Social Debiasing for Fair Multi-modal LLMs

Harry Cheng[1], Yangyang Guo[2], Qingpei Guo[3,4,*], Ming Yang[4], Tian Gan[1], Weili Guan[5], Liqiang Nie[5]
[1]Shandong University, [2]National University of Singapore, [3]University of Warwick, [4]Ant Group,
[5]Harbin Institute of Technology (Shenzhen)
xaCheng1996@gmail.com, guoyang.eric@gmail.com, qingpei.gqp@antgroup.com,
m.yang@antgroup.com, gantian@sdu.edu.cn, honeyguan@gmail.com, nieliqiang@gmail.com

Abstract

Multi-modal Large Language Models (MLLMs) have dramatically advanced the research field and delivered powerful vision-language understanding capabilities. However, these models often inherit deep-rooted social biases from their training data, leading to uncomfortable responses with respect to attributes such as race and gender. This paper addresses the issue of social biases in MLLMs by i) introducing a comprehensive counterfactual dataset with multiple social concepts (CMSC), which complements existing datasets by providing 18 diverse and balanced social concepts; and ii) proposing a counter-stereotype debiasing (CSD) strategy that mitigates social biases in MLLMs by leveraging the opposites of prevalent stereotypes. CSD incorporates both a novel bias-aware data sampling method and a loss rescaling method, enabling the model to effectively reduce biases. We conduct extensive experiments with four prevalent MLLM architectures. The results demonstrate the advantage of the CMSC dataset and the edge of CSD strategy in reducing social biases compared to existing competing methods, without compromising the overall performance on general multi-modal reasoning benchmarks.

1. Introduction

Multi-modal Large Language Models (MLLMs) have revolutionized the field of general-purpose vision-language understanding. Representative models, such as LLaVA [38], Qwen-VL [3], and Bunny [25], exhibit remarkable zero-shot performance and can be easily fine-tuned for diverse downstream applications.

Despite MLLMs' widespread use, it is imperative to recognize that these models can exhibit severe social biases with respect to attributes such as race and gender [18, 65]. Figure 1a illustrates one example that a biased MLLM is much more likely to associate the role of *nurse* with *female* rather than *male*, which may be stereotypical. Such biased predictions often occur unconsciously within MLLMs, making them difficult to detect and avoid by explicit rules. A key reason for this issue lies in the composition of the MLLM training data, which can contain content related to violence or racism [6]. These inappropriate samples may collectively demonstrate some stereotypes, resulting in subtle yet uncomfortable responses [8, 63].

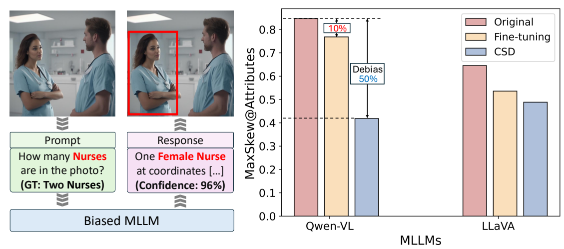

Figure 1. Social bias examples. **Left:** Gender-biased prediction on *Nurse* from LLaVA-7B[1]. **Right:** Maximum Skews across different human attributes on the FairFace dataset from Qwen-VL-7B and LLaVA-7B. As a metric for measuring model bias, a larger Skew implies a higher degree of social bias. Our CSD method outperforms existing approaches by a significant margin.

Existing studies on mitigating the social bias problem in MLLMs remain largely under-explored. A naive approach is to collect attribute-balanced vision-language counterfactual datasets [24, 34], which can be used to directly fine-tune biased models to predict fairer distributions. However, as demonstrated in Table 1, existing large-scale datasets are limited by their focus on a single social concept, such as occupation [27], without attending to multifaceted social stereotypes [62]. This significantly hampers the model in learning more comprehensive representations. Therefore, *we are motivated to construct a more extensive and diverse counterfactual dataset.* From a methodological perspective, directly fine-tuning on a counterfactual dataset assigns equal importance to instances receiving different social biases, resulting in sub-optimal debiasing performance (as demonstrated in Section 6.2). Analogy to a common solution in chemistry—neutralizing acidic water requires

*Corresponding author.
[1]We manually visualize the bounding boxes from LLaVA's text output.

Dataset	Venue	Type	#Images	Social Concepts
CoCo-Counterfactuals [34]	NeurIPS'23	General	34K	No annotations
FairFace [33]	WACV'21	Social	108K	No annotations
VisoGender [24]	NeurIPS'23	Social	0.6K	Occupation
PATA [55]	CVPR'23	Social	5K	Occupation
SocialCounterfactuals [27]	CVPR'24	Social	171K	Occupation
MM-Bias [31]	EACL'23	Social	3K	14 minorities
CMSC (Ours)	–	Social	60K	**18 balanced concepts**

Table 1. Comparison of CMSC with mainstream datasets used for model debiasing. Existing datasets are limited either by scale or by the number of social concepts they cover. Type – The category of biases the dataset focuses on.

adding an alkaline substance rather than plain water—*we are prompted to speculate: Can we leverage the opposite of the suffered social bias to build a fairer model?*

To address these two issues, we first construct a large-scale, high-quality Counterfactual dataset with Multiple Social Concepts (CMSC). As demonstrated in Table 1, our CMSC dataset narrows the gap in data scale and concept richness. We conduct extensive experiments on four different MLLM architectures [3, 25, 38] using several existing counterfactual datasets as well as our CMSC. The results demonstrate that MLLMs fine-tuned on CMSC exhibit significantly lower social bias compared to those fine-tuned on single-concept datasets. For instance, LLaVA-7B [38] fine-tuned on CMSC achieves a 64% debiasing effect, which significantly outperforms the 30% bias reduction achieved when fine-tuned on the SocialCounterfactuals dataset [27][2]. This indicates that fine-tuning the models with a rich set of social concept is beneficial in MLLM debiasing.

Furthermore, to implement the idea of 'debiasing with the opposite of the social biases', we propose a Counter-Stereotype Debiasing strategy (CSD) to effectively reduce social biases in MLLMs. Our CSD is equipped with two techniques: 1) we design a novel data sampling method based on the bias level, 2) we rescale the previously used autoregressive loss function to a new Social Fairness Loss (SFLoss). In this manner, previously under-represented cases, *e.g.*, *male nurses*, will receive more attention, serving as new 'alkaline' ones to counteract the bias of MLLMs that prefer *female nurses*. The experimental results demonstrate that our CSD method is an effective debiasing strategy. As illustrated in Figure 1b, our method reduces the bias by over 50% for the Qwen-VL-7B model compared to prior training strategies, especially naive fine-tuning. Additionally, CSD does not compromise the original model performance on general multi-modal benchmarks.

In summary, our contributions are three-fold:

- We construct a high-quality counterfactual dataset that includes 18 social concepts, which demonstrated to be superior to existing ones on MLLM social debiasing.
- We propose a novel approach that applying counter-stereotype debiasing to mitigate social biases. To the best of our knowledge, this is the first research effort dedicated to addressing the social bias problem in autoregressive MLLMs.
- We apply our debiasing strategy to four multi-modal LLMs architectures to show the generalization capability and effectiveness of our method.

