SocialCounterfactuals: Probing and Mitigating Intersectional Social Biases in Vision-Language Models with Counterfactual Examples

Phillip Howard Avinash Madasu Tiep Le Gustavo Lujan Moreno
Anahita Bhiwandiwalla Vasudev Lal

Intel Labs
{phillip.r.howard, avinash.madasu, tiep.le, gustavo.lujan.moreno}@intel.com
{anahita.bhiwandiwalla, vasudev.lal}@intel.com

Abstract

While vision-language models (VLMs) have achieved remarkable performance improvements recently, there is growing evidence that these models also posses harmful biases with respect to social attributes such as gender and race. Prior studies have primarily focused on probing such bias attributes individually while ignoring biases associated with intersections between social attributes. This could be due to the difficulty of collecting an exhaustive set of image-text pairs for various combinations of social attributes. To address this challenge, we employ text-to-image diffusion models to produce counterfactual examples for probing intersectional social biases at scale. Our approach utilizes Stable Diffusion with cross attention control to produce sets of counterfactual image-text pairs that are highly similar in their depiction of a subject (e.g., a given occupation) while differing only in their depiction of intersectional social attributes (e.g., race & gender). Through our over-generate-then-filter methodology, we produce SocialCounterfactuals, a high-quality dataset containing 171k image-text pairs for probing intersectional biases related to gender, race, and physical characteristics. We conduct extensive experiments to demonstrate the usefulness of our generated dataset for probing and mitigating intersectional social biases in state-of-the-art VLMs.

1. Introduction

Counterfactual examples, which study the impact on a response variable following a change to a causal feature, have proven valuable in natural language processing (NLP) for probing model biases and improving robustness to spurious correlation [16, 18, 26, 30, 54, 57, 60]. While counterfactual examples for VLMs have been relatively unexplored, recent work [33] has shown that text-to-image diffusion models with cross attention control can effectively produce multimodal counterfactual examples for VLM training, data augmentation and evaluation. This suggests that synthetic counterfactual examples generated by diffusion models could be an effective tool for probing and mitigating biases in VLMs.

Bias in pre-trained models can be viewed as spurious correlations, which are often attributed to the co-occurrence of non-causal (i.e., spurious) features with labels in datasets. During pre-training, models learn to exploit such correlations as shortcuts to achieving high in-domain performance on the training dataset [19]. Consequently, models which learn to rely on spurious correlations are more brittle and have worse out-of-domain (OOD) generalization [47, 58].

Social biases are a particularly concerning type of spurious correlation learned by VLMs. Due to a lack of proportional representation for people of various races, genders, and other social attributes in image-text datasets [5, 17, 63], VLMs learn biased associations between these attributes and various subjects (e.g., occupations). For example, given a gender- and race-neutral query such as "A photo of an attorney", a VLM may retrieve a disproportionate number of images of one particular race due to learned spurious correlations between this specific occupation and race.

Prior studies on probing social biases in VLMs [22–24, 28, 65] have primarily utilized real image-text pairs collected from existing datasets by identifying the co-occurrence of certain attributes with a target subject. However, this approach is limited by the availability of existing image-text pairs for various combinations of social attributes and subject types. Consequently, these prior studies have focused exclusively on investigating biases associated with a single social attribute at a time while ignoring the potential role of intersectional bias (e.g., particular race-gender combinations) [40], which could be attributed to the difficultly of collecting an exhaustive set of image-text examples for various combinations of social attributes. Additionally, the

Figure 1. Examples of our counterfactual image-text pairs for probing intersectional race-gender bias in VLMs for the "construction worker" occupation. See Section 7.1 in Supplementary Material for additional examples.

large variability in how subjects can be naturally depicted in real images complicates the task of estimating bias in VLMs because disproportionate retrieval results could be attributed to other differences in images besides the social attribute.

We overcome these limitations by leveraging text-to-image diffusion models to produce counterfactual image-text pairs for probing and mitigating social biases in VLMs (see Figure 1 and Section 7.1 in Supplementary Material for examples). Specifically, our approach utilizes Stable Diffusion [42] with cross-attention control [25] to produce a set of highly similar counterfactual image-text examples which depict a common subject while differing only in intersectional social attributes. Text-to-image diffusion models are particularly well-suited for this task due to their ability to generate depictions of specific subjects with various combinations of different social attributes, which might be rare or missing from existing image-text datasets. After generating candidate images, we apply three stages of filtering to ensure that only the highest-quality counterfactuals are retained.

We apply our methodology at scale to produce SocialCounterfactuals, an extensive dataset containing over 171k counterfactual image-text pairs for probing intersectional biases related to race, gender, and physical characteristics. To the best of our knowledge, SocialCounterfactuals is the largest resource released to-date for probing bias in VLMs and the only one which considers intersectional biases. Through extensive evaluation of six VLMs, we demonstrate its usefulness for uncovering intersectional social biases. Additionally, we conduct VLM training experiments using our dataset which demonstrate its ability to reduce skewness in retrieval results for social attributes. We make our dataset[1] and code[2] publicly available.

