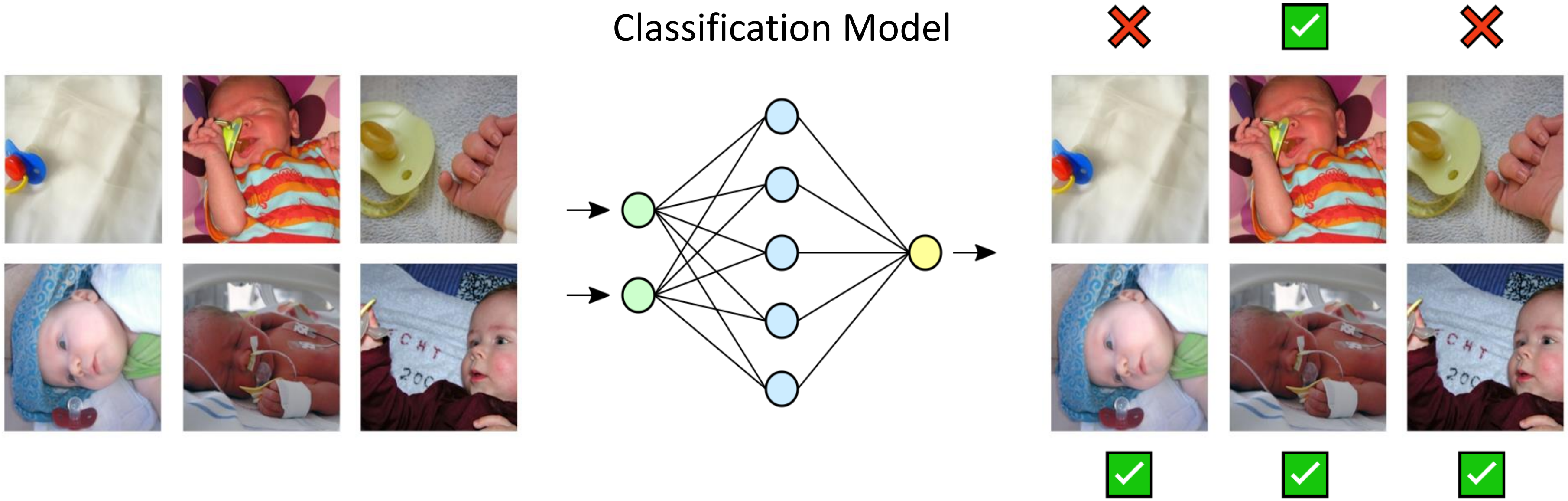
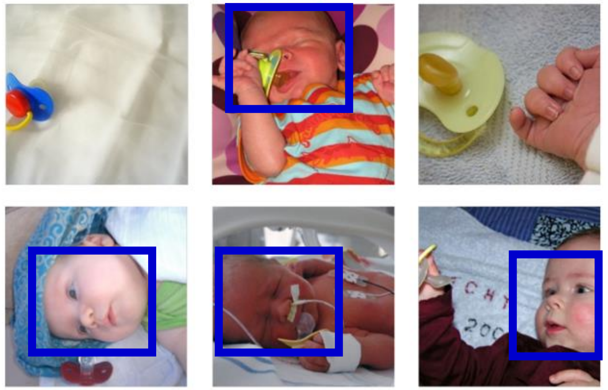
Chapter 4: Refining Skewed Perceptions in Vision-Language Models

Refining Skewed Perceptions in Vision-Language Models through Visual Representations.
Haocheng Dai, Sarang Joshi.
In submission.



The baby pacifier class in ImageNet

Classification Outcome



The baby pacifier class in ImageNet is <u>spuriously correlated</u> with the presence of babies.

When trying to identify hair color

	Non-blond Woman	Non-blond Man	Blond Woman	Blond Man
CelebA				
Training #	71629 (44%)	66874 (41%)	22880 (14%)	1387 (1%)
Validation #	8535	8276	2874	182
Accuracy	97.78%	99.86%	85.88%	36.99%

Liu et al. 2015

The blond hair class in CelebA is spuriously correlated with

How previous work resolve this?