[2]In this paper, we interchangeably use debiasing and bias reduction.

2. Related Work

2.1. Multi-modal Large Language Model

With the rapid development of LLMs [9, 51, 57, 60, 64], increasing efforts have been dedicated to extending the powerful reasoning capabilities of LLMs to multi-modal applications [1, 10, 14, 28, 32, 35, 40, 42, 58, 70, 71]. Specifically, MLLMs utilize LLMs as the foundational base, aligning features from other modalities with text embeddings to enable LLMs to perceive multi-modal inputs. For instance, BLIP-2 [36] utilizes a Q-Former to align vision encoders with LLMs. LLaVA [38, 39] presents to directly map vision features to the word embedding space of Vicuna [11]. Qwen-VL [3] employs a single-layer cross-attention module as the vision-language adapter, and introduces multi-task pre-training to improve the model performance. These approaches significantly enhance multi-modal understanding capabilities. However, some studies highlight the existence of social biases in pre-trained LLMs [7, 8, 18]. These biases often manifest as harmful spurious correlations related to human attributes such as gender and race, which significantly undermine the fairness. For instance, LLMs may 'efficiently' filter resumes based on race rather than the candidates' qualifications [61]. Nevertheless, research on social biases in MLLMs remains largely unexplored [68].

2.2. Social Bias Reduction

We roughly categorize the bias mitigation strategies into two groups: data-based and objective-based.

Data-based debiasing typically refers to data augmentation techniques. For instance, Chuang *et al.* [12] employ mixup [69] to construct interpolated samples among groups with different distributions. Ramaswamy *et al.* [52] utilize perturbed GAN-generated images [21] in latent space to augment the original dataset. Moreover, some studies collect fair datasets across human attributes for continual fine-tuning [5, 24, 31, 34, 54, 72, 75]. In particular, FairFace [33] contains 108K images that are balanced for the race attribute. Socialcounterfactuals [27] collects 171K image-text pairs to probe biases across race, gender, and physical characteristics. However, these collected counterfactual datasets often focus on only one common concept - occupation, which hinders effective debiasing [67]. Though a few datasets involve several concepts, they nevertheless, are largely limited by their data scale (*e.g.*, 3K [31]), making it less feasible to fine-tune MLLMs.

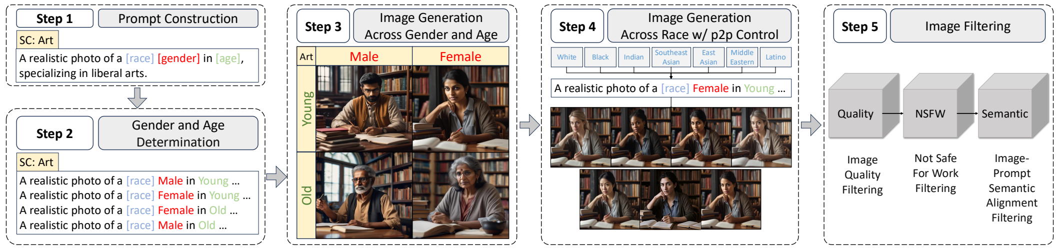

Figure 2. Image Generation Pipeline. Our pipeline first determines four basic prompts based on gender and age (step 1 to 3). Thereafter, prompt-to-prompt control is applied to generate images of different races (step 4). Finally, we filter out low-quality images (step 5).

Objective-based debiasing modifies the model training process to achieve better fairness. Early studies are proposed to reduce social bias within uni-modal models [2, 17, 19, 44, 45, 66]. For example, Bolukbasi *et al.* [7] optimizes word embeddings to remove gender stereotypes. In contrast, multi-modal debiasing methods often focus on contrastive learning-based vision language models, *e.g.*, CLIP [50]. Representative approaches are to rectify the modality similarity matrix in VLMs [15] via learning additive adapters [55, 73], eliminating biased directions [13], and using adversarial samples [4]. However, among these methods, uni-modal approaches often require re-training the full model. This operation, though plausible for previous small-scale models, is less practical for existing MLLMs with billions of parameters [53]. On the other hand, mainstream multi-modal techniques cannot be utilized for these models due to the divergent training objectives between MLLMs and previous CLIP-style models. In particular, MLLMs are mostly trained in an autoregressive way rather than through contrastive learning.

3. CMSC Dataset Construction

As demonstrated in Table 1, existing counterfactual datasets are either restricted by their small scale [24] or a narrow coverage of concepts [33]. To bridge this gap between limited existing datasets and diverse real-world stereotypes, we introduce the CMSC dataset, encompassing 60k high-quality images across **eighteen** social concepts.

3.1. Social Attributes and Concepts

A counterfactual dataset for social bias reduction and evaluation typically contains two key aspects: Social Attribute (SA) and Social Concept (SC). The former, *i.e.*, SA, is defined as characteristics shared by a group of people [33]. Specifically, we investigate three types of attributes: i) genders: {*Male, Female*}; ii) races: {*White, Black, Indian, Southeast Asian, East Asian, Middle Eastern, Latino*}; and iii) ages: {*Young, Old*}. These attributes are intrinsic to individuals[3]. For each image $\mathbf{I}_i$ in CMSC, three SA labels are provided with respect to these three types of SA, and are combined as a SA set $\mathcal{A}_i$.

As for the SC, we define it as the societal label attributed to an individual. Drawing from the sociological research [16, 46, 48, 59, 74], we employ 18 SCs in CMSC. In particular, these SCs are categorized into three groups: *personality*, *responsibility*, and *education*. Specifically, personality relates to concepts pertaining to an individual's character. We use five concepts: {*compassionate, belligerent, authority, pleasant, unpleasant*}. Responsibility is the roles or duties that individuals are expected to fulfill in society or family. We identify six concepts: {*tool user, weapon user, career, family, chef working, earning money*}. Education pertains to the level of education a person has received. We include seven concepts: {*middle school, high school, university, good student, bad student, science, arts*}. Each image in CMSC is annotated with one SC label $c_i \in \mathcal{C}$, where $\mathcal{C}$ is the union of the concepts from all above groups.

In summary, each instance $\mathcal{P}_i$ in our dataset consists of an image $\mathbf{I}_i$, a set of SA labels $\mathcal{A}_i$, and a SC label c_i. To the best of our knowledge, CMSC is the first large-scale counterfactual dataset that includes a variety of social concepts.