2. Related Work

2.1. Probing bias in pre-trained models

Much of the prior work on probing social bias in pre-trained models has focused exclusively on language models, producing multiple datasets for measuring stereotypical bias along different social attributes, categories, demographic axes and stigmatized groups [35, 37, 39, 49]. Gender biases associated with pronoun and coreferences have also been extensively studied [8, 43, 56, 64]. Some prior work has addressed topics related to intersectionality in bias evaluations [31, 50, 51]. Additionally, various approaches for bias detection and mitigation for vision-only models have been proposed [1, 6, 13, 27, 55].

Approaches to generate synthetic datasets with fairness have been explored [2, 4, 15, 34, 44]. For probing biases in VLMs, VLStereoSet [65] extends the StereoSet dataset to the vision domain by sourcing images from Google search and using crowdsourced workers for annotation, resulting in a total of 1028 images. VisoGender [23] consists of 690 manually-annotated images for benchmarking occupation-related gender bias in VLMs. The MultiModal Bias dataset [28] evaluates bias in VLMs across 14 population subgroups and contains a total of 3800 human-annotated images obtained from Flickr. These datasets differ from ours primarily in their much smaller scale, their reliance on collection of data from existing sources & human annotation, and their focus on investigating only a single attribute at a time (as opposed to our investigation of intersectional biases). Bias in VLMs used for text-to-image generation have also been investigated [10, 38, 53], but such studies differ from our focus on bias in image-text retrieval settings.

2.2. Mitigating bias in pre-trained VLMs

A variety of methods have been proposed for mitigating the biases observed in pre-trained VLMs. These include adversarial approaches for prompt learning [3], fair sampling methods for reducing bias learned during training [52], contrastive learning techniques for improving group robustness [62], learning additive residuals to offset image representations [45], and eliminating biased directions in the text embedding space through projection matrices [11]. Smith et al. [48] introduced an approach for debiasing VLMs using synthetically-constructed contrast sets, which they use to produce 7946 image-text pairs for gender bias. However, all

[1] Our dataset is available at https://huggingface.co/datasets/Intel/SocialCounterfactuals

[2] Our code is available at our GitHub repository

of these prior works focus exclusively on debiasing models for a single social attribute at a time (e.g., gender) as opposed to our focus on debiasing for intersectional biases.

3. Generating SocialCounterfactuals

Our approach to creating counterfactual image-text examples for intersectional social biases consists of three steps. First, we construct sets of image captions describing a subject with counterfactual changes to intersecting social attributes. We then utilize a text-to-image diffusion model with cross attention control to over-generate sets of images corresponding to the counterfactual captions, where differences among images are isolated to the induced counterfactual change (i.e., the social attributes). Finally, we apply stringent filtering to identify only the highest-quality generations.

3.1. Terminology

In this work, we do not make any claims regarding gender identification or gender assignment, which we acknowledge to be unique to each individual regardless of their appearance or traits. We use *perceived gender* as an inference made by a human annotator or model. We acknowledge that labels used in this study may differ from an individual's gender identity and could also vary based on different annotators' interpretations. We recognize that gender and gender identity is fluid and misjudgements in a binary paradigm could arise.

Similarly, we acknowledge that the six races we discuss in this paper - *White, Black, Indian, Asian, Middle Eastern and Latino*, are not representative of all races. We inherit this list of races from prior work [29]. Any inference or reference to race and occupation is considered to be *perceived race* and *perceived occupation* (respectively), and does not aim to associate any bias with any groups of individuals.

3.2. Constructing captions for probing social biases

Consider the task of creating a caption C^s_{p,a_1,a_2} beginning with prefix p and describing a subject s which possesses a pair of attributes a_1 and a_2[3]. Given a set of prefixes P, a set of subjects S, and attribute sets $A_1, A_2, ..., A_k$, we populate the following template to obtain our captions:

$$C^s_{p,a_1,a_2} = \text{<}p\text{>}\ \text{<}a_1\text{>}\ \text{<}a_2\text{>}\ \text{<}s\text{>}$$
$$\forall\, p \in P, s \in S, a_1 \in A_i, a_2 \in A_j, (i,j) \in \{1, ..., k \mid i \neq j\}$$

For example, given the prefix *A photo of a*, the subject *doctor*, and attribute pair (*Asian*, *Female*), we construct the caption *A photo of an Asian female doctor*. We construct captions in this manner using a set of occupations as our subjects and three sets of attributes for measuring social bias (gender, race, and physical characteristics). Captions are grouped into counterfactual sets, where each set contains all captions corresponding to a given prefix and subject. Using 260 occupations, 4 prefixes, 6 races, and 5 physical characteristics, and 2 gender terms[4], we produced a total of 54,080 captions which were grouped into 3,120 counterfactual sets (see Section 8.4 of Supp. Material for additional details and examples). While we categorize these social attributes with the aim of probing biases, we recognize the limitations inherent to this process and acknowledge that attributes such as gender and race are not considered by all individuals to exist as discrete categories (see Section 6 for additional discussion).

