Mitigating Test-Time Bias for Fair Image Retrieval

Fanjie Kong[1], *Shuai Yuan*[1], *Weituo Hao*[2], *Ricardo Henao*[1,3]

[1] Duke University [2] TikTok Inc. [3] King Abdullah University of Science and Technology

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

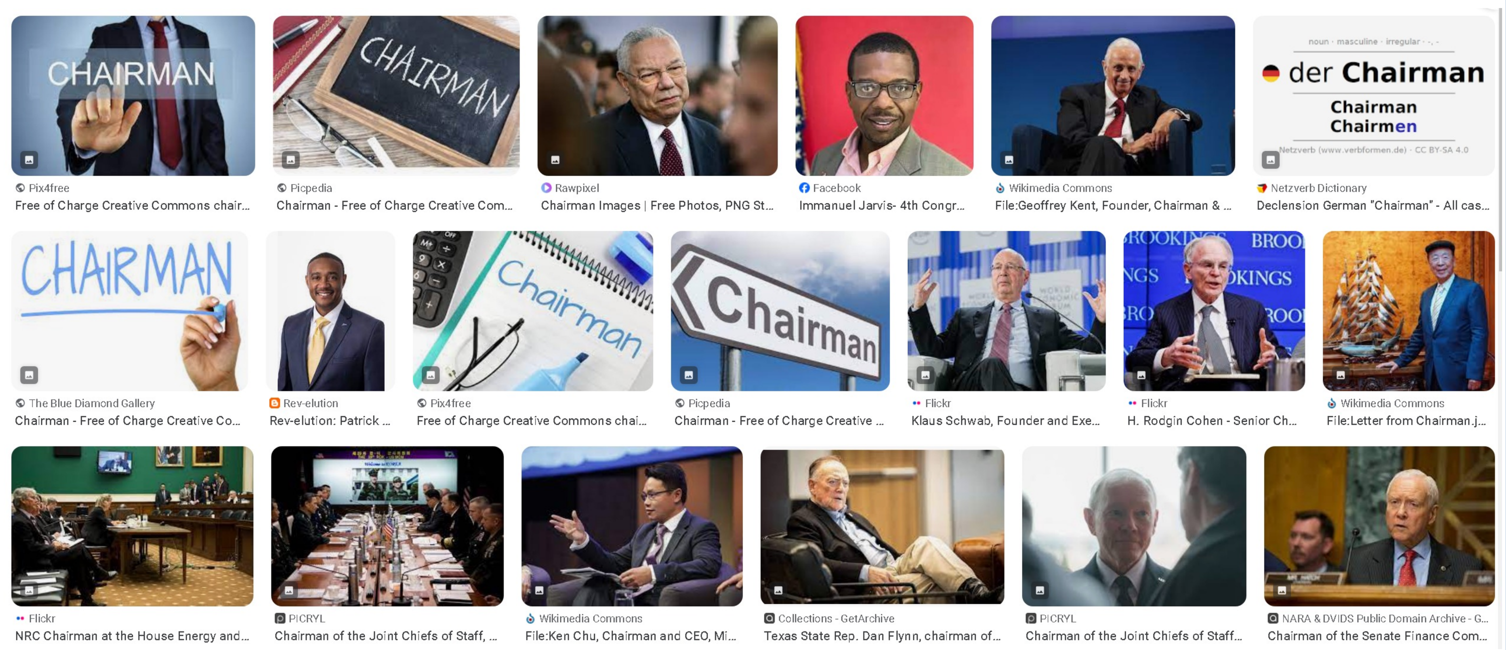


Fig 1. Google image search results for the gender-neutral query "chairman".

Gender bias are prevalent in web image search results when utilizing gender-neutral queries. Creating ***a fair and unbiased web image search system*** is crucial for fostering social equity and preventing the perpetuation of gender stereotypes.

