FairLLaVA: Fairness-Aware Parameter-Efficient Fine-Tuning for Large Vision-Language Assistants (Supplementary)

Mahesh Bhosale [1] Abdul Wasi [1] Shantam Srivastava [1] Shifa Latif [2] Tianyu Luan [3]
Mingchen Gao [1] David Doermann [1]
Xuan Gong [4]

[1]University at Buffalo [2]University of Kashmir [3]Accenture [4]Harvard Medical School

1. Cross-Sectional Analysis

In the main paper, evaluation for a given demographic attribute is performed by aggregating over all subgroups of the remaining attributes. For example, when reporting results for gender, we include samples from all age groups and all race categories. This aggregate evaluation can make the effectiveness of the attribute-specific DAC appear less direct. For instance, in Table 2 in the main paper, FairLLaVA trained to debias Age or Gender sometimes obtain higher ES scores on Race than the FairLLaVA explicitly trained for Race; similarly, the Race-debiased FairLLaVA can outperform the Gender-debiased FairLLaVA on Gender ES. Results in Table 2 in the main paper are not contradictory. Two factors explain why a model debiased for one attribute can sometimes obtain a higher ES score on another attribute than the model explicitly trained for that attribute: i) demographic attributes in MIMIC-CXR are correlated (e.g., age distributions differ across sex and race), so mitigating one source of bias can indirectly reduce disparities in another. ii) ES reflects both fairness and utility. It increases not only when subgroup disparities decrease, but also when the underlying task metric remains high. Consequently, a variant may achieve a higher ES on a non-target attribute if it preserves overall performance better, even when another variant reduces that attribute-specific gap more strongly. Thus, ES should be interpreted as a balance between gap reduction and task performance, not as a direct measure of isolated debiasing.

To isolate the effect of debiasing a specific attribute, we additionally report controlled fairness gaps by fixing the other demographic attributes. Concretely, when evaluating age-related disparities, we compare age groups only within the same race-gender subgroup slice, rather than mixing samples across different races or genders. This reduces confounding from correlated demographics and provides a cleaner view of whether the method is truly reducing bias for the intended attribute. We call this cross-sectional analysis. As shown in Tab. 1 upper rows, under this analysis, the gap decreases the most when the corresponding attribute is explicitly debiased, confirming the intended effect. We also compare the FairLLaVA-All variant with the strongest baseline of LLaVA-Rad.

Method	Race ↓		Age ↓		Gender ↓	
	RG-F1	GREEN	RG-F1	GREEN	RG-F1	GREEN
FairLLaVA-Race	2.98	4.59	9.06	16.88	1.41	1.84
FairLLaVA-Age	4.12	6.37	8.17	14.88	1.61	3.03
FairLLaVA-Gender	2.89	6.01	9.29	16.04	1.18	1.81
LLaVA-Rad	3.59	6.30	10.38	17.28	1.17	3.54
FairLLaVA-All	**3.26**	**4.78**	**7.33**	**12.89**	**1.08**	**2.74**

Table 1. **Cross-sectional fairness analysis.** RG-F1 is an abbreviation of RadGraph-F1. To isolate the effect of debiasing each demographic attribute, subgroup gaps are computed while holding the remaining attributes fixed (e.g., comparing age groups within the same race–gender slice). The targeted-attribute variants cause most reduction in the gap as compared to other-attribute variants, as intended. FairLLaVA-All also holds strong under this analysis as compared to the strong baseline of LLaVA-Rad.

2. Individual Performance and Counterfactual Fairness Gaps

To assess spurious demographic reliance beyond aggregate subgroup metrics, we perform a counterfactual fairness analysis at the individual level. The goal is to measure the fairness gap when the protected attribute varies while the underlying clinical evidence is kept as similar as possible. Concretely, for each sample, we retrieve its nearest match from a different demographic subgroup in the latent feature space, subject to two constraints: (i) the pair

Method	Race ↓		Age ↓		Gender ↓	
	RG-F1	GREEN	RG-F1	GREEN	RG-F1	GREEN
LLaVA-Rad	13.71	24.45	27.44	21.44	17.93	21.60
FairLLaVA	6.65	23.08	20.34	17.65	16.92	18.90

Table 2. **Counterfactual fairness gaps.** FairLLaVA also reduces the individual counterfactual fairness gaps on MIMIC-CXR dataset.

must share the same CheXpert label set, so that the clinical findings are matched, and (ii) the latent similarity must exceed a threshold of 0.7, ensuring that the paired samples are visually and semantically close. For example, a female study with pleural effusion is matched to the nearest male study with the same CheXpert label i.e. pleural effusion. We then compute the fairness gap across such matched pairs. Since the paired samples are aligned in clinical content, a large gap indicates that the model is relying on demographic cues beyond the disease evidence, whereas a smaller gap suggests reduced spurious dependence on the protected attribute. We used a BiomedCLIP-CXR [14] as feature extractor. As seen in Tab. 2, FairLLaVA consistently lowers these counterfactual gaps relative to LLaVA-Rad across Race, Age, and Gender, indicating improved robustness to demographic variation under matched clinical evidence. This analysis complements population-level ES metrics by providing an individual-level fairness signal, and can also serve as an initial signal to decide which demographic attributes to de-bias and choose λ weights.

