FairGE: Gated Expert Routing for Intersectional Fairness in Medical Foundation Models

Yuchen Shao[1⋆], Yutong Xie[2⋆], Jian Chen[1], and Qi Chen[3]

[1] South China University of Technology, Guangzhou, China
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[3] Australian Institute for Machine Learning, Adelaide University, Adelaide, Australia
123shaoyuchen@gmail.com, ellachen@scut.edu.cn, yutong.xie678@gmail.com, qi.chen04@adelaide.edu.au

Abstract. Medical foundation models are increasingly used as generic visual backbones for downstream diagnosis, yet existing fairness studies in medical imaging mostly focus on a single protected attribute (e.g., age or sex) and rarely examine intersectional subgroups defined by multiple demographics. Compared to single-attribute settings, multi-attribute fairness introduces additional challenges: (i) intersectional subgroup data sparsity, where some demographic combinations contain very few samples, making subgroup-specific adaptation unstable and prone to overfitting; and (ii) sensitive attribute leakage (proxy discrimination), where learned representations retain correlated demographic information and allow models to rely on demographic shortcuts rather than clinically relevant visual features. To address this, we propose Fair Gated-Expert adaptation with multi-attribute gating (**FairGE**), a fairness-aware adaptation framework for medical foundation models. FairGE integrates two complementary components: a Multi-attribute Gated distribution-aware Mixture-of-Experts (MA-dMoE) module that employs shared experts with multi-attribute gating to enable intersectional adaptation while avoiding overfitting on small subgroups, and a Multi-attribute Purification (MAP) module that combines a lightweight MLP with multiple adversarial attribute classifiers (one for each sensitive attribute) trained via gradient reversal to prevent the model from encoding sensitive demographic information in the final representation. Experiments on two representative medical benchmarks, HAM10000 and Harvard-FairVLMed, show that FairGE consistently reduces both marginal and intersectional disparities, improves worst-group performance, and preserves overall AUC. Code is available at `https://github.com/yuchenshao749/FairGE`.

Keywords: Medical foundation models · Fairness · Intersectional fairness · Mixture-of-Experts · Medical multimodal large language models.

1 Introduction

Foundation models have recently transformed medical image analysis, enabling a single pretrained model to support many downstream tasks with modest task-

⋆ Y. Shao & Y. Xie contributed equally. Corresponding authors: Q. Chen & J. Chen

specific data [12, 26]. This trend also extends to medical multimodal large language models (MLLMs), where a visual encoder is paired with a language model for clinical reasoning and report generation [1, 8, 21]. As such models are increasingly used as generic visual feature extractors in diagnostic pipelines, demographic fairness becomes a central concern: unequal performance across patient groups can undermine clinical trust and exacerbate existing health disparities [13, 17].

Motivated by these concerns, recent work has begun to evaluate and improve fairness in medical imaging foundation models. FairMedFM [6] benchmarks a broad range of vision foundation models across datasets and sensitive attributes under common usage modes (e.g., zero-shot, linear probing, and parameter-efficient fine-tuning). FairCLIP [10] introduces the Harvard-FairVLMed glaucoma benchmark with fine-grained demographic annotations and studies fairness of CLIP/BLIP-2 style vision-language models [9, 15]. At the method level, prior fairness-aware adaptation work includes distribution-aware Mixture-of-Experts (dMoE) [14] for medical image segmentation and MoE-based fairness designs for medical vision-language models [3, 4, 18, 22], highlighting the promise of attribute-aware routing for fairer medical AI systems.

Despite this progress, existing studies are still largely centered on single-attribute fairness, such as considering only one of age, gender, or race. In real-world clinical deployment, however, each patient is characterized by multiple demographics simultaneously, and fairness risks often concentrate in small intersectional groups defined by attribute combinations (e.g., age-sex subgroups) [16, 17, 25]. This makes multi-attribute fairness a more practical and important target. For this setting, two challenges are particularly critical: (1) intersectional subgroup data sparsity, where some demographic combinations contain very few samples, making subgroup-specific adaptation unstable and prone to overfitting; and (2) sensitive attribute leakage (proxy discrimination), where learned visual representations retain correlated demographic information and allow models to rely on demographic shortcuts rather than clinically relevant visual features.

