Appendix

A Impact of Demographic Group Classifier on Debiased Results

Table 3: Group sensitivities and sensitivity ratios (ρ) for demographic attributes predicted by different classifiers on Occupation 1 - Gender and Occupation 2 - Race.

	Gender			Race		
Method	Male Sensitivity	Female Sensitivity	ρ	Light skin Sensitivity	Dark skin Sensitivity	ρ
PBM - Supervised Learning	0.97	0.88	1.10	**0.93**	**0.84**	1.11
PBM - Word Embedding	**0.98**	0.94	1.04	0.84	0.78	**1.08**
PBM - Zero-shot Prompt	**0.98**	**0.97**	**1.01**	0.88	0.81	1.09

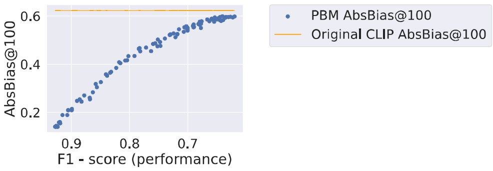

Figure 5: Relationship between the performance of the demographic group classifier (F1-score) and the retrieval bias (AbsBias@100) when utilizing PBM.

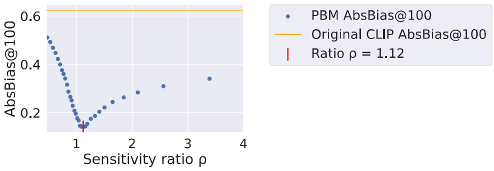

Figure 6: Relationship between the bias of the demographic group classifier (ratio of male sensitivity to female sensitivity) and the retrieval bias (AbsBias@100) when utilizing PBM.

The demographic group classifier is an important module of our proposed method PBM. The debiasing result is intricately linked to the demographic group classifier's accuracy and prediction bias towards different demographic groups. Figure 5 showcases the relationship between the demographic group classifier's performance and the ensuing retrieval bias, by artificially introducing noise to the demographic group (logit) predictions via Gaussian noise with a standard deviation ranging from 0 (no noise) to 1. These results underscore that better group classifier performance yields lower bias, that bias converges to that of the original CLIP as the group classifier gets worse, and importantly, that the bias after PBM will be no worse than that of the original CLIP.

Further, Table 3 shows the individual demographic group sensitivities under three different scenarios, from which we can see that the group classifier is i) able to achieve good classification sensitivity (no lower than 0.81 and 0.90 in average), likely because demographic image attributes (gender and skin tone) are typically captured in images, and ii) that different scenarios exhibit different degrees of bias as measured by the group sensitivity ratio, which must be close to 1 for the model to be unbiased. Table 3 reveals sensitivity discrepancies among different attributes. To delve deeper into the influence of classifier bias on PBM outcomes, we present the retrieval bias as a function of the sensitivity ratio in Figure 6. This is achieved by altering the gender classification threshold from 0 (maximizing male sensitivity) to 1 (minimizing male sensitivity). From Figure 6, we can conclude that the classifier bias does affect retrieval bias, however, only severely for more extreme sensitivity ratios, which is fortunately not the case in our results as shown in Table 3.

B Bias-recall Trade-off Strategies

In Figure 4, we exhibit the bias-recall trade-off curves for MI-clip, Adversarial Training, and various PBM methods. Here, we outline the missing details to achieve these trade-offs.

For adversarial learning, the trade-off is controlled by adjusting the adversarial loss weights between 0 and 1.0. In MI-clip, we modify the clipped dimensions from 10 to 500 (CLIP output dimension is 512). Regarding PBM methods, a trade-off parameter is introduced via a stochastic variable θ, which denotes the likelihood of choosing a fair subset at any given time, instead of simply opting for the image with the top similarity score. Each curve is plotted by interpolating 10 points of the corresponding trade-off parameters.

