Fair-MoE: Medical Fairness-Oriented Mixture of Experts in Vision-Language Models

Peiran Wang[1,2] ⋆, Linjie Tong[1,2] ⋆, Jian Wu[1], Jiaxiang Liu[1] ✉, and Zuozhu Liu[1] ✉

[1] Zhejiang University
wujian2000@zju.edu.cn, {jiaxiang.21, zuozhuliu}@intl.zju.edu.cn
[2] University of Illinois at Urbana-Champaign
{peiranw3, linjiet2}@illinois.edu

Abstract. Fairness is an important principle in medical ethics. Vision Language Models (VLMs) have shown significant potential in the medical field due to their ability to leverage both visual and linguistic contexts, reducing the need for large datasets and enabling the performance of complex tasks. However, the exploration of fairness within VLM applications remains limited. Applying VLMs without a comprehensive analysis of fairness could lead to concerns about equal treatment opportunities and diminish public trust in medical deep learning models. To build trust in medical VLMs, we propose **Fair-MoE**, a model specifically designed to ensure both fairness and effectiveness. Fair-MoE comprises two key components: the **F**airness-**O**riented **M**ixture **o**f **E**xperts (FO-MoE) and the **F**airness-**O**riented **L**oss (FOL). FO-MoE is designed to leverage the expertise of various specialists to filter out biased patch embeddings and use an ensemble approach to extract more equitable information relevant to specific tasks. FOL is a novel fairness-oriented loss function that not only minimizes the distances between different attributes but also optimizes the differences in the dispersion of various attributes' distributions. Extended experiments show that Fair-MoE improves both fairness and accuracy across all four attributes. Code is made publicly available at https://github.com/LinjieT/Fair-MoE-Medical-Fairness-Oriented-Mixture-of-Experts-in-Vision-Language-Models.

Keywords: Fairness · Mixture of Experts · Vision-Language Models.

1 Introduction

Fairness is a key principle of medical ethics [5,24]. It requires that diagnostic systems avoid systematically disadvantaging specific groups based on inherent or acquired characteristics [22,20,21]. A key fairness concern in medical field is diagnostic disparity [2,33,29], where certain groups receive lower accuracy, leading to poorer health outcomes and exacerbating healthcare inequities. As

⋆ Equal contribution. ✉ Corresponding author.

Vision-Language Models (VLMs) become increasingly prevalent in medical image analysis [15,16,3,7,25,17,14,13], they offer a promising solution by integrating medical image-report pairs directly, bypassing the need for time-intensive manual annotation. VLMs process visual and textual data simultaneously, leveraging both to tackle complex tasks. However, despite their success in disease diagnosis, VLMs still inherit and propagate fairness issues, often embedding biases due to imbalances in training datasets or insufficient representation of specific demographic groups [20,6,12,28]. Biases inherent in both images and text make addressing fairness in VLMs particularly challenging [20]. Ensuring fairness in VLM-based diagnostics is therefore crucial to prevent unintended harm and promote equitable healthcare outcomes.

While fairness research in vision-only models has progressed, there is a critical gap in datasets for evaluating and mitigating bias in VLMs [10,11,19,31,32,34]. The recent release of the Harvard-FairVLMed dataset, the only fair-target VLM dataset, offers a unique opportunity to address these challenges in medical VLMs by providing image-text pairs on glaucoma, ground truths, and protected attributes [20]. While FairCLIP [20] has been proposed as a fairness benchmark using this dataset, it retains CLIP's original architecture and minimizes Sinkhorn distance [23], limiting its ability to effectively learn fair information. This gap necessitates a new VLM model that improves both accuracy and fairness, focusing on extracting relevant information while ignoring biased information. A Mixture of Experts (MoE)-based approach offers a promising solution [8], as it leverages multiple experts to collaboratively process domain-specific information, enhancing learning capacity and harnessing fairness to obtain bias-free features. Recent advances in MoE [26,18,1,35,9] have further enhanced learning capabilities. The superior learning ability of MoE demonstrates potential in facilitating fair and relevant information extraction while mitigating bias and filtering out irrelevant information. However, applying MoE to fair VLMs for medical diagnostics is underexplored, representing a significant opportunity for development [4,27].

To address the aforementioned issues and pursue more equitable VLMs, we propose the first MoE-based model for fair medical vision language model: the Fair Medical Vision Language Mixture of Experts (Fair-MoE). This model comprises two key modules: Fairness-Oriented Mixture of Expert (FO-MoE) and Fairness-Oriented Loss (FOL). FO-MoE is the first MoE designed for fair VLMs, utilizing expert capacity to filter bias in patch embeddings. This enhances the model's ability to extract fair, task-relevant information while minimizing bias. Unlike other fairness losses that focus solely on distance between protected attributes, FOL introduces a novel fair load balance loss function, considering both distance and dispersion between these attributes. In this way, it can not only guarantee fairness but also enhance the learning ability of the MoE. The main contributions can be summarized as follows:

- We propose Fair-MoE, a new framework combining FO-MoE and FOL to advance fairness and eliminate human biases in medical VLMs.

