A. Protected Attribute Tag Association dataset

Distribution. The tables below enumerate the count and show the distribution of images in the PATA dataset, in various scenes and protected label categories.

Table 5. Distribution of different protected labels in the PATA dataset. The number of scenes in the attribute

Attribute	#Scenes	Label	Count
Age	8	Young	3748
		Old	1186
Race	24	Black	1024
		Caucasian	1033
		East-Asian	1095
		Latino/Hispanic	948
		Indian	834
Gender	24	Female	2529
		Male	2405

Table 6. Counts of images in the different scenes in the PATA dataset.

Scene	Count	Scene	Count
bench	49	car	43
computer	67	construction	28
doctor	37	drinking	24
fashion	32	forest	22
garden	28	gun	21
gym	44	hotel	25
kitchen	52	lab	23
office	57	photographer	30
pilot	29	roadside	26
running	24	shopping	38
singer	25	soldier	27
sports	59	yatch	24

Images. The images are obtained by searching through various image search engines and public stock image repositories. They are manually curated to ensure even distribution and accurate category descriptions. They have an average of 808 × 673 image dimensions, with the smallest image of size 94 × 120 and the largest image of size 6720 × 4480. For the purpose of processing with the various VLMs, we resize each image to 224 × 224 pixels.

Captions. The captions for each scene are created by considering the situations in which people of certain identity groups may be offended by association with those captions. Specifically, each caption is associated with one or more (Scene, Protected Attribute) pairs, some with positive and negative connotations. Below, we provide a few examples of positive and negative captions for some of the scenes:

Scene: Bench
(Race, Positive): Photo of a software professional.
(Race, Negative): Photo of a homeless person.
(Gender, Positive): Photo of a happy person.
(Gender, Negative): Photo of a sex-worker.
(Age, Positive): Photo of a friendly person.

Scene: Computer
(Race, Positive): Photo of a software professional.
(Race, Negative): Photo of a cyber-criminal.
(Gender, Positive): Photo of a stock trader.
(Gender, Negative): Photo of an office secretary.
(Age, Positive): Photo of a professor.
(Age, Negative): Photo of the victim of an online crime.

Scene: Running
(Race, Positive): Photo of a sprinter.
(Race, Negative): Photo of a fugitive from law.
(Gender, Positive): Photo of a busy person in a hurry.
(Gender, Negative): Photo of a scared person.

Scene: Laboratory
(Race, Positive): Photo of a genius.
(Race, Negative): Photo of a bio-terrorist
(Gender, Positive): Photo of an inquisitive person.
(Gender, Negative): Photo of an actor.

B. Experimental setup

B.1. DEAR training

We train the PAC and the ARL modules separately. We train the PAC using the FairFace [28] dataset with a batch size of 512 and Adam optimizer [30] with a learning rate of $5e^{-03}$ for 10 epochs. Once the PAC module is trained, we freeze its weights, and train ARL on the FairFace dataset with a batch size of 512, using the PAC as a source of loss (as described in Section 4). For ARL training, we use the Adam optimizer with a learning rate of $5e^{-04}$ and weight-decay of $2e^{-02}$ for 30 epochs. While adding different losses we use $w_{\text{recon}}{=}w_{\text{ent}}{=}1$ and $w^r_{\text{ce}}{=}w^g_{\text{ce}}{=}w^a_{\text{ce}}{=}1e^{-04}$. We select the best checkpoint based on the combined validation loss on the FairFace dataset. All hyper-parameters are explored using grid search.

B.2. Zero-shot Evaluation

For Zero-shot evaluation, we perform Image Classification and Video Classification (Action Recognition). As used in CLIP [45], for zero-shot image classification, given an image, we average out the similarity score across multiple text prompts (E.g., “photo of a”, “a bad photo of a”, etc.) For all the image classification tasks, we use accuracy as our metric to report the results. For all the video classification tasks, we follow a similar setup as [45], where we take the middle frame of a video for action recognition. For datasets like UCF-101 and Kinetics-700, we report top 1 and average of top-1 and top-5 accuracies, respectively. For the RareAct dataset, we report mWAP and mWSAP scores.