To address these challenges, we propose Fair Gated-Expert adaptation with multi-attribute gating (**FairGE**), a fairness-aware adaptation framework for frozen medical foundation model visual encoders. FairGE consists of two complementary modules tailored to multi-attribute fairness. First, we introduce a Multi-attribute Gated distribution-aware Mixture-of-Experts (MA-dMoE) module that inserts lightweight dMoE adapters into the last encoder blocks and uses shared experts with multi-attribute gating: for each demographic attribute (e.g., age and sex) we maintain a separate attribute-specific router and, for a given patient, aggregate the routing distributions from all active attributes. This improves intersectional adaptation while avoiding overfitting from training separate adapters for sparse intersectional groups. Second, we introduce a Multi-attribute Purification (MAP) module that applies a compact Purifier MLP to the pooled visual feature and trains multiple adversarial attribute classifiers (one for each sensitive attribute) via gradient reversal to suppress sensitive demographic information in the final representation.We validate FairGE on two

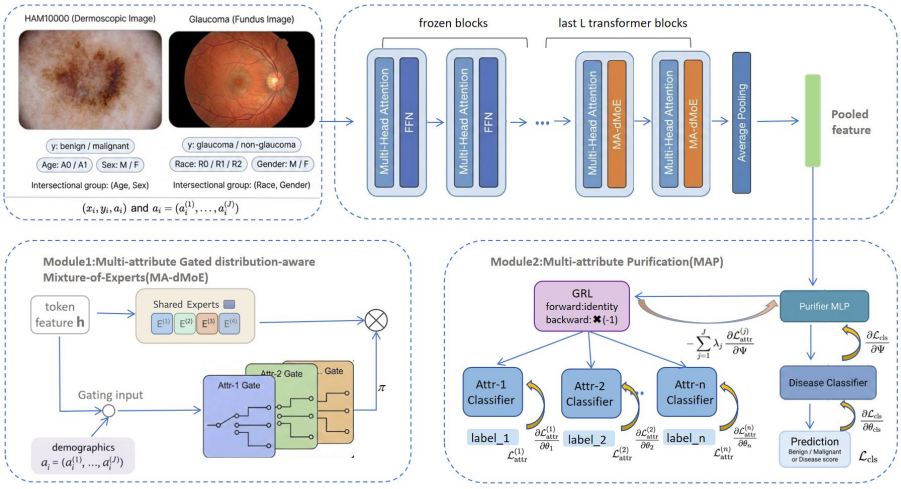

Fig. 1. Overview of **FairGE** with two modules: MA-dMoE inserts multi-attribute gated dMoE adapters into the last few visual-encoder blocks for intersectional adaptation, and MAP applies a Purifier MLP with adversarial attribute classifiers to suppress demographic leakage before disease classification.

representative benchmarks derived from HAM10000 dermoscopy images and the Harvard-FairVLMed glaucoma dataset.

Our contributions are summarized as follows:

- We propose FairGE, a fairness-aware adaptation framework to address subgroup and intersectional fairness under multiple sensitive attributes.
- FairGE integrates MA-dMoE, which combines shared experts with multi-attribute gating to improve subgroup adaptation, and MAP, which uses a lightweight Purifier MLP and multiple adversarial attribute classifiers to suppress sensitive demographic leakage from the final visual representation.
- We conduct comprehensive experiments on two medical imaging benchmarks, showing consistent gains in overall and intersectional subgroup AUC, improved worst-group performance, and a strong fairness–AUC trade-off.

2 Method

Figure 1 illustrates our FairGE. Given a medical image and its demographic attributes (e.g., age and sex), we extract visual features using a pretrained medical foundation model encoder. FairGE then applies two complementary modules: (1) MA-dMoE, inserted into the last few transformer blocks to improve intersectional adaptation via shared experts and multi-attribute demographic gating; and (2) MAP, which applies a compact Purifier MLP and multiple adversarial attribute classifiers to suppress sensitive demographic information in the final representation. The purified feature is finally fed to a disease classifier.