Without knowing group label

Just Train Twice (JTT)
Liu et al., 2021

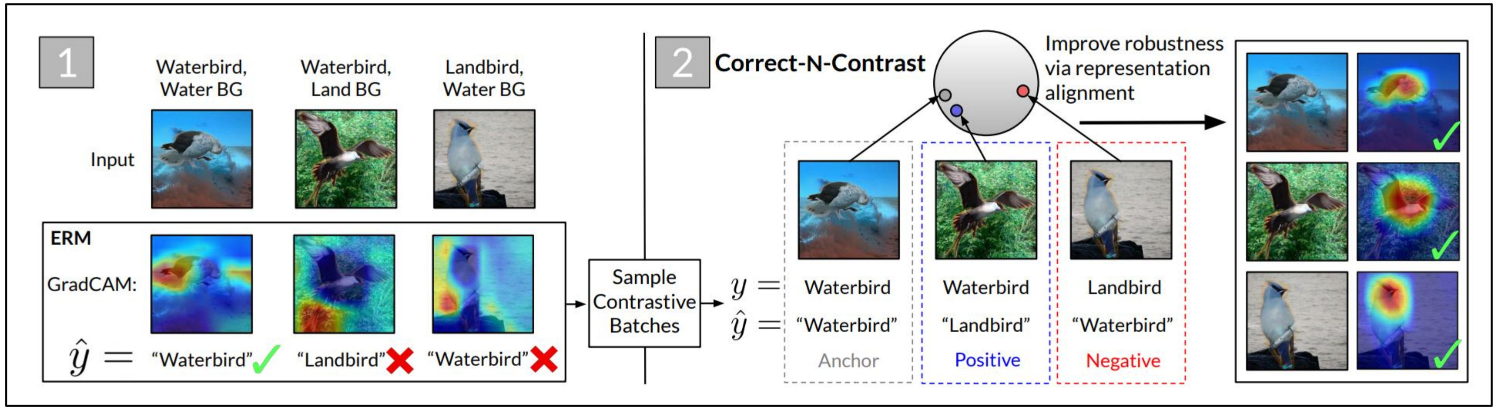
1. Identification

$$E = \{(x_i, y_i) \text{ s.t. } f_{\text{id}}(x_i) \neq y_i\}.$$

2. Upweighting

$$J_{\text{up-ERM}}(\theta, E) = \left(\lambda_{\text{up}} \sum_{(x,y)\in E} \ell(x,y;\theta) + \sum_{(x,y)\notin E} \ell(x,y;\theta) \right),$$

Correct-n-Contrast (CnC)
Zhang et al., 2022



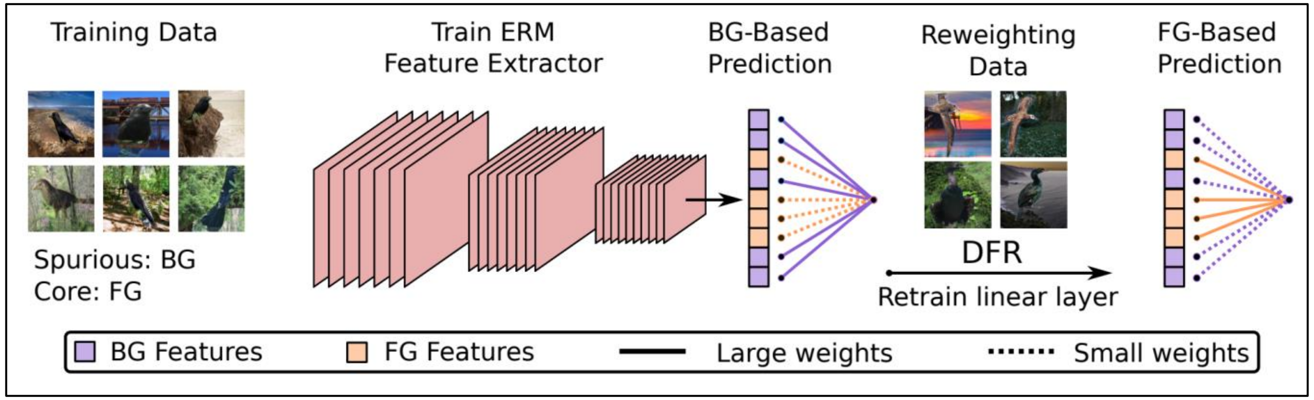
How previous work resolve this?