Fairness Objective

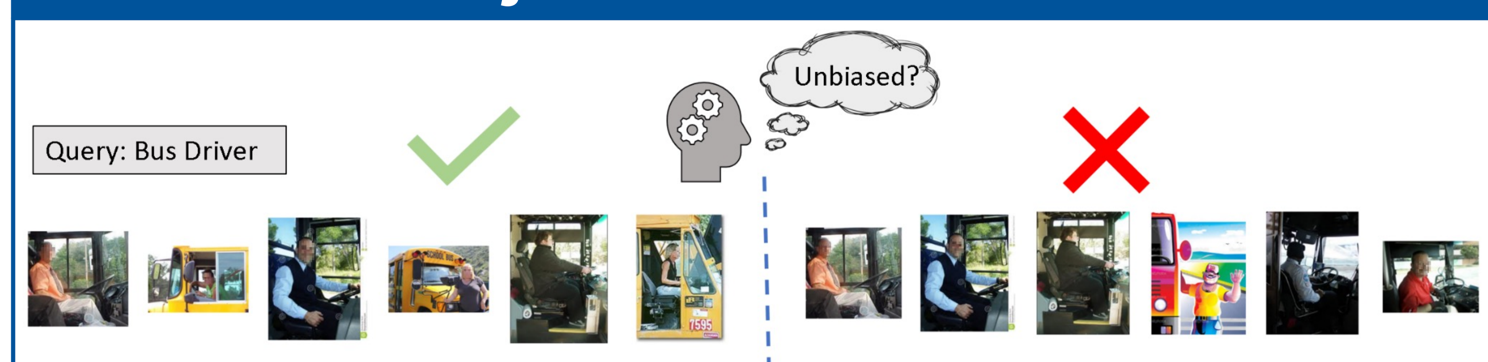


Fig 2. Equal vs. unequal gender representation in image search results.

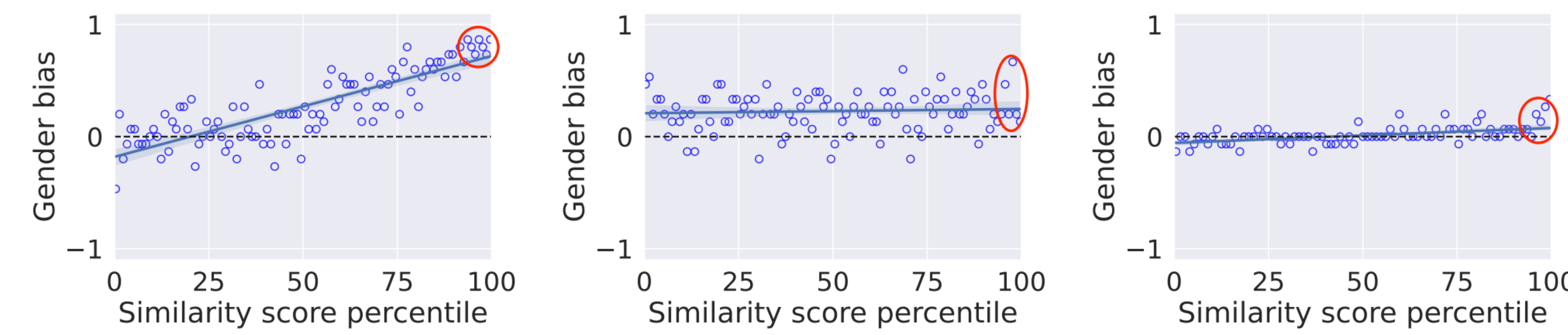
- We adhere to ***equal representation*** as our fairness objective[1]. Equal representation for all demographic groups of interest attempts to obscure the influence of any inherent biases.
- ***Equal representation*** is satisfied on the retrieval image set V_c corresponding to neutral queries c with respect to binary demographic attributes $g(v)$ if

$$\mathbb{E}_{V_c \sim P}[\mathbb{E}_v[\mathbb{1}(g(v)=+1)]] = \mathbb{E}_{V_c \sim P}[\mathbb{E}_v[\mathbb{1}(g(v)=-1)]]$$

[1] Matthew Kay, et al. 2015. Unequal representation and gender stereotypes in image search results for occupations.