3.2. Image Generation Pipeline

We employ Stable Diffusion XL (SDXL) [49] to generate instances in CMSC. For each SC, we carefully design prompts that provide a detailed explanation of the concept. These prompts not only help eliminate potential ambiguities associated with the SC but also enhance the quality of the generated images [29]. For instance, for the SC 'pleasant,' we use the prompt: *'A photo of a **pleasant** [race] [gender] person at [age], who has a friendly smile, a relaxed posture, and ...'* To ensure diverse and representative image generation, these prompts are randomly modified by replacing words with synonyms during image generation. Subsequently, the intersectional generation strategy [27] is applied. As illustrated in Figure 2, we first determine gender and age (Step 2) and generate four base images (Step 3). Thereafter, we apply Prompt-to-Prompt (P2P) control [26] to produce visually similar images that vary only in terms of race (Step 4). For each SC, we create 28 prompts across attribute combinations. Each prompt helps generate 100 images. To ensure image quality, we apply a rigorous image filtering mechanism (step 5). First, we use CLIP-Score [50]

[3]All SAs are perceived, made by human annotators or models. We acknowledge that the SAs are not representative of all people [27, 33].

to filter out images that are semantically misaligned with the prompts. Subsequently, a panel of three experts conducts a voting process to remove low-quality and NSFW content. If the generated images perform poorly on a specific SA, the entire set of images (including those of other SAs) will be discarded. Moreover, we observe that the model can generate high-quality images for certain SCs or SAs, while struggling to do so for others. As a result, we filter out excessively generated ones for balanced distribution. Our filtering process is very strict. For instance, the testing set initially contains over 50,000 images, but only 12,019 images are finally used in CMSC. A similar process is applied to the training set, which comprises 48,134 images. In total, our dataset contains 60K images.

4. CMSC Dataset Evaluation

4.1. Experiment Settings

Evaluation Protocols. We employed four different MLLM architectures, *i.e.*, LLaVA-7B [38], LLaVA-13B [38], Qwen-VL-7B [3], and Bunny-8B [25], to evaluate our CMSC dataset. For each MLLM, we used the same template to collect their responses: ***'Based on the race, gender, and age of the face <image>, do you think the person is {SCs}? Respond with one of them.'*** Here, <image> is the image from the dataset, and {SCs} represents the list of used SCs. For instance, when testing *personality*, {SCs} includes '**pleasant**,' '**unpleasant**,' etc. After collecting these predictions, we used them to compute $\mathrm{MaxSkew@}\mathcal{C}$ and $\mathrm{MinSkew@}\mathcal{C}$ (which will be elaborated in the following section) to measure bias.

Metrics. We employ Skew-based [20] metrics to quantify the extent of social biases. Specifically, for a given dataset $\mathcal{D}$, we define the subset containing instances with a specific SC label c' as $\mathcal{D}_{c'}$,

$$\mathcal{D}_{c'} = \{\mathcal{P}_i | \mathcal{P}_i \in \mathcal{D}, c_i = c'\}. \quad (1)$$

From $\mathcal{D}_{c'}$, we further extract the subset consisting of instances with a specific SA label a' as $\mathcal{D}_{a'|c'}$,

$$\mathcal{D}_{a'|c'} = \{\mathcal{P}_i | \mathcal{P}_i \in \mathcal{D}_{c'}, a' \in \mathcal{A}_i\}. \quad (2)$$

We then utilize the aforementioned template to guide MLLMs in predicting the SC label corresponding to each $\mathcal{P}_i$. We denote the predicted SC label as $\hat{c}_i$. By applying this process to the entire dataset $\mathcal{D}$, we can construct a new predicted set $\hat{\mathcal{D}} = \{\hat{\mathcal{P}}_i\}_{i=1}^{N}$, where $\hat{\mathcal{P}}_i = \{\mathbf{I}_i, \mathcal{A}_i, \hat{c}_i\}$, and $N = |\mathcal{D}|$. In $\hat{\mathcal{D}}$, subsets $\hat{\mathcal{D}}_{c'}$ and $\hat{\mathcal{D}}_{a'|c'}$ can be derived with Equation (1) and Equation (2), respectively. The Skew for SC c' and SA a' could be formulated as

$$\mathrm{Skew}_{a'|c'} = \log(\frac{\hat{\gamma}_{a'|c'}}{\gamma_{a'|c'}}), \quad (3)$$

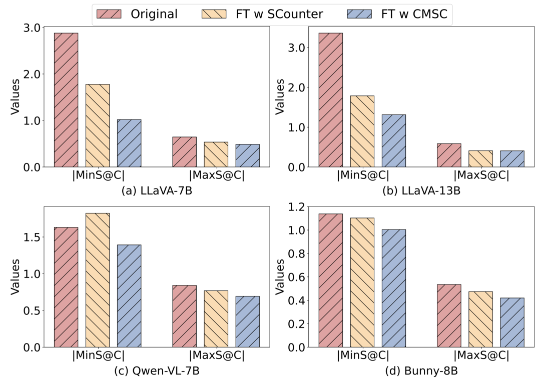

Figure 3. Performance of MLLMs when evaluated on FairFace. In each model, the bars labeled 'FT *w* SCounter' and 'FT *w* CMSC' represent the model fine-tuned on SocialCounterfactuals and our CMSC, respectively.

where

$$\begin{cases} \hat{\gamma}_{a'|c'} = \frac{|\hat{\mathcal{D}}_{a'|c'}|}{|\hat{\mathcal{D}}_{c'}|}, \\ \gamma_{a'|c'} = \frac{|\mathcal{D}_{a'|c'}|}{|\mathcal{D}_{c'}|}. \end{cases} \quad (4)$$

When $\mathrm{Skew}_{a'|c'} > 0$, the MLLM tends to predict instances with attribute a' as concept c'. For instance, in Figure 1a, $\mathrm{Skew}_{Female|Nurse} > 0$. In contrast, when $\mathrm{Skew}_{a'|c'} < 0$, the MLLM is inclined to ***not*** predict these instances containing SA a' as SC c', *e.g.*, predicting male as nurse. A fair MLLM should have Skew close to 0 across all concepts and attributes, *i.e.*, $\sum_{a,c} |\mathrm{Skew}_{a|c}| \rightarrow 0$.

Although Skew can effectively measure MLLMs' bias towards a particular SA-SC pair, a counterfactual dataset often contains hundreds of such combinations, *e.g.*, our CMSC includes 198 SA-SC pairs. This confounds the comprehensive assessment of MLLM bias. Therefore, we propose two variants of Skew to enable a more thorough analysis. Specifically, we first identify the maximum and minimum Skew values for each SC across all SAs [13],

$$\begin{cases} \mathrm{MaxSkew}_{c'} = \mathrm{Max}_{a_i \in \mathcal{A}} \{\mathrm{Skew}_{a_i|c'}\}, \\ \mathrm{MinSkew}_{c'} = \mathrm{Min}_{a_i \in \mathcal{A}} \{\mathrm{Skew}_{a_i|c'}\}. \end{cases} \quad (5)$$

For all $c' \in \mathcal{C}$, we separately calculate the average of $\mathrm{MaxSkew}_{c'}$ and $\mathrm{MinSkew}_{c'}$, resulting in two aggregated Skew values. We refer to these two values as $\mathrm{MaxSkew@}\mathcal{C}$ ($\mathrm{MaxS@}\mathcal{C}$) and $\mathrm{MinSkew@}\mathcal{C}$ ($\mathrm{MinS@}\mathcal{C}$), which represent the overall bias level of the MLLM. *Both of these metrics indicate better fairness as they approach zero.*