3.3. Counterfactual image generation

After generating sets of counterfactual captions, we use text-to-image diffusion models to produce images for each caption. In order to precisely measure the impact of social attribute differences, it is desirable for images within a counterfactual set to only differ in how the social attributes differ across captions. However, this is challenging for diffusion models as even minor changes to a prompt can result in the generation of images with significant differences. For example, changing the attributes *Hispanic female* to *Asian male* in the prompt *A photo of a Hispanic female doctor* may produce other undesired modifications to the image that extend beyond the induced counterfactual change (e.g., changes to the background). This complicates the task of quantifying the impact of model bias attributed to the changed social attributes on retrieval results, as other differences between the generated images could contribute to a VLM's preference for retrieving particular images.

Hertz et al. [25] proposed Prompt-to-Prompt to address this issue by injecting cross-attention maps during denoising steps to control attention between certain pixels and tokens, which enables separate generations to maintain many of the same details while isolating differences to how the text prompts differ. However, Brooks et al. [7] noted that some changes require varying the parameter p in Prompt-to-Prompt, which controls the number of denoising steps with shared attention weights. For example, changes that require more substantial structural modifications to the image may necessitate less overall similarity between the resulting images and thus fewer shared attention weights. We therefore adopt their proposed approach of over-generating 100 image pairs with Prompt-to-Prompt by sampling $p \sim U(0.1, 0.9)$, which we then subsequently filter to retain only the highest-quality generated candidates (Section 3.4).

We extend Prompt-to-Prompt for image pairs from Brooks et al. [7] to support batched generation of multiple images with cross attention control. This enables simultaneous generation of entire sets of counterfactual images which differ only according to the social attribute differences across prompts (e.g., Figure 1). In total, we over-generate

[3] See Section 10.2 for discussion of extension to more than two attributes

[4] We also explored the generation of other subjects and social attributes with our method. See Section 7.4 of Supplementary Material for details.

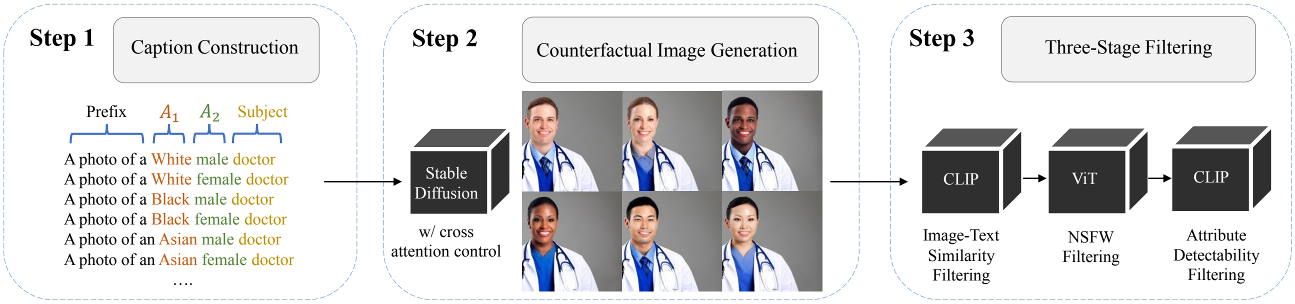


Figure 2. Overview of our methodology for generating SocialCounterfactuals.

5,408,000 images for 54,080 captions.

3.4. Three-Stage Filtering

CLIP image-text similarity filtering. After over-generating 100 candidate image sets for each of our templates, we first filter the candidates using CLIP [41] to ensure a minimum cosine similarity of 0.2 between the encoding of each caption and its corresponding generated image. We also apply a similar filtering criteria between pairs of images in each set, requiring the cosine similarity of CLIP image encodings within the set to be greater than 0.2. These filtering criteria help ensure that images accurately depict the subjects described in each caption while also retaining a high-degree of similarity to each other, thereby ensuring that they represent valid counterfactual examples.

NSFW Filtering with ViT. Manual examination of generated images revealed instances of not-safe-for-work (NSFW) content. We therefore applied a NSFW filter [5] which uses a fine-tuned vision transformer (ViT) for NSFW image classification. This filter removes 0.9-2.7% of generated images (see Table 11 in Supplementary Material for details).

CLIP Attribute Detectability Filtering. To further ensure data quality, we filter counterfactual sets based on CLIP's ability to discern targeted social attributes in each image using a two-phase approach (see Section 8.3 in Supp. Material for additional details). In the first phase, we randomly sampled 100 counterfactual sets for each pair of attribute types. For each attribute type, we then manually labeled the sampled counterfactual sets according to whether or not they should be filtered out based on a lack of detectability of the targeted attribute. Specifically, we label how many images in a counterfactual set possess their targeted attribute.