With knowing group label

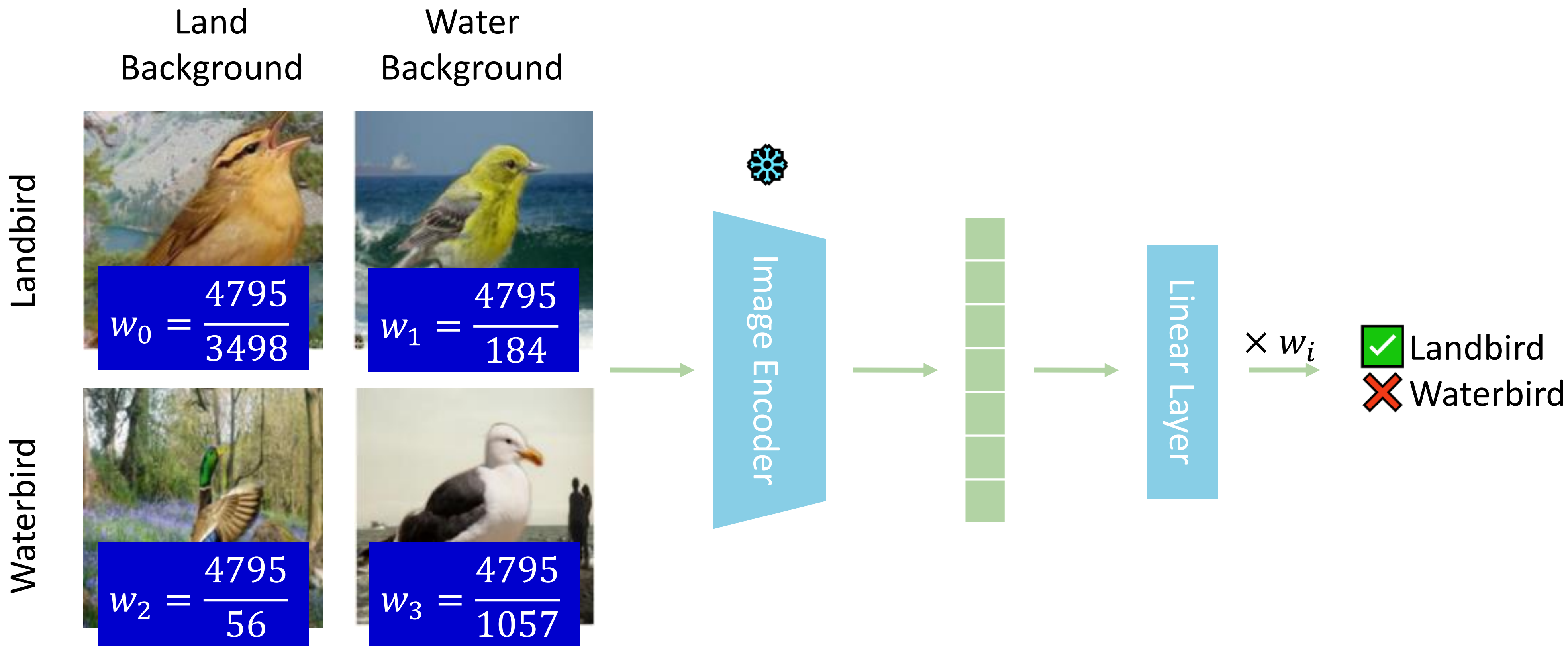
Group DRO
Sagawa et al.,
2019

$$\hat{\theta}_{\mathrm{DRO}} := \underset{\theta \in \Theta}{\arg\min} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim \hat{P}_g} [\ell(\theta; (x, y))] \right\}$$

Deep Feature Reweighting (DFR)
Izmailov et al., 2022



Deep Feature Re-weighting (DFR)



$$w_i = \frac{Total\ training\ \#}{G_i\ training\ \#}$$

It's essential to know the group to which each sample belongs, hence this method is consider supervised.

Preliminaries

ERM (Empirical Risk Minimization) aims to minimize the average loss over a training dataset by solving the optimization problem:

$$\min_{w} \boxed{\frac{1}{n}} \sum_{i=1}^{n} \ell(f_w(x_i), y_i)$$

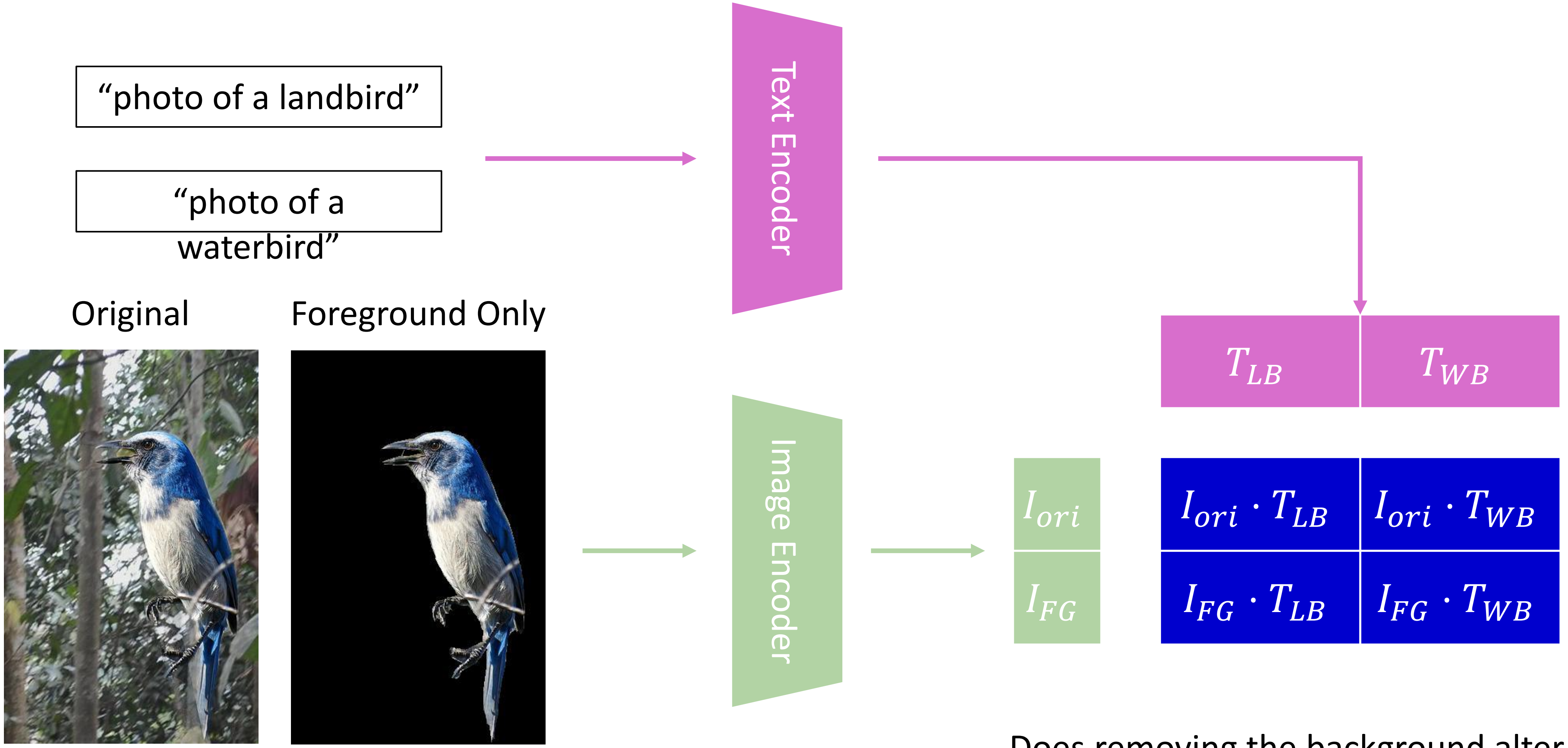
weights are invariant to the group the sample belongs to

