Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search

Jialu Wang, Yang Liu, Xin Eric Wang
Department of Computer Science and Engineering
University of California, Santa Cruz
{faldict,yangliu,xwang366}@ucsc.edu

Abstract

Internet search affects people's cognition of the world, so mitigating biases in search results and learning fair models is imperative for social good. We study a unique gender bias in image search in this work: the search images are often gender-imbalanced for gender-neutral natural language queries. We diagnose two typical image search models, the specialized model trained on in-domain datasets and the generalized representation model pre-trained on massive image and text data across the internet. Both models suffer from severe gender bias. Therefore, we introduce two novel debiasing approaches: an in-processing fair sampling method to address the gender imbalance issue for training models, and a post-processing feature clipping method base on mutual information to debias multimodal representations of pre-trained models. Extensive experiments on MS-COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) benchmarks show that our methods significantly reduce the gender bias in image search models.

1 Introduction

Internet information is shaping people's minds. The algorithmic processes behind modern search engines, with extensive use of machine learning, have great power to determine users' access to information (Eslami et al., 2015). These information systems are biased when results are systematically slanted in unfair discrimination against protected groups (Friedman and Nissenbaum, 1996).

Gender bias is a severe fairness issue in image search. Figure 1 shows an example: given a gender-neutral natural language query "a person is cooking", only 2 out of 10 images retrieved by an image search model (Radford et al., 2021) depict females, while equalized exposure for male and female is expected. Such gender-biased search results are harmful to society as they change people's cognition and worsen gender stereotypes (Kay et al., 2015). Mitigating gender bias in image search is imperative for social good.

In this paper, we formally develop a framework for quantifying gender bias in image search results, where text queries in English[1] are made gender-neutral, and gender-balanced search images are expected for models to retrieve. To evaluate model fairness, we use the normalized difference between masculine and feminine images in the retrieved results to represent gender bias. We diagnose the gender bias of two primary families of multimodal models for image search: (1) the specialized models that are often trained on in-domain datasets to perform text-image retrieval, and (2) the general-purpose representation models that are pre-trained on massive image and text data available online and can be applied to image search. Our analysis on MS-COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) datasets reveals that both types of models lead to serious gender bias issues (e.g., nearly 70% of the retrieved images are masculine images).

To mitigate gender bias in image search, we propose two novel debiasing solutions for both model families. The specialized in-domain training methods such as SCAN (Lee et al., 2018) often adopt contrastive learning to enforce image-text matching by maximizing the margin between positive and negative image-text pairs. However, the gender distribution in the training data is typically imbalanced, which results in unfair model training. Thus we introduce a fair sampling (*FairSample*) method to alleviate the gender imbalance during training without modifying the training data.

Our second solution aims at debiasing the large, pre-trained multimodal representation models, which effectively learn pre-trained image and text representations to accomplish down-stream applications (Bachman et al., 2019; Chen et al., 2020a,c; Gan et al., 2020; Chen et al., 2020d; Rad-

[1]This study is conducted on English corpora. We will assume the text queries are all English queries hereafter.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1995–2008
November 7–11, 2021.

Figure 1: Gender bias in image search. We show the top-10 retrieved images for searching "a person is cooking" on the Flickr30K (Young et al., 2014) test set using a state-of-the-art model (Radford et al., 2021). Despite the gender-neutral query, only 2 out of 10 images are depicting female cooking.

ford et al., 2021). We examine whether the representative CLIP model (Radford et al., 2021) embeds human biases into multimodal representations when they are applied to the task of image search. Furthermore, we propose a novel post-processing feature clipping approach, *clip*, that effectively prunes out features highly correlated with gender based on their mutual information to reduce the gender bias induced by multimodal representations. The *clip* method does not require any training and is compatible with various pre-trained models.

We evaluate both debiasing approaches on MS-COCO and Flickr30K and find that, on both benchmarks, the proposed approaches significantly reduce the gender bias exhibited by SCAN and CLIP models when evaluated on the gender-neutral corpora, yielding fairer and more gender-balanced search results. In addition, we evaluate the similarity bias of the CLIP model in realistic image search results for occupations on the internet, and observe that the post-processing methods mitigate the discrepancy between gender groups by a large margin.

Our contributions are four-fold: (1) we diagnose a unique gender bias in image search, especially for gender-neutral text queries; (2) we introduce a fair sampling method to mitigate gender bias during model training; (3) we also propose a novel post-processing clip method to debias pre-trained multimodal representation models; (4) we conduct extensive experiments to analyze the prevalent bias in existing models and demonstrate the effectiveness of our debiasing methods.

2 Gender Bias in Image Search

In an image search system, text queries may be either gender-neutral or gender-specific. Intuitively, when we search for a gender-neutral query like "a person is cooking", we expect a fair model returning approximately equal proportions of images depicting men and women. For gender-specific queries, an unbiased image search system is supposed to exclude images with misspecified gender information. This intention aligns with seeking more accurate search results and would be much different from the scope of measuring gender bias in gender-neutral cases. Therefore, we focus on identifying and quantifying gender bias when only searching for gender-neutral text queries.