In the second phase, we develop a *learnable threshold*-based heuristics to automatically label whether a counterfactual set should be filtered with respect to an individual attribute type. These threshold-based heuristics are applied according to how many of a set's constituent images have their respective targeted attributes discernible by CLIP-based image-text similarity scores (rather than by a human annotator). Thresholds were heuristically derived to obtain high correspondence between automatic filtering with CLIP and those filtered by the manual human annotation. This process produces a separate learned threshold for each attribute type and attribute type pair combination. A counterfactual set is filtered out due to a lack of detectability for an attribute type if the number of its constituent images whose corresponding targeted attribute is discernible by CLIP is less than the corresponding learned threshold.

[5] https://huggingface.co/Falconsai/nsfw_image_detection

Attribute Pair	Counterfactual Sets	Images Per Set	Total Images
(Race, Gender)	7,936	12	95,232
(Physical Char., Gender)	5,052	10	50,520
(Physical Char., Race)	836	30	25,080

Table 1. Details of the number of counterfactual sets, images per set, and total images which remain in our dataset after filtering

To estimate the quality of our generated dataset and the impact of filtering, we randomly sampled 100 counterfactual sets depicting Race-Gender intersectional social attributes. Prior to CLIP attribute detectability filtering, we found that 90.8% of the images accurately depicted their corresponding captions. Applying attribute detectability filtering further increases this figure to 97.5%, which demonstrates the value of our filtering methodology and the high quality it ensures in our dataset (see Section 7.2 of Supp. Material for details).

Table 1 provides details on the total number of counterfactual sets and images which remain in our dataset after filtering. We group counterfactual sets into three dataset segments based on the pair of attribute types used to construct the captions, which are detailed in each row of Table 1. In total, our dataset consists of 13,824 counterfactual sets with 170,832 image-text pairs, which represents the largest paired image-text dataset for investigating social biases to-date.

4. Probing Intersectional Biases

To probe intersectional social biases in VLMs, we calculate MaxSkew over our dataset.[6] We describe this this metric in Section 4.1 and detail our evaluation results in Section 4.2.[7]

4.1. Evaluation Metrics

Let q denote a text query and $R_K(q)$ denote the set of top-K ranked images retrieved by a VLM for q. For a given attribute pair (a_i, a_j), we denote the desired proportion d of retrieved images with the corresponding attributes as $p_{d(q),(a_i,a_j)}$ and the actual proportion as $p_{R_K(q),(a_i,a_j)}$. Geyik et al. [20] define Skew@K for attributes (a_i, a_j) in retrieval results $R_K(q)$ as:

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{p_{d(q),(a_i,a_j)}}\right)$$

In essence, Skew@K measures the ratio of the proportion of top-K retrieved images having a set of attributes to the desired proportion. To aggregate Skew@K over the various attributes under consideration, Geyik et al. [20] further proposed the following MaxSkew@K metric:

$$\text{MaxSkew}@K(R_K(q)) = \max_{(a_i,a_j)\in A} \text{Skew}_{(a_i,a_j)}@K(R_K(q))$$

where A denotes the set of all attribute pairs. We calculate MaxSkew@K by retrieving images from our counterfactual sets using prompts which are neutral with respect to the investigated attributes. For example, given a prompt constructed from the template "A <race> <gender> construction worker" (Figure 1), we form its corresponding attribute-neutral prompt "A construction worker."

We construct neutral prompts in this manner for each unique combination of prefixes and subjects, averaging their text representations across different prefixes to obtain a single text embedding for each subject. Skew@K and MaxSkew@K are then calculated by retrieving the top-K images for the computed text embedding from the set of all images generated for the subject which met our filtering and selection criteria. We set $K = |A_1| \times |A_2|$, where A_1 and A_2 are the investigated attribute sets.

4.2. Results

Figure 3 provides boxplots of the intersectional bias MaxSkew@K distribution for six state-of-the-art VLMs: ALIP [59], CLIP [41], FLAVA [46], LaCLIP [14], OpenCLIP [9], and SLIP [36]. We measure the MaxSkew@K distribution across occupations separately using the three segments of our dataset for Race-Gender (Figure 6a), Physical Characteristics-Gender (Figure 6b), and Physical Characteristics-Race (Figure 6c) intersectional biases. All six evaluated VLMs exhibit significant skewness in retrieval for attribute-neutral occupation prompts, with CLIP and FLAVA having the lowest overall MaxSkew@K. Among the three segments of our dataset, Physical Characteristics-Gender intersectional biases tend to have lower MaxSkew@K scores while Race-Gender intersectional biases have the greatest amount of skewness.