3. Hyper-parameters

Sensitivity Total loss is given by eq.10 which has many hyper-parameters. Here we study their sensitivity to the fairness gap. We train the DAC using a class-frequency-weighted cross-entropy loss to mitigate class imbalance, and use λ to weight the DAC-based MI minimization term. Fig. 1 (a) shows that increasing the three DAC λ values results in only marginal changes in the equity-scaled metric (ES-M) on MIMIC-CXR. We further examine the λ for the LM loss in Fig. 1 (b). In both cases, the overall performance remains largely stable, indicating that the added components do not introduce significant sensitivity.

Hyper-parameter Search In Tab. 3, we vary $(\lambda_{\text{race}}, \lambda_{\text{age}}, \lambda_{\text{gender}})$ by increasing one attribute weight at a time to study its effect on MIMIC-CXR. While all three attributes are debiased jointly, the largest ES gain for an attribute is achieved when its corresponding λ is assigned the highest value. In the main manuscript, we use $(0.2, 0.6, 0.1)$ to approximately reflect the relative LLaVA-Rad baseline gap ratios $(\Delta_{\text{race}} : \Delta_{\text{age}} : \Delta_{\text{gender}})$.

$(\lambda_r, \lambda_a, \lambda_g)$	Race ↑		Age Group ↑		Gender ↑	
	RG-F1	GREEN	RG-F1	GREEN	RG-F1	GREEN
$(0.2, 0.6, 0.2)$	2.77	3.22	3.85	3.34	18.98	12.92
$(0.6, 0.2, 0.2)$	5.62	3.60	3.42	1.83	22.45	12.23
$(0.2, 0.2, 0.6)$	5.30	3.18	3.24	1.85	24.74	13.77

Table 3. **Effect of varying attribute-specific MI weights.** $(\lambda_r, \lambda_a, \lambda_g)$ on equity-scaled performance.

Dataset	Sex↑	Age↑		Race↑
	AUC	MAE	AgeGrp Acc	AUC
MIMIC-CXR	0.9549	6.9715	0.7265	0.8901
PadChest	0.9841	6.2339	0.8023	–

Table 4. **Missing demographic attribute prediction** on out of domain radiology datasets using TorchXRayVision [4].

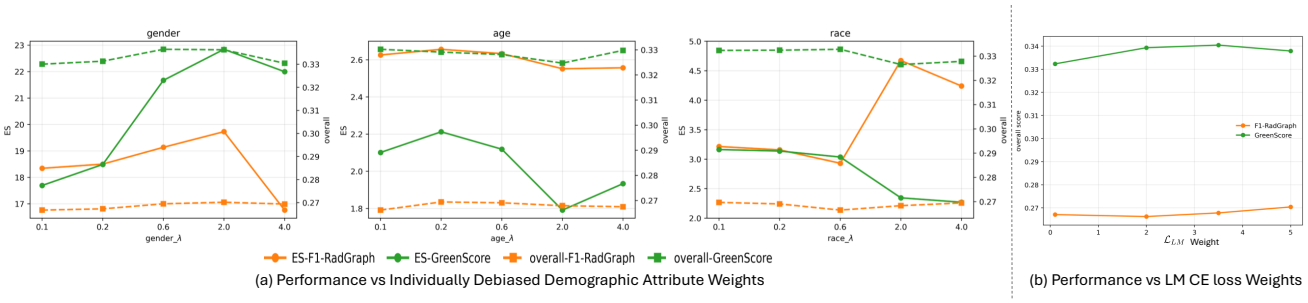
Similarly, we set the values for PadChest as (0.0, 0.3, 0.2) and for HAM10000 as (0.0, 0.6, 0.2).

4. Handling Missing Labels

Requiring demographic labels could be a limiting factor as we discussed in Sec. Limitations in the main paper. in this section, we show that this limitation can be effectively addressed. Recall, FairLLaVA does not need labels at inference time. For training, missing attributes can be predicted reliably on zero-shot radiology datasets: as shown in Tab. 4, TorchXRayVision [4] demographic predictors, trained on CheXpert [6] and NIH ChestX-ray14 [7], achieve high AUC and low age MAE (in years) in our evaluation on both MIMIC-CXR and PadChest datasets.