2.1 Problem Statement

Given a text query provided by the users, the goal of an image search system is to retrieve the matching images from the curated images. In the domain of multi-modality, given the dataset $\{(v_n, c_n)\}_{n=1}^{N}$ with N image-text pairs, the task of image search aims at matching every image v based on the providing text c. We use $\mathcal{V} = \{v_n\}_{n=1}^{N}$ to denote the image set and $\mathcal{C} = \{c_n\}_{n=1}^{N}$ to denote the text set. Given a text query $c \in \mathcal{C}$ and an image $v \in \mathcal{V}$, image retrieval models often predict the similarity score $S(v, c)$ between the image and text. One general solution is to embed the image and text into a high-dimensional representation space and compute a proper distance metric, such as Euclidean distance or cosine similarity, between vectors (Wang et al.,

2014). We take cosine similarity for an example:

$$S(v,c) = \frac{\vec{v}\cdot\vec{c}}{\|\vec{v}\|\|\vec{c}\|}$$
$$\text{s.t. } \vec{v} = \mathsf{image\_encoder}(v) \quad (1)$$
$$\vec{c} = \mathsf{text\_encoder}(c)$$

The image search system outputs a set of top-K retrieved images $\mathcal{R}_K(c)$ with the highest similarity scores. In this work, we assume that when evaluating on test data, $\forall c \in \mathcal{C}$, the text query c is written in gender-neutral language.

2.2 Measuring Gender Bias

The situations of image search results are complex: there might be no people, one person, or more than one person in the images. Let $g(v) \in \{\text{male}, \text{female}, \text{neutral}\}$ represent the gender attribute of an image v. Note that in this study gender refers to biological sex Larson, 2017. We use the following rules to determine $g(v)$: $g(v)$ = male when there are only men in the image, $g(v)$ = female when there are only women in the image, otherwise $g(v)$ = neutral.

Portraits in image search results with different gender attributes often receive unequal exposure. Inspired by Kay et al. (2015) and Zhao et al. (2017), we measure gender bias in image search by comparing the proportions of masculine and feminine images in search results. Given the set of retrieved images $\mathcal{R}_K(c)$, we count the images depicting males and females

$$N_{\text{male}} = \sum_{v\in\mathcal{R}_K(c)} \mathbb{1}[g(v) = \text{male}],$$
$$N_{\text{female}} = \sum_{v\in\mathcal{R}_K(c)} \mathbb{1}[g(v) = \text{female}],$$

and define the gender bias metric as:

$$\Delta_K(c) = \begin{cases} 0, & \text{if } N_{\text{male}} + N_{\text{female}} = 0 \\ \frac{N_{\text{male}} - N_{\text{female}}}{N_{\text{male}} + N_{\text{female}}}, & \text{otherwise} \end{cases} \quad (2)$$

We don't take absolute values for measuring the direction of skewness, i.e., if $\Delta_K(c) > 0$ it skews towards males. Note that a similar definition of gender bias $\frac{N_{\text{male}}}{N_{\text{male}}+N_{\text{female}}}$ in Zhao et al. (2017) is equivalent to $(1 + \Delta(c))/2$. But our definition of gender bias considers the special case when none of the retrieved images are gender-specific, i.e., $N_{\text{male}} + N_{\text{female}} = 0$. For the whole test set, we measure the mean difference over all the text queries:

$$\text{Bias@}K = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \Delta_K(c) \quad (3)$$

3 Mitigating Gender Bias in Image Search

There are two fashions of multimodal models for the image search task. One is to build a specialized model that could embed image and text into representation vectors with measurable similarity scores. The other is to use general-purpose image-text representations pre-trained on sufficiently big data and compute a particular distance metric. We focus on two representative models, SCAN (Lee et al., 2018) and CLIP (Radford et al., 2021), for both fashions. For the first fashion, we propose an in-processing learning approach to ameliorate the unfairness caused by imbalanced gender distribution in training examples. This approach builds on contrastive learning but extends with a *fair sampling* step. The in-processing solution requires full training on in-domain data examples. For the second fashion, we propose a post-processing feature *clipping* technique to mitigate bias from an information-theoretical perspective. This approach is compatible with pre-trained models and is light to implement without repeating training steps.

3.1 In-processing Debiasing: Fair Sampling

Image search models in the first fashion are often trained under the contrastive learning framework (Le-Khac et al., 2020). For our in-processing debiasing approach, we now explain the two primary components, contrastive learning and fair sampling, within our context.

Contrastive Learning We start by formally introducing the standard contrastive learning framework commonly used in previous works (Lee et al., 2018; Chen et al., 2020b) for image-text retrieval. Given a batch of N image-text pairs $\mathcal{B} = \{(v_n, c_n)\}_{n=1}^{N}$, the model aims to maximize the similarity scores of matched image-text pairs (positive pairs) while minimizing that of mismatched pairs (negative pairs). The representative SCAN model (Lee et al., 2018), denoted as $S(v,c)$ outputting a similarity score between image and text, is optimized with a standard hinge-based triplet loss:

$$\mathcal{L}_{i-t} = \sum_{(v,c)\in\mathcal{B}} [\gamma - S(v,c) + S(v,\tilde{c})]_+ \quad (4)$$
$$\mathcal{L}_{t-i} = \sum_{(v,c)\in\mathcal{B}} [\gamma - S(v,c) + S(\tilde{v},c)]_+ \quad (5)$$

where γ is the margin, $\tilde{v}$ and $\tilde{c}$ are negative examples, and $[\cdot]_+$ denotes the ramp function. $\mathcal{L}_{i-t}$

corresponds to image-to-text retrieval, while $\mathcal{L}_{t-i}$ corresponds to text-to-image retrieval (or image search). Common negative sampling strategy includes selecting all the negatives (Huang et al., 2017), selecting hard negatives of highest similarity scores in the mini-batch (Faghri et al., 2018), and selecting hard negatives from the whole training data (Chen et al., 2020b). Minimizing the margin-based triplet loss will make positive image-text pairs closer to each other than other negative samples in the joint embedding space.