WGA (Worst Group Accuracy) refers to the lowest accuracy among different subgroups within a dataset, which defined as:

$$\min_{g \in G} \frac{1}{|S_g|} \sum_{(x_i, y_i) \in S_g} 1(f_w(x_i) = y_i)$$

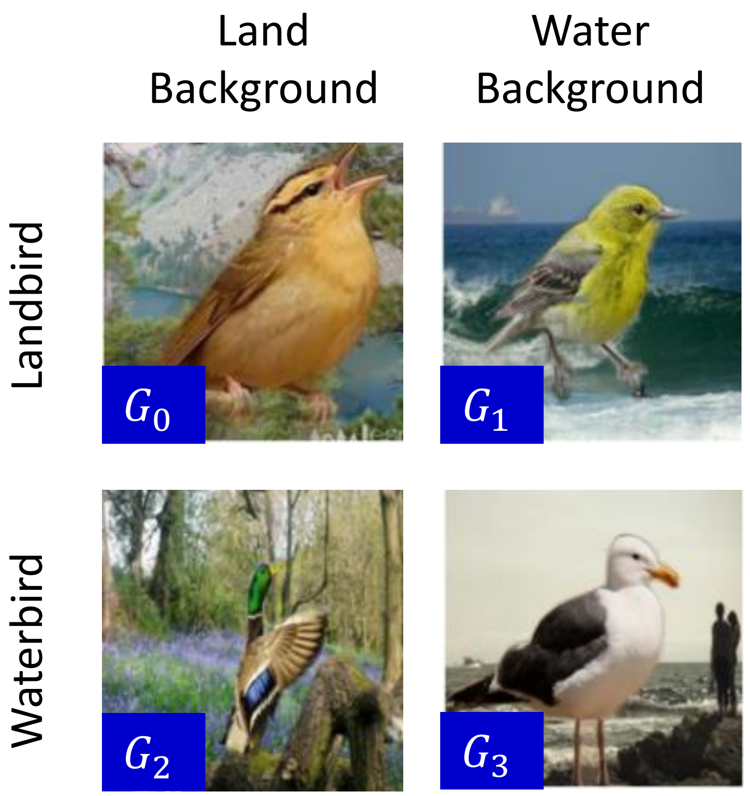
where G represents the set of groups, S_g is the set of samples in group g, and 1 is the indicator function for correct predictions.

