FairCLIP: Harnessing Fairness in Vision-Language Learning

Supplementary Material

1. Experimental Setup

1.1. Pre-Training

We use the widely-adopted VL methods – CLIP [7] and BLIP2 [4] – for our analysis. For the natural pre-trained variants, we use the official checkpoints provided by CLIP and BLIP2. For the medical pre-trained variants, we pre-train both methods on our Harvard-FairVLMed dataset after initializing from the official checkpoints. We fine-tune CLIP for 10 epochs using the Adam [3] optimizer, with a learning rate (lr) of 1e-5. The hyperparameters β_1 and β_2 are configured at 0.1, along with a weight decay of 6e-5 and batch size of 32. These specific hyper-parameters were selected after extensive tuning to achieve optimal performance with CLIP. FairCLIP uses the same aforementioned hyper-parameters, leveraging a batch size $|\mathcal{B}_a|$ of 32 to draw samples from each group. The Sinkhorn loss is integrated with CLIP's original loss function, using a weight of 1e-7. For BLIP2, we primarily focus on the vision-language representation learning stage (i.e., Stage 1) and use the official ViT-L/14 model from CLIP with FP32 precision as the frozen vision encoder. Following the official implementation, we use AdamW [5] as the optimizer, with β_1, β_2, and weight decay set to 0.9, 0.98, and 0.05, respectively. We also use a cosine lr decay with a max, min, and warmup lr of 1e-4, 1e-5, and 1e-6, respectively. Moreover, we apply random resized cropping (224×224) and random horizontal flipping to the fundus images, whereas we utilize the BLIP caption augmentations to pre-process the clinical notes, with maximum words set to 50. Finally, all models are pre-trained on the paired fundus images and clinical notes from Harvard-FairVLMed using a batch size of 32 for 50 epochs on a single V100 GPU. These pre-trained CLIP and BLIP2 models are then used for the subsequent linear probing and zero-shot evaluation.

1.2. Metrics

To comprehensively understand the balance between model performance and fairness, we use multiple metrics for evaluation, including Demographic Parity Difference (DPD) [1, 2], Difference in Equalized Odds (DEOdds) [1], Area Under the Receiver Operating Characteristic Curve (AUC), Equity-Scaled AUC [6], and Group-wise AUC. Particularly, DPD and DEOdds are widely used fairness metrics that focus on the fairness of the model's predictions, ensuring that no group is systematically advantaged or disadvantaged. In contrast, AUC is a mainstream performance metric used in medical scenarios. Group-wise AUC is an intuitive and straightforward metric to reveal the discrepancy between groups. In safety-critical medical applications, neither fairness nor performance alone is sufficient as the sole measurement criterion. Hence, ES-AUC is an effective metric that efficiently balances both performance and fairness. It offers a holistic evaluation, facilitating the analysis of the trade-off between these two essential criteria. ES-AUC is defined as:

$$\text{ES-AUC} = \frac{\text{AUC}}{1 + \sum_a^{\mathcal{A}} |\text{AUC} - \text{AUC}_a|}$$

where $\mathcal{A}$ can be $\{\text{Asian}, \text{Black}, \text{White}\}$, $\{\text{Female}, \text{Male}\}$, $\{\text{Non-Hispanic}, \text{Hispanic}\}$, or $\{\text{English}, \text{Spanish}, \text{Others}\}$. A higher ES-AUC score indicates that the model achieves not only greater performance but also simultaneously improves model equity.

2. Results

2.1. CLIP vs. FairCLIP

In addition to the zero-shot comparison of CLIP and FairCLIP presented in Section 5.3, Table S1 demonstrates their end-to-end fine-tuning results, further validating the effectiveness of FairCLIP. Once again, the performance is evaluated using the metrics DPD, DEOdds, AUC, ES-AUC, and group-wise AUC.