Fair Sampling One major issue in the contrastive learning framework is that the gender distribution in a batch of image-text pairs is typically imbalanced. Hence, the negative samples will slant towards the majority group, leading to systematic discrimination. To address this problem, we propose a fair sampling strategy. We split the batch of image-text pairs into masculine and feminine pairs based on the image's gender attribute:

$$\mathcal{V}_{\text{male}} = \{v \mid g(v) = \text{male}, (v,c) \in \mathcal{B}\}$$
$$\mathcal{V}_{\text{female}} = \{v \mid g(v) = \text{female}, (v,c) \in \mathcal{B}\}$$
$$\mathcal{V}_{\text{neutral}} = \{v \mid g(v) = \text{neutral}, (v,c) \in \mathcal{B}\}$$

For every positive image and text pair $(v,c) \in \mathcal{B}$, we identify the gender information contained in the query c. If the natural language query is gender-neutral, we sample a negative image from the set of male and female images with probability $\frac{1}{2}$, respectively. Otherwise, we keep the primitive negative sampling selection strategy for keeping the model's generalization on gender-specific queries. Let $\mathcal{B}^*$ be the batch of gender-neutral image-text pairs, the image search loss with fair sampling is:

$$\begin{aligned}\mathcal{L}_{t-i}^{fair} = \sum_{(v,c)\in\mathcal{B}^*} \Big(&\frac{1}{2}\mathbb{E}_{\bar{v}\in\mathcal{V}_{\text{male}}}[\gamma - S(v,c) + S(\bar{v},c)]_+ \\ &+ \frac{1}{2}\mathbb{E}_{\bar{v}\in\mathcal{V}_{\text{female}}}[\gamma - S(v,c) + S(\bar{v},c)]_+\Big) \\ &+ \sum_{(v,c)\in\mathcal{B}/\mathcal{B}^*} [\gamma - S(v,c) + S(\tilde{v},c)]_+ \quad (6)\end{aligned}$$

Empirically, we find that if we thoroughly apply the Fair Sampling strategy, the recall performance drops too much. To obtain a better tradeoff, we use a weight α to combine the objectives

$$\alpha\mathcal{L}_{t-i}^{fair} + (1-\alpha)\mathcal{L}_{t-i}$$

as the final text-to-image loss function. We do not alter the sentence retrieval loss $\mathcal{L}_{i-t}$ during training for preserving generalization.

Algorithm 1 *clip* algorithm

Require: Index set $\Omega = \{1, ..., d\}$, number of clipped features $0 \le m < d$
$\mathcal{Z} \leftarrow \emptyset$;
for $i = 1$ to d **do**
  Estimate mutual information $I(V_i; g(V))$;
end for
for $j = 1$ to m **do**
  $z \leftarrow \arg\max\{I(V_i; g(V)) : i \in \Omega/\mathcal{Z}\}$;
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{z\}$;
end for
return Index set of clipped features $\mathcal{Z}$

3.2 Post-processing Debiasing: Feature Clipping based on Mutual Information

Pre-training methods have shown promising zero-shot performance on extensive NLP and computer vision benchmarks. The recently introduced CLIP model (Radford et al., 2021) was pre-trained on an enormous amount of image-text pairs found across the internet to connect text and images. CLIP can encode image and text into d-dimensional embedding vectors, based on which we can use cosine similarity to quantify the similarity of image and text pairs. In this work, we find that the pre-trained CLIP model reaches the state-of-the-art performance but exhibits large gender bias due to training on uncurated image-text pairs collected from the internet. Although Radford et al. (2021) released the pre-trained CLIP model, the training process is almost unreproducible due to limitations on computational costs and massive training data.

In order to avoid re-training of the CLIP model, we introduce a novel post-processing mechanism to mitigate the representation bias in the CLIP model. We propose to "clip" the dimensions of feature embeddings that are highly correlated with gender information. This idea is motivated by the fact that an unbiased retrieve implies the independence between the covariates (active features) and sensitive attributes (gender) (Barocas et al., 2019). Clipping the highly correlating covariates will return us a relatively independent and neutral set of training data that does not encode hidden gender bias.

The proposed *clip* algorithm is demonstrated in Algorithm 1, and we explain the key steps below. Let $\Omega = \{1, ..., d\}$ be the full index set. We use $V = V_\Omega = [V_1, V_2, ..., V_d]$ to represent the variable of d-dimensional encoding image vectors and $g(V) \in \{\text{male}, \text{female}, \text{neutral}\}$ to represent the corresponding gender attribute. The goal is to output the index set $\mathcal{Z}$ of clipped covariates that reduce the dependence between representations $V_{\Omega/\mathcal{Z}}$

and gender attributes $g(V)$. We measure the correlation between each dimension V_i and gender attribute $g(V)$ by estimating their mutual information $I(V_i; g(V))$ (Gao et al., 2017):

$$I(V_I; g(V)) = D_{\text{KL}}(\mathbb{P}_{(V_i, g(V))} \| \mathbb{P}_{V_i} \otimes \mathbb{P}_{g(V)}) \quad (7)$$

where D_{KL} is the KL divergence (Kullback and Leibler, 1951), $\mathbb{P}_{(V_i, g(V))}$ indicates the joint distribution, $\mathbb{P}_{V_i}$ and $\mathbb{P}_{g(V)}$ indicate their marginals. Next, we greedily clip m covariates with highest mutual information, and construct $(d-m)$-dimensional embedding vectors $V_{\Omega/\mathcal{Z}}$. m is a hyper-parameter that we will experimentally find to best trade-off accuracy and the reduced gender bias, and we show how the selection of m affects the performance in Section 5.3. To project text representations, denoted by variable C, into the same embedding space, we also apply the index set $\mathcal{Z}$ to obtain clipped text embedding vectors $C_{\Omega/\mathcal{Z}}$.