2.1 Multi-attribute Gated distribution-aware Mixture-of-Experts

As shown in Figure 1, let $\mathcal{D} = \{(x_i, y_i, \mathbf{a}_i)\}_{i=1}^{N}$ be a labeled dataset, where x_i is a medical image, $y_i \in \{0, 1\}$ is the disease label, and $\mathbf{a}_i = (a_i^{(1)}, \ldots, a_i^{(J)})$ is a vector of J discrete demographic attributes (e.g., age group and sex), with $a_i^{(j)} \in \mathcal{A}^{(j)}$. We build on a pretrained ViT-style visual encoder and insert dMoE adapters into the last L transformer blocks, while keeping earlier blocks frozen.

For an adapted block ℓ, let $\mathbf{h}_{i,t}^{(\ell)} \in \mathbb{R}^d$ denote the token at position t (after self-attention and before the FFN). We replace the standard FFN with an MA-dMoE layer with E shared experts $\{E_\ell^{(e)}\}_{e=1}^{E}$ and token-wise sparse routing weights $\pi_{i,t,\cdot}^{(\ell)} \in \mathbb{R}^E$ (Noisy Top-k routing [14]):

$$\tilde{\mathbf{h}}_{i,t}^{(\ell)} = \mathbf{h}_{i,t}^{(\ell)} + \sum_{e=1}^{E} \pi_{i,t,e}^{(\ell)} \, E_\ell^{(e)}\left(\mathbf{h}_{i,t}^{(\ell)}\right). \tag{1}$$

The experts are shared across all demographic groups; only the routing differs. This shared-expert design is parameter-efficient and helps avoid overfitting when some intersectional subgroups are under-represented.

To make routing aware of multiple demographics, for each attribute $j \in \{1, \ldots, J\}$ and each group $g \in \mathcal{A}^{(j)}$, we define a group-specific router $G_\ell^{(j,g)}$. For sample x_i, the active router for attribute j is $G_\ell^{(j,a_i^{(j)})}$, which produces a token-wise expert distribution $\pi_{i,t}^{(\ell,j)} \in \mathbb{R}^E$ after Noisy Top-k routing and softmax. We then combine the attribute-specific gates using non-negative weights $\{\alpha_j\}_{j=1}^{J}$:

$$\tilde{\boldsymbol{\pi}}_{i,t}^{(\ell)} = \sum_{j=1}^{J} \alpha_j \, \pi_{i,t}^{(\ell,j)}, \qquad \pi_{i,t,e}^{(\ell)} = \frac{\tilde{\pi}_{i,t,e}^{(\ell)}}{\sum_{e'=1}^{E} \tilde{\pi}_{i,t,e'}^{(\ell)}}. \tag{2}$$

The fused weights $\pi_{i,t,e}^{(\ell)}$ are used in Eq. (1). This allows FairGE to condition expert usage on multiple attributes jointly (e.g., age and sex) without training a separate adapter for each intersectional subgroup.

Let $\mathbf{H}_i \in \mathbb{R}^{T \times d}$ denote the final token sequence output by the adapted encoder. We obtain the pooled visual representation by

$$\mathbf{z}_i = \mathrm{Pool}(\mathbf{H}_i) \in \mathbb{R}^d. \tag{3}$$

2.2 Multi-attribute Purification

Although multi-attribute gating improves subgroup-specific adaptation, the pooled feature may still encode sensitive demographic information and induce shortcut-based predictions. To mitigate this, we apply a lightweight Purifier MLP $p_\psi : \mathbb{R}^d \to \mathbb{R}^d$ to obtain a purified feature

$$\mathbf{z}_i' = p_\psi(\mathbf{z}_i), \tag{4}$$

and use a disease classifier head $g_{\theta_{cls}} : \mathbb{R}^d \to [0,1]$ for the main task:

$$\hat{y}_i = g_{\theta_{\text{cls}}}(\mathbf{z}'_i), \qquad \mathcal{L}_{\text{cls}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \text{BCE}(\hat{y}_i, y_i). \tag{5}$$