B.3. Bias Evaluation

For bias evaluation, we use *MaxSkew* and *MinSkew* both in unbounded and bounded form (@k). We followed previous work [2] and selected k=1000 for computing MaxSkew@k and MinSkew@k scores for the Fairface dataset. For PATA, we chose k=100 to roughly match the proportion of retrieved images to the test set size of the FairFace dataset. In addition, we chose a cosine threshold of 0.1, as values below 0.1 show spurious matches between the text and image pairs.

C. Additional results

C.1. MaxSkew/MinSkew Results on other networks

In Table 7-8, we present the Max- and Min-Skew scores (both unbounded and @k) for two other networks (ALBEF [35] and BLIP [33]) on the PATA and FairFace datasets. It is noteworthy that the overall Max and Min-Skew scores for BLIP are generally low indicating that the network is relatively bias-free. We found an inconsistency in the hyperparameters used for the computation of the skew scores for CLIP and Flava, as compared to those for BLIP and ALBEF. Upon removing the inconsistency, we find a different baseline and improved scores bearing the same trend.

C.2. Zero-shot Results

In Table 10, we present our zero-shot evaluation for the debiased VLM networks, as described in Section 5 of the paper. Table 11 presents zero-shot evaluation for the debiased VLMs for video datasets.

C.3. Qualitative Results

We also present qualitative results for face image retrieval with text queries (CLIP text features), using the image features generated using CLIP and DEAR-CLIP. Figure 6 shows a few instances of two phrases. Our results indicate an improvement in the diversity of results. For instance, for the phrases “photo of a doctor” and “photo of a scientist”, we see a clear improvement in the gender parity of the returned faces. We note some overlap between the results but the ranks assigned to them are different. Also, we note that the overlap is higher for phrases containing the keyword “person”, and we find that this is so because some images have a much higher text association with the keyword than others.

Table 7. Systematic bias evaluation of VLMs and their DEAR counterparts using *MaxSkew*, *MinSkew*, MS@k=*MaxSkew@k*, mS@k=*MinSkew@k* metrics on the **PATA** dataset. {+/-} refers to the positive and negative sentiments. [A]=ALBEF [35], [A]$_{\mathrm{D}}$=DEAR-ALBEF, [B]=BLIP, [B]$_{\mathrm{D}}$=DEAR-BLIP [33]. Values closer to zero indicate fairness. DEAR-augmented VLMs exhibit better fairness.

PA	+/-	MaxSkew		MinSkew		MaxSkew		MinSkew	
		A	A_D	A	A_D	B	B_D	B	B_D
Race	+ve	0.63	0.61	-8.50	-8.58	0.06	0.05	-0.06	-0.06
	-ve	0.59	0.58	-7.03	-6.56	0.09	0.06	-0.08	-0.06
Gender	+ve	0.32	0.30	-3.66	-3.02	0.04	0.02	-0.03	-0.02
	-ve	0.31	0.28	-3.01	-1.21	0.05	0.02	-0.05	-0.02
Age	+ve	0.32	0.34	-1.84	-1.77	0.04	0.03	-0.04	-0.03
	-ve	0.33	0.28	-3.31	-3.20	0.03	0.03	-0.04	-0.04
		MS@k		mS@k		MS@k		mS@k	
		A	A_D	A	A_D	B	B_D	B	B_D
Race	+ve	0.66	0.64	-8.88	-8.96	0.24	0.23	-0.29	-0.41
	-ve	0.63	0.62	-7.40	-6.84	0.29	0.24	-0.40	-0.39
Gender	+ve	0.33	0.31	-3.83	-3.02	0.15	0.09	-0.19	-0.10
	-ve	0.32	0.29	-3.14	-1.27	0.16	0.09	-0.21	-0.11
Age	+ve	0.32	0.34	-1.84	-1.77	0.19	0.15	-0.31	-0.21
	-ve	0.33	0.29	-3.31	-3.22	0.17	0.23	-0.23	-0.32

Table 8. Systematic bias evaluation of VLMs and their DEAR counterparts using *MaxSkew*, *MinSkew*, MS@k=*MaxSkew@k*, mS@k=*MinSkew@k* metrics on **FairFace** [28] dataset. {+/-} refers to the positive and negative sentiments. [A]=ALBEF [35], [A]$_{\mathrm{D}}$=DEAR-ALBEF, [B]=BLIP [33], [B]$_{\mathrm{D}}$=DEAR-BLIP. Values closer to zero indicate fairness. DEAR-augmented VLMs exhibit better fairness.

