuming access to the original data and sufficient compute resources), it may still be infeasible to capture all proxies for demographic bias (Hendricks et al., 2018) because it is possible that the data necessary to combat bias has not been curated yet (Weidinger et al., 2021). Through model-level adjustments, Hendricks et al. (2018) train an image captioning model to confidently predict gender when there is gender evidence and to be cautious in its absence.

domain adaptation of pretrained models. For specific-domain downstream tasks, it is desirable to adapt pretrained models to have less bias without degrading their feature quality. Prompting has become the de-facto domain adaptation technique for VL models (Zhou et al., 2021; Ju et al., 2021), as well as large language models (Shin et al., 2020; Liu et al., 2021). Learning input tokens (prompt learning) to reduce bias is an effective technique that requires minimal training data and prevents overfitting (Zhu et al., 2021). Similarly, Zhai et al. (2021) show that optimizing over only the text encoder and freezing the image encoder is superior to full finetuning and improves generalization. To counteract feature degradation from bias reduction by prompt learning, we employed joint training with an ITC loss, inspired by Li et al. (2021).

6 Limitations and Ethical Consideration

Our methods and findings are subject to some limitations, as well as some ethical considerations of how bias and fairness are operationalized.

assumptions on computational restrictions. Our methods rest on two assumptions about the setting of the downstream application, namely that (i) the VL model is too large to be pretrained from scratch within the computational budget, and (ii) there is no access to the original training dataset. In the absence of those assumptions, we strongly encourage employing ethical dataset curation practices as well as including fairness considerations in the initial training of the model. However, in the case where our assumptions hold, our method provides a cheap, simple yet effective method for debiasing VL models.

context-dependency of the debiasing goal. One limitation in the applicability of our debiasing method comes from the fact that any “desired distribution" of age, gender, ethnicity or other identity factor is related to (and may have to stem from) the context in which the model is developed or deployed. For example, the demographic distribution of ethnicities and their lived experiences varies across countries or regions so when debiasing VL models, different sensitive attributes and text prompts may be more or less relevant. Our bias measurement and mitigation techniques can be applied to any set of sensitive attribute queries and text prompts but defining how these relate to bias is a normative, subjective and contextual question.

lack of intersectional analysis. Due to practical constraints on available dataset labels, our experiments have only investigated social bias with respect to gender and ethnicity attributes. We encourage future research on more attributes, as well as intersectional analysis of how biases stack together (e.g., age and gender together may display much larger bias than either in isolation). However, we expect our mitigation and measurement techniques to work with similar efficacy and efficiency in intersectional experiments.

focus on representational harms. We primarily focus on representational harms, i.e., the harms which arise from unjust, inequitable portrayals across demographic groups. The problematic entrenchment of harmful norms is clear if marginalized groups are more highly associated with negative, criminal or non-human traits, while societally-dominant groups are associated with positive traits such as being ‘smart’, ‘good’ or ‘kind’. These representational harms can appear in common downstream use cases of VL models including image captioning or image search, with a potential mechanism for concomitant allocational harms. For example, an individual applying for a certain job may be discouraged if all faces returned by Google search on the position do not match their own identity or a recruiter may be influenced towards unfairly prioritizing applicants from the well-represented demographic. We do not explicitly test allocational harms and suggest future research should explore both general and case-specific settings by engaging multiple stakeholders and affected communities (Weidinger et al., 2021).

sole focus of bias in face images. Face datasets were used in original research on implicit bias (Greenwald et al., 1998) and have been adopted widely for bias in machine learning contexts, especially in the computer vision community. This motivated our use of face datasets in the subdomain of VL. Note that many well-known large face image datasets present privacy and representational issues, and that FairFace (Kärkkäinen and Joo, 2021) thus serves an important role in ethical bias research due to its synthetic nature. However, focusing only on face datasets encodes only a narrow presentation of social bias. In reality, social, cultural and historical biases extend far beyond face images, and includes associations on cultural artifacts, practices and geographic localities. We encourage future work on broader presentations of bias and harms in addition to those captured from captioning face datasets.

code of ethics. Our method can be applied to reduce representational harm in search queries. Our methods avoid using costly and environmentally-damaging training procedures. We use the privacy-preserving dataset FairFace which avoids potential unconsensual use of face images, but UTKFace may entail privacy risks. We do not employ human annotators in any capacity.