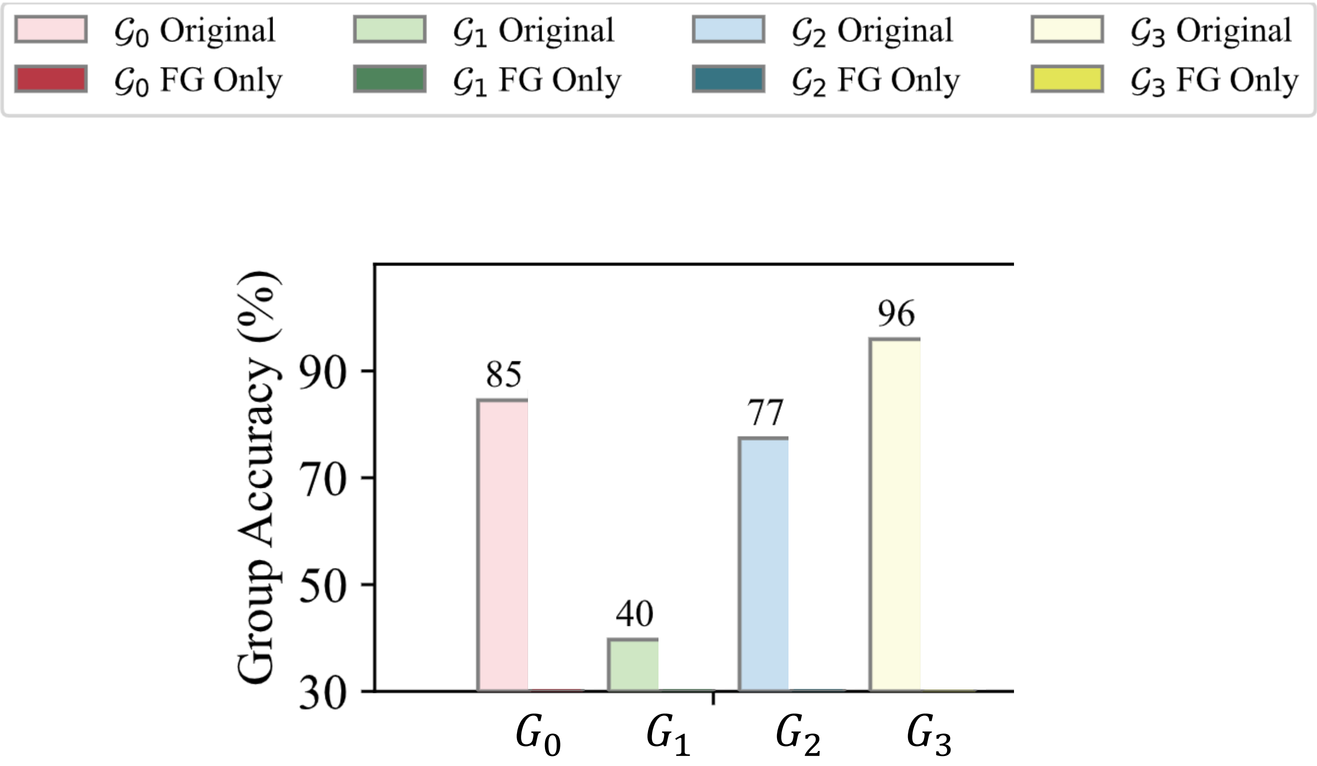
CLIP: Zero-shot Classification



Does removing the background alter the system's predicted category?

Spurious correlation confuses the VLMs





Conclusion 1

Visual representations in current VLMs are entangled with spurious features that significantly impair classification performance.

Question: Can we remove the spurious feature in visual representation via text representation?