C Datasets

Occupation 1 (Kay et al., 2015) Occupation 1 comprises the top 100 Google Image Search results for 45 gender-neutral occupation terms, such as "chef", "librarian", "primary school teacher", *etc*. Each image within this dataset is annotated with a crowd-sourced gender attribute (either Male or Female) that characterizes the person depicted in the image. The entire Occupation 1 dataset is exclusively utilized for evaluating gender debiasing effect, as shown in Table 1.

Occupation 2 (Celis and Keswani, 2020) Occupation 2 includes the top 100 Google Image Search results for 96 occupations. Each image in the dataset comes annotated with a gender attribute and a race attribute (represented by skin-tone, namely, Fair Skin and Dark Skin). Notably, the gender attribute and race attribute also include a N/A category, where the annotators have chosen the option of "Not applicable" or "Cannot determine" for the gender or skin-tone depicted in the image. Consequently, we treat the image labeled with N/A as a neutral example that does not contribute to the bias of retrieval, since the user cannot perceive gender or racial information from the image. Different from Occupation 1, gender attributes in Occupation 2 are categorized as {Male, Female, N/A}, while race attributes are classified as {Fair Skin, Dark Skin, N/A}. The enitre Occupation 2 dataset is only used for evaluation on mitigating gender and race bias, as the results shown in Table 1.

MS-COCO (Lin et al., 2014) The first large-scale image-text dataset is MS-COCO captions dataset, which is partitioned into 113,287 training images, 5,000 validation images, and 5,000 test images. Each image is accompanied by five corresponding captions. Our experimental setup aligns with the methodology detailed by Wang et al. (2021a). Only the first caption of each image is used for evaluation. Further, they ensure all captions are gender-neutral by identifying and replacing or removing gender-specific words with corresponding neutral terms, with the help of predefined word banks (Zhao et al., 2017; Hendricks et al., 2018).

Flickr30k (Plummer et al., 2015) The second large-scale image-text dataset employed in our experiment is Flickr30K, which contains 31,000 images obtained from Flickr. Adhering to the partitioning scheme presented in Plummer et al. (2015), we allocate 1,000 images each for validation and testing, with the remaining images designated for training. We obtain the ground truth of gender attributes of images in Flickr30k in the same way as MS-COCO (Wang et al., 2021a), as we detect the gender-specific words in the caption to determine the gender attributes of its corresponding image.

D Baseline Models

Random Select To simulate an ideal scenario, where image features bear no dependency to gender (and race) attributes, for a neutral query c, we randomly select K candidates from the true relevant image set V_c^*, with replacement . As for each query c, the size of the relevant image set $|V_c^*|$ is at most 100. Using sampling with replacement simulate the situation that the gender attribute distribution is fixed, and irrelevant to retrieval algorithm. We report AbsBias@K for reference. The Recall@K is omitted since the value is meaningless, as we only sample from the true relevant image set V_c^*.

CLIP (Radford et al., 2021) We consider OpenAI's CLIP ViT-B/16 (Radford et al., 2021) as the VL model for all debiasing methods. Specifically, the image encoder $f_\phi(\cdot)$ is a Vision Transformer (ViT) (Dosovitskiy et al., 2020) comprising 12 transformer blocks of width 768, with 12 self-attention heads in each block. ViT processes images of size 224×224 by dividing them into 16×16 patches and outputs 512-dimensional image features by linear projection. The text encoder $f_\psi(\cdot)$ is a standard text transformer (Vaswani et al., 2017) with masked self-attention, consisting of 12 transformer blocks of width 512 and 8 self-attention heads in each block, with a linear projection layer at the end as well. The CLIP ViT-B/16 model is loaded with pre-trained weights provided by OpenAI (Radford et al., 2021). All the following debiasing methods use this pre-trained CLIP ViT-B/16.