The clipped image and text representations, denoted by $\vec{v}^*$ and $\vec{c}^*$, will have a relatively low correlation with gender attributes due to the "loss" of mutual information. Then we compute the cosine similarity between image and text by substituting $\vec{v}^*$ and $\vec{c}^*$ into Equation (1):

$$S(v, c) = \frac{\vec{v}^* \cdot \vec{c}^*}{\|\vec{v}^*\| \|\vec{c}^*\|} \quad (8)$$

Finally, we rank the images based on the cosine similarity between the clipped representations.

4 Experimental Setup

4.1 Datasets

We evaluate our approaches on the standard MS-COCO (Chen et al., 2015) and Flickr30K (Young et al., 2014) datasets. Following Karpathy and Fei-Fei (2017) and Faghri et al. (2018), we split MS-COCO captions dataset into 113,287 training images, 5,000 validation images and 5,000 test images.[2] Each image corresponds to 5 human-annotated captions. We report the results on the test set by averaging over five folds of 1K test images or evaluating the full 5K test images. Flickr30K consists of 31,000 images collected from Flickr.[3] Following the same split of Karpathy and Fei-Fei (2017); Lee et al. (2018), we select 1,000 images for validation, 1,000 images for testing, and the rest of the images for training.

Identifying Gender Attributes of Images Sensitive attributes such as gender are often not explicitly annotated in large-scale datasets such as MS-COCO and Flickr30K, but we observe that implicit gender attributes of images can be extracted from their associated human-annotated captions. Therefore, we pre-define a set of masculine words and a set of feminine words.[4] Following Zhao et al. (2017) and Burns et al. (2018) we use the ground-truth annotated captions to identify the gender attributes of images. An image will be labeled as "male" if at least one of its captions contains masculine words and no captions include feminine words. Similarly, an image will be labeled as "female" if at least one of its captions contains feminine words and no captions include masculine words. Otherwise, the image will be labeled as "gender-neutral".

4.2 Models

We compare the fairness performance of the following approaches:

- **SCAN** (Lee et al., 2018): we use the official implementation for training and evaluation[5].
- **FairSample**: we apply the fair sampling method proposed in Section 3.1 to the SCAN framework and adopt the same hyper-parameters suggested by Lee et al. (2018) for training.
- **CLIP** (Radford et al., 2021): we use the pre-trained CLIP model released by OpenAI.[6] The model uses a Vision Transformer (Dosovitskiy et al., 2021) as the image encoder and a masked self-attention Transformer (Vaswani et al., 2017) as the text encoder. The original model produces 500-dimensional image and text vectors.
- **CLIP-clip**: we apply the feature pruning algorithm in Section 3.2 to the image and text features generated by the CLIP model. We set $m = 100$ and clip the image and text representations into 400-dimensional vectors.

Note that SCAN and FairSample are trained and tested on the in-domain MS-COCO and Flickr30K datasets, while the pre-trained CLIP model is directly tested on MS-COCO and Flickr30K test sets without fine-tuning on their training sets (same for CLIP-clip as it simply drops CLIP features).

[2] The data is available at cocodataset.org.

[3] The data is available at http://bryanplummer.com/Flickr30kEntities/.

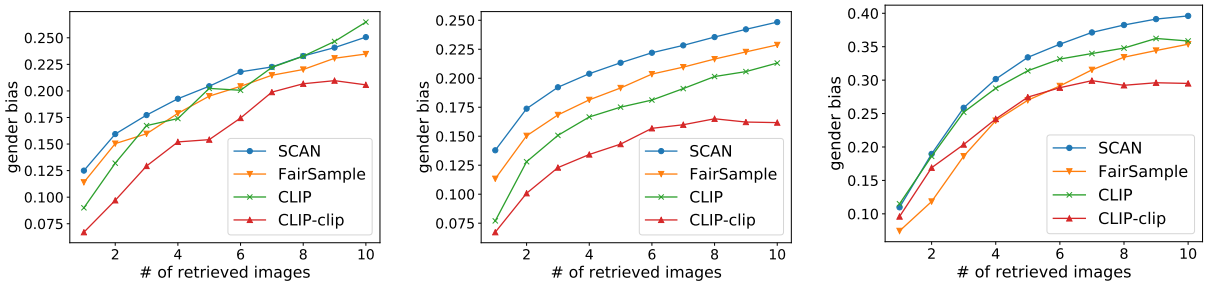
[4] We show the word lists in Appendix A.

[5] The code is available at https://github.com/kuanghuei/SCAN.

[6] The pre-trained model is available at https://github.com/openai/CLIP.

Before Pre-processing	After Pre-processing
A **man** with a red helmet on a small moped on a dirt road.	A **person** with a red helmet on a small moped on a dirt road.
A little **girl** is getting ready to blow out a candle on a small dessert.	A little **child** is getting ready to blow out a candle on a small dessert.
A **female** surfboarder dressed in black holding a white surfboard.	A surfboarder dressed in black holding a white surfboard.
A group of young **men and women** sitting at a table.	A group of young **people** sitting at a table.

Table 1: Samples of the constructed gender-neutral captions. For evaluation, we convert gender-specific captions to gender-neutral ones by replacing or removing the gender-specific words.



(a) MS-COCO 1K Test Set. (b) MS-COCO 5K Test Set. (c) Flick30K Test Set.

Figure 2: Gender bias analysis with different top-K results.

4.3 Evaluation

Gender-Neutral Text Queries In this study, we focus on equalizing the search results of gender-neutral text queries. In addition to the existing gender-neutral captions in the test sets, we pre-process those gender-specific captions to construct a purely gender-neutral test corpus to guarantee a fair and large-scale evaluation. For every caption, we identify all these gender-specific words and remove or replace them with corresponding gender-neutral words. We show some pre-processing examples in Table 1.