5. Variance in Equity Scaled Metric

Figure 2 reports 95% confidence intervals obtained via bootstrap resampling (n=1000) on MIMIC-CXR dataset. We observe that ES scores, as well as the underlying subgroup gaps, can exhibit large variance. This is expected because i) fairness gaps are calculated as difference between maximum and minimum subgroup performance, which are extreme values on both ends that inherently have high variance ii) fairness metrics are computed over smaller demographic subgroups, where class imbalance and limited sample counts can amplify estimation noise, iii) moreover, ES depends jointly on both the subgroup gap and the overall task performance, so uncertainty in either term amplifies into the ES score. For this reason, ES should be interpreted together with its confidence interval. All the quantitative results in this work, therefore, report median values. Despite this variability, on average, FairLLaVA (Fig. 2) shows con-



(a) Performance vs Individually Debiased Demographic Attribute Weights

(b) Performance vs LM CE loss Weights

Figure 1. **Hyper Parameters Sensitivity** (a) Varying the contribution of each attribute-specific MI term to the total loss on MIMIC-CXR leads to only minor changes, indicating stable overall performance across attributes. (b) Varying the contribution of language model loss $\mathcal{L}_{LM}$ leads to minor changes in overall performance.

sistently stronger performance across most demographic attributes.

6. Prevalence Trends

To further assess whether FairLLaVA suppresses clinically meaningful subgroup-specific disease patterns, we compare disease prevalence in ground-truth and model-generated reports on the test set. We obtain disease labels by applying the CheXpert labeler to both the reference reports and FairLLaVA-generated reports from the test split. For each CheXpert finding and each demographic subgroup (gender, race-major, and age-group), prevalence is computed as the fraction of samples labeled positive. We then measure the prevalence shift as,

$$\Delta = p_{\text{pred}} - p_{\text{ref}}, \tag{1}$$

where p_{ref} and p_{pred} denote the reference and generated prevalence, respectively.

To focus on clinically meaningful and statistically reliable patterns, we retain only subgroup-finding pairs with subgroup size $N > 50$ and reference prevalence $p_{\text{ref}} \geq 0.7$. We then rank these pairs by descending reference prevalence in Tab. 5. As shown in Tab. 5, FairLLaVA largely preserves these strong prevalence patterns, indicating that debiasing does not simply erase important population-level disease signals.

7. Subgroup Size–Performance Correlation

In this section we present metrics for each subgroup of the demographic attributes "Age", "Race" and "Gender" in MIMIC-CXR and "Age", "Gender" in PadChest in Fig. 3 and Fig. 4 respectively. We detail distribution of the samples across these subgroups in Fig. 5 (a) for MIMIC CXR and Fig. 5 (b) for PadChest. Datasets are quite unbalanced for some demographic attributes, especially for "Age" and "Race". However, as pointed out in the main paper, lower sample count does not automatically mean lower performance, indicating naive classical frequency based methods might not work. For example "White" is the most frequent race in MIMIC-CXR dataset, however, it does not perform the best on any of the baselines as seen in Fig. 3. But performance for the "Black" race is almost always the best, despite it being more than double less represented in the dataset. Similar results are seen on the PadChest dataset in Fig. 4, for example in "Age" 65+ never gets highest performance on any baseline despite having significantly larger count of samples (on neither clinically oriented GREEN score nor classical BLEU-4 score).

8. Fairness Gaps and Overall Performance

We report Equity Scaled Metric (ES-M) in the main paper as balance between the gaps between demographic groups as well as the absolute quality of the generated reports.

In Tab. 6, we show both fairness gaps and overall performance on the MIMIC-CXR dataset. For more general-purpose as well as medical MLLMs such as MedGemma-4B/27B [9], Qwen2.5-7B [13], DeepSeek-VL2 [5], LaVA-Rad [2], the fairness gaps are small, but this comes with substantially lower overall performance compared to the top performing models. For example, Qwen2.5-7B attains the lowest GREEN gaps for race and age and the second-lowest for gender, yet its overall GREEN score (16.10) is less than half of FairLLaVA's 34.32. This illustrates why gap-only metrics can be misleading: a model that is uniformly weak across all groups can appear "fair" while offering limited clinical utility, which motivates our use of ES-Metrics as a more comprehensive evaluation. Among classical fairness methods, reweighting and oversampling we see distinct tradeoffs, they reduce some gaps but either leave others relatively large or noticeably degrade overall performance. The adversarial style fairness solution [10] has significantly lower performance especially in clinically important metrics such as overall GREEN score that drops to 9.36, nearly a four-fold decrease relative to FairLLaVA (34.32), consistent with catastrophic forgetting of clinically meaningful information. In contrast, FairLLaVA substantially lowers dis-