CelebA

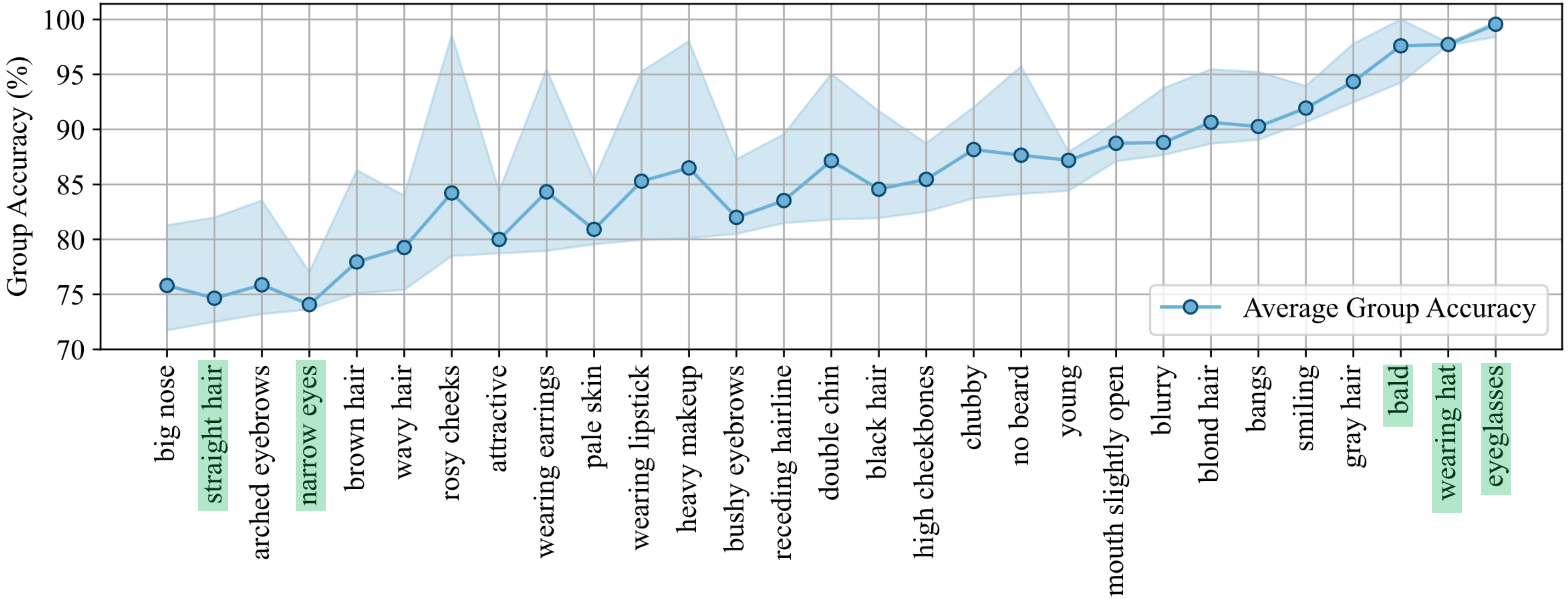
Female | Male

Non-blond Hair

Blond Hair

Attribute Name	Group 0 (Female w/o Attr)	Group 1 (Male w/o Attr)	Group 2 (Female w/ Attr)	Group 3 (Male w/ Attr)
Arched Eyebrows	54932	64560	39577	3701
Attractive	29920	49247	64589	19014
Bags Under Eyes	84963	44527	9546	23734
Bald	94500	64557	9	3704
Bangs	75612	62473	18897	5788
Big Lips	65962	57595	28547	10666
Big Nose	84954	39475	9555	28786
Black Hair	75725	48139	18784	20122
Blond Hair	71629	66874	22880	1387
Blurry	90109	64299	4400	3962
Brown Hair	71706	57872	22803	10389
Bushy Eyebrows	87757	51627	6752	16634
Chubby	93392	59989	1117	8272
Double Chin	93620	61579	889	6682
Eyeglasses	92354	59895	2155	8366
Gray Hair	93563	62311	946	5950
Heavy Makeup	32157	68058	62352	203
High Cheekbones	41836	47289	52673	20972
Mouth Slightly Open	44938	39346	49571	28915
Narrow Eyes	83877	60024	10632	8237
No Beard	117	26874	94392	41387
Oval Face	63330	53339	31179	14922
Pale Skin	89199	66566	5310	1695
Pointy Nose	60774	57150	33735	11111
Receding Hairline	89502	60228	5007	8033
Rosy Cheeks	84200	68045	10309	216
Smiling	43688	41002	50821	27259
Straight Hair	76848	51975	17661	16286
Wavy Hair	52289	58499	42220	9762
Wearing Earrings	65206	67202	29303	1059
Wearing Hat	92112	62619	2397	5642
Wearing Lipstick	18516	67817	75993	444
Wearing Necklace	75984	67022	18525	1239
Young	11167	24815	83342	43446

The upper bound of CLIP visual representation with only linear transformation



CLIP's visual representations are fine grained.

For each attribute classification experiments, we freeze the image encoder and only train a linear classification layer via DFR.

As a supervised method, DFR usually signifies the peak performance that a linear layer can attain.

Conclusion 2

The previous experiment on DFR shows that linear layer is sufficient to extract key features for various downstream tasks.

Question: Since we are studying vision language model, can we use the language embedding and a linear layer to debias the visual representation?

How pure is the CLIP language representation?

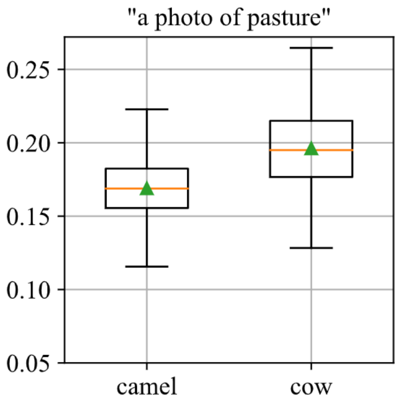
Does the representation of “a photo of pasture” only contain pasture feature?

Ideally, they should have about the same distance …

We collect camel photos that are free of pasture.

We collect cow photos that are free of pasture.