7 Conclusion

This paper establishes a framework for measuring and mitigating bias in VL models. Firstly, we demonstrate that ranking metrics (specifically *MaxSkew* and *NDKL*) are effective bias measures. We report these metrics for a range of pretrained VL models for gender and racial bias in photos of faces. Our results confirm previous findings in other domains that (i) more pretraining data correlates with lower model bias, and (ii) training models with SSL can reduce bias. Secondly, we demonstrate a supervised adversarial debiasing method of VL models via learned "debiasing" tokens on 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 feature degradation being wholly mitigable using joint training with an ITC loss on small publicly-available image datasets.

Future work could include (i) debiasing during the pretraining stage, with SSL showing a promising avenue in that regard, or (ii) defining a wider diversity of attributes such as removing the harmful assumption of binary gender or considering intersectional biases. We encourage researchers in VL to continue to investigate bias in their models, be transparent in documenting model weaknesses using metrics like those proposed in this paper, and seek to apply relatively cheap and easy debiasing protocols like ours.

Our code, models and debiasing tokens are publicly-available[3] for the community to use in the hope that progress can be made towards the safer and fairer use of this technology in society.

8 Acknowledgements

The authors would like to thank Wonsuk Yang, Mohamed Baioumy, Vit Ruzicka, Jakob Mökander and Laura Weidinger for their helpful feedback. We also thank our anonymous reviewers for their contributions. 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.).

References

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. 2018. 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*, pages 0–0.

[3] See https://github.com/oxai/debias-vision-lang.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. Association for Computational Linguistics.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

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

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

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, pages 214–226.

Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Evertrove. Evertrove - the semantic image api. Accessed: 2022-03-05.

Sidney Fussell. 2020. An algorithm that 'predicts' criminality based on a face sparks a furor.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. 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*, pages 2221–2231.

Hila Gonen and Yoav Goldberg. 2019. 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)*, pages 609–614.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. 2022. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. 2019. Bias correction of learned generative models using likelihood-free importance weighting. *Advances in neural information processing systems*, 32.

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

Donna Haraway. 2004. *The Haraway Reader*, volume 53. Routledge.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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)*, pages 771–787.

Xiao Hu, Haobo Wang, Somesh Dube, Anirudh Vegesana, Kaiwen Yu, Yung-Hsiang Lu, and Ming Yin. 2018. Discovering biases in image datasets with the crowd. pages 2015–2017.

Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K Thiruvathukal, and Ming Yin. 2020. Crowdsourcing detection of sampling biases in image datasets. In *Proceedings of The Web Conference 2020*, pages 2955–2961.

HuggingFace. Hugging face inference api. Accessed: 2022-03-05.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2021. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *Proc. of the 16th European Chapter of the Association for Computational Linguistics (EACL)*.

Kimmo Kärkkäinen and Jungseock Joo. 2021. 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)*.

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. *Conference on Human Factors in Computing Systems*.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Hannah Rose Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A Dreyer, Aleksandar Shtedritski, and Yuki M Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*.

Alex Krizhevsky. 2009. *Learning Multiple Layers of Features from Tiny Images*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter*, 23(1):14–23.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *NAACL*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, pages 622–628. Association for Computational Linguistics (ACL).

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2021. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens Amanda Askell, Peter Welinder Paul

Christiano, Jan Leike, and Ryan Lowe. 2020. Training language models to follow instructions with human feedback. *ACL*.

Joon Sung Park, Michael S. Bernstein, Robin N. Brewer, Ece Kamar, and Meredith Ringel Morris. 2021. *Understanding the Representation and Representativeness of Age in AI Data Sets*, page 834–842. Association for Computing Machinery, New York, NY, USA.

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. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. *WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning*, page 2443–2449. Association for Computing Machinery, New York, NY, USA.

Ryan Steed and Aylin Caliskan. 2021. 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.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73.

Angelina Wang, Arvind Narayanan, and Olga Russakovsky. 2020. 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.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021a. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *EMNLP*.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021b. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2018. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. Technical report.

Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114:246–257.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. *arXiv preprint arXiv:2112.04359*.

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

Xiaolin Wu and Xi Zhang. 2016. Automated inference on criminality using face images. *ArXiv*, abs/1611.04135.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pages 11492–11501. PMLR.

Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17, New York, NY, USA. Association for Computing Machinery.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.

Andrew Zhai and Hao-Yu Wu. 2018. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2021. Lit: Zero-shot transfer with locked-image text tuning.