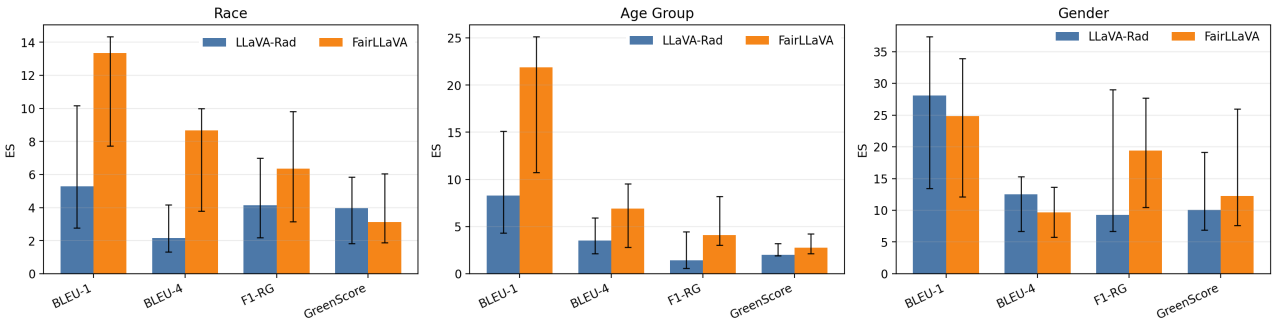


Figure 2. **95% Confidence Intervals** of ES metric on MIMIC-CXR with bootstrap resampling (n=1000)

Group	Finding	N	p_{ref}	p_{pred}	Δ
Age Group					
0-44	Pleural Effusion	87	0.851	0.874	0.023
44-65	Pleural Effusion	899	0.790	0.810	0.020
65+	Pleural Effusion	1014	0.764	0.824	0.060
0-44	Pneumothorax	87	0.736	0.736	0.000
Gender					
F	Pleural Effusion	896	0.823	0.839	0.017
M	Pleural Effusion	1104	0.745	0.804	0.060
Race					
Hispanic or Latino	Pleural Effusion	51	0.863	0.882	0.020
Black or African American	Pleural Effusion	427	0.824	0.874	0.049
Hispanic or Latino	Pneumothorax	51	0.784	0.804	0.020
Asian	Support Devices	77	0.779	0.701	-0.078
White	Pleural Effusion	1381	0.775	0.812	0.038
Asian	Pleural Effusion	77	0.701	0.701	0.000

Table 5. **Prevalence preservation under FairLLaVA** for subgroup-finding pairs with reference prevalence $p_{\text{ref}} \geq 0.7$ and subgroup size $N > 50$. We report the reference prevalence p_{ref}, generated prevalence p_{pred}, and prevalence shift $\Delta = p_{\text{pred}} - p_{\text{ref}}$. High-prevalence findings are largely preserved across demographic groups, with generally small shifts in prevalence.

parities yet keeping comparable overall performance to the best performing LLaVA-Rad, yielding a more balanced fairness–utility trade-off (demonstrated in ES-Metrics tables in the main paper).

Similarly, Tab. 7 on PadChest dataset shows FairLLaVA maintains overall performance that is comparable to, and in some cases exceeds, the best-performing LLaVA-Rad model, while substantially reducing fairness gaps. It clearly outperforms other fairness approaches in terms of demographic fairness across Age and Gender, as well as overall evaluation metrics.

9. Implementation Details

9.1. Preprocessing for HAM10000

For HAM10000, QA data were generated using a concept-grounded synthesis pipeline built on both a language model and a vision–language model using SelfSynthx [11]. We first used OpenAI GPT-4o (`gpt-4o-2024-08-06`) to extract class-level dermoscopic concepts from the diagnosis labels (MEL, NV, BCC, AKIEC, BKL, DF, VASC), after mapping each code to its full clinical name to reduce ambiguity. This produced a label-to-concept bank of dermatology-relevant visual descriptors. We then generated image-level candidate descriptions using LLaVA-1.5-7B (`llava-1.5-7b-hf`, served via vLLM), and scored

Figure 3. Sub-Group performance of baselines across "Race", "Age" and "Gender" subgroups on MIMIC-CXR dataset on GREEN and BLEU-4 metric. We observe that the high number of counts in the train dataset does not correlate with the increased performance. Please also check Fig. 5