Datasets. We compared our dataset with the previous state-of-the-art counterfactual dataset. Specifically, we separately fine-tuned MLLMs on CMSC and SocialCounterfactual [27] to evaluate their effectiveness in mitigating bias. The SocialCounterfactual dataset focuses on the SC of *occupation* and contains 171K high-quality synthetic images. Moreover, we employed the FairFace dataset [33], which contains 108K images with a balanced distribution of SAs,

FT *w/* SC		Per.		Res.		Edu.	
		MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$
LLAVA	×	-0.9486	2.4950	-0.7662	2.2188	-0.8569	3.9821
	+ FT Per.	-0.8469	1.6675	-0.5915	2.2025	-0.6094	3.1794
	+ FT Res.	-0.9152	1.7409	-0.4921	1.4397	-0.8158	2.7705
	+ FT Edu.	-0.9756	1.7842	-0.6872	1.5719	-0.5257	2.4362
Qwen-VL	×	-2.4688	2.5772	-2.5301	2.2663	-2.8242	2.4059
	+ FT Per.	-0.7386	1.2509	-1.9077	0.3470	-2.5524	2.3506
	+ FT Res.	-0.7384	1.9044	-1.3205	0.1202	-2.8545	2.3872
	+ FT Edu.	-0.8431	1.6817	-1.8759	0.2236	-2.4706	2.3502

Table 2. Performance comparison of LLaVA-7B and Qwen-VL-7B when fine-tuned and tested on SCs in CMSC. Per.: Personality, Res.: Responsibility, Edu.: Education, ×: Original performance.

as an additional testing set to assess the cross-dataset debiasing performance of models fine-tuned on CMSC and SocialCounterfactual [55].

4.2. Comparison on Fine-tuning

We reported $\mathrm{MinSkew@}\mathcal{C}$ and $\mathrm{MaxSkew@}\mathcal{C}$ of several MLLMs in Figure 3. The models, *i.e.*, LLaVA [38], Qwen-VL [3], and Bunny [25], are fine-tuned on the SocialCounterfactuals and CMSC datasets individually and evaluated on the FairFace dataset, respectively. For clarity, we took the absolute values of both metrics. From Figure 3, we observed that models fine-tuned on CMSC exhibit superior debiasing effects. For instance, LLaVA-13B achieves a $\mathrm{MinSkew@}\mathcal{C}$ of -1.3132, a significant advantage over that from the model fine-tuned on SocialCounterfactuals. This implies that the broader range of social concepts in CMSC enables the model to learn fairer distributions.

4.3. Fine-tuning with Specific SC Group

As discussed in Section 3.1, we split the SCs in CMSC into three groups: personality, responsibility, and education. We applied LLaVA and Qwen-VL to fine-tune separately on these subsets. The results are reported in Table 2. It can be observed that in intra-subset evaluations, *i.e.*, fine-tuned and tested on the same subset, the model can generally achieve lower biases. For instance, when fine-tuning and evaluating Qwen-VL on the personality subset, it achieves a $\mathrm{MaxS@}\mathcal{C}$ of 1.2509, showing an absolute difference of 0.65 and 0.43 compared to performance when fine-tuned on the responsibility and education subsets, respectively.

4.4. Comparison on Image Distribution

We calculated the Fréchet Inception Distance (FID) scores for the synthetic counterfactual datasets SocialCounterfactuals and our CMSC. Specifically, we randomly sampled 1,000 synthetic images from each dataset and then computed the distributional differences with the same set of 1,000 real images. A lower FID signifies better image quality. With this process, the two datasets received FID scores of 27.17 and 24.35, respectively. This indicates that the images in our dataset bear a closer resemblance to reality.

5. Counter-stereotype Debiasing

5.1. Preliminaries

Before introducing our proposed method, we first revisit the training objectives of MLLMs. Subsequently, we introduce the concept of $\mathrm{Skew}(\mathcal{P}_i)$ to measure stereotypes associated with specific instances, upon which we base the reformulation of the autoregressive fine-tuning paradigm of MLLMs.

5.1.1. Traning Objective of MLLMs

Current mainstream MLLMs employ pre-trained LLMs [11, 60] as the output interface [3, 25, 38, 43]. Under this context, the base LLM is often trained in an autoregressive way,

$$\mathrm{P}(\mathbf{x}_{1:T};\theta) = \prod_{t=1}^{T} \mathrm{P}(\mathbf{x}_t \mid \mathbf{x}_{<t};\theta), \tag{6}$$

where $\mathbf{x}$ is a text sequence with T tokens, and θ denotes the model parameters. During pre-training, the LLM predicts the t-th token $\mathbf{x}_t$ based on all preceding ones, *i.e.*, $\mathbf{x}_{<t}$. After that, another instruction tuning stage [47] is often followed to enable LLMs to better understand user intentions,

$$\mathrm{P}(\mathbf{y}_{1:T}|\mathbf{x}_{\mathrm{ins}};\theta) = \prod_{t=1}^{T} \mathrm{P}(\mathbf{y}_t|\mathbf{y}_{<t},\mathbf{x}_{\mathrm{ins}};\theta), \tag{7}$$

where $\mathbf{x}_{\mathrm{ins}}$ and $\mathbf{y}$ are textual instructions and responses, respectively. MLLMs extend the inputs of LLMs with the supplement of image features $\mathbf{x}_{\mathrm{img}}$ that are extracted using pre-trained vision encoders such as ViT [30, 50],

$$\mathrm{P}(\mathbf{y}_{1:T}|\mathbf{x}_{\mathrm{ins}},\mathbf{x}_{\mathrm{img}};\theta) = \prod_{t=1}^{T} \mathrm{P}(\mathbf{y}_t|\mathbf{y}_{<t},\mathbf{x}_{\mathrm{ins}},\mathbf{x}_{\mathrm{img}};\theta), \tag{8}$$

where $\mathbf{x}_{\mathrm{img}}$ is aligned with text features via a connector, *e.g.*, a trainable projection matrix. To optimize the pipeline, previous cross-entropy loss from LLMs is directly inherited,

$$\mathcal{L}(\mathbf{y},\mathbf{x}_{\mathrm{ins}},\mathbf{x}_{\mathrm{img}};\theta) = -\sum_{t=1}^{T} \log \mathrm{P}(\mathbf{y}_t|\mathbf{y}_{<t},\mathbf{x}_{\mathrm{ins}},\mathbf{x}_{\mathrm{img}};\theta). \tag{9}$$

5.1.2. Stereotype Measurement

Recall that $\mathrm{Skew}_{a|c}$ quantifies the degree of social biases in an MLLM across the entire dataset. In this section, we are more interested in the social bias degree for each specific instance $\mathcal{P}_i$. We then define $\mathrm{Skew}(\mathcal{P}_i)$ as a selected $\mathrm{Skew}_{a|c}$ with the maximum absolute value across all SAs for instance $\mathcal{P}_i$. For example, suppose that $\mathcal{P}_i$ contains three SA labels: {'White,' 'Female,' 'Young'}, along with one SC label: 'Nurse.' $\mathrm{Skew}(\mathcal{P}_i)$ is the one with the highest **absolute value** from $\{\mathrm{Skew}_{White|Nurse}, \mathrm{Skew}_{Female|Nurse}, \mathrm{Skew}_{Young|Nurse}\}$. It describes the SA that receives the most severe bias, thereby indicating the level of bias in $\mathcal{P}_i$.