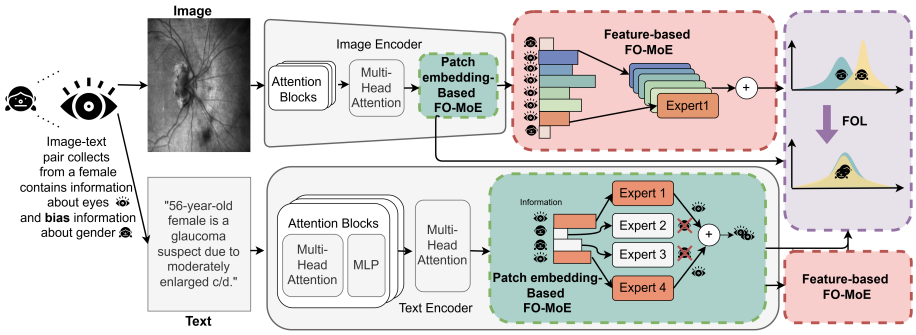

Fig. 1. An illustration of Fair-MoE. MoE-based architecture with FO-MoE enables model to extract fair information. Model is trained through contrastive learning with novel fair loss FOL added.

- To enhance task relevance while mitigating bias in extracted features, we propose FO-MoE module, the first fairness-oriented MoE module explicitly designed for medical VLM applications.
- To address biased attributes by considering their feature space distance and dispersion, we propose FOL, a novel fairness-aware loss.

2 Method

Fig. 1 demonstrates pipeline of Fair-MoE. Firstly, in the Image and Text Encoder, attention blocks are stacked to extract features from text and image. Multi-head attention computed in the last attention block is fed into FO-MoE consisting of Patch Embedding-based MoE and Feature-Based MoE to obtain bias-free text and image features. Finally, the similarity between bias-free text and image features and the loss function FOL are used to further optimize the distance and dispersion of bias attributes in the feature space.

2.1 Fairness-Oriented Mixture of Expert (FO-MoE)

To enhance learning ability and avoid learning biased content directly from images, we propose FO-MoE in the image and text encoder. We replace the MLP layer in the last attention block of each encoder with a **Patch Embedding-based MoE** layer to filter out biased patch embeddings and place a **Feature-based MoE** layer after the encoders to extract fair and relevant to specific tasks information. Each MoE layer comprises multiple experts, which are MLPs designed to capture and learn distinct aspects of information from the inputs. Additionally, the input passes through a gating mechanism, also implemented as a MLP, which assigns a weight to each expert. The weight indicates the likelihood that an expert should process the input, and output is aggregated by weighted summing outputs of experts.

Patch Embedding-based MoE: The input to the patch embedding-based MoE, denoted as $I^1 \in \mathbb{R}^{(N+1)\times D}$, consists of $N+1$ patch embeddings of dimension D, extracted from the input image or text. A gating mechanism, $G^1 : \mathbb{R}^{(N+1)\times D} \to \mathbb{R}^{(N+1)\times M^1}$, assigns a weight to each embedding for each of the M^1 experts. To enhance learning efficiency using the sparse MoE approach [26], and to ensure fair and task-relevant information extraction, each patch embedding is routed to the K_1 experts with the highest assigned weights. To mitigate bias, a capacity parameter C, defining the maximum number of embeddings an expert can process, is introduced. The weight matrix aggregating expert outputs is given by: $\hat{W}^1 = \text{Top}_c(\text{Top}_r(W^1, K_1), \frac{C(N+1)K_1}{M^1})$, where $W^1 = \text{softmax}(G^1(I^1))$, and $\text{Top}_c(\cdot, k)$ and $\text{Top}_r(\cdot, k)$ denote the operations that retain the k largest elements in each column and row, respectively. The workflow of Patch Embedding-based MoE is illustrated in Fig. 2.

Feature-based MoE: Feature vector $I_0^2 \in R^D$ will be sent to a Feature-based MoE with M^2 experts that further eliminates biased information to get the fair feature. The structure of Feature-based MoE is shown in Fig 2. The output $W^2 = Top_r(softmax(G^2(I_0^2)), k^2)$ that keeps highest k^2 weights from gates $G^2 : R^D \to R^{M^2}$ is used to aggregate outputs from experts to obtain a more fair visual feature vector $I^3 = \sum_{b=0}^{M^2-1} \hat{W}_b^2 E_b^2(I_0^2)$. $\hat{W}_b^2$ is a scalar indicates b th element in $\hat{W}_b^2$. $E_b^2(x)$ denotes bth experts in Feature-based MoE.

