Supplementary Materials for Joint Vision-Language Social Bias Removal for CLIP

Haoyu Zhang, Yangyang Guo*, Mohan Kankanhalli
National University of Singapore

1. Additional Related Work

Vision-Language Alignment plays an important role in the intersection of Computer Vision and Natural Language Processing due to its diverse application in downstream tasks such as Visual Question Answering (VQA) [1], image captioning [3] and image-text retrieval [15]. Early work aims to map image and text features to a common latent space for specific tasks [1, 3] using CNNs [10] and RNNs [17], respectively. The recent pervasiveness of Transformers [6] has brought a paradigm shift to the pre-training with large-scale interleaved image-text pairs followed by downstream fine-tuning. Specifically, some single-stream Vision-Language Pre-Trained Models (VL-PTMs) such as UNITER [4] and ViLT [9] utilise a single transformer module to model the intra-modal and cross-modal interactions. In contrast, models such as CLIP [16], ALIGN [8] and ALBEF [11] leverage two separate intra-modal transformers to better capture the intra-modal interaction in each modality. By pre-training with objectives such as image-text matching [16], masked language modelling [11] and masked image modelling [4], VL-PTMs are enabled to align the visual information and linguistic concepts. Most recently, some pioneering efforts have been devoted to aligning image features with Large Language Models using Q-Former [12] or linear transformation [13]. These models achieve significant performance results in both V-L understanding and generation.

2. Preliminary Experiment Details

2.1. Details for Bias Visualization

We utilize a set of biased text prompts with the template "This is a photo of a(n) $\{w_{a_i}\}$ $\{w_{c_j}\}$.", where w_{a_i} is a word describing a sensitive attribute from the attribute list $A = \{w_{a_1}, \ldots, w_{a_m}\}$, and w_{c_j} is a word describing a neutral concept from the concept list $C = \{w_{c_1}, \ldots, w_{c_n}\}$. Each attribute set A corresponds to a specific type of social bias. For gender, we use $A = \{$"male", "female"$\}$, and for age, we use $A = \{$"young", "middle-aged", "old"$\}$, and for skin tone, we use $A = \{$"light-skinned", "dark-skinned"$\}$. For all three types of biases, we use the same set of neutral concepts, $C = \{$"dancer", "lawman", ..., "astronaut"$\}$ which contains 52 common occupations that are supposedly neutral with regards to gender, age and skin tone. The 52 occupations are original labels from the FACET fairness dataset. To produce biased text prompts, for each set A, we swap the w_{a_i} while fixing the w_{c_j} to produce prompt pairs/triplets such as "This is a photo of a *male* teacher." and "This is a photo of a *female* teacher." We then obtain CLIP text embeddings of these biased text prompts and use t-SNE to visualize the distribution of biased text embeddings in a two-dimensional plot. Similarly, to visualize the biases in the image embeddings, we use a set of biased image prompts consisting of pairs/triplets of images with the same neutral concepts from the list of 52 occupations with different sensitive attributes. Different from the biased text prompts obtained by swapping the attribute keyword in the template, the biased image prompts are randomly sampled from the FACET datasets.

2.2. Details for Embedding Association Tests

We follow both SEAT [14] and IEAT [19] to use the same statistical method to measure the direction and magnitude of biases by calculating the association between specific sensitive attributes and neutral concepts. Let X and Y denote two sets representing two sensitive attributes opposite to each other (such as "male" and "female"), where $X = \{x_1, \ldots, x_m\}$, $Y = \{y_1, \ldots, y_m\}$, and each x or y is a text (or image) embedding related to the sensitive attribute that X or Y represents, such as "boy" (related to "male") or "girl" (related to "female"). Moreover, let A and B denote two sets corresponding to two neutral concepts respectively (such as "career" and "family"), where $A = \{a_1, \ldots, a_n\}$, $B = \{b_1, \ldots, b_n\}$, and each a or b, similar to x or y, is a text (or image) embedding related to the neutral concept that A or B represents, such as "executive" (related to "career") or "home" (related to "family"). SEAT and IEAT use the test statistic $s(X, Y, A, B)$ to measure the differential association of the sensitive attributes X and Y with the neutral

*Corresponding author.