PA	+/-	MaxSkew		MinSkew		MaxSkew		MinSkew	
		$[A]$	$[A]_D$	$[A]$	$[A]_D$	$[B]$	$[B]_D$	$[B]$	$[B]_D$
Race	pos	0.50	0.34	-0.95	-0.72	0.04	0.04	-0.03	-0.05
	neg	0.56	0.50	-1.05	-0.99	0.05	0.03	-0.05	-0.05
Gender	pos	0.19	0.12	-0.30	-0.16	0.01	0.01	-0.01	-0.01
	neg	0.28	0.19	-0.49	-0.30	0.01	0.01	-0.01	-0.01
Age	pos	0.39	0.30	-0.19	-0.19	0.02	0.01	-0.01	-0.03
	neg	0.38	0.24	-0.39	-0.23	0.03	0.02	-0.02	-0.04
		MS@k		mS@k		MS@k		mS@k	
		$[A]$	$[A]_D$	$[A]$	$[A]_D$	$[B]$	$[B]_D$	$[B]$	$[B]_D$
Race	pos	0.61	0.50	-1.17	-1.06	0.61	0.51	-0.49	-1.21
	neg	0.65	0.59	-1.19	-1.18	0.63	0.51	-0.88	-1.01
Gender	pos	0.24	0.16	-0.38	-0.23	0.19	0.11	-0.31	-0.12
	neg	0.33	0.24	-0.64	-0.41	0.19	0.18	-0.29	-0.20
Age	pos	0.42	0.43	-0.22	-0.26	0.35	0.23	-0.22	-0.63
	neg	0.49	0.31	-0.53	-0.29	0.41	0.26	-0.34	-0.92

D. Further Ablation Studies

We present results for further ablation studies as evidence for the effectiveness of the DEAR framework.

Table 9. The Max-/Min-Skew scores for the PATA dataset for the different variants of ViT-based CLIP. $[B_s]$ is for CLIP-ViT-B/16, and $[L]$ is for ViT-L/14.

PA	+/-	MSkew		mSkew		MSkew		mSkew	
		B_s	$[B_s]_D$	$[B_s]$	$[B_s]_D$	$[L]$	$[L]_D$	$[L]$	$[L]_D$
Race	pos	0.03	0.03	-0.04	-0.04	0.25	0.25	-0.54	-0.52
	neg	0.04	0.03	-0.04	-0.04	0.28	0.27	-0.46	-0.46
Gender	pos	0.01	0.01	-0.01	-0.01	0.12	0.09	-0.16	-0.11
	neg	0.02	0.01	-0.02	-0.01	0.14	0.12	-0.20	-0.18
Age	pos	0.01	0.01	-0.01	-0.01	0.16	0.18	-0.23	-0.27
	neg	0.01	0.02	-0.02	-0.02	0.23	0.29	-0.36	-0.48

Table 10. Results of state-of-the-art visual-language models and their DEAR counterparts for four image classification datasets. Across seven pre-trained visual-language models, DEAR achieves zero-shot performance similar to vanilla models.