To suppress multi-attribute sensitive attribute leakage, we attach one adversarial classifier for each sensitive attribute. For attribute j, let $h^{(j)}_{\theta_j} : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{A}^{(j)}|}$ be an attribute predictor (e.g., age classifier, sex classifier), and let $\mathcal{R}(\cdot)$ denote a gradient reversal layer (GRL [2]), which acts as the identity in the forward pass and multiplies the backward gradient by -1. The attribute loss is

$$\mathcal{L}^{(j)}_{\text{attr}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \text{CE}\Big(h^{(j)}_{\theta_j}(\mathcal{R}(\mathbf{z}'_i)), a^{(j)}_i\Big). \tag{6}$$

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \sum_{j=1}^{J} \lambda_j \mathcal{L}^{(j)}_{\text{attr}}, \tag{7}$$

where $\lambda_j \geq 0$ is a weighting coefficient for the j-th adversarial branch. As shown in Figure 1, the GRL preserves the input in the forward pass and reverses the sign of the gradient in the backward pass. Consequently, each attribute classifier $h^{(j)}_{\theta_j}$ is optimized to predict the corresponding sensitive attribute $a^{(j)}_i$, whereas the gradient propagated to the upstream feature extractor (including the adapted encoder and the Purifier MLP) has the opposite sign, thereby encouraging the final representation $\mathbf{z}'_i$ to discard sensitive-attribute information. The coefficient λ_j controls the relative strength of the j-th adversarial signal in the overall objective.

In all experiments, we initialize the visual encoder from the medical foundation model, train only the last L dMoE-adapted blocks (experts and routers), the Purifier, the disease classifier, and the adversarial attribute classifiers, and keep earlier encoder blocks frozen.

3 Experiments

3.1 Datasets and Experimental Setup

We evaluate on two benchmarks with demographic annotations. HAM10000 [20] contains 9948 dermoscopic images of pigmented skin lesions; we follow prior work and cast it as a binary benign vs. malignant classification task with age and sex as protected attributes. We construct patient-level train/val/test splits, where age and sex are considered during the splitting process to encourage demographic coverage across splits. Harvard-FairVLMed [10] provides 10,000 fundus images with rich demographic annotations; we construct a binary glaucoma classification task from its glaucoma label and use race and gender as protected attributes.

Table 1. Sample counts of intersectional demographic subgroups in HAM10000 and Harvard-FairVLMed. For HAM10000, A0/A1 denote age ≤ 60 / > 60.

Dataset	Group	Male	Female	Total
HAM10000	A0	3525	3624	7149
	A1	1875	924	2799
	Total	5400	4548	9948
Harvard-FairVLMed	Asian	393	426	819
	Black	658	833	1491
	White	3318	4372	7690
	Total	4369	5631	10000

Table 1 reports the number of samples in each intersectional subgroup used in our experiments. For HAM10000, we summarize age–sex intersections, and for Harvard-FairVLMed, we summarize race–gender intersections.

We first use linear probing with a frozen backbone as a diagnostic baseline to reveal fairness properties of the pretrained visual representation. We then evaluate our proposed FairGE, which performs parameter-efficient adaptation with MA-dMoE and MAP.

We report overall AUC and accuracy on the test sets. For fairness, we evaluate intersectional age-sex subgroups using the max subgroup AUC gap (AUC-gap) and equity-scaled AUC (ES-AUC) [11], and additionally report a difference-in-equalized-odds score (DEOdds) [5] on HAM10000.

3.2 Results on Fairness-Aware Adaptation

We first consider a simple baseline called "Base", which removes any explicit fairness intervention (*i.e.,* without MA-dMoE and MAP for our method). Besides, we compare FairGE against three groups of fairness baselines: (1) Traditional fairness methods: group reweighting (Reweight), and an adversarial classifier (AdvClassifier) that predicts sensitive attributes from intermediate features [7, 24]; (2) Recent debiasing methods: FairSDE [19] and a VAE-based debiasing approach [28] applied to foundation model embeddings; and (3) CLIP-based fairness methods: on HAM10000, we additionally compare with CMAC-MMD and DualFairVL [23, 27], using their finetuned visual backbones as frozen feature extractors and training a linear classifier on top.

Across both datasets, we observe a consistent pattern. On HAM10000 (Table 2), Base achieves strong overall AUC (90.56) but suffers from a large intersectional gap (AUC-gap 11.98). Traditional debiasing methods offer only modest gains for these groups and leave substantial disparities. In contrast, our FairGE adaptation markedly reduces the gap (AUC-gap 3.19), substantially improves ES-AUC (from 75.89 to 85.81), and increases overall AUC to 92.07.