Metrics As introduced in Section 2.2, we employ the fairness metric in Equation (3), Bias@K, to measure the gender bias among the top-K images. In addition, following standard practice, we measure the retrieval performance by Recall@K, defined as the fraction of queries for which the correct image is retrieved among the top-K images.

5 Debiasing Results

5.1 Main Results on MS-COCO & Flickr30K

We report the results comparing our debiasing methods and the baseline methods in Table 2.

Model Bias Although the pre-trained CLIP model is evaluated without fine-tuning, we observe that it achieves a comparable recall performance with the SCAN model on MS-COCO and dominates the Flickr30K dataset. However, both models suffer from severe gender bias. Especially, the Bias@10 of the SCAN model on Flickr30K is 0.3960, meaning nearly 70% of the retrieved gender-specific images portray men and only 30% portray women. Similarly, the CLIP model achieves 0.2648 gender bias on MS-COCO 1K test set, indicating about 6.4 out of 10 retrieved images portray men while about 3.6 out of 10 portray women. Given that all of the testing text queries are gender-neutral, this result shows that severe implicit gender bias exists in image search models.

Debiasing Effectiveness As shown in Table 2, both the in-processing sampling strategy *FairSample* and the post-processing feature pruning algorithm *clip* consistently mitigate the gender bias on test data. For instance, among the top-10 search images, SCAN with FairSample reduces gender bias from 0.3960 to 0.3537 (decreased by 10.7%) on Flickr30K. Using the clipped CLIP features for image search (CLIP-clip), the gender bias drops from 0.2648 to 0.2057 (22.3%) on MS-COCO 1K, from 0.2131 to 0.1611 (24.4%) on MS-COCO 5K, and from 0.3586 to 0.2951 (17.7%) on Flickr30K. For the tradeoff, CLIP-clip sacrifices the recall performance slightly (from 93.6% Recall@10 to 91.3% on Flickr30K). On the other hand, SCAN with FairSample even achieves a comparable recall performance with SCAN.

Dataset	Method	Gender Bias↓			Recall↑		
		Bias@1	Bias@5	Bias@10	Recall@1	Recall@5	Recall@10
COCO1K	SCAN	.1250	.2044	.2506	47.7	82.0	**91.0**
	FairSample	.1140	.1951	.2347	**49.7**	**82.5**	90.9
	CLIP	.0900	.2024	.2648	48.2	77.9	88.0
	CLIP-clip	**.0670**	**.1541**	**.2057**	46.1	75.2	86.0
COCO5K	SCAN	.1379	.2133	.2484	25.4	54.1	67.8
	FairSample	.1133	.1916	.2288	26.8	**55.3**	**68.5**
	CLIP	.0770	.1750	.2131	**28.7**	53.9	64.7
	CLIP-clip	**.0672**	**.1474**	**.1611**	27.3	50.8	62.0
Flickr30K	SCAN	.1098	.3341	.3960	41.4	69.9	79.1
	FairSample	**.0744**	**.2699**	.3537	35.8	67.5	77.7
	CLIP	.1150	.3150	.3586	**67.2**	**89.1**	**93.6**
	CLIP-clip	.0960	.2746	**.2951**	63.9	85.4	91.3

Table 2: Results on MS-COCO (1K and 5K) and Flickr30K test sets. We compare the baseline models (SCAN (Lee et al., 2018) and CLIP (Radford et al., 2021)) and our debiasing methods (FairSample and CLIP-clip) on both the gender bias metric Bias@K and the retrieval metric Recall@K.

5.2 Gender Bias at Different Top-K Results

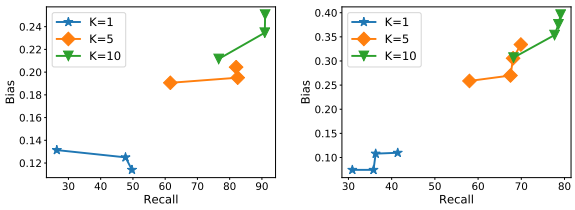
We plot how gender bias varies across different values of K (1-10) for all the compared methods in Figure 2. We observe that when $K < 5$, the gender bias has a higher variance due to the inadequate retrieved images. When $K \geq 5$, the curves tend to be flat. This result indicates that Bias@10 is more recommended than Bias@1 for measuring gender bias as it is more stable. It is also noticeable that CLIP-clip achieves the best fairness performance in terms of Bias@10 consistently on all three test sets compared to the other models.

5.3 Tradeoff between Recall and Bias

There is an inherent tradeoff between fairness and accuracy in fair machine learning (Zhao and Gordon, 2019). To achieve the best recall-bias tradeoff in our methods, we further examine the effect of the controlling hyper-parameters: the weight α in FairSampling and the number of clipped dimensions m in CLIP-clip.

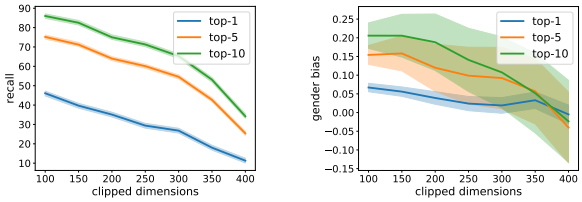
Figure 3 demonstrates the recall-bias curve with the fair sampling weight $\alpha \in [0, 1]$. Models of higher recall often suffer higher gender bias, but the fairness improvement outweighs the recall performance drop in FairSample models. For example, the model fully trained with fair sampling ($\alpha = 1$) has the lowest bias and drops the recall performance the most—it relatively reduces 22.5% Bias@10 but only decreases 10.9% Recall@10 on Flickr30K. We choose $\alpha = 0.4$ for the final model, which has a better tradeoff in retaining the recall performance.