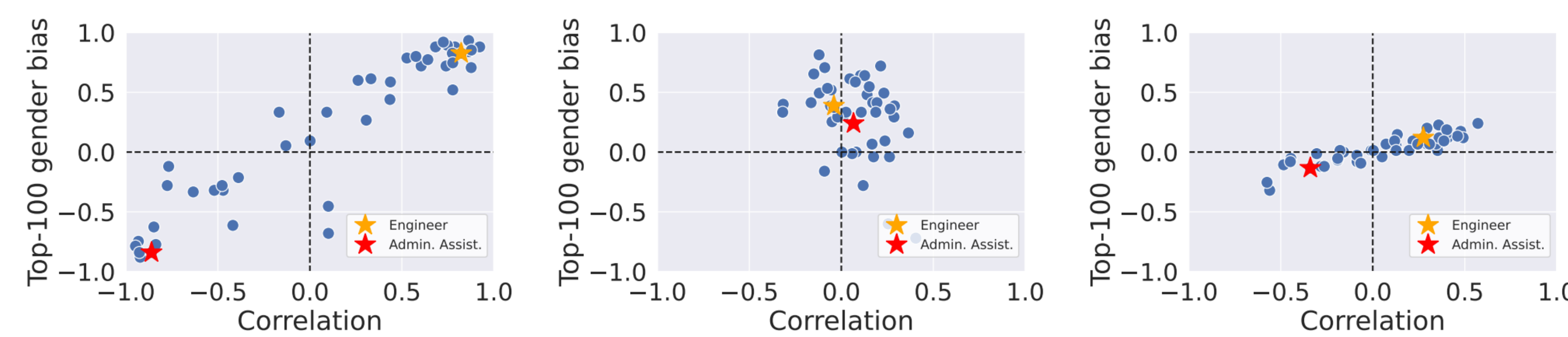
Test-time Bias Analysis

Even if the model is fair when treating each image, the biased candidate image pool will still propagate its distribution bias into retrieval results at test time.



(a) CLIP (no debiasing) (b) MI debiasing (c) PBM

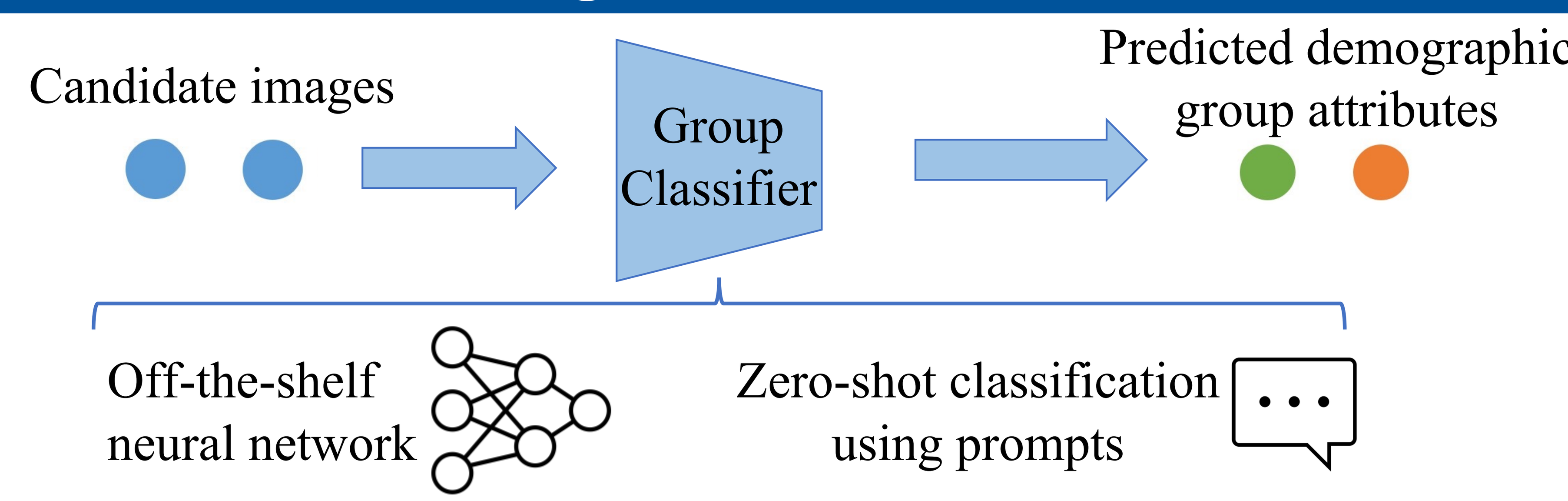
Fig 3. Similarity scores against gender bias in 1% quantile increments.