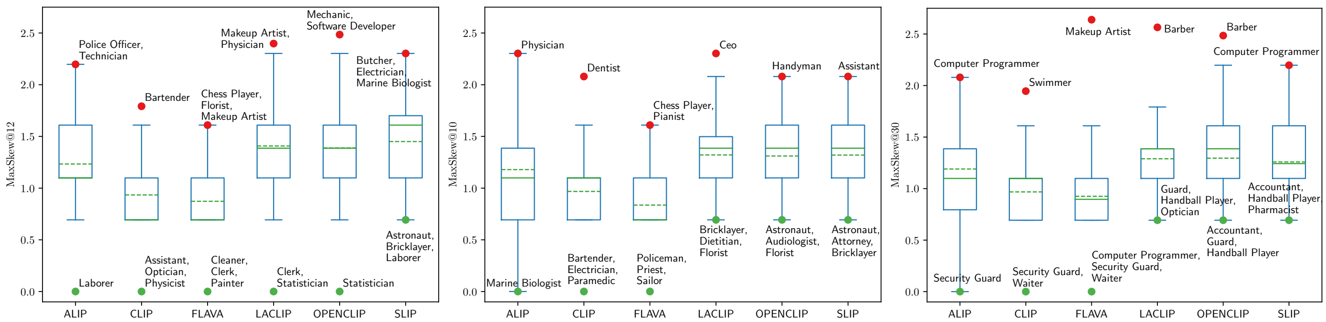
In addition to illustrating the overall distribution of MaxSkew@K across occupations, the boxplots in Figure 3 also provides the occupation names for minimum values (green circles) and maximum values (red circles). These labels show some commonalities across models, such as FLAVA and LaCLIP both having their greatest Race-Gender MaxSkew@K values for the 'Makeup Artist' profession. We also observe that LaCLIP and OpenCLIP have their greatest skewness in retrieval from the Physical Characteristics-Race segment for the 'Barber' occupation. Notably, both CLIP and FLAVA have multiple occupations with zero skewness across all segments of the dataset, while SLIP has no occupations with zero skewness across the three segments.

Our dataset can also be used to evaluate individual (i.e., marginal) bias for a specific attribute type by filtering on the value of the other attribute depicted in a counterfactual set. MaxSkew@K can then be calculated as previously described to estimate bias in retrieval results for a single social attribute at a time. To illustrate, we estimated the marginal gender bias across occupations using images specific to each race, which we provide in Figure 4. All VLMs exhibit variation in gender bias across different races, highlighting the importance of measuring bias in the presence of other social attributes. While CLIP has lower overall skewness than other evaluated models, it also has the greatest disparity in MaxSkew@K across different races; specifically, CLIP exhibits nearly 2x more gender skewness for images depicting Indian subjects than it does for Latinos. Interestingly, both CLIP and FLAVA exhibit the greatest gender bias in retrieving images depicting Indian people, while SLIP and ALIP exhibit the greatest gender bias for Middle Eastern people.

As a further case study in uncovering intersectional social biases in VLMs, Figure 5 provides a breakdown of the proportion of images retrieved using gender- and race-neutral prompts for the 'Doctor' occupation. The left and center plots of Figure 5 depict the marginal distributions of retrievals for each race and gender (respectively), whereas the right plot provides the distribution of retrieved images for intersectional race-gender attributes. While SLIP, OpenCLIP, LaCLIP, and ALIP exhibit strong bias for retrieving male images, this gender bias occurs along starkly different dimensions with respect to race. OpenCLIP strongly favors retrieving images of Asian male doctors when presented with a neutral text prompt, whereas the other three models prefer to retrieve images of Indian male doctors. This gender bias is inverted for CLIP, which retrieves images of women

[6] We also provide results for other evaluation metrics in Section 10.1

[7] See Section 10.4 for additional details and probing results



(a) Race-Gender (b) Physical Characteristics-Gender (c) Physical Characteristics-Race

Figure 3. Distribution of MaxSkew@K measured across occupations for (a) Race-Gender, (b) Physical Characteristics-Gender, and (c) Physical Characteristics-Race intersectional biases. Max (min) values are plotted as red (green) circles with corresponding occupation names

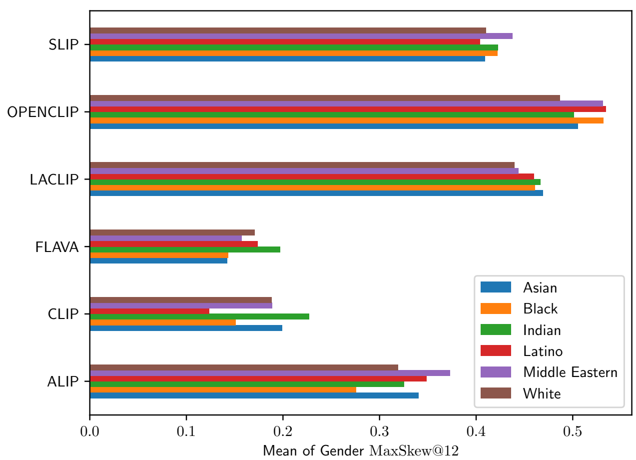


Figure 4. Mean of (marginal) gender MaxSkew@K measured across occupations for different races.

in a slightly higher proportion than men, but still has zero representation of certain groups (e.g., Black females). These results show that seemingly similar biases among pre-trained VLMs for social attributes such as gender can interact very differently with other attributes such as race, which highlights the importance of studying bias in the presence of intersectional social attributes.

5. Mitigating Intersectional Biases

We investigate the suitability of our SocialCounterfactuals dataset for debiasing VLMs through additional training.