On Harvard-FairVLMed (Table 3), we observe similar long-tail behavior across race–gender intersections: Base attains reasonable overall AUC (76.03) but again exhibits a large subgroup gap (AUC-gap 14.76). FairGE consistently

Table 2. Fairness results on HAM10000 (binary benign vs. malignant). We report overall AUC, the maximum AUC gap across the four age–sex intersectional subgroups (AUC-gap), equity-scaled AUC (ES-AUC), DEOdds, and AUC for each subgroup. Here, A0/A1 denote age ≤ 60 / > 60, and M/F denote male/female.

Method	Overall AUC	Fairness metrics			age-sex intersection			
		AUC-gap	ES-AUC	DEOdds	A0-M	A0-F	A1-M	A1-F
Base	90.56±0.84	11.98±1.36	75.89±1.47	20.63±1.81	94.16	87.79	85.98	82.18
Reweight	90.40±0.92	10.79±1.18	76.29±1.52	17.70±1.64	93.79	87.43	85.67	83.00
AdvClassifier	90.44±0.76	11.22±1.27	76.64±1.43	17.94±1.58	93.70	87.74	86.36	82.48
FairSDE	90.59±0.69	8.71±1.04	77.10±1.29	27.17±1.91	93.77	85.58	86.82	85.06
VAE-debias	88.81±1.12	5.13±0.88	80.07±1.16	15.69±1.38	90.39	85.50	85.26	86.34
DualFairVL	89.93±0.95	11.17±1.45	75.35±1.67	27.83±1.96	93.75	87.64	82.58	84.04
CMAC-MMD	91.09±0.73	5.88±0.94	83.62±1.08	21.50±1.72	92.57	91.56	86.69	88.51
FairGE (ours)	**92.07±0.61**	**3.19±0.57**	**85.81±0.92**	**10.77±1.05**	**92.53**	**90.42**	**89.32**	**89.64**

Table 3. Fairness results on Harvard-FairVLMed glaucoma classification. We report overall AUC, AUC-gap, ES-AUC, and AUC for the six race–gender intersectional subgroups. Here, R0/R1/R2 denote Asian/Black/White, and M/F denote male/female.

Method	Overall AUC	Fairness metrics		Race–gender intersection					
		AUC-gap	ES-AUC	R0-M	R0-F	R1-M	R1-F	R2-M	R2-F
Base	76.03±0.86	14.76±1.42	61.31±1.58	77.36	80.39	78.26	65.63	79.51	73.83
Reweight	75.88±0.91	13.35±1.27	61.49±1.46	73.56	78.91	73.96	67.11	80.46	73.10
AdvClassifier	76.80±0.78	14.90±1.53	62.17±1.39	78.52	82.09	77.44	67.19	80.72	74.45
FairSDE	76.86±0.83	15.80±1.68	61.32±1.62	73.78	81.86	76.23	66.06	80.40	74.57
VAE-debias	76.06±1.04	12.56±1.21	61.59±1.34	82.32	78.91	74.00	69.76	79.93	73.90
FairGE (ours)	**79.08±1.13**	**7.53±0.82**	**68.03±1.06**	**78.09**	**77.21**	**75.35**	**73.93**	**81.46**	**76.96**

improves fairness, reducing the AUC-gap to 7.53 and raising ES-AUC from 61.31 to 68.03, together with an overall AUC gain to 79.08.

3.3 Ablation Studies

We conduct ablation studies on HAM10000 to examine the contribution of each component in FairGE. Specifically, we evaluate four ablated variants: (1) *w/o MA-dMoE*, which removes the dMoE adapters and keeps only the lightweight Purifier MLP on top of the frozen visual representation; (2) *w/o group-gating*, which keeps the expert FFNs but replaces demographic-specific gates with a single shared gate over the same expert pool; (3) *w/o MAP*, which removes the feature purification module and keeps only the dMoE-based adaptation; and (4) *FairGE-SA*, which disables multi-attribute gating and uses single-attribute (age-based) gating only. For fair comparison, all variants use the same training protocol and evaluation metrics as in Table 2.