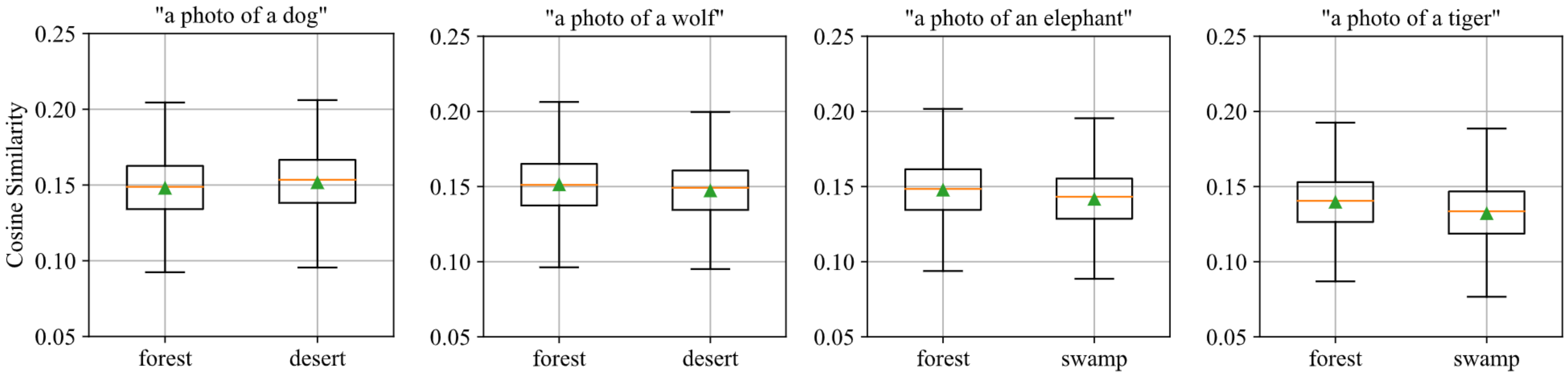
Cosine Similarity



In reality, "a photo of pasture"
are much closer to cow photos

"a photo of pasture"

For non-spurious correlated text and image pairs

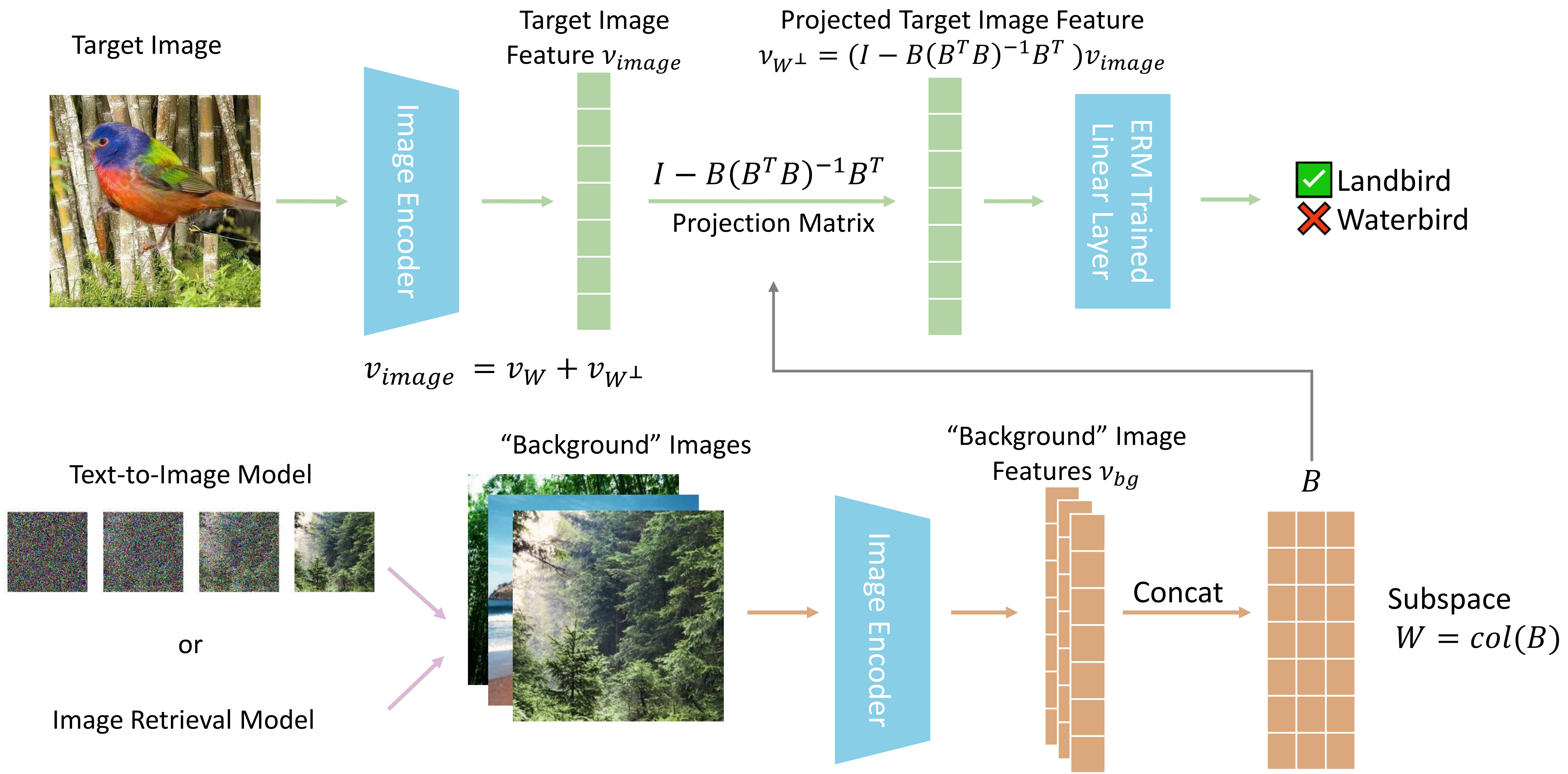


Conclusion 3

We find that CLIP's text embeddings are contaminated by diverse elements, making text embeddings impractical for debiasing the model.