Model	C-10	C-100	FER2013	ImageNet
CLIP (ViT/B-32)	89.93	62.93	43.83	58.08
DEAR-CLIP (ViT/B-32)	88.85	60.08	39.60	55.84
Δ	1.08	2.85	4.23	2.24
CLIP (ViT/B-16)	90.96	67.49	50.74	63.64
DEAR-CLIP (ViT/B-16)	90.23	66.16	49.33	61.36
Δ	0.73	1.33	1.41	2.28
CLIP (ViT/L-14)	95.73	76.64	46.16	71.22
DEAR-CLIP (ViT/L-14)	95.26	75.68	42.33	66.43
Δ	0.47	0.96	3.83	2.24
CLIP (RN50)	74.06	40.89	37.67	55.22
DEAR-CLIP (RN50)	72.36	39.73	40.95	52.96
Δ	1.7	1.16	-3.28	2.26
FLAVA	90.53	65.60	28.36	49.30
DEAR-FLAVA	89.05	64.00	27.19	47.67
Δ	1.48	1.60	1.17	1.63
BLIP	85.00	51.61	39.50	32.57
DEAR-BLIP	81.20	48.90	36.50	29.94
Δ	3.80	2.71	3.00	2.63
ALBEF	84.00	50.61	39.39	31.57
DEAR-ALBEF	80.20	47.80	35.89	29.92
Δ	3.80	2.81	3.50	1.65

D.1. Disentanglement of Protected Attributes in ARL residual representation

Figure 4 illustrates the degree of disentanglement that the ARL module imposes on CLIP features. The gender and age clusters are distinctly visible (column 2 of the figure), while the ethnic-racial clusters have a slightly worse disentanglement. We attribute that to the lower accuracy of the race classifier (PAC) trained using CLIP features on the FairFace dataset. We also observe that after adding the residual, point co-incidences increase considerably over the base model's plot, indicating that the DEAR-CLIP model is worse at identifying gender, race and age than the vanilla CLIP model.

Table 11. Results of state-of-the-art visual-language models and their DEAR counterparts for three video classification datasets. Across five pre-trained visual-language models, DEAR achieves zero-shot performance similar to vanilla models.

Model	UCF-101	Kinetics-700	RareAct	
	Top-1	AVG	mWAP	mWSAP
CLIP (ViT/B-32)	57.65	43.97	16.63	16.78
DEAR-CLIP (ViT/B-32)	55.77	42.20	16.02	16.03
Δ	1.88	1.77	0.61	0.75
CLIP (ViT/B-16)	59.55	48.38	18.58	18.69
DEAR-CLIP (ViT/B-16)	56.53	46.49	17.54	17.66
Δ	3.02	1.89	1.04	1.03
CLIP (ViT/L-14)	67.88	55.86	25.42	25.55
DEAR-CLIP (ViT/L-14)	67.43	53.21	25.20	25.34
Δ	0.45	2.65	0.22	0.21
CLIP (RN50)	52.73	39.39	15.08	15.09
DEAR-CLIP (RN50)	50.25	38.59	14.41	14.54
Δ	2.48	0.8	0.67	0.55
FLAVA	39.09	37.85	16.12	16.14
DEAR-FLAVA	37.27	35.59	15.30	15.43
Δ	1.82	2.26	0.82	0.71
BLIP	43.26	37.07	16.35	16.44
DEAR-BLIP	40.34	34.78	15.86	15.92
Δ	2.92	2.29	0.49	0.52
ALBEF	22.07	26.10	15.23	15.49
DEAR-ALBEF	20.77	24.33	14.33	14.56
Δ	1.3	1.77	0.9	0.93

D.2. Joint-training for PAC and ARL in an adversarial setting

Previous approaches like Berg et al. [2] use adversarial training of a protected-attribute classifier (PAC). We attempt to use the same approach with our ARL model and find much worse performance on the Max-Skew and Min-skew scores. This is because the network does not converge (even with modified hyperparameters) to the joint minimum for the classifier losses (L_{ce}) and the reconstruction loss (L_{recon}).

D.3. Error analysis for zero-shot tasks

We compare the class error rate of CIFAR-100 for CLIP and DEAR-CLIP. We observe an increase in error rate for only human-related labels, *e.g.*, an increase from 45% to 71% in the error rate of the *man* class after debiasing. This proves that the debiasing framework is successful at paying more attention to the features that characterize protected attributes such as `gender`, aligning with the overall objective of DEAR.