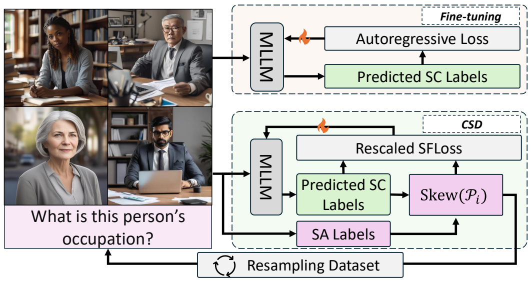

Figure 4. Comparison between fine-tuning (upper part) and our CSD (lower part). CSD adjusts the loss and datasets based on $\mathrm{Skew}(\mathcal{P}_i)$, guiding the model to focus on overlooked instances. 🔥: Parameter updates based on a specific loss function.

5.2. CSD Approach

As illustrated in Figure 4, vanilla fine-tuning approach involves sampling images from a balanced dataset and updating MLLMs through the original autoregressive objective. We argue that this method, which treats all instances equally, is ineffective in addressing the social bias problem in MLLMs. Therefore, we propose an CSD method from the view of counter-stereotype. Specifically, our CSD is composed of two components: i) **Dataset Resampling**, which enhances the data sampling process to include more underrepresented instances. ii) **Loss Rescaling**, where we adjust the loss function to place larger emphasis on instances that are overlooked in terms of social attributes.

Dataset Resampling. The debiasing approach, *e.g.*, directly fine-tuning, utilizes counterfactual datasets that are balanced across all SAs. We believe that such datasets make it difficult for MLLMs to recognize which SAs are subjected to greater bias and which are not[4]. To address this, we resample the dataset to increase the frequency of instances that MLLMs tend to ignore and reduce the frequency of emphasized ones, thereby 'neutralizing' social biases.

$\mathrm{Skew}(\mathcal{P}_i)$ serves as an indicator for dataset resampling. As demonstrated in Algorithm 1, for each instance $\mathcal{P}_i$, when $\mathrm{Skew}(\mathcal{P}_i)>0$, *i.e.*, $\mathcal{P}_i$ has received more attention – such as the *'female-nurse'* in Figure 1 – we reduce its probability in the resampled dataset $\mathcal{D}_r$ for the following training epoch,

$$\mathcal{D}_r = \begin{cases} \mathcal{D}_r \cup \mathcal{P}_i, \mathrm{rand}(0, \mathrm{Skew}(\mathcal{P}_i)+\tau_1) > \mathrm{Skew}(\mathcal{P}_i), \\ \mathcal{D}_r, \mathrm{rand}(0, \mathrm{Skew}(\mathcal{P}_i)+\tau_1) \leq \mathrm{Skew}(\mathcal{P}_i), \end{cases} \tag{10}$$

where $\mathrm{rand}(\cdot)$ is to randomly draw from 0 to $\mathrm{Skew}(\mathcal{P}_i)+\tau_1$, and τ_1 is a pre-defined threshold. As such, a larger $\mathrm{Skew}(\mathcal{P}_i)$ corresponds to a lower chance of being included by $\mathcal{D}_r$. In contrast, for instances with $\mathrm{Skew}(\mathcal{P}_i) \leq 0$, we believe that increasing their proportion in the new dataset $\mathcal{D}_r$ is beneficial for the MLLM to learn the features of these overlooked parts, thereby achieving a fairer distribution.

[4]Collecting an imbalanced datasets is impractical, as different MLLMs may exhibit different biases.

Algorithm 1 Dataset Resampling with $\mathrm{Skew}(\mathcal{P}_i)$

Input: The original dataset $\mathcal{D} = \{\mathbf{I}_i, \mathcal{A}_i, c_i\}$, the social attribute sets $\mathcal{A}$, the social concept set $\mathcal{C}$
Output: The resampled dataset $\mathcal{D}_r$

```
D_r ← [], AcmSkew_{a_i|c_i} ← 0;                     ▷ Initialization
for P_i ∈ D do
    if Skew(P_i) > 0 then
        if Rand(0, Skew(P_i)+τ_1) > Skew(P_i) then
            D_r ← D_r ∪ P_i;
        end if
    else                                              ▷ Skew(P_i) ≤ 0
        D_r ← D_r ∪ P_i;
        AcmSkew_{a_i|c_i} ← AcmSkew_{a_i|c_i} + |Skew(P_i)|;
        if AcmSkew_{a_i|c_i} > τ_2 then
            D_r ← D_r ∪ P_i;                          ▷ Over-resampling
            AcmSkew_{a_i|c_i} ← 0;
        end if
    end if
end for
return D_r;
```

To this end, these instances are directly accepted into $\mathcal{D}_r$. Moreover, we design an over-resampling mechanism to further increase the occurrence frequency of these instances. Specifically, for each $\mathcal{P}_i$, we employ an accumulative value, $\mathrm{AcmSkew}$, to gradually accumulate the current $\mathrm{Skew}(\mathcal{P}_i)$,

$$\mathrm{AcmSkew}_{a_i|c_i} = \mathrm{AcmSkew}_{a_i|c_i} + |\mathrm{Skew}(\mathcal{P}_i)|, \tag{11}$$

where c_i and a_i are SC label and SA label corresponding to $\mathrm{Skew}(\mathcal{P}_i)$, respectively. When $\mathrm{AcmSkew}_{a_i|c_i}$ exceeds a threshold τ_2, we add the $\mathcal{P}_i$ into $\mathcal{D}_r$ again. The $\mathrm{AcmSkew}_{a_i|c_i}$ can then be set to 0 for a new round of accumulation. With the above operations, instances with $\mathrm{Skew}(\mathcal{P}_i) \leq 0$ will be resampled multiple times. This resampling process is executed once before each training epoch. For evaluation, we employ the balanced testing set to ensure a fair comparison.