Zhang, Zhifei, Song, Yang, Qi, and Hairong. 2017. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14830–14840.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai Wei Chang. 2019. 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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to Prompt for Vision-Language Models. *arXiv preprint arXiv:2109.01134*.

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Xiaogang Wang, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2021. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. *arXiv preprint arXiv:2112.01522*.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43–76.

A Word Embedding Association Test (WEAT)

The Word Embedding Association Test (Caliskan et al., 2017) 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

$$\begin{aligned} s(C_1, C_2, A_1, A_2) = \sum_{c_1 \in C_1} & s(c_1, A_1, A_2) \\ - \sum_{c_2 \in C_2} & s(c_2, A_1, A_2), \end{aligned} \tag{4}$$

where, with μ denoting the arithmetic mean,

$$\begin{aligned} s(w, A_1, A_2) = \mu_{a_1 \in A_1} & \cos(w, a_1) \\ - \mu_{a_2 \in A_2} & \cos(w, a_2) \end{aligned} \tag{5}$$

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

$$P_{r_i}[s(C_1^i, C_2^i, A_1, A_2) > s(C_1, C_2, A_1, A_2)] \tag{6}$$

Finally, the effect size, i.e., the normalized measure of the separation between the associations of the targets and attributes, (Caliskan et al., 2017) is defined as

$$\frac{\mu_{c_1\, \in C_1} s(c_1, A_1, A_2) - \mu_{c_2\, \in C_2} s(c_2, A_1, A_2)}{\sigma_{c\, \in C_1 \cup C_2} s(c, A_1, A_2)} \tag{7}$$

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.

B Ranking metrics

The following outlines the mathematical implementation of three bias metrics. 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 measures the difference between the desired proportion of image attributes in τ_T^k and the actual proportion (Geyik et al., 2019). 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 statistical parity. 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

$$Skew_A@k(\tau_T) = \ln \frac{p_{\tau_T,T,A}}{p_{d,T,A}} \tag{8}$$

This measurement gives an indication of possible representational bias (Weidinger et al., 2021), 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, and (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 Geyik et al. (2019), 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

$$MaxSkew_{@}k(\tau_T) = \max_{A_i \in \mathcal{A}} Skew_{A_i}@k(\tau_T) \tag{9}$$

This signifies the "*largest unfair advantage*" (Geyik et al., 2019) 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

$$\mathit{NDKL}(\tau_T) = \frac{1}{Z} \sum_{i=1}^{|\tau_y|} \frac{1}{\log_2(i+1)} d_{KL}(D_{\tau_T^i} || D_T) \tag{10}$$

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 $A \in \mathcal{A}$). Thus, this aggregation overcomes the first disadvantage of *Skew*, however, *NDKL* is non-negative, and so it cannot distinguish between two "opposite-biased" search procedures.

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

In Fig. 4 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.

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 the syntax of templates, model architecture and dataset.

We note that *WEAT* measured on UTKFace has an 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. This may be explained by the fact that *WEAT* was primarily designed for single word embeddings, while we are using long prompts. May et al. (2019) 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.

Table 5: 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

D Performance effects of learnable text token initialization

In Tab. 5 we show the effects on zero-shot performance when adding zero-initialized text tokens to the text queries, before any debiasing 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 debiased model is not due to the adversarial learning but rather the learnable text tokens which shift the distribution of the text query.

E Debiasing

Prepending learnable text tokens. 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.

F Experimental protocol

Debiasing implementation.

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 that we debias over, dim(A). We train with the *Adam* optimizer (Kingma and Ba, 2015)