Table 1. List of attribute-concept sets used in our embedding association tests. N_a^v and N_c^v denote the number of images used in the sets representing attributes (X and Y) and concepts (A and B) respectively to generate image embeddings, whereas N_a^t and N_c^t denote the number of sentences used in the sets representing attributes and concepts to generate text embeddings.

Test	Attribute X	Attribute Y	Concept A	Concept B	N_a^v	N_c^v	N_a^t	N_c^t
Weight	Thin	Fat	Pleasant	Unpleasant	55	10	8	10
Skin Tone	Light	Dark	Pleasant	Unpleasant	55	7	8	5
Race	European American	African American	Pleasant	Unpleasant	55	6	25	32
Age	Young	Old	Pleasant	Unpleasant	55	6	8	8
Gender-Science	Male	Female	Science	Liberal Arts	21	40	8	8
Gender-Career	Male	Female	Career	Family	21	40	8	8

concepts A and B, which is defined by

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B), \quad (1)$$

where $s(w, A, B)$ is defined as

$$s(w, A, B) = \text{mean}_{a \in A} sim(w, a) - \text{mean}_{b \in B} sim(w, b), \quad (2)$$

where $sim(.,.)$ denotes the cosine similarity between two embeddings. $s(w, A, B)$ quantifies the differential association of w (a text or image embedding corresponding to an attribute (X or Y) with neutral concepts A and B based on cosine similarities. The test statistic $s(X, Y, A, B)$ further aggregates and compares the differential association across all texts or images related to the attribute and its opposite attribute to measure the overall differential association of attributes with concepts. It represents possible biases, e.g. concept A is biased towards attribute X. We adopt a partial list of X, Y, A, B pairs in SEAT and IEAT to measure various common social biases. Each pair of X, Y, A, B is a bias test, and six tests are conducted on CLIP image and text encoders respectively. The details of all six tests are shown in Table. 1, and for each of these tests, SEAT or IEAT provides a collection of sentences or images corresponding to the attributes and concepts.

For each test, we first test the significance of the association represented by $s(X, Y, A, B)$ using a permutation test over all possible equal-size partitions $\{(X_i, Y_i)\}_i$ of the set $X \cup Y$ with the null hypothesis that no biased association exists. A two-sided p-value for this null hypothesis is calculated by:

$$\Pr[|s(X, Y, A, B)| < |s(X_i, Y_i, A, B)|]. \quad (3)$$

We follow the setup of SEAT and IEAT to consider any biased association with a p-value smaller than 0.1 as significant. The significant biases can be further quantified in terms of direction and magnitude by calculating the effect size, d, which is defined by

$$d = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std}_{w \in X \cup Y} s(w, A, B)}. \quad (4)$$

A positive effect size suggests that the concept of A is more biased towards the attribute of X. The absolute value of the effect size indicates the magnitude of the bias towards the bias direction. To explore text bias, we use the CLIP text encoder to generate text embeddings of X, Y, A, B for each test, and calculate the corresponding p-values and effect sizes. Similarly, for image bias, we obtain image embeddings of X, Y, A, B to calculate p-values and effect sizes.

3. Additional Preliminary Experiment Results

In addition to the bias visualization in Section 3.1 of our main paper, we provide another visualization of skin tone, gender and age biases in CLIP with a different backbone (ViT-B/16) in Fig. 1. The visualization results in the main paper were obtained from the ViT-B/32 backbone. We observe similar bias patterns across different backbones. Specifically, social biases exist in both text and image modalities. All three types of biases in the image modality are obvious, as most image embedding pairs/triplets with the same concepts but different attributes are distributed far from each other. On the other hand, in the text modality, the skin tone bias is the most evident one, with the biased embeddings forming two clusters based on different skin tones.

Moreover, to further validate the robustness of our conclusion in Section 3.1, we randomly sample additional sets of images with the same concepts (occupations) but different social attributes for embedding and t-SNE plotting and observe the bias pattern, as shown in Fig. 2. This is to reduce the effect of additional visual features (*e.g.*, image background) on the t-SNE plots. We choose not to take the average of embeddings because for each image, its corresponding embedding values may have different ranges. Taking the average of these embeddings may lead to the loss of critical information. Based on the visualization with randomly sampled images, we show that the bias patterns in image embeddings still exist regardless of different backgrounds in randomly sampled images.