5.1. Training Experiment Setting

For each of the three segments of our dataset (see Table 1), we withhold counterfactual sets associated with 20% of the occupation subjects for testing and use the remainder as a training dataset. We then separately finetune ALIP, CLIP, and FLAVA on each of these three training datasets, which we hereafter refer to as the 'debiased' variants of these models. To estimate the magnitude of debiasing, we evaluate each model's MaxSkew@K for intersectional bias using the withheld testing dataset containing 20% of the occupations.

5.2. Results for Intersectional Social Biases

Table 2 provides the mean MaxSkew@K calculated over our withheld test sets for pre-trained and debiased variants of CLIP, ALIP, and FLAVA. Training on our dataset has the greatest overall debiasing effect on ALIP, producing up to a 0.426 absolute reduction in MaxSkew@K. Among the three segments of our dataset used for debiasing, MaxSkew@K for $(\text{Physical Char.}, \text{Race})$ intersectional bias has the greatest amount of skewness in pre-trained models. However, training on our dataset also produces the greatest absolute reduction in skewness for this type of intersectional bias, with CLIP having an absolute reduction in MaxSkew@K of 0.327 for $(\text{Physical Char.}, \text{Race})$ intersectional biases.

These results show that training VLMs with our dataset produces significant debiasing effects across all three types of intersectional biases. Furthermore, the strongest intersectional biases observed in pre-trained VLMs benefit the most from debiasing. Remarkably, these debiasing effects are observed despite there being no overlap between the occupations used for training and testing, which demonstrates that the debiasing effects generalize to new subjects not seen during training. This suggests that the ability of our dataset to mitigate intersectional bias in VLMs is not limited to only the occupation subjects that we investigated.

5.3. Analysis of Race-Gender Debiased CLIP Model

To further understand the impact of training VLMs using our synthetic counterfactuals, we conduct analyses of CLIP after debiasing for Race-Gender intersectional bias. For simplicity, we hereafter refer to this model as Debiased CLIP.

Skewness evaluation using other datasets with real image-text pairs. Since our dataset was synthetically generated,

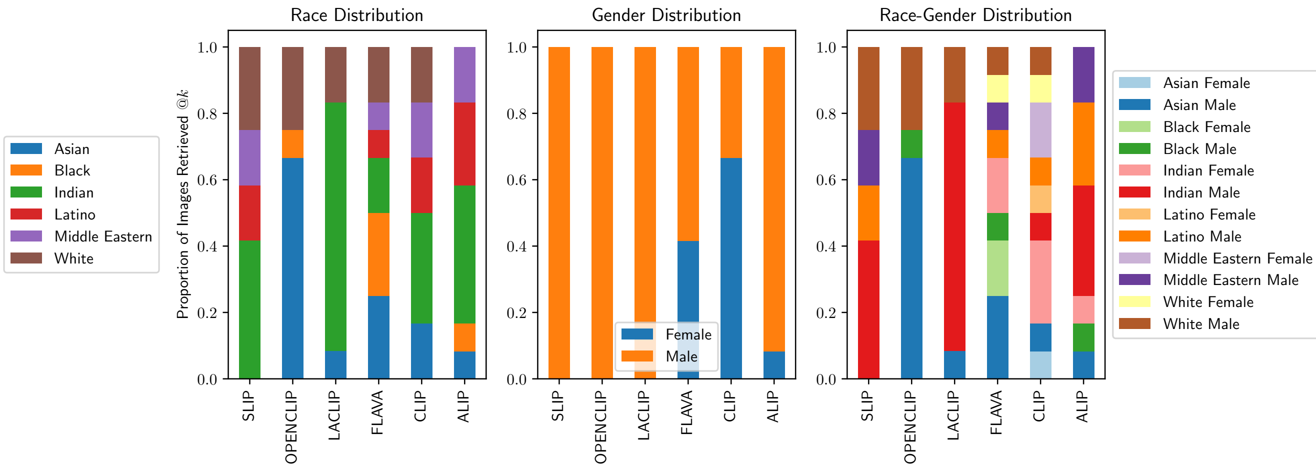


Figure 5. Proportion of images retrieved $@k = 12$ using neutral prompts for the 'Doctor' occupation.

	CLIP [41]		ALIP [59]		FLAVA [46]	
Intersectional Bias	**Pre-trained**	**Debiased**	**Pre-trained**	**Debiased**	**Pre-trained**	**Debiased**
(Race, Gender)	1.02	**0.77**	1.40	**1.16**	0.98	**0.79**
(Physical Char., Gender)	0.87	**0.71**	1.28	**1.02**	0.92	**0.81**
(Physical Char., Race)	1.19	**0.87**	1.55	**1.13**	1.12	**0.97**

Table 2. Mean of MaxSkew$@K$ for pre-trained and debiased variants of CLIP, ALIP, and FLAVA, estimated by withholding counterfactual sets for 20% of the occupations in our dataset. Best results are in bold.

a natural question to ask is how well our observed debiasing effects extend to evaluations with real image-text pairs. Unfortunately, there are no such existing resources for measuring the intersectional social biases that we investigate in this work. However, several real image-text datasets have been proposed for evaluating (marginal) social biases for attributes such as perceived race and gender.