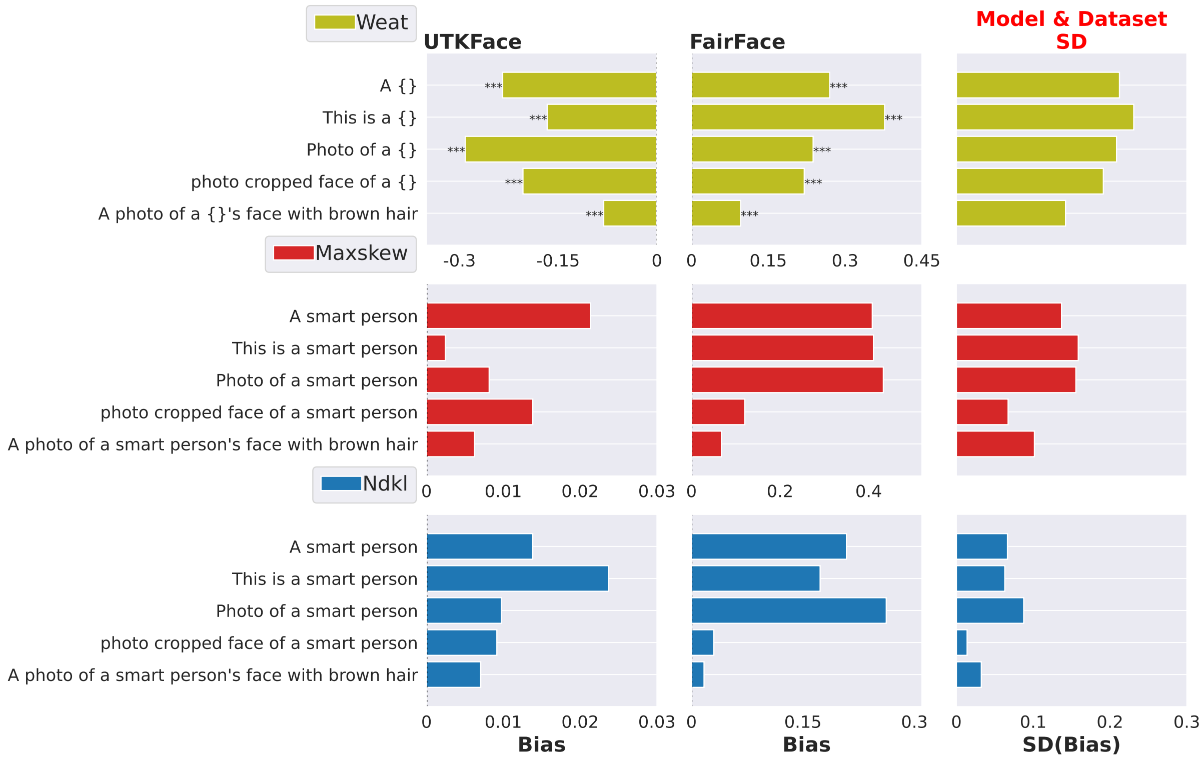

Figure 4: Bias measures across different combinations of minor syntactic changes, models (RN50, ViT$_{B/16}$, ViT$_{B/32}$), and datasets (FairFace validation set and UTKFace). Bias is measured for gender, and we use the *WEAT* pairwise adjectives concept sets from Caliskan et al. (2017). Standard deviation of bias measurement is taken over all combinations of model architecture and datasets, for other results we use ViT$_{B/32}$.

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 performance as tested on CIFAR100 (Krizhevsky, 2009)[4] or Flickr-1k (Young et al., 2014) 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. We experimented with other implementations of prompt learning than prepending tokens (e.g. appending or adding learned embeddings, and different initializations, e.g. zero-pad, embedding of common token from training corpus, and uniformly random), but these variations showed different feature and bias metric results only at start of training, and no significant change in results. As the number of learned tokens impacted feature quality, we chose 2 tokens as a reasonable trade-off (more tokens giving lower feature quality). For ITC joint training we used $\lambda = 0.05$ with image-text batches from the Flickr30K training set, unless otherwise specified.

G Harmful Zero-Shot Misclassification

We follow the protocol of Agarwal et al. (2021) by using CLIP to classify images from the Fair-

[4] Chosen over *IN1K*$_{acc}$ monitoring due to its smaller scale.

Table 6: **Harmful misclassification rate** of FairFace validation images into criminal and non-human categories, by FairFace ethnicity group. We compare between the CLIP Audit paper (Agarwal et al., 2021), 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 (Agarwal et al., 2021)	✗	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 (Agarwal et al., 2021)	✗	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

Table 7: **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 ViT$_{B/16}$ architecture.

Debias	**Bias Measures** ↓		**ZS Performance** ↑	
Adaptation	*MaxSkew@1000*	*NDKL*	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 finetuning	0.688(+195%)	0.664(+543%)	18.6(-78%)	6.6(-90%)

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

Tab. 6 shows the results directly taken from Agarwal et al. (2021) alongside results from our implementation with the pretrained baseline CLIP ViT$_{B/16}$. Our gender-debiased model trained on FairFace has a lower misclassification rate into crime-related 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%).

H Additional Results

In Tab. 7 we show the result of finetuning over different parts of the model as well as pure prompt learning, all with pure adversarial training. The strong regularization from having few learned embeddings keeps the feature quality at an acceptable level, and finetuning larger parts of the model lowered model performance to an unacceptable level very quickly during training.