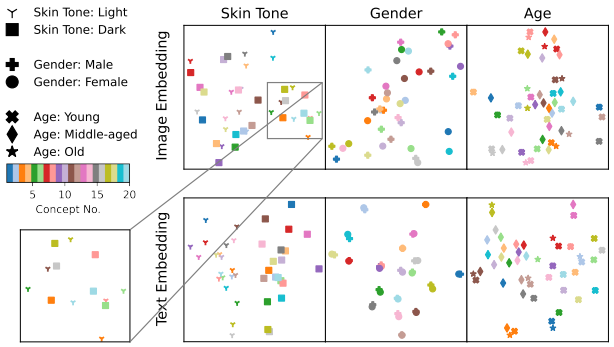

Figure 1. Additional t-SNE visualization results of different social biases in the image (the top row) and text (the bottom row) embeddings of ViT-B/16 backbone.

4. Additional Experimental Settings

4.1. Fairness Metrics

To calculate this metric, we first calculate the MaxSkew@k for a specific text prompt t, denoted as MaxSkew@$k(t)$. The text prompt t has a neutral concept (e.g. "kind"), and it is used to retrieve k most similar images from a pool of images. Each image has a sensitive attribute, $a \in \mathcal{A}$, where $\mathcal{A}$ is a set of sensitive attributes such as genders, races, and ages. The MaxSkew@$k(t)$ is calculated as:

$$\text{MaxSkew@}k(t) = \max_{a \in \mathcal{A}} \ln \left(\frac{p_{t,a}}{p_a} \right), \tag{5}$$

where $p_{t,a}$ denotes the proportion of images with attribute a in the k images retrieved using t, and p_a is equal to $\frac{1}{|\mathcal{A}|}$, representing the desired distribution which has $\frac{1}{|\mathcal{A}|}$ proportion of each attribute, assuming equality of opportunity definition of fairness [7]. The MaxSkew@$k(t)$ being 0 indicates absolute fairness regarding the concept in t because it suggests that the proportion of different groups in the retrieved images is equal to $\frac{1}{|\mathcal{A}|}$, meaning that every image with specific attributes has the equal opportunity to be chosen.

Similarly, the Normalised Discounted Cumulative KL-Divergence (NDKL), NDKL@k metric is also calculated by averaging over NDKL@$k(t)$ for different text prompts t calculated based on the retrieval of k images, as mentioned above, and it measures how the distribution of sensitive attributes in the retrieved k images differs from the ideal distribution of sensitive attributes obeying equality of opportunity, defined by

$$\text{NDKL@}k(t) = \frac{1}{Z} \sum_{i=1}^{k} \frac{1}{\log_2(i+1)} D_{KL}(P_{v_t^i} \parallel P_v), \tag{6}$$

where $D_{KL}(P \parallel Q) = \sum_j P(j) \ln \frac{P(j)}{Q(j)}$ refers to the KL-divergence of distribution P with respect to distribution Q, Z is a normalisation factor, $P_{v_t^i}$ is the discrete distribution