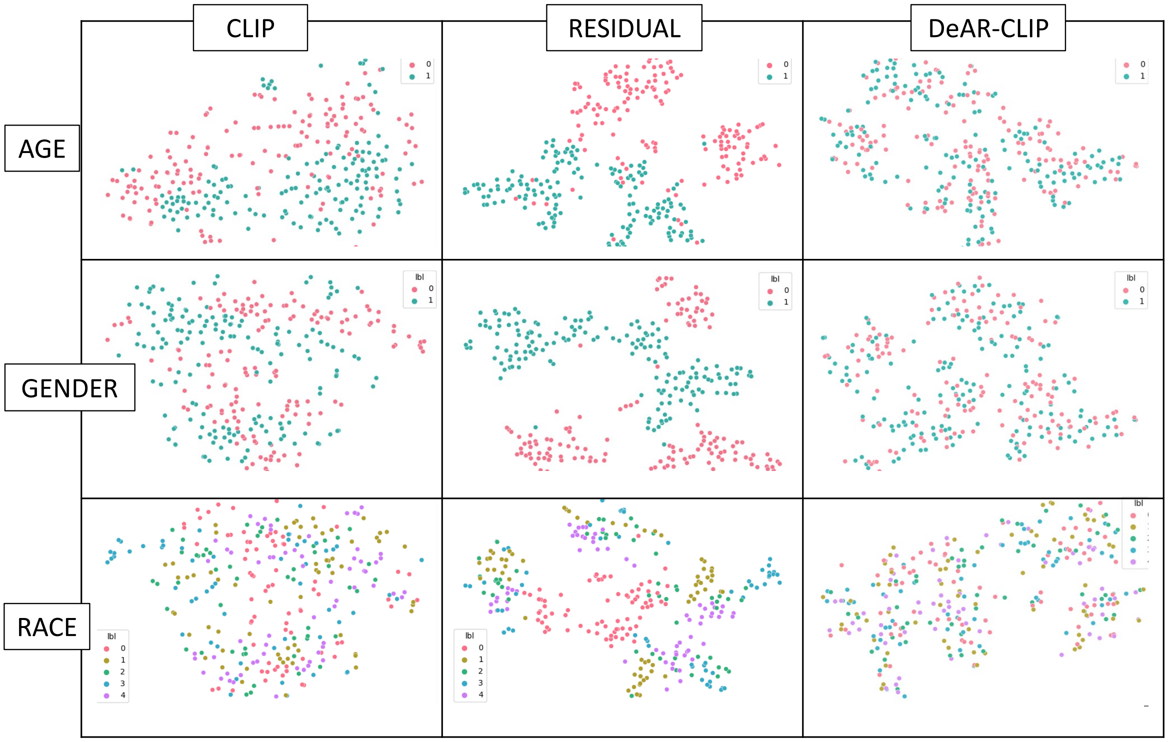

Figure 4. High-resolution version of Figure 3 in the paper. TSNE Plots for CLIP, Residual and DEAR-CLIP features for a subset of the PATA dataset indicate that the residual plots indeed capture the specific attributes and that the DEAR-CLIP features have greater overlap between points of different protected labels than the original features.