As shown in Table 4, all components contribute to both accuracy and AUC-based intersectional fairness. Removing the dMoE adapters (*w/o MA-dMoE*) causes the largest drop in overall and subgroup AUC, confirming expert-based adaptation as the main driver of performance gains. Collapsing demographic-

Table 4. Component ablation on HAM10000 (binary benign vs. malignant). We report the same metrics as Table 2: overall AUC, AUC-gap, ES-AUC, DEOdds, and AUC for each age-sex intersectional subgroup. All values are reported in percentage form.

Method	AUC	AUC-gap	ES-AUC	DEOdds	A0-M	A0-F	A1-M	A1-F
Base	90.56	11.98	75.89	20.63	94.16	87.79	85.98	82.18
w/o MA-dMoE	90.78	9.42	78.08	18.63	93.16	88.20	86.51	83.74
w/o group-gating	91.05	7.16	80.67	15.84	92.84	88.92	87.47	85.68
w/o MAP	91.34	5.38	82.65	13.68	92.30	89.42	88.13	86.92
FairGE-SA	91.54	4.22	83.94	12.23	92.02	89.76	88.48	87.80
FairGE	**92.07**	**3.19**	**85.81**	**10.77**	**92.53**	**90.42**	**89.32**	**89.64**

Table 5. Effect of the number of adapted blocks L on HAM10000 data.

L	AUC-gap	ES-AUC
2	5.16	81.98
4	4.72	83.06
6	**4.22**	**83.94**
8	5.43	82.52
10	6.78	81.14

Table 6. Effect of age/sex routing weights $(\alpha_{\text{age}}, \alpha_{\text{sex}})$ on HAM10000 data.

$(\alpha_{\text{age}}, \alpha_{\text{sex}})$	AUC-gap	ES-AUC
(1.0, 0.0)	4.22	83.94
(0.9, 0.1)	3.90	84.86
(0.8, 0.2)	**3.19**	**85.81**
(0.7, 0.3)	3.62	84.24
(0.6, 0.4)	4.11	83.25

specific gates into a shared gate (*w/o group-gating*) still improves over Base but remains worse than the gated variants, underscoring the value of demographic-aware routing. Dropping the Purifier MLP (*w/o MAP*) leads to a modest yet consistent degradation in fairness metrics, indicating complementary debiasing effects beyond dMoE routing. Among the gated variants, *FairGE* with multi-attribute gating (age and sex) achieves the smallest AUC-gap and highest ES-AUC, outperforming the single-attribute *FairGE-SA*.

3.4 Discussion

To better understand the design choices in our FairGE adapter, we analyze (i) the number L of last transformer blocks adapted with dMoE in the single-attribute variant FairGE-SA, and (ii) the age/sex weights $(\alpha_{\text{age}}, \alpha_{\text{sex}})$ in the multi-attribute gating of FairGE. All experiments in this section are conducted on HAM10000.

As shown in Table 5, when applying FairGE-SA to different numbers of last blocks, fairness metrics improve as we increase the number of adapted blocks from L=2 to L=6, while further expanding adaptation to 8 or 10 blocks yields diminishing or even slightly worse AUC-gap and ES-AUC. We then use L=6 as a default for FairGE, which strikes a good balance between capacity and stability.

Table 6 studies the multi-attribute gating in FairGE. Since age exhibits larger disparities than sex in our main results, we restrict to settings with $\alpha_{\text{age}} \geq 0.6$.

A higher weight to age is generally beneficial, and the best trade-off is achieved at $(\alpha_{\text{age}}, \alpha_{\text{sex}}) = (0.8, 0.2)$, which we adopt in our default FairGE configuration.

4 Conclusion

We investigated subgroup and intersectional fairness in medical image classification with frozen medical foundation model encoders on HAM10000 and Harvard-FairVLMed, and proposed **FairGE**, a lightweight fairness-aware adaptation framework built on MA-dMoE and MAP. Across both datasets, FairGE improves the overall fairness–AUC trade-off and yields more balanced performance across intersectional subgroups. Ablations further show that expert-based adaptation, multi-attribute gating, and feature purification provide complementary benefits. Future work will extend FairGE to more medical foundation models, broader protected attributes, and additional clinical tasks to further evaluate its generalizability and fairness in real-world deployment.