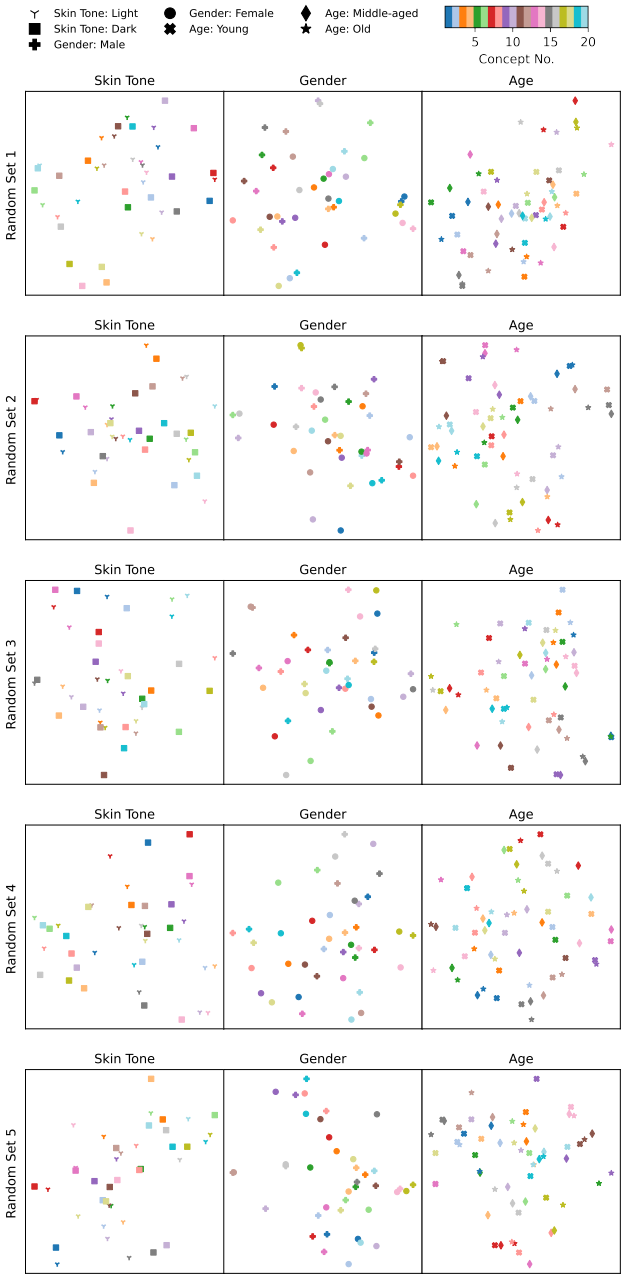

Figure 2. Additional t-SNE visualization of social biases in five sets of randomly sampled image embeddings of CLIP ViT-B/32.

of the sensitive attributes in the top i images retrieved by the text prompt t, and P_v is the desired distribution which has $\frac{1}{|\mathcal{A}|}$ proportion of each attribute. Different from the MaxSkew@k which indicates the most significant unfairness, the NDKL@k calculates a weighted average of the unfairness in the distribution of sensitive attributes in the k retrieved images, providing insights about the overall unfairness. A smaller NDKL@k suggests a higher level of fairness.

Table 2. Race debiasing results of three approaches trained on FairFace. ABLE is calculated based on in-domain fairness. We marked the best numbers pertaining to fairness results with **bold**, and those pertaining to the V-L alignment results on IN1K and Flickr with underline. † implies equal or better V-L performance over the original CLIP.

Methods	Backbone	In-Domain		Out-of-Domain		IN1K		Flickr		ABLE (%)↑
		FairFace		UTKFace		Acc. (%)↑		R@5 (%)↑		
		MS↓	NDKL↓	MS↓	NDKL↓	Top-1	Top-5	TR	IR	
Original CLIP	ViT-B/16	0.528	0.182	0.575	0.137	68.31	91.83	96.4	85.5	63.31
CLIP-clip		0.500	0.179	0.578	0.170	65.12	89.82	93.9	83.1	62.82
Biased-prompts		0.485	0.195	0.587	0.213	66.99	90.71	93.6	85.6†	64.16
Ours		**0.353**	**0.125**	**0.378**	**0.069**	68.07	91.64	96.5†	83.8	**69.14**
Original CLIP	ViT-B/32	0.579	0.181	0.680	0.215	63.39	88.83	94.7	83.5	59.49
CLIP-clip		0.673	0.219	0.725	0.261	62.63	88.51	94.1	82.3	56.22
Biased-prompts		0.582	0.256	**0.306**	0.149	61.79	87.45	91.5	83.2	58.70
Ours		**0.342**	**0.111**	0.440	**0.110**	62.63	88.43	94.3	82.7	**66.58**
Original CLIP	ViT-L/14	0.571	0.236	0.626	0.184	75.55	94.57	97.2	87.2	64.64
CLIP-clip		0.535	0.210	0.593	0.157	73.60	93.40	94.9	82.3	65.24
Biased-prompts		0.617	0.256	**0.515**	0.174	74.42	94.00	94.8	87.7†	62.57
Ours		**0.454**	**0.174**	0.523	**0.137**	75.13	94.34	97.2†	87.0	**68.83**
Original CLIP	ViT-H/14	0.538	0.206	0.548	0.169	77.95	95.19	99.5	94.1	66.77
CLIP-clip		0.514	**0.192**	0.588	0.188	77.75	95.17	99.0	93.7	67.61
Biased-prompts		0.529	0.195	**0.434**	0.180	77.31	95.06	99.0	93.3	66.87
Ours		**0.498**	0.200	0.513	**0.141**	77.63	95.16	99.5†	93.8	**68.18**