MI-*clip* (Wang et al., 2021a) MI-*clip* (Wang et al., 2021a) clips the fixed number of output dimensions of the image encoder in CLIP to reduce the mutual information between image features and demographic attribute distribution. For MI-*clip* in Table 1, we clip 312 dimensions of output image features. These 312 dimensions were chosen by examining the reduction in bias and while maintaining retrieval performance. We also show a trade-off between bias reduction and retrieval performance of MI-*clip* with the number of clipped dimensions from 100 to 500 on the Occupation 1 dataset in Figure 4.

Adversarial Training (Edwards and Storkey, 2015) For adversarial training, we use the same minimax problem setup as in (Edwards and Storkey, 2015). The encoder is the original CLIP image encoder. The decoder is realized by a ViT with 8 vision transformer blocks, and the adversarial predictor is a 3-layer MLP. Training employs the same loss function in Edwards and Storkey (2015), where the final loss is the sum of the cost of reconstructing v from $f_\phi(v)$, a measure of dependence between $f_\phi(v)$ and $g(v)$ and the error of target task (*i.e.,* image-text aligning loss $\mathcal{L}_{\text{NT-Xent}}$ (Radford et al., 2021)). We assign different weights to the loss of dependence measuring loss, in order to demonstrate the trade-off between bias reduction and retrieval performance. We report the adversarial learning results when the weight of measuring dependence loss is 0.7, and the weights for the other two losses are both 0.15. Additional weight combinations were considered and shown in Figure 4.

Debias Prompt (Berg et al., 2022) The Debias Prompt method (Berg et al., 2022) also leverages an adversarial learning framework. However, instead of just fine-tuning the image and text encoders, they also prepended zero-initialized learnable prompts before inputting query tokens. Considering that their debias-prompt model is already debiased for gender and race attributes, we directly evaluate their pre-trained model (sourced from their github repository `https://github.com/oxai/debias-vision-lang`) on the Occupation 1 and Occupation 2 datasets.

CLIP-FairExpec (Mehrotra and Celis, 2021) We tailor FairExpec (*not originally proposed* for TBIR) to our task by integrating it with CLIP and our proposed gender predictor $\hat{g}(\cdot)$. We refer to this adaptation as CLIP-FairExpec in our experiments.

Using binary gender as an example for simplicity, our CLIP-FairExpec treats the image-text similarity output as the utility score for each image. Then, the objective of the optimization is to maximize the total similarity scores for selecting K images corresponding to a query c. The noise estimate q in the original FairExpec is derived from the probability output of our attribute predictor $\hat{g}(v)$. We use the probability output from the attribute predictor $\hat{g}(v)$ as the noise estimate q. Also, there is a constraint on the sum of the noise estimates q such that the sum is at least $L - \delta K$ and at most $U - \delta K$, where L and U is the lower bound and upper bound for the sum of the noise estimate, respectively. $\delta \in (0, 1)$ is a noise tolerance level, that controls how much the constraints can be violated due to the presence of noise. Since our fairness objective is equal representation for each gender attribute class, we wish the sum of noise estimate for each class of gender attribute is equal. Hence, we set the $L = U = K/2$. In order to force our model to prioritize minimizing bias over maintaining performance, we choose a very small $\delta = 0.001$. We select K images from $\mathcal{V}$ with respect to a neutral query c based on the above constrained optimization problem. Further, each selection is solved by the GUROBI solver (Gurobi Optimization, LLC, 2023). Upon solving for all selections for queries in C, we compute the AbsBias@K and Recall@K presented in Table 1.

SCAN (Lee et al., 2018) We consider the Stacked Cross Attention Network (SCAN) (Lee et al., 2018) as an alternative VL model to CLIP. SCAN is a specialized in-domain training model, so it is trained on the MS-COCO training dataset and tested on the MS-COCO test dataset. Similarly, for the experiments with Flickr30k, the model is trained on the Flickr30k training set and then tested on the Flickr30k test set. We use official implementation of SCAN from `https://github.com/kuanghuei/SCAN`.