Question: Since text representation is more biased than we thought, can we debias vision language model using visual representation?

Can we debias via visual representations?



foreground

from *CUB* dataset

As we know that +

background

from *Places* dataset

=

A sample in *Waterbird* dataset

Hence, for each *Waterbird* sample,
you can find the “Background” source via 2 perspectives:

Class | **Sub-class**

- Land
  - Broadleaf
  - Bamboo forest
- Water
  - Natural lake
  - Ocean

Debiasing result using different source of “background” vector

For a target image like this  , the background vector used in debiasing can be:

Corresponding class text: “a photo of land background”
Corresponding subclass text: “a photo of broadleaf”

A random image from *Places* dataset
A random image from nature
A random image from either *bamboo forest* or *broadleaf* (within class)
A random image from *broadleaf* (within subclass)

The corresponding background

How do different sources of “background” vector impact the debiasing framework?

Projection Head Source	“Background” Vector Source	“Background” Vector #	CLIP ViT WG↑	CLIP ViT Avg↑	CLIP ResNet WG↑	CLIP ResNet Avg↑
ERM	† no projection, original Waterbirds	n/a	72.27%	97.83%	61.37%	96.62%
	† random images from Places	1	70.09%	96.49%	61.84%	94.92%
		3	70.09%	96.31%	62.15%	94.71%
		10	71.81%	96.06%	63.08%	94.08%
	† random images from nature	1	77.73%	97.33%	62.93%	95.61%
		3	78.97%	97.23%	66.20%	94.11%
		10	81.46%	96.26%	61.53%	91.17%
		20	82.40%	95.03%	62.77%	90.15%
	¶ random images within class	1	81.93%	96.20%	73.52%	93.60%
		3	86.29%	95.47%	78.82%	91.74%
		10	87.07%	93.45%	73.99%	89.86%
	¶ random images within subclass	1	84.27%	95.84%	74.30%	94.09%
		3	87.54%	94.15%	79.75%	92.05%
		10	87.85%	93.35%	72.90%	89.82%
	¶ corresponding background	n/a	88.16%	96.71%	79.28%	93.83%
	† no projection, background removed	n/a	91.12%	97.69%	87.23%	96.25%
DFR	† no projection, original Waterbirds	n/a	85.67%	97.45%	80.37%	94.19%

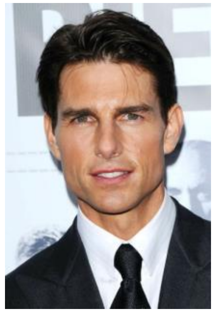
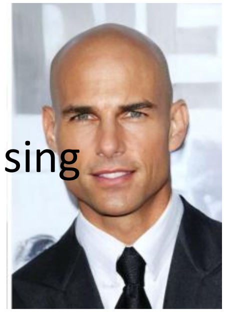
More related “background” vectors

† Unsupervised debiasing

¶ Supervised debiasing

How do different sources of “background” vector impact the debiasing framework?

When we are trying to distinguish the hair color, but the attribute is spuriously correlated to gender:

Ideally, we hope to debias  using  . But practically, we can only debi 

Projection Head Source	“Background” Vector Source	CLIP ViT WG↑	CLIP ViT Avg↑	CLIP ResNet WG↑	CLIP ResNet Avg↑
ERM	† no projection	47.22%	94.78%	38.89%	95.29%
	† irrelevant text	61.67%	93.95%	50.56%	94.99%
	¶ opposite gender text	61.67%	93.79%	45.56%	94.99%
	¶ corresponding gender text	68.33%	93.76%	52.22%	95.05%
	† an irrelevant image	58.89%	93.81%	55.56%	94.38%
	¶ an opposite gender image	66.67%	85.45%	66.11%	87.98%
	† a male and female image	79.37%	86.21%	81.11%	87.43%
	¶ a corresponding gender image	83.88%	87.60%	83.33%	87.76%
DFR	† no projection	89.38%	90.70%	89.77%	91.38%

More related “background” vectors

Conclusions

1. We show that VLMs like CLIP rely on non-causal spurious features for decision-making, yet linear probing is sufficient to extract key features for various downstream tasks.

2. We find that CLIP's text embeddings are contaminated by diverse elements, making text embeddings impractical for debiasing the model.

3. We demonstrate that using visual embeddings from CLIP to distill visual representations is highly effective.