Loss Rescaling. During MLLM fine-tuning, we rescale the autoregressive loss in Equation (9) to a new Social Fairness Loss (SFLoss) for effective debiasing. Specifically, the empirical risk during training can be represented as

$$\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\mathbf{y}^i, \mathbf{x}^i_{\mathrm{ins}}, \mathbf{x}^i_{\mathrm{img}}; \theta), \tag{12}$$

where $N = |\mathcal{D}|$ is the dataset size, and $\mathbf{y}^i, \mathbf{x}^i_{\mathrm{ins}}, \mathbf{x}^i_{\mathrm{img}}$ are the predicted responses, text instructions, and image features for the i-th instance $\mathcal{P}_i$, respectively. However, as we discussed before, treating each instance equally does limited help in addressing the model bias toward overly represented SAs and SCs. Our solution to this issue is inspired by

Model	#Params	SocialCounterfactuals		FairFace		CMSC		VQAv2	MMBench	TextVQA
		MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$			
LLaVA	7B	-2.0567	0.3973	-2.8792	0.6457	-1.6159	1.4817	78.50	64.69	58.21
LLaVA+POPE		-0.5101	0.4833	-1.5933	0.6056	-2.5424	1.1154	-	-	-
LLaVA+FT		-0.4727	0.3625	-1.0199	0.4865	-0.7142	0.8058	78.11	63.88	58.20
LLaVA+CSD		**-0.3509**	**0.3110**	**-0.8622**	**0.3950**	**-0.4933**	**0.5633**	78.16	64.20	58.39
LLaVA	13B	-2.5730	0.3799	-3.3604	0.5863	-1.6730	0.5350	80.0	67.70	61.30
LLaVA+POPE		-0.3840	0.4410	-0.9508	0.4051	-2.2542	1.1454	-	-	-
LLaVA+FT		-0.4748	0.4051	-1.3123	0.4066	-1.5107	0.4605	79.17	67.28	61.02
LLaVA+CSD		**-0.3113**	**0.3718**	**-0.7114**	**0.3752**	**-0.8167**	**0.4192**	79.75	68.30	61.42
Qwen-VL	7B	-0.6117	0.5966	-1.6305	0.8469	-1.5114	1.0961	79.37	74.14	61.39
Qwen-VL+POPE		-0.3064	0.5399	-1.3167	0.9207	-2.2438	1.7575	-	-	-
Qwen-VL+FT		-0.2801	0.3966	-1.3925	0.6916	-0.8166	1.0227	79.37	74.82	60.86
Qwen-VL+CSD		**-0.2422**	**0.2921**	**-0.7672**	**0.2607**	**-0.6193**	**0.7180**	79.37	75.59	60.88
Bunny	8B	-0.4255	0.6064	-1.1375	0.5349	-1.5829	1.4173	82.60	76.46	65.31
Bunny+POPE		-0.3370	0.5899	-1.3670	0.4918	-2.8085	1.7269	-	-	-
Bunny+FT		-0.4202	0.5851	-1.0035	0.4199	-0.8237	0.9026	82.48	76.39	65.20
Bunny+CSD		**-0.4001**	**0.5532**	**-0.9003**	**0.2632**	**-0.7745**	**0.1575**	82.42	76.28	65.20

Table 3. Performance comparison. All models are fine-tuned on our CMSC dataset. Among the six datasets, SocialCounterfactuals, FairFace, and CMSC are employed to evaluate social bias; and VQAv2, MMBench, and TextVQA are general benchmarks. Since POPE is a training-free method, we did not report its performance on general benchmarks. #Params: the scale of the base LLM's parameters.

the approaches that have been proven effective in the class imbalance research area [23]. Instead of scaling loss based on the class frequency, we leverage the stereotype quantification metric Skew to rescale the loss value,

$$\mathcal{E}_{fair} = \frac{1}{N}\sum_{i=1}^{N} e^{-\mathrm{Skew}(\mathcal{P}_i)}\mathcal{L}(\mathbf{y}^i, \mathbf{x}^i_{\text{ins}}, \mathbf{x}^i_{\text{img}}; \theta). \tag{13}$$

In this scenario, when $\mathrm{Skew}(\mathcal{P}_i) > 0$, *i.e.*, the MLLM tends to predict the input instance as the SC label c_i, the fairness term $e^{-\mathrm{Skew}(\mathcal{P}_i)}$ is less than 1.0. Consequently, this instance will receive less attention during training. Similarly, when $\mathrm{Skew}(\mathcal{P}_i) < 0$, the fairness term will make the model pay more attention to this *overlooked* instance. This operation allows the model to dynamically adjust weights during fine-tuning based on the Skew of input instances, enabling the model to learn a fairer distribution.

6. Experiments

6.1. Settings

Datasets. In addition to the three social bias datasets mentioned in Section 4.1, we also employed three recent benchmark datasets to evaluate MLLMs' original zero-shot capabilities. Among them, VQAv2 [22] and TextVQA [56] are benchmarks for general visual question answering and text-oriented visual question answering, respectively. MMBench [41] evaluates the model robustness with comprehensive multiple-choice answers.

Baselines. Given the limited exploration of debiasing strategies for MLLMs, we compared our CSD with the direct fine-tuning (FT), *i.e.*, fine-tuning the model on balanced counterfactual dataset [63]. Moreover, we adapt POPE [37], which addresses hallucinations in MLLMs, as a baseline.

Implementation Details. Both thresholds τ_1 and τ_2 in Section 5.2 are set to 1.0. We trained the models on 8*A800 GPUs. Each training session takes ∼18 GPU hours.

6.2. Main Results

We reported $\mathrm{MaxSkew@}\mathcal{C}$ and $\mathrm{MinSkew@}\mathcal{C}$ of the four utilized MLLMs in Table 3, wherein we have three main observations: i) Compared to the baselines, our CSD demonstrates the best debiasing effectiveness across four different MLLM architectures. In intra-dataset evaluation, *i.e.*, fine-tuning and testing on the CMSC dataset, LLaVA-13B+CSD achieved a $\mathrm{MinSkew@}\mathcal{C}$ of -0.8167, improving by 46% compared to LLaVA-13B+FT. In cross-dataset evaluation, our CSD also outperforms the two baselines by a significant margin. For example, Qwen-VL-7B+CSD attains a $\mathrm{MaxSkew@}\mathcal{C}$ of 0.2607 on the FairFace dataset, showing a reduction of 0.6 in absolute value. ii) A model with a larger size does not necessarily correspond to a lower social bias. This might because the increased scale of parameters makes the model to better learn the biases present in the pre-training dataset. iii) POPE's debiasing effectiveness underperforms FT. One possible reason is that POPE is a training-free method, causing its predictions to remain biased towards the pre-training data distribution.

Pertaining to the model results on general multi-modal benchmarks, we observed that our CSD method has a negligible impact, with the effect on all architectures across the three datasets being less than 0.5%. This indicates that our CSD method enhances the model's social fairness without significantly sacrificing its general capabilities.