(a) CLIP (no debiasing) (b) MI debiasing (c) PBM

Fig 4. Averaged top-100 gender bias against similarity-bias correlation.

Post-hoc Bias Mitigation



During testing, we first predict the demographic groups of candidate images and then create a fair retrieval bag based on the prediction.

Algorithm 1 Post-hoc Bias Mitigation (PBM).

Input: Text query c, retrieval size K, image database $\mathcal{V}^{\text{test}}$, similarity measure from pre-trained vision-language models $S(\cdot,\cdot)$, and gender prediction model $\hat{g}(\cdot)$.
Output: Image retrieval bag $V_{c,K}$.

Split $\mathcal{V}^{\text{test}}$ into $\mathcal{V}^{\text{test}}_{+1}$, $\mathcal{V}^{\text{test}}_{-1}$ and $\mathcal{V}^{\text{test}}_{\text{N/A}}$ using the gender prediction model $\hat{g}(\cdot)$;
Let $V_{c,K} = \varnothing$;
while $|V_{c,K}| < K$ **do**
  $v_{+1} = \arg\max_{v \in \mathcal{V}^{\text{test}}_{+1}} S(v,c)$; $v_{-1} = \arg\max_{v \in \mathcal{V}^{\text{test}}_{-1}} S(v,c)$; $v_{\text{N/A}} = \arg\max_{v \in \mathcal{V}^{\text{test}}_{\text{N/A}}} S(v,c)$;
  if $[S(v_{+1},c) + S(v_{-1},c)]/2 > S(v_{\text{N/A}},c)$ **then**
    $V_{c,K} \leftarrow V_{c,K} \cup \{v_{+1}, v_{-1}\}$; $\mathcal{V}^{\text{test}}_{+1} \leftarrow \mathcal{V}^{\text{test}}_{+1} \setminus \{v_{+1}\}$; $\mathcal{V}^{\text{test}}_{-1} \leftarrow \mathcal{V}^{\text{test}}_{-1} \setminus \{v_{-1}\}$;
  else
    $V_{c,K} \leftarrow V_{c,K} \cup \{v_{\text{N/A}}\}$; $\mathcal{V}^{\text{test}}_{\text{N/A}} \leftarrow \mathcal{V}^{\text{test}}_{\text{N/A}} \setminus \{v_{\text{N/A}}\}$;
  end if
end while
return $V_{c,K}$

Experiment Results

- Web image search dataset: Occupation 1 & 2
- Large-scale image-text dataset: COCO and Flickr30k

Tab 1. Debiasing Results from Occupation 1 & 2 datasets

Method	Occupation 1 - Gender		Occupation 2 - Gender		Occupation 2 - Race	
	AbsBias@100 (↓)	Recall@100(↑)	AbsBias@100(↓)	Recall@100(↑)	AbsBias@100(↓)	Recall@100(↑)
Random Selection	.6370	-	.3155	-	.4171	-
CLIP Original (Radford et al., 2021)	.6231	58.3	.3566	46.2	.5002	46.2
MI-*clip* (Wang et al., 2021a)	.3769	47.0	.2539	42.2	.4099	42.3
Adversarial Training (Edwards and Storkey, 2015)	.2316	44.0	.2603	37.8	.4880	43.3
Debias Prompt (Berg et al., 2022)	.6373	**59.3**	.3564	**46.2**	.4946	**50.2**
CLIP-FairExpec (Mehrotra and Celis, 2021)	.2498	47.0	.2619	44.2	.2788	34.7
PBM - Zero-shot Embedding	.0969	49.8	.1150	42.1	.3133	40.2
PBM - Zero-shot Prompt	**.0560**	46.1	**.0443**	42.5	.2571	36.0
PBM - Supervised Classifier	.1404	50.3	.1171	42.1	**.0955**	37.9
PBM - Ground-truth Gender and Skin-tone	.0000	49.1	.0000	42.4	.0000	41.3