During the evaluation of fairness, we use different t to retrieve images and calculate the corresponding MaxSkew@$k(t)$ and NDKL@$k(t)$, and take the average of the fairness metrics over all t to obtain the final mean MaxSkew@k and NDKL@k fairness scores. We follow existing work [2, 18] to set k as 1,000.

4.2. Datasets

We split the FairFace and UTKFace datasets into train, validation and test sets with a ratio of 8:1:1 for training and evaluation. For FACET which is evaluation-only, we use 10% of the data for evaluation to maintain a similar scale to the test sets of FairFace and UTKFace. We discard the images with the race label "Others" in the UTKFace train set when training the model to debias race. Moreover, to align the age groups with those in the FACET dataset, we merge the age labels in FairFace and UTKFace datasets into three groups: "young", "middle-aged" and "old".

4.3. Preparation of Biased Triplets During Training

At the training stage, we utilize a batch of biased triplets consisting of two oppositely biased text prompts and a biased image, denoted as (t_i, t_i', v_i). The biased image v_i is directly drawn from the fairness datasets such as FairFace and UTKFace. For each v_i sampled, we first produce its corresponding biased text prompts t_i by filling in keywords in a fixed template "This is a photo of a $\{age\}$ $\{race\}$ $\{gender\}$." Specifically, the keywords *age*, *race* and *gender* are the annotated labels of v_i. We then create the counterfactual text embedding t_i' by altering one of the keywords (for debiasing one type of bias only). For example, if t_i is "This is a photo of a middle-aged white male.", its counterfactual prompt t_i' is "This is a photo of a middle-aged white *female*." for debiasing the gender bias. When debiasing biases with more than two directions, such as age and race, we randomly choose a label from the remaining labels to produce the counterfactual prompt. For universal debiasing, where we aim to remove multiple biases simultaneously, we make counterfactual text prompts by switching multiple keywords corresponding to each of the biases.

4.4. Implementation of Compared Baselines

CLIP-clip removes the most biased dimensions with the largest amount of mutual information, only keeping m dimensions in image and text embeddings. We follow previous work [2] to choose m from [256, 400, 490] for ViT-B/16 and ViT-B/32, from [384, 600, 735] for ViT-L/14, and from [512, 800, 980] for ViT-H/14, respectively.

Biased-prompts calculates a calibrated projection matrix based on biased text prompts to debias text embeddings. The complete set of biased prompts is not released in the original paper of Biased-prompts, we therefore re-implement this work using suggested prompts [5]. We set the weighting hyperparameter to be 1000, following the default value used in the original paper.

4.5. Implementation of Our Method

We freeze the CLIP model and only train the bias alignment module for 100 epochs with a batch size of 512. Adam optimizer is used with a learning rate chosen from

Table 3. Race debiasing results of three approaches trained on UTKFace.