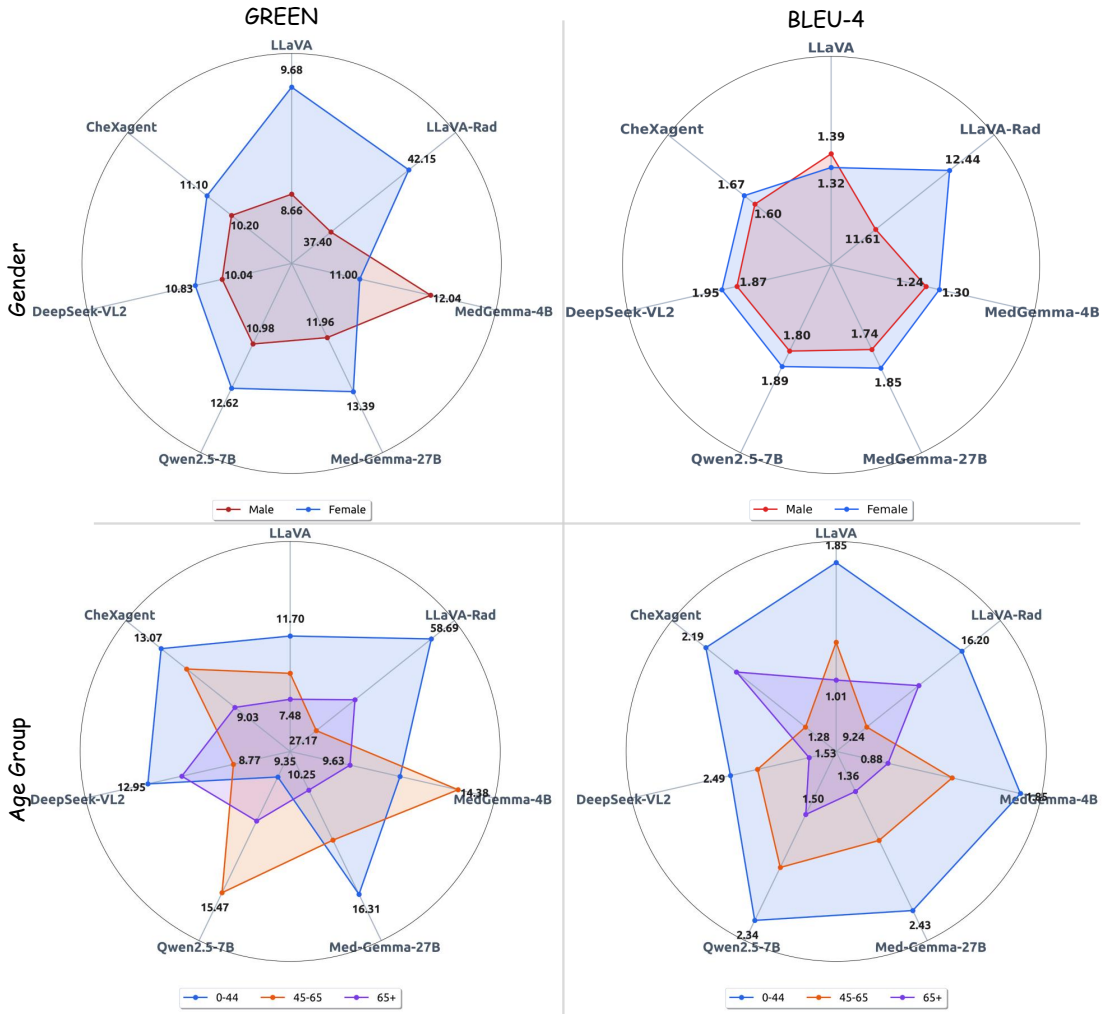


Figure 4. Sub-Group performance of baselines across "Age" and "Gender" subgroups on PadChest dataset on GREEN and BLEU-4 metric. We observe that the high number of counts in the train dataset does not correlate with the increased performance. Please also check Fig. 5