As shown in Figure 4, we set the range of the clipping dimension m between 100 and 400 on MS-COCO 1K. We find that clipping too many covariates (1) harms the expressiveness of image and text representations (Recall@1 drops from 46.1% to 11.3%, Recall@5 drops from 75.2% to 25.4%, and Recall@10 drops from 86.0% to 34.2%), and (2) causes high standard deviation in gender bias. In light of the harm on expressiveness, we select $m = 100$ for conventional use.



(a) MS-COCO 1K test set (b) Flickr30K test set

Figure 3: The Pareto frontier of recall-bias tradeoff curve for FairSample on MS-COCO 1K and Flickr30K.

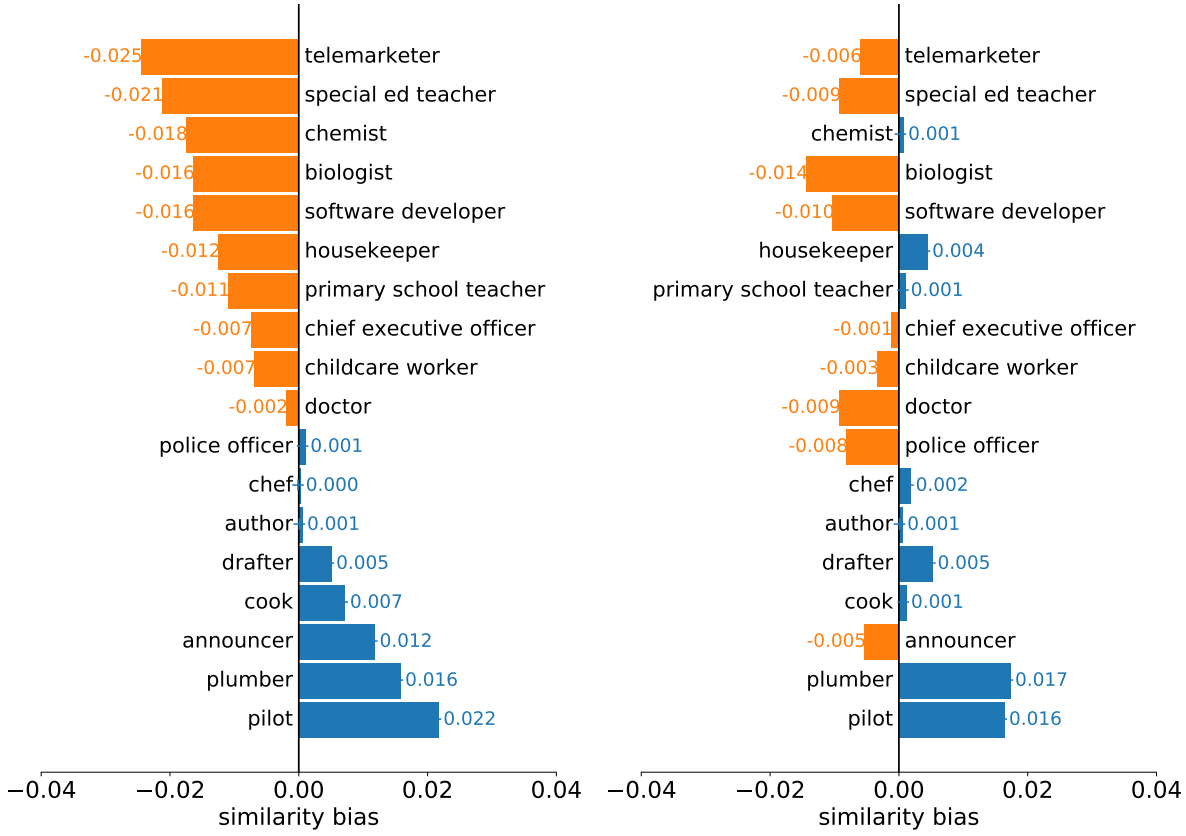


(a) recall (b) gender bias

Figure 4: Effect of the number of clipped dimensions m on performance of recall and bias on MS-COCO 1K.

5.4 Evaluation on Internet Image Search

The aforementioned evaluation results on MS-COCO and Flickr30K datasets are limited that they rely on gender labels extracted from human



(a) CLIP (b) CLIP-clip

Figure 5: Gender bias evaluation of internet image search results on `occupations` (Kay et al., 2015). We visualize the similarity biases on 18 occupations. ■ indicates the occupation is biased towards males and ■ indicates it is biased towards females. The clip algorithm mitigates gender bias for a variety of occupations.

captions. In this sense, it is important to measure the gender biases on a benchmark where the gender labels are identified by crowd annotators. To this end, we further evaluate on the `occupation` dataset (Kay et al., 2015), which collects top 100 Google Image Search results for each gender-neutral occupation search term.[7] Each image is associated with the crowd-sourced gender attribute of the participant portrayed in the image. Inspired by Burns et al. (2018) and Tang et al. (2020), we measure the gender bias by computing the difference of expected cosine similarity between male and female occupational images. Given an occupation o, the similarity bias is formulated as

$$\text{Bias} = \mathbb{E}_{v \in \mathcal{V}^o_{\text{male}}} S(v, o) - \mathbb{E}_{v \in \mathcal{V}^o_{\text{female}}} S(v, o) \quad (9)$$

where $\mathcal{V}^o_{\text{male}}$ and $\mathcal{V}^o_{\text{female}}$ are the sets of images for occupation o, labeled as "male" and "female".

[7]The data is available at https://github.com/mjskay/gender-in-image-search.