In terms of the racial subgroups, both CLIP and FairCLIP show varied performance. For instance, in the ViT-B/16 setting, CLIP achieves a lower DPD (5.85) compared to FairCLIP (11.38), indicating a better balance in outcomes across races. However, FairCLIP (ViT-L/14) outperforms CLIP in AUC and group-wise AUC for the Asian and Black groups, suggesting a more equitable performance across these racial categories. For the gender subgroups, FairCLIP consistently outperforms CLIP in both DPD and DEOdds, indicating a more balanced performance between the male and female subgroups. The AUC scores are also higher for FairCLIP, with the ViT-B/16 achieving an AUC of 81.88 and a higher group-wise AUC for both genders. In terms of ethnicity, FairCLIP generally achieves higher AUC scores than CLIP. Notably, FairCLIP (ViT-L/14) shows a significant improvement in ES-AUC (79.08) and group-wise AUC for Hispanic groups. Lastly, for the language subgroups, FairCLIP shows a slightly better performance in terms of AUC and group-wise AUC for English and Spanish speakers. However, both models struggle with the "Others" language group, with FairCLIP (ViT-L/14) showing a notable improvement in ES-AUC (74.44).

Overall, similar to the results presented in Section 5.3, we observe that our proposed method FairCLIP consistently

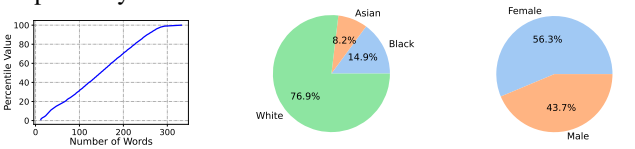
Table S1. End-to-end fine-tuning results of CLIP vs. FairCLIP, reporting the mean and standard deviation across three random seeds.

Attribute	Model	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑	Group-wise AUC ↑		
						Asian	**Black**	**White**
Race	CLIP (ViT-B/16)	**5.85** ± 3.39	10.68 ± 3.75	81.19 ± 0.44	75.07 ± 1.36	84.82 ± 1.77	77.15 ± 1.17	81.73 ± 0.61
	FairCLIP (ViT-B/16)	11.38 ± 4.23	10.53 ± 3.10	81.70 ± 0.34	**76.85** ± 0.64	83.30 ± 1.09	77.35 ± 0.87	82.07 ± 0.28
	CLIP (ViT-L/14)	7.39 ± 1.98	10.59 ± 1.64	80.21 ± 1.43	75.37 ± 1.03	82.04 ± 2.26	76.29 ± 1.73	80.89 ± 1.42
	FairCLIP (ViT-L/14)	8.67 ± 4.32	**8.84** ± 5.24	**81.80** ± 0.19	76.70 ± 1.74	**84.87** ± 1.05	**78.52** ± 1.37	**82.17** ± 0.41
						Female	**Male**	
Gender	CLIP (ViT-B/16)	1.89 ± 1.65	6.78 ± 2.88	81.19 ± 0.44	77.47 ± 0.51	78.96 ± 0.30	83.78 ± 0.96	
	FairCLIP (ViT-B/16)	**1.72** ± 0.36	**5.59** ± 0.12	**81.88** ± 0.30	**78.46** ± 0.31	**79.84** ± 0.25	**84.20** ± 0.33	
	CLIP (ViT-L/14)	1.85 ± 0.95	6.73 ± 1.39	80.21 ± 1.43	76.39 ± 1.60	77.92 ± 1.56	82.93 ± 1.25	
	FairCLIP (ViT-L/14)	2.26 ± 1.28	7.58 ± 2.59	81.07 ± 0.78	77.36 ± 0.27	78.86 ± 0.44	83.66 ± 1.27	
						Non-Hispanic	**Hispanic**	
Ethnicity	CLIP (ViT-B/16)	**9.57** ± 2.34	**11.35** ± 5.03	81.19 ± 0.44	76.09 ± 1.44	81.43 ± 0.53	74.68 ± 2.07	
	FairCLIP (ViT-B/16)	12.80 ± 2.01	14.49 ± 3.15	**81.47** ± 0.15	78.22 ± 1.44	**81.61** ± 0.21	77.42 ± 1.86	
	CLIP (ViT-L/14)	12.15 ± 3.21	15.08 ± 3.18	80.21 ± 1.43	75.79 ± 1.45	80.45 ± 1.45	74.61 ± 1.59	
	FairCLIP (ViT-L/14)	10.47 ± 0.96	13.62 ± 2.15	**81.47** ± 0.58	**79.08** ± 1.16	81.57 ± 0.64	**78.52** ± 1.54	
						English	**Spanish**	**Others**
Language	CLIP (ViT-B/16)	13.12 ± 3.49	22.10 ± 3.77	81.19 ± 0.44	70.12 ± 1.71	81.64 ± 0.43	**80.59** ± 5.00	70.62 ± 3.25
	FairCLIP (ViT-B/16)	15.29 ± 1.83	**21.14** ± 4.88	**81.71** ± 0.28	71.74 ± 1.26	**82.21** ± 0.30	79.36 ± 1.89	70.63 ± 0.29
	CLIP (ViT-L/14)	**10.95** ± 5.92	26.58 ± 9.41	80.21 ± 1.43	70.77 ± 1.64	80.61 ± 1.42	78.12 ± 3.96	71.00 ± 1.48
	FairCLIP (ViT-L/14)	15.81 ± 4.49	25.18 ± 11.78	81.22 ± 0.42	**74.44** ± 1.22	81.41 ± 0.36	**80.59** ± 4.38	**75.65** ± 0.88