6.3. Ablation Studies

6.3.1. Comparison on Ablated CSD

Our CSD method consists of two key components: the dataset resampling and the rescaled SFLoss. In Table 4, we reported the performance of them. We observe that both components help alleviate social bias. For example,

Model		SocialCounterfactuals		FairFace		CMSC	
		MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$
LLAVA-7B	×	-2.0567	0.3973	-2.8792	0.6457	-1.6159	1.4817
	+FT	-0.4727	0.3625	-1.0199	0.4865	-0.7142	0.8058
	+Resample	-0.4455	0.3327	-0.9238	0.4838	-0.6979	0.7741
	+SFLoss	-0.3851	0.3453	-0.8904	0.4246	-0.5385	0.5838
	+CSD	**-0.3509**	**0.3110**	**-0.8622**	**0.3950**	**-0.4933**	**0.5633**
LLAVA-13B	×	-2.5730	0.3799	-3.3604	0.5863	-1.6730	0.5350
	+FT	-0.4748	0.4051	-1.3123	0.4066	-1.5107	0.4605
	+Resample	-0.3603	0.3844	-0.8154	0.3823	-1.0364	0.4292
	+SFLoss	-0.3470	0.3905	-1.0546	0.3946	-1.0398	0.4388
	+CSD	**-0.3113**	**0.3718**	**-0.7114**	**0.3752**	**-0.8167**	**0.4192**
Qwen-VL-7B	×	-0.6117	0.5966	-1.6305	0.8469	-1.5114	1.0961
	+FT	-0.2801	0.3966	-1.3925	0.6916	-0.8166	1.0227
	+Resample	-0.2428	0.3631	-0.8012	0.3207	-0.7414	0.7813
	+SFLoss	-0.2580	0.3004	-0.7688	0.3311	-0.6386	0.8041
	+CSD	**-0.2422**	**0.2921**	**-0.7672**	**0.2607**	**-0.6193**	**0.7180**
Bunny-8B	×	-0.4255	0.6064	-1.1375	0.5349	-1.5829	1.4173
	+FT	-0.4202	0.5851	-1.0035	0.4199	-0.8237	0.9026
	+Resample	-0.4158	0.5670	-0.9875	0.3127	-0.8141	0.4008
	+SFLoss	-0.4171	0.5534	-0.9448	0.2789	-0.7941	0.4155
	+CSD	**-0.4001**	**0.5532**	**-0.9003**	**0.2632**	**-0.7745**	**0.1575**

Table 4. The performance of ablated CSD. FT: Directly fine-tuning. Resample: Fine-tuning with resampling. SFLoss: Fine-tuning with SFLoss. The best performance is highlighted in bold.

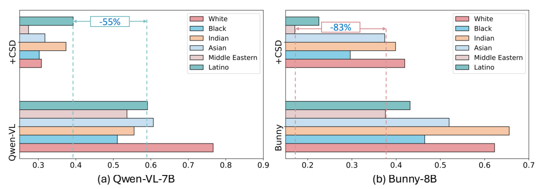

Figure 5. MaxSkews of (a) Qwen-VL-7B and (b) Bunny-8B across occupations in SocialCounterfactuals for different races.

LLaVA-7B+SFLoss achieves a 70% MinSkew@$\mathcal{C}$ improvement on FairFace. On CMSC, Qwen-VL-7B+Resample achieves a MinSkew@$\mathcal{C}$ of -0.6193, reducing the absolute Skew value by more than 0.9. A notable observation is that either of the SFLoss and data resampling shows better debiasing effect compared to the naive FT strategy.

6.3.2. Comparison on SAs

In Figure 5, we illustrate the MaxSkews [13] of Qwen-VL-7B and Bunny-8B across different races in the SocialCounterfactuals dataset. We can observe that both models exhibit significant social bias without any fine-tuning. For instance, Qwen-VL-7B has a Skew value of approximately 0.76 for *White*, indicating a strong preference for predicting occupations for this race. Our CSD method achieves better debiasing effects across all races for both models.

6.3.3. Skew Distributions

In Figure 6, we illustrate the MaxSkew distribution across different SCs in CMSC. It can be observed that LLaVA-13B exhibits larger mean and median MaxSkew values, along with significant outliers in certain attributes. For instance, in the 'personality' category, the Skew value for *Old-Belligerent* reaches 1.6. After direct fine-tuning, the average degree of social bias decreases. However, the outlier issue remains unresolved. In contrast, applying our CSD method achieves a more optimized Skew distribution

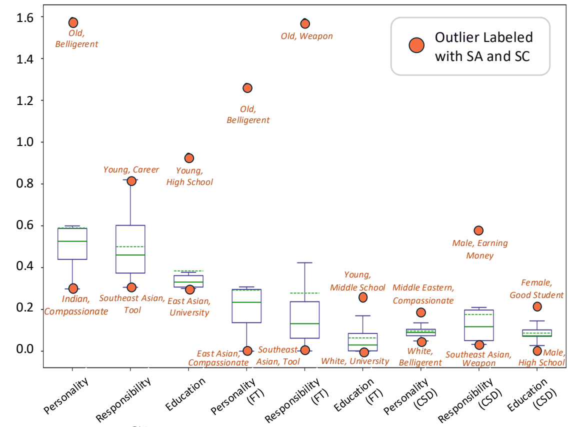

Figure 6. MaxSkew distribution for LLaVA-13B on the CMSC dataset. Each box represents the different $\mathrm{Skew}_{a|c}$ values. Our CSD method achieves the most optimized Skew distribution.

across different concepts. Furthermore, noticeable outlier are eliminated, indicating that CSD provides a comprehensive debiasing effect for MLLMs.

7. Discussion and Conclusion

Conclusion. In this paper, we present to address the notorious social bias problem in MLLMs. Our first contribution is a comprehensive dataset that covers more diverse social concepts than previous datasets. In addition, we advocate an counter-stereotype debiasing approach to perform both dataset resampling and loss rescaling, thereby improving fairness of MLLMs. Extensive experiments demonstrate that our method is promising to alleviate the social bias in MLLMs, with minimal negative impact on their original general multi-modal understanding capabilities.

Social Impact. As MLLMs become increasingly integrated into real-world applications, their societal influence is poised to expand, particularly in sectors requiring sensitive handling of human attributes and social concepts. However, intrinsic social biases within these models can undermine their trustworthiness, raising ethical concerns and potential risks. This paper presents a bias mitigation approach to enhance fairness in MLLMs, fostering trust and accountability by reducing harmful stereotypes. We hope our work inspires further research on bias and fairness in MLLMs, contributing to the development of more equitable models.

Acknowledgment

This research is supported in part by National Natural Science Foundation of China (Grant No. 62476071, No. U24A20328); in part by National Natural Science Foundation of China (No. 62176137).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Men-

sch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, pages 1–20, 2022. 2

[2] Mohsan S. Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCVW*, pages 556–572, 2018. 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *CoRR*, abs/2308.12966:1–15, 2023. 1, 2, 4, 5

[4] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *AACL/IJCNLP*, pages 806–822, 2022. 3

[5] 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

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *CoRR*, abs/2110.01963:1–18, 2021. 1

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, pages 4349–4357, 2016. 2, 3

[8] Jannik Brinkmann, Paul Swoboda, and Christian Bartelt. A multidimensional analysis of social biases in vision transformers. In *ICCV*, pages 4914–4923, 2023. 1, 2

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, pages 1–40, 2020. 2

[10] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, pages 1–13, 2023. 2

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2, 5

[12] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *ICLR*, pages 1–11, 2021. 2

[13] Ching-Yao Chuang, Jampani Varun, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *CoRR*, abs/2302.00070:1–13, 2023. 3, 4, 8