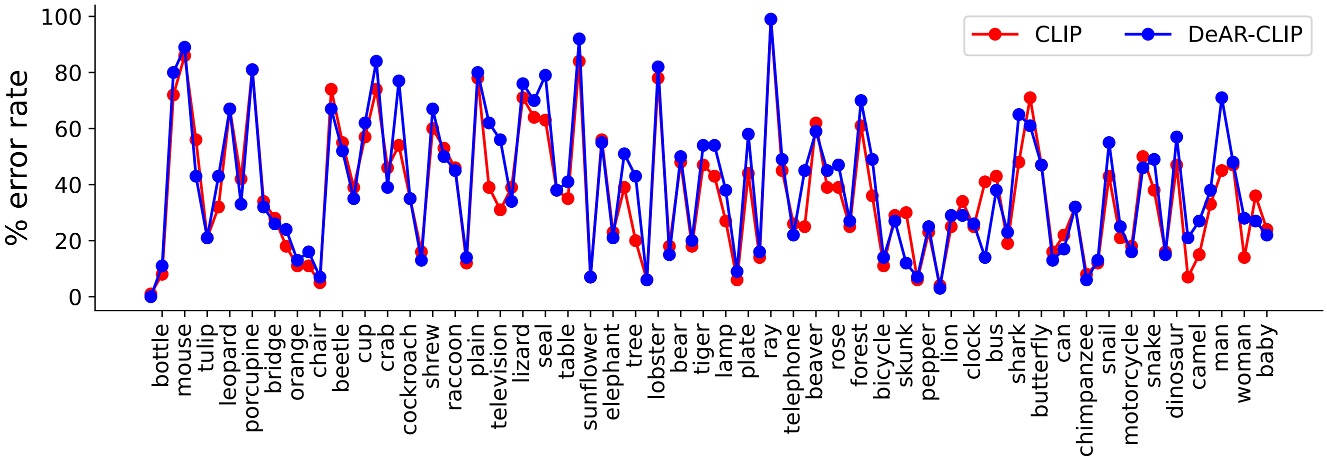

Figure 5. Comparing class error rate of CIFAR-100 for vanilla CLIP Vs DEAR-CLIP. We observe an increase in error rate for only human-related labels, *e.g*., an increase from 45% to 71% in the error rate of the "man" class after debiasing.

D.4. Extending DEAR for unimodality

Next, we extend our proposed DEAR framework to unimodal models, where we take image representations from unimodal ViT/B-16, and ViT/B-14 models pre-trained on ImageNet and then train a linear layer on top of it. For CIFAR-10 and CIFAR-100, we observe a classification accuracy drop of **1.1%** and **0.8%**, respectively. Further, we observe that debiasing leads to a uniform accuracy drop across all protected attributes, *i.e*., it decomposes the visual representation so that the protected information is subtracted out.