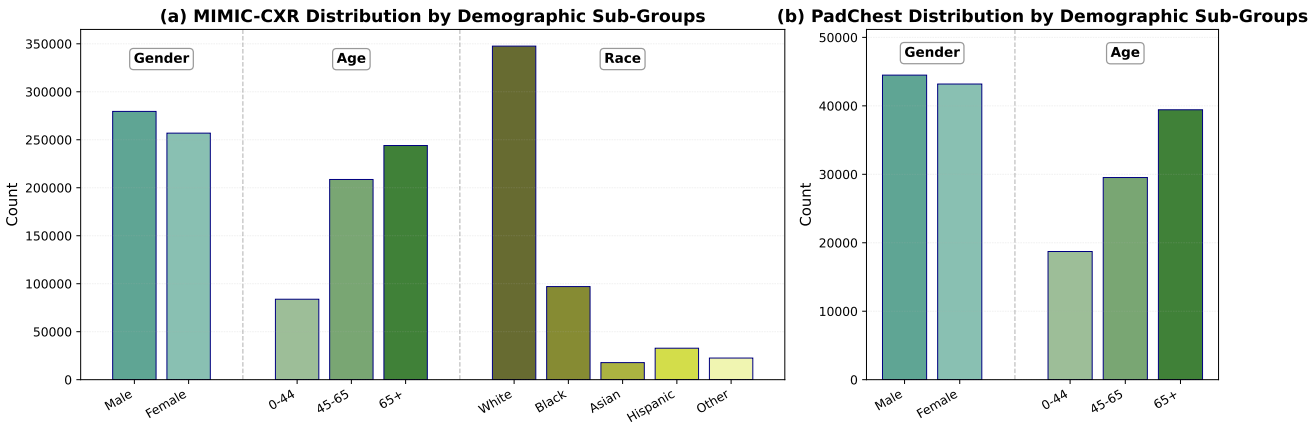


Figure 5. Distribution of counts of demographic subgroups in MIMIC-CXR and PadChest dataset train splits. Some demographic group counts are highly imbalanced, (a) MIMIC-CXR and (b) PadChest. F1-RG is an abbreviation of RadGraph-F1.

Method	Race ↓				Age Group ↓				Gender ↓				Overall ↑			
	BLEU-1	BLEU-4	RadGraph-F1	GREEN	BLEU-1	BLEU-4	RadGraph-F1	GREEN	BLEU-1	BLEU-4	RadGraph-F1	GREEN	BLEU-1	BLEU-4	RadGraph-F1	GREEN
LLaVA-Rad	6.22	6.20	6.20	8.02	3.61	3.39	19.97	16.91	0.36	0.23	2.23	2.29	38.17	15.40	29.80	35.82
MedGemma-4B	6.21	0.69	4.70	2.96	4.37	0.39	5.98	5.61	0.20	0.04	0.45	1.67	17.02	1.61	10.54	20.02
MedGemma-27B	8.03	0.06	4.13	4.73	3.56	0.56	8.33	5.70	0.02	0.00	1.18	1.51	18.15	1.82	14.46	22.45
Qwen2.5-7B	1.43	0.22	2.66	1.88	1.97	0.31	5.44	3.92	0.08	0.04	0.23	0.32	18.29	1.80	10.18	16.098
DeepSeek-VL2	2.17	0.21	2.73	3.34	3.32	0.44	3.90	4.64	0.14	0.07	0.31	0.51	12.93	1.78	8.12	14.16
Reweighting-All	11.87	3.44	5.27	5.50	2.17	1.40	8.58	14.93	0.99	0.11	1.10	0.04	23.30	7.69	20.76	24.38
Resampling-All	3.22	4.90	7.66	15.25	1.93	3.03	7.49	16.05	0.21	0.23	0.60	1.15	36.75	13.69	25.55	34.61
Adv. MLP Classifier-All	4.37	0.74	3.51	3.00	2.12	0.70	1.63	11.32	0.41	0.20	0.56	0.31	21.04	1.88	10.61	9.36
FairLLaVA-All	1.61	0.61	3.50	9.97	0.59	1.01	5.60	12.29	0.40	0.45	0.47	1.80	34.85	13.92	28.52	34.32

Table 6. Fairness Gaps (First three main columns across Race, Age, Gender) and Overall performance (last column) on MIMIC CXR dataset. Highlights tradeoff between Overall-Performance and Fairness-Gaps. Fairness gaps lower the better, Overall performance higher the better.

Method	Age Group ↓				Gender ↓				Overall ↑			
	BLEU-1	BLEU-4	RadGraph-F1	GREEN	BLEU-1	BLEU-4	RadGraph-F1	GREEN	BLEU-1	BLEU-4	RadGraph-F1	GREEN
LLaVA-Rad	9.81	6.96	3.91	31.52	0.02	0.83	0.19	4.76	25.01	12.02	15.35	39.78
MedGemma-4B	3.70	0.97	0.81	4.73	0.37	0.07	0.13	1.04	12.31	1.27	4.37	11.52
MedGemma-27B	4.16	1.07	1.14	6.06	0.53	0.11	0.09	1.43	13.94	1.79	5.03	12.67
Qwen2.5-7B	4.47	0.84	2.28	6.12	0.61	0.09	0.24	1.64	12.45	1.84	5.11	11.80
DeepSeek-VL2	3.92	0.96	1.84	4.18	0.47	0.08	0.19	0.79	11.77	1.91	4.20	10.44
Reweighting-All	16.71	10.29	7.15	33.50	2.19	0.65	1.98	6.42	14.02	7.42	14.37	37.84
Resampling-All	10.35	7.57	4.01	32.91	0.68	0.31	0.01	6.04	23.72	11.26	14.34	38.67
FairLLaVA-All	8.93	6.65	3.45	30.53	0.01	0.82	0.04	4.63	25.11	12.03	15.66	40.03

Table 7. Fairness Gaps (First two main columns across Age, Gender) and Overall performance (last column) on PadChest dataset. Highlights tradeoff between Overall-Performance and Fairness-Gaps.

their relevance to the concept bank using `e5-large-v2` embeddings with an InfoNCE-style selection step to retain the most informative concepts for each image.