Tab 2. Debiasing Results from COCO and Flickr30k datasets

Dataset	Method	Gender Bias			Recall		
		Bias@1 (↓)	Bias@5(↓)	Bias@10(↓)	Recall@1(↑)	Recall@5(↑)	Recall@10(↑)
COCO-1k	SCAN (Lee et al., 2018)	.1250	.2044	.2506	47.7	82.0	**91.0**
	FairSample (Wang et al., 2021a)	.1140	.1951	.2347	**49.7**	**82.5**	90.9
	CLIP (Radford et al., 2021)	.0900	.2024	.2648	48.2	77.9	88.0
	MI-*clip* (Wang et al., 2021a)	.0670	.1541	.2057	46.1	75.2	86.0
	Our PBM	**.0402**	**.0961**	**.1082**	37.3	73.6	84.8
COCO-5k	SCAN (Lee et al., 2018)	.1379	.2133	.2484	25.4	54.1	67.8
	FairSample (Wang et al., 2021a)	.1133	.1916	.2288	26.8	**55.3**	**68.5**
	CLIP (Radford et al., 2021)	.0770	.1750	.2131	**28.7**	53.9	64.7
	MI-*clip* (Wang et al., 2021a)	.0672	.1474	.1611	27.3	50.8	62.0
	Our PBM	**.0492**	**.1006**	**.1212**	22.3	50.5	61.9
Flickr30K	SCAN (Lee et al., 2018)	.1098	.3341	.3960	41.4	69.9	79.1
	FairSample (Wang et al., 2021a)	.0744	.2699	.3537	35.8	67.5	77.7
	CLIP (Radford et al., 2021)	.1150	.3150	.3586	**67.2**	**89.1**	**93.6**
	MI-*clip* (Wang et al., 2021a)	.0960	.2746	.2951	63.9	85.4	91.3
	Our PBM	**.0360**	**.1527**	**.1640**	41.2	85.3	92.6

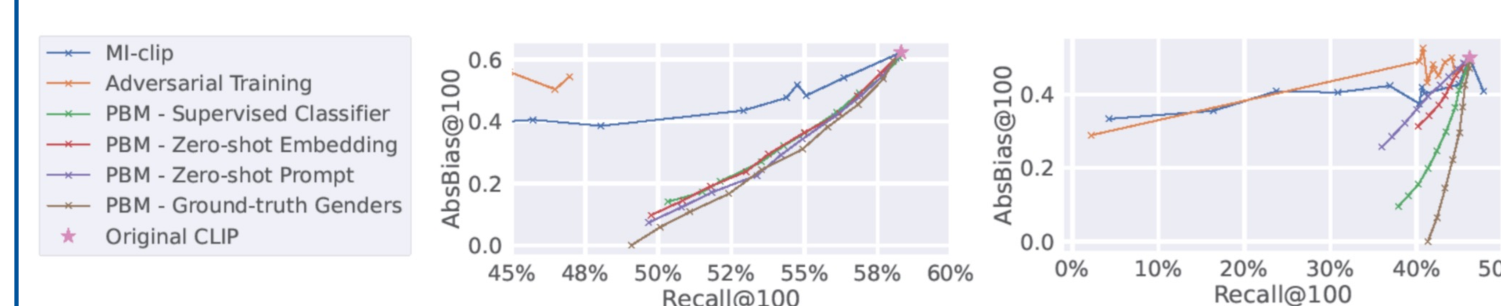


Fig 5. Trade-off between performance and bias for debiasing gender attributes within Occupation 1 (Middle) and race attributes using Occupation 2 (Right).

PBM significantly reduces bias while maintaining high retrieval accuracy. Furthermore, PBM does not require retraining of model weights, thereby avoiding intensive computation for debiasing large models.

Additional Resources

Full Paper

Code

Video