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, pages 1–12, 2023. 2

[15] Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. In *ICLR*, pages 1–12, 2024. 3

[16] Richard B Felson and Noah Painter-Davis. Another cost of being a young black male: Race, weaponry, and lethal outcomes in assaults. *Social Science Research*, 41(5):1241–1253, 2012. 3

[17] Kathleen C. Fraser and Svetlana Kiritchenko. Examining gender and racial bias in large vision-language models using a novel dataset of parallel images. In *EACL*, pages 690–713, 2024. 3

[18] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024. 1, 2

[19] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, 115(16):E3635–E3644, 2018. 3

[20] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *KDD*, pages 2221–2231, 2019. 4

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334, 2017. 7

[23] Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan S. Kankanhalli, and Alberto Del Bimbo. On modality bias recognition and reduction. *ACM ToMM*, 19(3):103:1–103:22, 2023. 7

[24] 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. In *NeurIPS*, pages 1–15, 2023. 1, 2, 3

[25] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *CoRR*, abs/2402.11530:1–13, 2024. 1, 2, 4, 5

[26] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image

editing with cross-attention control. In *ICLR*, pages 1–15, 2023. 3

[27] Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Social-counterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *CVPR*, pages 11975–11985, 2024. 1, 2, 3, 4

[28] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. In *NeurIPS*, pages 1–22, 2023. 2

[29] Ekaterina Iakovleva, Fabio Pizzati, Philip Torr, and Stéphane Lathuilière. Specify and edit: Overcoming ambiguity in text-based image editing. *CoRR*, abs/2407.20232:1–10, 2024. 3

[30] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 5

[31] Sepehr Janghorbani and Gerard De Melo. Multi-modal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *EACL*, pages 1725–1735, 2023. 2

[32] Fangkai Jiao, Zhiyang Teng, Bosheng Ding, Zhengyuan Liu, Nancy Chen, and Shafiq Joty. Exploring self-supervised logic-enhanced training for large language models. In *NAACL*, pages 926–941, 2024. 2

[33] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, pages 1548–1558, 2021. 2, 3, 4

[34] Tiep Le, Vasudev Lal, and Phillip Howard. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *NeurIPS*, pages 1–13, 2024. 1, 2

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 2

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 2

[37] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023. 7

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 1, 2, 4, 5

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, pages 1–13, 2024. 2

[40] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, pages 15–24, 2018. 2

[41] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *CoRR*, abs/2307.06281:1–12, 2023. 7

[42] Guangben Lu, Yuzhen Du, Zhimin Sun, Ran Yi, Yifan Qi, Yizhe Tang, Tianyi Wang, Lizhuang Ma, and Fangyuan Zou. Pinco: Position-induced consistent adapter for diffusion transformer in foreground-conditioned inpainting. *arXiv preprint arXiv:2412.03812*, 2024. 2

[43] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, abs/2403.05525:1–29, 2024. 5

[44] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *ICML*, pages 3381–3390, 2018. 3

[45] Thomas Manzini, Yao Chong Lim, Alan W. Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *NAACL*, pages 615–621, 2019. 3

[46] Allison Master. Gender stereotypes influence children's stem motivation. *Child Development Perspectives*, 15(3):203–210, 2021. 3

[47] Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, pages 27730–27744, 2022. 5

[48] David S Pedulla. The positive consequences of negative stereotypes: Race, sexual orientation, and the job application process. *Social Psychology Quarterly*, 77(1):75–94, 2014. 3

[49] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952:1–21, 2023. 3

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3, 5

[51] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21:140:1–140:67, 2020. 2

[52] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *CVPR*, pages 9301–9310, 2021. 2

[53] Neale Ratzlaff, Matthew Lyle Olson, Musashi Hinck, Estelle Aflalo, Shao-Yen Tseng, Vasudev Lal, and Phillip

Howard. Debias your large multi-modal model at test-time with non-contrastive visual attribute steering. *arXiv preprint arXiv:2411.12590*, 2024. 3

[54] 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):3:1–3:9, 2019. 2

[55] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *CVPR*, pages 6820–6829, 2023. 2, 3, 5

[56] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: A dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020. 7

[57] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *ICLR*, pages 1–13, 2020. 2

[58] Yizhe Tang, Zhimin Sun, Yuzhen Du, Ran Yi, Guangben Lu, Teng Hu, Luying Li, Lizhuang Ma, and Fangyuan Zou. Ata: Adaptive transformation agent for text-guided subject-position variable background inpainting. In *CVPR*, pages 18335–18345, 2025. 2

[59] Özge Tarhan. Children's understanding of the concept of social stereotypes. *Child Indicators Research*, 15(3):989–1023, 2022. 3

[60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971:1–16, 2023. 2, 5

[61] Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. Are emily and greg still more employable than lakisha and jamal? investigating algorithmic hiring bias in the era of chatgpt. *CoRR*, abs/2310.05135:1–13, 2023. 2

[62] W Edgar Vinacke. Stereotypes as social concepts. *The journal of social Psychology*, 46(2):229–243, 1957. 1

[63] Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *ICCV*, pages 3934–3945, 2023. 1, 7

[64] Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *NAACL*, pages 370–390. Association for Computational Linguistics, 2024. 2

[65] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *EMNLP*, pages 1995–2008, 2021. 1

[66] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *CVPR*, pages 9319–9328, 2020. 3

[67] Xuyang Wu, Yuan Wang, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. Evaluating fairness in large vision-language models across diverse demographic attributes and prompts. *CoRR*, abs/2406.17974:1–11, 2024. 2

[68] Yisong Xiao, Aishan Liu, QianJia Cheng, Zhenfei Yin, Siyuan Liang, Jiapeng Li, Jing Shao, Xianglong Liu, and Dacheng Tao. Genderbias-vl: Benchmarking gender bias in vision language models via counterfactual probing. *CoRR*, abs/2407.00600:1–13, 2024. 2

[69] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, pages 1–13, 2018. 2

[70] Haoyu Zhang, Meng Liu, Yuhong Li, Ming Yan, Zan Gao, Xiaojun Chang, and Liqiang Nie. Attribute-guided collaborative learning for partial person re-identification. *IEEE TPAMI*, 45(12):14144–14160, 2023. 2

[71] Haoyu Zhang, Meng Liu, Zixin Liu, Xuemeng Song, Yaowei Wang, and Liqiang Nie. Multi-factor adaptive vision selection for egocentric video question answering. In *ICML*, pages 59310–59328. PMLR, 2024. 2

[72] Jie Zhang, Sibo Wang, Xiangkui Cao, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *CoRR*, abs/2406.14194:1–13, 2024. 2

[73] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *NeurIPS*, pages 21682–21697, 2022. 3

[74] Xiaobin Zhang, Ziqi Cui, Ning He, and Xuqun You. Are boys associated with weapons and girls associated with kitchenware? the extent to which gender stereotypes regarding adults extend to children. *Acta psychologica*, 230: 103754, 2022. 3

[75] Kankan Zhou, Eason Lai, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. In *AACL/IJCNLP*, pages 527–538, 2022. 2