Using the selected concepts, we synthesized diverse question types, including short diagnostic, explanatory, and reasoning-style forms, and generated candidate answers with the same VLM, LLaVA-1.5-7B. To ensure correctness, candidate QA pairs were filtered for diagnosis consistency using exact label mention or fuzzy string matching against the ground-truth HAM10000 diagnosis; when consistency was weak, a conservative fallback answer was used. The final output for each image was a curated `new_QA` set containing diagnosis-consistent, concept-supported question-answer pairs. Please refer to SelfSynthx [11] for more details.

9.2. Preprocessing on PadChest

For the English translated version of the PadChest [1, 12] dataset we follow [2] to pre-process the whole report summary into standard radiology report sections such as "Findings", "Indication" and "Impression". Indication section briefly states the reason why the study was ordered and the clinical question it aims to address (e.g., symptoms, suspected diagnosis). Findings section provides a description of what is observed in the images, without overall judgment. Impression includes interpretive summary of the key findings, likely diagnoses, and any critical recommendations. This work is focused on the producing the findings section of the radiology report. We also remove mentions of dates of previous studies each report references. We prompt [8] for pre-processing. Example of such prompt is given in Fig. 6.

9.3. Implementation of Baselines

For a fair comparison, all the reweighting, resampling, Adv. classifier baselines follow exactly similar instantiation of the Language Model and in-domain Image Embedder as FairLLaVA. We use Vicuna-7b-v1.5 [3] as our base language model and BioMedCLIP [14] as the image encoder. Adv. MLP classifier uses same FairLLaVA DAC architecture of 2 layer MLP and uses middle three layers (14, 16, 20) for debiasing for one epoch. Before debiasing, Adv-MLP classifier is pretrained on LLaVA-Rad for three epochs achieving overall demography classification accuracy of 72%. All the baselines are either trained on 8 NVIDIA RTX-A6000 GPUs or 2 NVIDIA A100 with effective batch size of 96 for one epoch on all datasets. We use the same seed as used in FairLLaVA to control for randomness in weight initialization and otherwise in the implementation.

10. Additional Ablations

To characterize the effect of the proposed modules for demographic information minimization, we compare two

saama OpenBioLLM70B

You are an expert medical assistant AI capable of modifying clinical documents to user specifications. You make minimal changes to the original document to satisfy user requests. You never add information that is not already directly stated in the original document.

Extract three sections from the input radiology report: 'Indication', 'Findings' and 'Impression'.Leave an extracted section as null if it does not exist in the original report. The output should be in JSON format. An Indication section can refer to the History, Indication or Reason for Study sections in the original report. Remove any information not directly observable from the current imaging study. For instance, remove any patient demographic data, past medical history, or comparison to prior images or studies. The generated 'Findings' and 'Impression' sections should not reference any changes based on prior images, studies, or external knowledge about the patient. Rewrite such comparisons as a status observation based only on the current image or study. Remember to remove any numbering or bullets.

Examples of inputs and expected outputs:
INPUT:
REPORT: There is an increased cardiothoracic index, suggesting possible cardiomegaly. A calcified granuloma is observed at the base of the left lung. There is an image compatible with chronic inflammatory changes, likely bronchiectasis, in the retrocardiac region. Also noted is a double-curved dorsolumbar scoliosis.

OUTPUT:
{"INDICATION":"Progressive shortness of breath and suspected congestive heart failure-evaluation of cardiac size and pulmonary status."
"FINDINGS":"Increased cardiothoracic ratio. Calcified granuloma at the base of left lung. Double-curved dorsolumbar scoliosis".
"IMPRESSION":"Findings suggestive of Cardiomegaly Chronic inflammatory changes seen."}

Please do below report
INPUT:
REPORT:Findings are consistent with COPD, with prominence of the right hilum suggestive of inflammatory changes on the right side. Evidence of an old fracture in the left ribs. Mild compression of the left side of the trachea, possibly related to a goiter, suprasternal structure, or other adjacent anatomical changes.