Methods	Backbone	In-Domain UTKFace MS↓	In-Domain UTKFace NDKL↓	Out-of-Domain FairFace MS↓	Out-of-Domain FairFace NDKL↓	IN1K Acc. (%)↑ Top-1	IN1K Acc. (%)↑ Top-5	Flickr R@5 (%)↑ TR	Flickr R@5 (%)↑ IR	ABLE (%)↑
Original CLIP	ViT-B/16	0.575	0.137	0.528	0.182	68.31	91.83	96.4	85.5	61.71
CLIP-clip		0.578	0.170	0.693	0.237	65.73	89.85	91.5	79.2	60.54
Biased-prompts		0.587	0.213	0.485	0.195	66.99	90.71	93.6	85.6†	60.75
Ours		**0.523**	**0.103**	**0.462**	**0.149**	67.63	91.52	96.1	84.4	**63.19**
Original CLIP	ViT-B/32	0.680	0.215	0.579	0.181	63.39	88.83	94.7	83.5	56.30
CLIP-clip		0.795	0.399	0.855	0.330	62.69	88.28	93.3	81.5	52.50
Biased-prompts		**0.306**	**0.149**	0.582	0.256	61.79	87.45	91.5	83.2	**67.20**
Ours		0.613	0.155	**0.515**	**0.153**	62.12	88.35	94.7†	82.3	57.87
Original CLIP	ViT-L/14	0.626	0.184	0.571	0.236	75.55	94.57	97.2	87.2	62.63
CLIP-clip		0.635	0.179	**0.519**	**0.199**	75.09	94.31	96.4	86.3	62.14
Biased-prompts		**0.515**	**0.174**	0.617	0.256	74.42	94.00	94.8	87.7†	**66.29**
Ours		0.620	0.184	0.563	0.214	75.06	94.34	96.7	86.6	62.67
Original CLIP	ViT-H/14	0.548	0.169	0.538	0.206	77.95	95.19	99.5	94.1	66.39
CLIP-clip		0.525	0.233	0.657	0.238	77.76	95.22†	99.5†	93.7	66.69
Biased-prompts		0.434	0.180	**0.529**	**0.195**	77.31	95.06	99.0	93.3	70.50
Ours		**0.415**	**0.149**	0.761	0.397	77.71	95.18	99.5†	93.8	**71.40**

$[2 \times 10^{-6}, 5 \times 10^{-6}]$. We also apply early stopping based on validation results. The bias alignment module is implemented with two multilayer perceptrons consisting of one hidden layer and a ReLU activation function, respectively. The hidden layer size has the range $[0.5d, d, 2d]$, where d denotes the embedding dimension of the backbone. During training, the α combining the counterfactual debiasing loss and the bias alignment loss is selected from $[0.1, 0.3, 0.5, 0.7, 0.9]$.

5. Additional Experimental Results

5.1. Race Debiasing Results

We present additional results for race debiasing in Table 2 and Table 3. Similar to the results of gender and age debiasing, our method also achieves a better trade-off between debiasing and V-L alignment compared to baselines for race debiasing. This highlights the generalizability of our method across bias types.

5.2. COCO Retrieval Results

We further evaluated our debiased models' V-L performance on the COCO retrieval dataset (shown in Table 4). We found that our method only causes a minor performance drop, highlighting our method's capability to maintain V-L task performance after debiasing.

Table 4. COCO results (Recall@5) for original and debiased CLIP on ViT-B/16. Universal refers to mitigating all three types of social biases.

Original TR↑	Original IR↑	Gender TR↑	Gender IR↑	Age TR↑	Age IR↑	Race TR↑	Race IR↑	Universal TR↑	Universal IR↑
76.6	58.3	75.6	56.8	76.5	57.3	75.7	56.7	76.2	56.5

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 1

[2] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *ACL*, pages 806–822. Association for Computational Linguistics, 2022. 4

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *CoRR*, 2015. 1

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, page 104–120. Springer-Verlag, 2020. 1

[5] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *CoRR*, 2023. 4

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186. Association for Computational Linguistics, 2019. 1

[7] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, page 3323–3331, 2016. 3

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation

learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1

[9] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594. PMLR, 2021. 1

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pages 2278–2324, 1998. 1

[11] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pages 9694–9705, 2021. 1

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. JMLR.org, 2023. 1

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[14] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *NAACL*, pages 622–628. Association for Computational Linguistics, 2019. 1

[15] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 1

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[17] David E. Rumelhart and James L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, 1987. 1

[18] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *CVPR*, pages 6820–6829, 2023. 4

[19] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *FAccT*, page 701–713. Association for Computing Machinery, 2021. 1