Acknowledgments. This research was funded by the National Natural Science Foundation of China (Grant No. 62376099).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, Q., Xie, Y., Wu, B., Chen, X., Ang, J., To, M.S., Chang, X., Wu, Q.: Act like a radiologist: radiology report generation across anatomical regions. In: Proceedings of the Asian Conference on Computer Vision. pp. 1–17 (2024)
2. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
3. Germino, J., Moniz, N., Chawla, N.V.: Fairness-aware mixture of experts with interpretability budgets. In: International Conference on Discovery Science. pp. 341–355. Springer (2023)
4. Germino, J., Moniz, N., Chawla, N.V.: Fairmoe: counterfactually-fair mixture of experts with levels of interpretability. Machine Learning **113**(9), 6539–6559 (2024)
5. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
6. Jin, R., Xu, Z., Zhong, Y., Yao, Q., Dou, Q., Zhou, S.K., Li, X.: Fairmedfm: fairness benchmarking for medical imaging foundation models. Advances in Neural Information Processing Systems **37**, 111318–111357 (2024)
7. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems **33**(1), 1–33 (2012)
8. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

10. Luo, Y., Shi, M., Khan, M.O., Afzal, M.M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., et al.: Fairclip: Harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12289–12301 (2024)
11. Luo, Y., Tian, Y., Shi, M., Pasquale, L.R., Shen, L.Q., Zebardast, N., Elze, T., Wang, M.: Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. IEEE Transactions on Medical Imaging **43**(7), 2623–2633 (2024)
12. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
13. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
14. Oh, Y., Jin, P., Park, S., Kim, S., Yoon, S., Kim, K., Kim, J.S., Li, X., Li, Q.: Distribution-aware fairness learning in medical image segmentation from a control-theoretic perspective. arXiv preprint arXiv:2502.00619 (2025)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
16. Ramachandranpillai, R., Sampath, K., Mohammad, A., Alikhani, M.: Fairness at every intersection: Uncovering and mitigating intersectional biases in multimodal clinical predictions. arXiv preprint arXiv:2412.00606 (2024)
17. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine **27**(12), 2176–2182 (2021)
18. Sharma, S., Henderson, J., Ghosh, J.: Feamoe: fair, explainable and adaptive mixture of experts. arXiv preprint arXiv:2210.04995 (2022)
19. Tan, X., Wang, Y., Pham, T.H., Zhang, P., Zhang, X.: Achieving fairness without harm via selective demographic experts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 39331–39339 (2026)
20. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 180161 (2018)
21. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. Nejm Ai **1**(3), AIoa2300138 (2024)
22. Wang, P., Tong, L., Wu, J., Liu, J., Liu, Z.: Fair-moe: Medical fairness-oriented mixture of experts in vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 186–196. Springer (2025)
23. Xia, Y., Ma, B., He, J., Wang, Z., Dou, Q., Xia, Y.: Toward robust medical fairness: Debiased dual-modal alignment via text-guided attribute-disentangled prompt learning for vision-language models. arXiv preprint arXiv:2508.18886 (2025)
24. Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S.K.: Addressing fairness issues in deep learning-based medical image analysis: a systematic review. npj Digital Medicine **7**(1), 286 (2024)
25. Yang, Y., Liu, Y., Liu, X., Gulhane, A., Mastrodicasa, D., Wu, W., Wang, E.J., Sahani, D., Patel, S.: Demographic bias of expert-level vision-language foundation models in medical imaging. Science Advances **11**(13), eadq0305 (2025)

26. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical image analysis **91**, 102996 (2024)
27. Zhang, Y., Dunn, A.G., Naseem, U., Kim, J.: Intersectional fairness in vision-language models for medical image disease classification. arXiv preprint arXiv:2512.15249 (2025)
28. Zheng, G., Jacobs, M.A., Braverman, V., Parekh, V.S.: Towards fair medical ai: Adversarial debiasing of 3d ct foundation embeddings. arXiv preprint arXiv:2502.04386 (2025)