FairSample (Wang et al., 2021a) To mitigate bias during the training of SCAN, we implement the FairSample approach as recommended by Wang et al. (2021a). We maintain the same hyperparameters settings as Lee et al. (2018). To address the bias arising from the unbalanced gender distribution within training batches, FairSample is proposed in the following way: for every positive image-text pair (v, c) within a training batch, we first identify if the query c is gender-neutral or gender-specific. If the training query c is gender-neutral, a negative image is sampled from either the male or female image sets, each with a probability of 1/2. However, if the query is gender-specific, we maintain the original negative sampling strategy, thereby preserving the model's ability to generalize effectively on such queries.

E Post-hoc Bias Mitigation (PBM)

E.1 Engineering Details

PBM - Supervised Classifier We can determine the gender attributes with a pre-trained image classifier. Here, the image classifier is pre-trained on MS-COCO training set with gender attribute annotations from Zhao et al. (2021). The image classifier is a 3-layer multi-layer perceptron (MLP) as shown in Table 5, that takes the image representation from the original CLIP as input. We empirically show that the image classifier can be highly accurate even using a light-weight classification MLP. The F1-score for gender attribute prediction is 92.8%.

PBM - Zero-Shot Embedding We describe the first of the two types of zero-shot inference described in Section 3.5. For zero-shot inference based on the embedding approach, we choose the text embeddings for {"Unknown Gender", "Man", "Woman"} tokens to classify the gender attributes of images, and {"Unknown Skin", "Fair Skin", "Dark Skin"} for categorizing race attributes. The gender or race attribute of an image v is determined by which text embedding has the maximum similarity score to the image representation $f_\phi(v)$.

PBM - Zero-Shot Prompt For the second zero-shot inference described in Section 3.5, namely, the prompt method, we prepend adjectives to the text query c. We use {"", "Male", "Female"} for gender attributes and {"", "Fair-skinned", "Dark-skinned"} for race attributes. The gender attributes for each image retrieved by the query c is determined by which prompted query has the maximum similarity score to the image representation $f_\phi(v)$.

PBM - Ground-Truth Attribute (Gender or Skin-tone) We use the annotations in the dataset as the predicted attributes $\hat{g}(v)$ for reference. This shows the upper-bound performance of our method if all gender predictions are correct (known).

E.2 Additional PBM Results

In Table 4, we showcase the results of applying PBM to CLIp models that has been debiased by other approaches, such as MI-clip, Adversarial Learning, and Debias Prompt. When PBM is utilized in conjunction with other debiasing strategies, it exhibits a unique bias-recall trade-off, thus catering to a variety of application scenarios.

Table 4: Results of applying PBM - Supervised Learning on modified or fine-tuned CLIP.

Method	Occupation 1 - Gender		Occupation 2 - Race	
	AbsBias@100 (↓)	Recall@100(↑)	AbsBias@100(↓)	Recall@100(↑)
PBM	.1404	50.3	.0955	37.9
MI-*clip* - **PBM**	**.0780**	42.1	**.0737**	29.1
Adversarial Training - **PBM**	.1000	39.6	.0997	35.7
Debias Prompt - **PBM**	.1711	**52.1**	.1035	**40.6**

F Neural Network Architectures

We summarize the details of the neural networks employed in our experiments in Table 5. For the Image Encoder, the Patch Extraction (dimensions: 16,16) extracts 196 non-overlapping 16 × 16 patches from the 224×224 image. These extracted patches are subsequently flattened. The subsequent Positional and Linear Embedding (768) maps these patch vectors onto a 768-dimensional space and adds 2D positional embeddings of patches to the 768-dimensional vectors. Next, 12 Vision Transformer Blocks (768, 12) processes the 768-dimensional embeddings. Each of these blocks features 12 self-attention heads. Lastly, the output embedding is obtained from a unique classification token ([CLS]) that we add to the input sequence of patch embeddings. The output from [CLS] Token 1 × 768 is then reduced from 768 dimensions to 512 dimensions using a Linear Projection (512).