Figure 5 demonstrates the absolute similarity bias of CLIP and CLIP-clip on the `occupation` dataset for 18 occupations. We observe that the CLIP model exhibits severe similarity discrepancy for some occupations, including telemarketer, chemist, and housekeeper, while the *clip* algorithm alleviates this problem effectively. Note that for doctor and police officer, the CLIP-clip model exaggerates the similarity discrepancy, but the similarity bias is still less than 0.01. In general, CLIP-clip is effective for mitigating similarity bias and obtains a 42.3% lower mean absolute bias of the 100 occupations than the CLIP model (0.0064 *vs.* 0.0111).

6 Related Work

Fairness in Machine Learning A number of unfair treatments by machine learning models were reported recently (Angwin et al., 2016; Buolamwini and Gebru, 2018; Bolukbasi et al., 2016; Otterbacher et al., 2017), and the literature has observed a growing demand and interests in proposing defenses, including regularizing disparate impact (Za-

far et al., 2015) and disparate treatment (Hardt et al., 2016), promoting fairness through causal inference (Kusner et al., 2017), and adding fairness guarantees in recommendations and information retrieval (Beutel et al., 2019; Biega et al., 2018; Morik et al., 2020). The existing fair machine learning solutions can be broadly categorized as pre-processing (KamiranFaisal and CaldersToon, 2012; Feldman et al., 2015; Calmon et al., 2017), in-processing, and post-processing approaches. Pre-processing algorithms typically re-weight and repair the training data which captures label bias or historical discrimination (KamiranFaisal and CaldersToon, 2012; Feldman et al., 2015; Calmon et al., 2017). In-processing algorithms focus on modifying the training objective with additional fairness constraints or regularization terms (Zafar et al., 2017; Agarwal et al., 2018; Cotter et al., 2019). Post-processing algorithms enforce fairness constraints by applying a post hoc correction of a (pre-)trained classifier (Hardt et al., 2016; Calmon et al., 2017). In this work, the *fair sampling* strategy designed for the contrastive learning framework could be considered as an in-processing treatment, while the *clip* algorithm is in the post-processing regime that features an information-theoretical clipping procedure. Our contribution highlights new challenges of reducing gender bias in a multimodal task and specializes new in-processing and post-processing ideas in the domain of image search.

Social Bias in Multi-modality Implicit social bias related to gender and race has been discussed in multimodal tasks including image captioning (Burns et al., 2018; Tang et al., 2020), visual question answering (Manjunatha et al., 2019), face recognition (Buolamwini and Gebru, 2018), and unsupervised image representation learning (Steed and Caliskan, 2021). For example, Zhao et al. (2017) shows that models trained on unbalanced data can amplify bias, and injecting corpus-level Lagrangian constraints can calibrate the bias amplification. Caliskan et al. (2017) demonstrates the association between the word embeddings of occupation and gendered concepts correlates with the imbalanced distribution of gender in text corpora. There are also a series of debiasing techniques in this area. Bolukbasi et al. (2016) propose to surgically alter the embedding space by identifying the gender subspace from gendered word pairs. Manzini et al. (2019) extend the bias component removal approach to the setting where the sensitive attribute is non-binary. Data augmentation approaches remove the implicit bias in the training corpora and train the models on the balanced datasets (Zhao et al., 2018). Our work complements this line of research by examining gender bias induced by multimodal models in image search results. Our focus on gender bias in the gender-neutral language would offer new insights for a less explored topic to the community.

Gender Bias in Online Search Systems Our work is also closely connected to studies in the HCI community showing the gender inequality in online image search results. Kay et al. (2015) articulate the gender bias in occupational image search results affect people's perceptions of the prevalence of men and women in each occupation. Kay et al. (2015) compare gender proportions in occupational image search results and discuss how the bias affects people's perceptions of the prevalence of men and women in each occupation. Singh et al. (2020) examine the prevalence of gender stereotypes on various digital media platforms. Otterbacher et al. (2017) identify gender bias with character traits. Nonetheless, these works do not attempt to mitigate gender bias in search algorithms. Our work extends these studies into understanding how gender biases enter search algorithms and provides novel solutions to mitigating gender bias in two typical model families for image search.

7 Conclusion

In this paper, we examine gender bias in image search models when search queries are gender-neutral. As an initial attempt to study this critical problem, we formally identify and quantify gender bias in image search. To mitigate the gender bias perpetuating two representative fashions of image search models, we propose two novel debiasing algorithms in in-processing and post-processing manners. When training a new image search model, the in-processing *FairSample* method can be used to learn a fairer model from scratch. Meanwhile, the *clip* algorithm can be used for lightweight deployment of pre-trained representation models with accessible gender information.

Broader Impact

The algorithmic processes behind modern search engines, with extensive use of machine learning algorithms, have great power to determine users'

access to information (Eslami et al., 2015). Our research provides evidence that unintentionally using image search models trained either on in-domain image retrieval data sets or massive corpora across the internet may lead to unequal inclusiveness between males and females in image search results, even when the search terms are gender-neutral. This inequity can and do have significant impacts on shaping and exaggerating gender stereotype in people's minds (Kay et al., 2015).

This work offers new methods for mitigating gender bias in multimodal models, and we regard the algorithms proposed in this paper have the potentials to be deployed in real-world systems. We conjecture that our methods may contribute to driving the development of responsible image search engines with other fairness issues. For instance, we would encourage future works to understand and mitigate the risks arising from other social biases, like racial bias, in image search results. We would also encourage researchers to explore whether the methodology presented in this work could be generalized to quantify and mitigate other bias measures.