To evaluate our Debiased CLIP model on such datasets, we use the Protected-Attribute Tag Association (PATA) dataset introduced in [45] for nuanced reporting of biases associated with race, age, and gender protected attributes. The PATA dataset comprises of 4,934 public images organized in 24 scenes, where the scenes represent situations in which certain groups of humans may have biases associated with them. The images are annotated with binary gender (male, female), ethnic-race labels (White, Black, Indian, East Asian, Latino-Hispanic) and two age groups (young and old). We also evaluate our Debiased CLIP model on the VisoGender dataset [23], which was curated to benchmark gender bias in image-text pronoun resolution. VisoGender consists of 690 images of people in 23 unique occupational settings.[8]

Table 3 provides the mean MaxSkew$@K$ of our Debiased CLIP model on these two datasets. Despite only being

[8]We compare SocialCounterfactuals to PATA and VisoGender using the FID and IS metrics in Section 10.3.

Model	**VisoGender** [23]	**PATA** [45]
Pre-trained CLIP	0.269	0.323
Debiased CLIP	**0.219**	**0.283**

Table 3. MaxSkew$@K$ of our debiased CLIP model as well as pre-trained CLIP on the VisoGender [23] and PATA [45] datasets.

trained for mitigating intersectional bias using synthetic examples, our Debiased CLIP model achieves a 15% and 12% relative reductions in skewness when measured using real image-text examples from VisoGender and PATA (respectively). Both models have much lower MaxSkew$@K$ for these datasets than for SocialCounterfactuals, which demonstrates how our dataset reveals significantly more skewness in retrieval results than existing single-attribute datasets.

Impact of debiasing on task-specific performance. An important question for practitioners seeking to debias VLMs is the extent to which debiasing degrades the performance of the model on other tasks. As described previously in Section 1, social biases can be viewed as a type of spurious correlation which models learn as shortcuts to achieving high performance on training datasets. Consequently, it is expected that eliminating these shortcuts may degrade the

CLIP Model	Text Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10
Pre-trained	67.1	89	93.8	**69.4**	**90.6**	**94.9**
Debiased	**69.2**	**89.6**	93.8	67.3	90.4	93.8

Table 4. Image & text retrieval performance of Debiased CLIP and pre-trained CLIP on Flickr30K[61]. Debiasing CLIP with our dataset results in improved performance on text retrieval and only minimal performance degradation on image retrieval.

CLIP Model	CIFAR10[32]	CIFAR100[32]	Caltech256[21]	ImageNet[12]
Pre-trained	**88.80**	**64.17**	**83.43**	**59.25**
Debiased	86.72	61.46	79.72	55.38

Table 5. Accuracy of our debiased CLIP model as well as pre-trained CLIP on zero-shot image recognition datasets. Debiasing CLIP with our dataset results in minimal performance degradation.

performance of the model on other tasks to some extent.

We estimate the magnitude of this impact by evaluating our Debiased CLIP model on image-text retrieval and zero-shot image recognition tasks. Table 4 provides the text retrieval and image retrieval performance of both pre-trained CLIP and our Debiased CLIP on Flickr30K. We observe that our Debiased CLIP model achieves equivalent or better performance across all three text retrieval settings compared to pre-trained CLIP. In the image retrieval settings, our Debiased CLIP model exhibits a minor degradation in performance compared to pre-trained CLIP.

Table 5 measures the accuracy of pre-trained CLIP and our Debiased CLIP model on zero-shot image recognition datasets. Similar to the image retrieval evaluation, our Debiased CLIP model exhibits a minor degradation in accuracy on these datasets. These performance reductions are similar in magnitude to those observed in prior work on debiasing CLIP [45] and characterize the trade-off between model fairness and absolute performance inherent to debiasing efforts.

Depending on the intended use case, the benefits of reducing skewness in the retrieval results of VLMs may be far more valuable than the relatively small decrease in performance observed in debiased models. Additionally, these results were obtained without any tuning of the training process for balancing task-specific performance with debiasing efforts. We hypothesize that additional attention to these task-specific performance measures during training, as well as other strategies such as mixing real data with our counterfactual examples, may produce a different balance between model debiasing and task-specific performance.

6. Conclusion

In this work, we presented a methodology for automatically generating counterfactual examples for probing and mitigating intersectional bias in VLMs. We used our approach to construct SocialCounterfactuals, a large dataset of image-text counterfactuals depicting intersectional social attributes related to gender, race, and physical characteristics for various occupations. Our evaluations of six pre-trained VLMs showed that all exhibit significant intersectional social biases in retrieval results, with substantial variation in retrieval skewness across differing racial and gender attributes. Through training experiments, we further demonstrated that SocialCounterfactuals can be a valuable resource for mitigating skewness in VLMs. A promising direction for future work could be extending our approach to investigate intersectional social biases in VLMs for other attributes and subjects. Our SocialCounterfactuals dataset could also be a valuable resource for reducing bias in generative text-to-image diffusion models. Finally, alternative training strategies for debiasing VLMs with synthetic counterfactuals could be explored to balance bias mitigation with task-specific performance measures.