Figure 6. OpenBio-LLM Prompt used for preprocessing the PadChest Dataset to convert full report summary into standard radiology report section - "Findings", "Indication", "Impression".

training strategies for the DAC classifier when debiasing with respect to the *Age* attribute on the MIMIC-CXR dataset:

1. **Pretrained DAC only.** We first pretrain the DAC classifier for three epochs using only the DAC loss $\mathcal{L}_{DAC}$ while keeping the base LLaVA-Rad model frozen. In the subsequent debiasing stage, we freeze this classifier and optimize the report generator with $\mathcal{L}_{LM} + \mathcal{L}_{DIM}$ using the fixed DAC as an adversary. This corresponds to the row *FairLLaVA-Age ($\mathcal{L}_{DAC}$, then $\mathcal{L}_{LM} + \mathcal{L}_{DIM}$)* in Tab. 8.
2. **Joint DAC + MI training.** In the second setting, which we use in all main experiments, the DAC is *not* pretrained. Instead, it is trained jointly with the report generator during fairness-aware fine-tuning, with the combined objective $\mathcal{L}_{LM} + \mathcal{L}_{DIM} + \mathcal{L}_{DAC}$. This corresponds to the row *FairLLaVA-Age ($\mathcal{L}_{LM} + \mathcal{L}_{DIM} + \mathcal{L}_{DAC}$)*.

All ablations are run for one debiasing epoch. As shown in Tab. 8, joint training of the DAC yields consistently better equity-scaled metrics *and* higher overall performance. For the age-focused ES-metrics, ES-BLEU-4 improves from 2.48 to 5.47, ES-RadGraph-F1 from 3.50 to 3.75, and ES-GREEN from 2.08 to 2.36 when moving from the pretrained DAC to the jointly trained DAC. At the same time, overall report quality improves: BLEU-1 rises from 31.77 to 35.28, BLEU-4 from 10.48 to 13.96, RadGraph-F1 from 27.16 to 28.49, and GREEN from 33.06 to 34.12.

These results highlight a key limitation of using a strong pretrained attribute classifier as an adversary. Because the DAC is optimized in isolation and then frozen, its gradients during debiasing tend to aggressively remove age-related signal from the shared representation, including clinically relevant features, which leads to *catastrophic forgetting* of some concepts and a noticeable drop in overall performance. This behavior is previously noted as limitation of Adv. MLP classifier [10], where pretrained demographic classifiers can over-regularize the model. In contrast, jointly training the DAC with $\mathcal{L}_{DIM}$ and $\mathcal{L}_{LM}$ allows the model to gradually disentangle demographic information while being continually constrained by the primary reporting objective, resulting in a more balanced trade-off between equity-scaled performance and overall clinical utility.

11. Additional Qualitative Samples

In Fig. 7, we show qualitative examples where the LLaVA-Rad baseline correctly identifies a key condition for one demographic subgroup but fails to mention the same finding for another, suggesting a dependence on spurious correlations with subgroup membership. In contrast, FairLLaVA, trained with our debiasing objectives, more consistently recovers the clinically relevant findings across all subgroups by focusing on image evidence rather than demographic cues. In some cases, such as the White–vs–Black pair, the baseline also appears less confident in its descriptions (Gender: Male in Fig. 7), whereas FairLLaVA provides clearer and more definitive statements about the underlying pathology in Fig. 7.

References

[1] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. 7

[2] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *arXiv preprint arXiv:2403.08002*, 2024. 3, 7

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhang-

Method	Age Group ↑				Overall ↑			
	ES-BLEU-1	ES-BLEU-4	ES-RadGraph-F1	ES-GREEN	BLEU-1	BLEU-4	RadGraph-F1	GREEN
FairLLaVA-Age ($\mathcal{L}_{DAC}$ then$\mathcal{L}_{LM} + \mathcal{L}_{DIM}$)	3.83	2.48	3.50	2.08	31.77	10.48	27.16	33.06
FairLLaVA-Age ($\mathcal{L}_{LM} + \mathcal{L}_{DIM} + \mathcal{L}_{DAC}$)	**17.82**	**5.47**	**3.75**	**2.36**	**35.28**	**13.96**	**28.49**	**34.12**

Table 8. ES-M metrics for the Age Group attribute on the MIMIC-CXR dataset and overall performance. Joint training of the $\mathcal{L}_{DAC}$ with $\mathcal{L}_{DIM}$ improves equity-scaled scores while preserving or improving overall report quality.

Figure 7. Qualitative samples show LLaVA-Rad baseline produces unfair results for some subgroups ("Old", "White", "Male") of demographic attributes.

hao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 7

[4] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, pages 231–249. PMLR, 2022. 2

[5] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[6] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad

Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 2

[7] Anna Majkowska, Saurabh Mittal, Daniel F. Steiner, Joshua J. Reicher, Scott M. McKinney, G. E. Duggan, Kiran Eswaran, Po-Hsuan Cameron Chen, Yun Liu, S. Ram Kalidindi, Amy Ding, Greg S. Corrado, Daniel Tse, and Shravya Shetty. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020. 2

[8] Ankit Pal and Malaikannan Sankarasubbu. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024. 7

[9] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 3

[10] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6820–6829, 2023. 3, 8

[11] Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. *arXiv preprint arXiv:2502.14044*, 2025. 4, 7

[12] Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand, 2024. Association for Computational Linguistics. 7

[13] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[14] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 2, 7