Similarly in the Text Encoder, the initial phase involves Positional and Token Embedding (512). This step maps each token in the input text onto a 512-dimensional vector space and integrates positional embeddings into these vectors. Following this, the text encoder employs 12 Transformer Blocks

Table 5: The architecture of each component of CLIP and the MLP used in our experiments.

ImageEncoder(·)

Layer	**Type**
1	Patch Extraction(16, 16)
2	Positional and Linear Embedding(768)
4 - 15	Vision Transformer Blocks(768, 12)
16	[CLS] Token 1×768
17	Linear Projection (512)

TextEncoder(·)

Layer	**Type**
1	Positional and Token Embedding (512)
2 - 13	Transformer Blocks (512, 8)
14	[CLS] Token 1×512
15	Linear Projection (512)

MLP(·)

Layer	Type
1	fc-512 + BatchNorm + ReLU()
2	fc-512 + BatchNorm + ReLU()
3	fc-512 + BatchNorm + ReLU()
4	fc-n_class + Softmax()

(512, 8) to process these 512-dimensional embeddings. Each of these blocks contains 8 self-attention heads. Finally, the output embedding is derived from [CLS] Token 1×512. The subsequent Linear Projection (512) then maps the extracted text representation onto the multi-modal embedding space that aligns with the image embeddings.

G Computation Resources

All of our experiments ran on one NVIDIA TITAN Xp 12GB GPU with CUDA version 11.5.

H Code and Data Availability

Occupation 1 dataset is available at https://github.com/mjskay/gender-in-image-search.

Occupation 2 dataset can be downloaded from https://drive.google.com/drive/folders/1j9I5ESc-7NRCZ-zSDOC6LHjeNp42RjkJ.

MS-COCO dataset can be access through https://cocodataset.org/#home, and its crowd-sourced gender/racial annotations from https://princetonvisualai.github.io/imagecaptioning-bias/.

Flickr30k dataset can be access via https://shannon.cs.illinois.edu/DenotationGraph/. And the gender word banks to identify the gender attributes of Flickr30k's images is avaliable in the Appendix of the paper by Wang et al. (2021a).

I Broader Impact

The recent years constituting what can be called the model architecture unifying era, witnessed a seismic shift from small task-specific models to foundation models containing billions of parameters, with numerous applications deployed based on such large models. However, as artificial intelligence (AI) systems become more prevalent, the challenging question of fairness becomes more urgent. The concept of fairness in machine learning revolves around creating algorithms and models that DO NOT discriminate against certain groups based on gender, race, socioeconomic status, or any other potentially biasing factors. As machine learning algorithms are increasingly used in decision-making processes, from job applications, college admissions, to criminal justice and healthcare, subsets of the population who represent minorities may see unfavoring model performance compared to individuals in majority groups. Therefore it is imperative to develop unbiased machine learning systems such that decisions are made fairly and equitably. Our study concerning fair image retrieval, among many other fairness research works, can be used to inform policymakers about the potential risks and benefits of AI systems, potentially enacting new laws and regulations to ensure that these systems are utilized responsibly and ethically.

Specifically, the biased performance of a model is possibly caused by statistical skewness both in the training and testing sets. Existing methods mainly focus on enforcing independence between the model's output and sensitive attributes during training. However, much less effort has been made to mitigate bias during test-time, a potentially vital component of the debiasing procedure. Many machine learning systems are deployed in a setting where the biased testing set is almost guaranteed, and as such, may suffer from fairness concerns. Importantly, PBM is able to dissociate the ranking similarity from sensitive/protected attributes (*e.g.*, gender) thus reducing bias, meaning that image candidates share an equal chance to be retrieved even in an unbalanced testing set. We do not claim that PBM guarantees fairness, and there is always the risk that it may be misinterpreted or exploited, but we hope that PBM encourages a more inclusive approach to AI development.