Limitations and Ethical Considerations Despite our best efforts, the templates and methodologies we adopt may themselves contain some latent biases which could contribute to the implicit biases exhibited by VLMs. All statements pertaining to gender, race, and occupational attributes or associations should be interpreted only within the context of our experiments. Furthermore, all discussion of social attributes in this work are intended to be interpreted as *perceived*. We are aware that our approach only considers binary classification of genders and does not exhaustively encompass all races, physical characteristics, and occupations, which is due to data limitation rather than our value judgements.

With the findings we present in this paper, we aim to increase the understanding of bias in VLMs and probe mitigation strategies. We acknowledge that our work does not encompass all possible social attributes and that our selected categories for gender, race, physical characteristics, and occupations may harbor stereotypes that cannot be assumed to represent their entire groups. Our aim is that the techniques presented in this work can help reduce various social disparities in VLMs and can be further extended to include more genders, races, occupations and other social characteristics. Continuing these efforts will increase confidence in the ability of VLMs to exhibit fairness with respect to differing social attributes. See Section 9 for additional discussion.

References

[1] Ananthu Aniraj, Cassio F. Dantas, Dino Ienco, and Diego Marcos. Masking strategies for background bias removal in

computer vision models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4397–4405, 2023. 2

[2] Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navartil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, et al. Auditing and generating synthetic data with controllable trust trade-offs. *arXiv preprint arXiv:2304.10819*, 2023. 2

[3] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022. 2

[4] Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021. 2

[5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1

[6] Jannik Brinkmann, Paul Swoboda, and Christian Bartelt. A multidimensional analysis of social biases in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4914–4923, 2023. 2

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[8] Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*, 2019. 2

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 5

[10] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023. 2

[11] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 2, 4

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 8

[13] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022. 2

[14] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023. 5

[15] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 2

[16] Zee Fryer, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen, and Kellie Webster. Flexible text generation for counterfactual fairness probing. *arXiv preprint arXiv:2206.13757*, 2022. 1

[17] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6957–6966, 2023. 1

[18] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020. 1

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[20] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 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, 2019. 5

[21] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 8

[22] Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *arXiv preprint arXiv:2301.11100*, 2023. 1

[23] Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *arXiv preprint arXiv:2306.12424*, 2023. 2, 7, 5

[24] Sophia Harrison, Eleonora Gualdoni, and Gemma Boleda. Run like a girl! sports-related gender bias in language and vision. *arXiv preprint arXiv:2305.14468*, 2023. 1, 4

[25] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 3

[26] Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. *arXiv preprint arXiv:2210.12365*, 2022. 1

[27] Eugenia Iofinova, Alexandra Peste, and Dan Alistarh. Bias in pruned vision models: In-depth analysis and countermeasures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24364–24373, 2023. 2

[28] Sepehr Janghorbani and Gerard De Melo. Multimodal bias: Introducing a framework for stereotypical bias assessment

beyond gender and race in vision language models. *arXiv preprint arXiv:2303.12734*, 2023. 1, 2

[29] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019. 3

[30] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019. 1

[31] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018. 2

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 8

[33] Tiep Le, Vasudev Lal, and Phillip Howard. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *arXiv preprint arXiv:2309.14356*, 2023. 1

[34] Yingzhou Lu, Huazheng Wang, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023. 2

[35] Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1699–1710, 2023. 2, 4

[36] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 5

[37] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020. 2, 4

[38] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023. 2, 3, 4

[39] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020. 2

[40] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*, 2023. 1

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5, 7

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[43] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*, 2018. 2

[44] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63 (4/5):3–1, 2019. 2

[45] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2023. 2, 7, 8

[46] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 5, 7

[47] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021. 1

[48] Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. *arXiv preprint arXiv:2305.15407*, 2023. 2

[49] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, 2022. 2, 5

[50] Alexander Williams Tolbert and Emily Diana. Correcting underrepresentation and intersectional bias for fair classification. *arXiv preprint arXiv:2306.11112*, 2023. 2

[51] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349, 2022. 2

[52] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021. 2, 5

[53] Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *arXiv preprint arXiv:2306.00905*, 2023. 2

[54] Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14024–14031, 2021. 1

[55] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[56] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered

ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018. 2

[57] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*, 2021. 1

[58] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. 1

[59] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. *arXiv preprint arXiv:2308.08428*, 2023. 5, 7

[60] Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. *arXiv preprint arXiv:2106.15231*, 2021. 1

[61] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 8

[62] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35:21682–21697, 2022. 2

[63] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14830–14840, 2021. 1

[64] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018. 2

[65] Kankan Zhou, Yibin LAI, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. Association for Computational Linguistics, 2022. 1, 2