outperforms CLIP.

2.2. Dataset Analysis

To supplement the details for our Harvard-FairVLMed dataset presented in the main paper, here we provide additional analyses representing the distribution of words in the clinical notes (Figure S1a), and the prevalence of subjects across the race and gender attributes (Figure S1b and S1c), respectively.



(a) (b) (c)

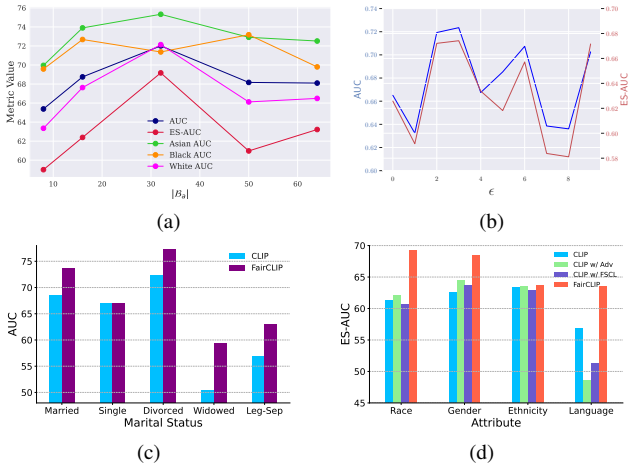
Figure S1. (a) Distribution of words in the clinical notes, (b) Prevalence of subjects across race, (c) Prevalence of subjects across gender.

2.3. Ablation Studies

In addition to the ablation studies presented in Section 5.4, we also study the effect of $|\mathcal{B}_a|$ in FairCLIP on model performance. From the results in Figure S2a, we observe that $|\mathcal{B}_a| = 32$ achieves desired performance.

Moreover, we also present detailed results for the clinical note summarization, vision vs. multimodal features, and natural vs. medical vision encoder ablation studies in Tables S2, S3, and S4, respectively. For a comprehensive discussion of these ablation studies, please refer to Section 5.4 in the main paper.

Furthermore, we present an ablation study on the effects of ϵ on model performance in Figure S2b. Also, we include additional fairness results based on marital status in Figure S2c. Lastly, we provide a comparison of FairCLIP against other fairness algorithms in Figure S2d.



(a) (b) (c) (d)

Figure S2. (a) Ablation study of using various $|\mathcal{B}_a|$ in FairCLIP, (b) Ablation study on the effects of ϵ on model performance, (c) Fairness results based on marital status, (d) Comparison of FairCLIP against other fairness algorithms.

Table S2. Impact of various LLM summarizations on the performance-fairness trade-off of BLIP2.