Our work has limitations. The gender bias measures and the debiasing methods proposed in this study require acquiring the gender labels of images. Our method for identifying the gender attributes of people portrayed in the images is limited: we make use of the contextual cues in the human-annotated captions from the image datasets. The accuracy of such a proxy-based method heavily relies on the coverage of gendered nouns and the inclusiveness of gendered language in the original human annotations. The corruption of gender labels, due to missing gendered words or inappropriate text preprocessing steps, may introduce biases we have not foreseen into the evaluated metrics. Additionally, the gendered word lists are collected from English corpora and may differ in other languages or cultures. It is possible that blind application of our methods by improperly acquiring the gender labels may create image search models that produce even greater inequality, which is very much discouraged. This limitation arises from the unavailability of such sensitive attributes in the source datasets. The lack of relevant data for studying gender bias in image search, and the concerns about how to acquire the gender attributes while preserving the privacy of people concerned, is itself an important question in this area. We believe this research would benefit when richer datasets become available.

Acknowledgements

The authors would like to thank anonymous reviewers for their constructive comments. This work is supported by the UC Santa Cruz Startup Funding, and the National Science Foundation (NSF) under grants IIS-2040800 and CCF-2023495.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A reductions approach to fair classification. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May*, 23:2016.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, volume 32, pages 15535–15545. Curran Associates, Inc.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Alex Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Zhe Zhao, L. Hong, Ed Huai hsin Chi, and Cristos Goodrow. 2019. Fairness in recommendation ranking through pairwise comparisons. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Asia J. Biega, K. Gummadi, and G. Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *ECCV*.

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

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020a. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR.

T. Chen, Jiajun Deng, and Jiebo Luo. 2020b. Adaptive offline quintuplet loss for image-text matching. *ArXiv*, abs/2003.03669.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020c. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Xinlei Chen, H. Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. L. Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020d. Uniter: Universal image-text representation learning. In *ECCV*.

Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, M. Gupta, S. You, and K. Sridharan. 2019. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *ArXiv*, abs/1809.04198.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "i always assumed that i wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *CHI 2015 - Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems - Proceedings, pages 153–162. Association for Computing Machinery. 33rd Annual CHI Conference on Human Factors in Computing Systems, CHI 2015 ; Conference date: 18-04-2015 Through 23-04-2015.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA. Association for Computing Machinery.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*.

Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. 2017. Estimating mutual information for discrete-continuous mixtures. In *Advances in Neural Information Processing Systems*, volume 30, pages 5986–5997. Curran Associates, Inc.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal LSTM. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7254–7262. IEEE Computer Society.

KamiranFaisal and CaldersToon. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*.

A. Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676.

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. *Unequal Representation and Gender Stereotypes in Image Search Results for Occupations*, page 3819–3828. Association for Computing Machinery, New York, NY, USA.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

P. H. Le-Khac, G. Healy, and A. F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*.

Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Varun Manjunatha, Nirat Saini, and Larry S. Davis. 2019. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. *Controlling Fairness and Bias in Dynamic Learning-to-Rank*, page 429–438. Association for Computing Machinery, New York, NY, USA.

Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. *Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results*, page 6620–6631. Association for Computing Machinery, New York, NY, USA.

A. Radford, J. W. Kim, Chris Hallacy, Aditya Ramesh, G. Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, J. Clark, G. Krüger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. Technical report, OpenAI.

Vivek K. Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2020. Female librarians and male computer programmers? gender bias in occupational images on digital media platforms. *J. Assoc. Inf. Sci. Technol.*, 71(11):1281–1294.

Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, New York, NY, USA.

Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, and X. Hu. 2020. Mitigating gender bias in captioning systems. *ArXiv*, abs/2006.08315.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

J. Wang, Yang Song, Thomas Leung, C. Rosenberg, J. Philbin, Bo Chen, and Y. Wu. 2014. Learning fine-grained image similarity with deep ranking. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393.

P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. 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.

M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *AISTATS*.

M. Zafar, I. Valera, M. G. Rodriguez, and K. Gummadi. 2015. Learning fair classifiers. *arXiv: Machine Learning*.

Han Zhao and Geoff Gordon. 2019. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Gender Word Lists

We show the word lists for identifying the gender attributes of a caption in Table 3.

feminine words	woman, women, female, girl, lady, mother, mom, sister, daughter, wife, girlfriend
masculine words	man, men, male, boy, gentleman, father, brother, son, husband, boyfriend
gender-neutral words	person, people, human, adult, baby, child, kid, children, guy, teenage, crowd

Table 3: Gender word lists. We identify the gender attributes of captions based on the occurrence of gender-specific words appeared in the sentences.

B Implementation Details

B.1 Computing Infrastructure

We use a GPU server with 4 NVIDIA RTX 2080 Ti GPUs for training and evaluation.

B.2 Computational Time Costs

We find that SCAN (Lee et al., 2018) and SCAN with fair sampling need about 20 hours for training 30 epochs on MS-COCO and 8-10 minutes for testing on 1K test set. In comparison, pre-trained CLIP (Radford et al., 2021) and CLIP-clip can be evaluated within 1 minutes on MS-COCO 1K test set.

C Qualitative Examples

We take a qualitative study on the image search results. We show the results of searching "a person riding a bike" in Figure 6. The first row presents the top-5 retrieved images for SCAN, the second row presents the top-5 retrieved images for SCAN+FairSample, the third row presents the top-5 retrieved images for CLIP, and the last row presents the top-5 retrieved images for CLIP-clip. While we notice that all the models retrieve relevant images, we find FairSample put images depicting females in a higher rank.

Figure 6: Qualitative analysis of gender bias in image search results. The text query is "a person riding a bike". The first row presents the top-5 retrieved images for SCAN, the second row presents the top-5 retrieved images for SCAN+FairSample, the third row presents the top-5 retrieved images for CLIP, and the last row presents the top-5 retrieved images for CLIP-clip.