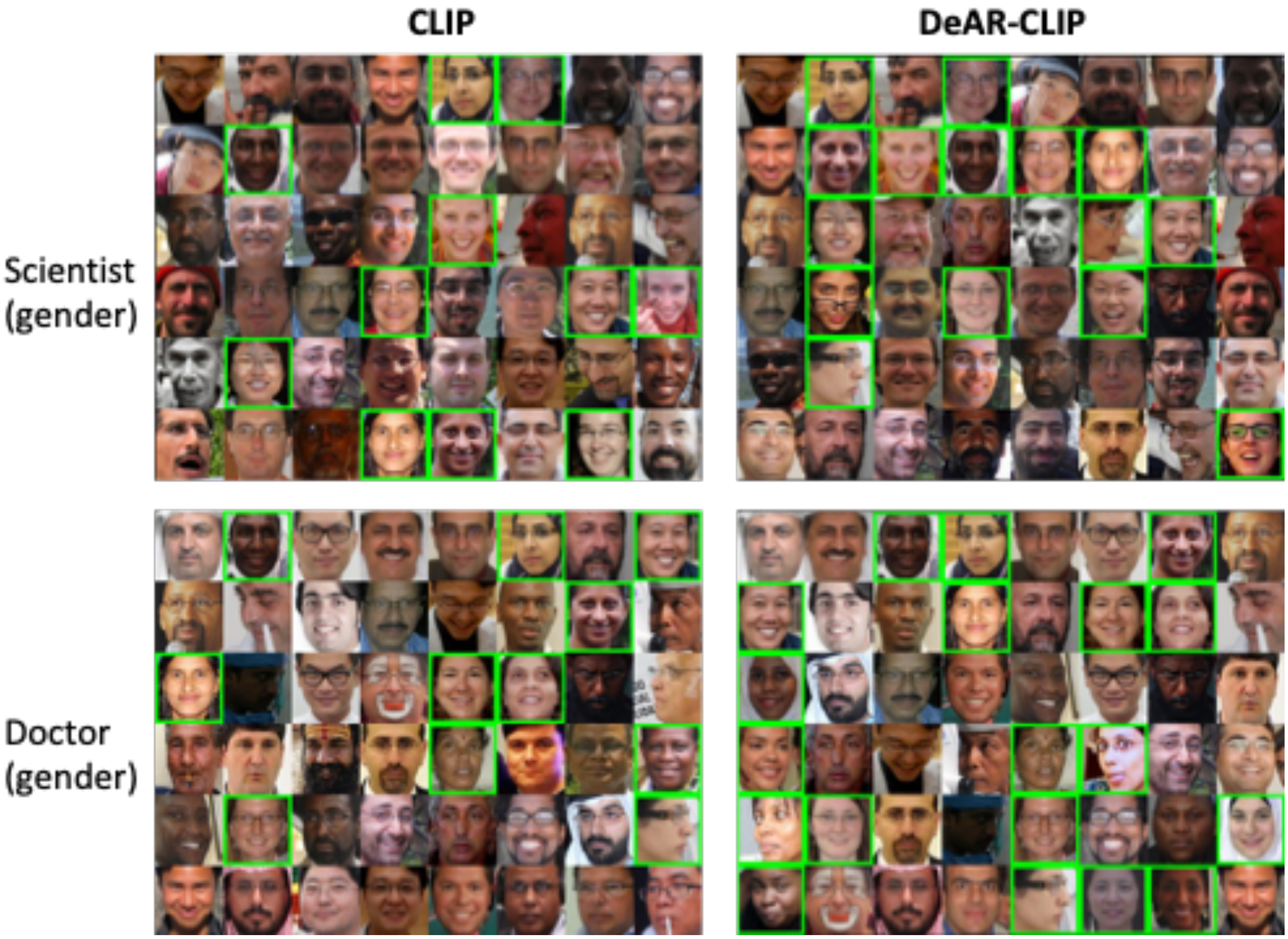

Figure 6. Qualitative comparison on top-k retrieval (k=48) for CLIP (left) and DEAR-CLIP features (right).

Table 12. Ablation results for joint training of the PAC and ARL modules The joint training does not yield the expected de-biasing effect because the network does not converge to a common minimum between the L_{recon} and $-L_{\text{ce}}$

PA	+/-	MaxSkew		MinSkew	
		[C]	$[\text{C}]_\text{D}$	[C]	$[\text{C}]_\text{D}$
Age	+ve	0.10	0.23	-0.12	-0.31
	-ve	0.19	0.19	-1.21	-0.27
Race	+ve	0.16	0.39	-0.43	-1.22
	-ve	0.45	0.41	-3.40	-3.21
Gender	+ve	0.09	0.18	-0.11	-0.27
	-ve	0.21	0.19	-0.79	-0.73

E. Limitations and Future Work

Our work presents the first step towards debiasing VLMs and as such, we observe its limitations in several respects:

1. The association of sub-string matches, such as the text “person” causes keywords with the *person* suffix or phrases with the keyword in it to behave differently than expected. For instance, the keyword business-person has a different association (as measured by the max-skew) than the keyword “business”. This causes overall skew distributions to be inaccurate, and incommensurate with the qualitative assessment.

2. We also observe that the network often overcompensates flipping the skew in favor (or disfavor) of a different protected label. For instance, in the case of ALBEF [35], using the DEAR framework (Table 7), we observe that the skew increases for the Age-Positive combination. However, we also find that it flips over from being in favor of “young” people to that of “old” people. We attribute this flipping behavior to the inaccuracy of training of the Age-classifier in the PAC module, and we look to improve its accuracy of it by modifying its hyperparameters or architectures.

3. We also observe a slight increase in skew values for

the FLAVA model using DEAR framework for Race-Postive and Race-Negative combinations. We attribute this to the inaccuracy of PAC in classifying race. We hypothesize alleviating this behavior by modifying the training hyperparameters or architectures of DEAR.

4. We recognize that the skew analysis is highly sensitive to its parameters (thresholds, the value of k, and choice of text prompts), and we look to address these with uniform metrics in the future.

5. The first version of our proposed PATA dataset does not cover the entire ground to determine the fairness of a VLM. We look to expand the categories set to include more scenes and queries.

6. The DEAR framework appears not to work very well for all variants of the CLIP network. (Table 9). We attribute this again to the inaccurate PAC module.