Attribute	Clinical Notes	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑	Group-wise AUC ↑		
						Asian	**Black**	**White**
Race	Original	8.36	11.28	80.13	73.76	82.08	74.35	81.03
	PMC-LLAMA	4.38	12.71	80.11	72.63	83.77	74.20	80.84
	MED42	6.26	14.49	80.36	73.59	82.93	74.60	81.23
	GPT-4	5.30	6.24	79.34	74.39	82.45	76.15	79.70
						Female	**Male**	
Gender	Original	2.34	6.56	80.13	75.22	77.13	83.66	
	PMC-LLAMA	1.72	7.92	80.11	74.20	76.43	84.39	
	MED42	0.72	4.20	80.36	76.19	77.84	83.31	
	GPT-4	3.18	9.51	79.34	73.49	75.67	83.64	
						Non-Hispanic	**Hispanic**	
Ethnicity	Original	16.26	20.59	80.13	77.18	80.28	76.46	
	PMC-LLAMA	14.96	16.17	80.11	76.24	80.24	75.17	
	MED42	16.32	18.25	80.36	76.98	80.49	76.11	
	GPT-4	16.55	15.83	79.34	76.40	79.44	75.59	
						English	**Spanish**	**Others**
Language	Original	11.47	39.13	80.13	69.88	80.64	83.52	69.36
	PMC-LLAMA	9.15	34.78	80.11	71.71	80.60	78.13	70.88
	MED42	21.78	22.28	80.36	69.76	80.70	72.16	73.71
	GPT-4	14.65	39.13	79.34	69.52	79.89	77.27	67.84

Table S3. Impact of vision-only and (vision + language) features on the performance-fairness trade-off of linear probing via BLIP2.

Attribute	V	L	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑	Group-wise AUC ↑		
							Asian	**Black**	**White**
Race	✓	✗	6.26	14.49	80.36	73.59	82.93	74.60	81.23
	✓	✓	7.78	5.35	82.16	79.20	80.84	80.27	82.68
							Female	**Male**	
Gender	✓	✗	0.72	4.20	80.36	76.19	77.84	83.31	
	✓	✓	1.12	3.87	82.16	79.56	80.66	83.93	
							Non-Hispanic	**Hispanic**	
Ethnicity	✓	✗	16.32	18.25	80.36	76.98	80.49	76.11	
	✓	✓	16.19	15.69	82.16	78.98	82.29	78.26	
							English	**Spanish**	**Others**
Language	✓	✗	21.78	22.28	80.36	69.76	80.70	72.16	73.71
	✓	✓	15.08	21.73	82.16	66.27	82.76	68.18	72.77

Table S4. Impact of using pre-trained vision encoders from natural (CLIP) and medical (PMC-CLIP) domains on the performance-fairness trade-off of BLIP2.

Attribute	Encoder Type	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑	Group-wise AUC ↑		
						Asian	**Black**	**White**
Race	CLIP	6.26	14.49	80.36	73.59	82.93	74.60	81.23
	PMC-CLIP	8.12	7.15	81.23	76.04	83.27	77.09	81.87
						Female	**Male**	
Gender	CLIP	0.72	4.20	80.36	76.19	77.84	83.31	
	PMC-CLIP	3.66	11.07	81.23	76.42	78.24	84.54	
						Non-Hispanic	**Hispanic**	
Ethnicity	CLIP	16.32	18.25	80.36	76.98	80.49	76.11	
	PMC-CLIP	15.20	15.33	81.23	77.28	81.43	76.32	
						English	**Spanish**	**Others**
Language	CLIP	21.78	22.28	80.36	69.76	80.70	72.16	73.71
	PMC-CLIP	10.31	22.28	81.23	70.53	81.73	76.70	71.08

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018. 1

[2] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. 1

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[6] Yan Luo, Yu Tian, Min Shi, Louis R. Pasquale, Lucy Q. Shen, Nazlee Zebardast, Tobias Elze, and Mengyu Wang. Harvard glaucoma fairness: A retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1