2.2 Fairness-Oriented Loss (FOL)

Optimizing the variance of weights to aggregate outputs from different experts enhances the learning capacity of the MoE by achieving load balance across the experts [18]. Furthermore, variance, as a measure of distribution dispersion, plays a critical role in fairness. By optimizing the dispersion differences between distributions of protected attribute groups, disparities in these distributions can be reduced. Building on this principle, we can improve existing fairness loss functions which focus on optimizing distance between distribution of protected attribute groups to decrease disparities among different distributions of protected attribute groups [20,30] by leveraging variance utilized to load balance loss to develop a fair loss (FOL), that takes both distance and dispersion into account.

In FOL, the weight output by a gate for a certain expert is selected as a random variable. Take Patch Embedding-based MoE in image encoder as an instance to compute FOL, all weights $\hat{W}^1$ computed from data sampled from the whole dataset and certain protected attribute group p are denoted as O_N, $O_{N|p}$, respectively. To optimize dispersion between different attribute's distribution, model should optimize difference of weight's variance of different attribute's distribution. Thus, loss for Patch Embedding-based MoE of image is $F_{EI} = \sum_{p\in P} \sum_{j=0}^{M^1-1} (Var(O_{N_j}) - Var(O_{N|p_j}))^2$ where O_{N_j}, $O_{N|p_j}$ denote jth column of O_N, $O_{N|p}$ which denote all expert j's weights. $Var(\cdot)_j$ means compute variance of input. P is a set of groups for certain attribute. In the same way, loss for Patch Embedding-based MoE of text F_{ET}, loss for Feature-based MoE of image F_{FI} and loss for Feature-based MoE of text F_{FT} can be gotten. Finally, FOL

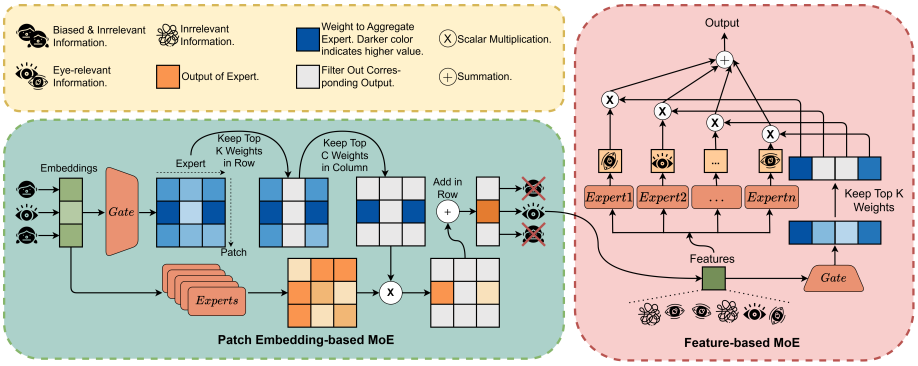

Fig. 2. The illustration of Patch Embedding-based MoE and Feature-based MoE's structure.

is defined as $FOL = F_{EI} + F_{ET} + F_{FI} + F_{FT} + L_{distance}$, where $L_{distance}$ is Sinkhorn distance Loss [23].

Table 1. Main Results of Fair-MoE. The green text highlights our method.

Attr.	Model	ES-AUC	AUC	DPD	EOD
Race	CLIP/b16	62.67±3.15	67.70±3.13	14.57±3.77	18.47±5.12
	CLIP/l14	66.83±2.19	70.63±2.98	11.69±3.85	15.13±2.66
	FairCLIP/b16	61.17±1.87	67.47±1.16	10.16±10.05	11.44±11.07
	FairCLIP/l14	67.53±4.26	71.57±2.94	16.01±5.87	17.03±3.74
	Fair-MoE/b16	69.63±1.21	71.93±0.90	7.25±5.13	7.43±3.04
	Fair-MoE/l14	**72.53**±1.07	**73.93**±0.97	**2.63**±0.65	**4.25**±0.75
Gender	CLIP/b16	63.30±2.73	67.70±3.13	2.79±1.49	7.52±4.78
	CLIP/l14	66.30±2.63	70.63±2.98	3.13±2.60	7.56±3.54
	FairCLIP/b16	64.43±1.86	68.47±2.26	2.50±1.47	4.98±3.74
	FairCLIP/l14	67.37±1.62	70.80±1.84	2.11±1.81	5.24±1.46
	Fair-MoE/b16	68.07±0.96	71.97±1.16	**1.91**±1.02	**3.53**±0.90
	Fair-MoE/l14	**69.97**±3.39	**74.97**±2.90	2.94±1.60	7.33±2.55
Ethnicity	CLIP/b16	64.87±2.26	70.63±0.90	7.53±2.96	14.83±3.01
	CLIP/l14	64.13±1.58	69.37±1.04	8.74±0.41	9.13±0.69
	FairCLIP/b16	61.43±1.05	67.33±1.33	10.54±1.52	17.93±4.01
	FairCLIP/l14	64.23±1.11	69.23±0.92	15.37±2.17	15.77±3.17
	Fair-MoE/b16	65.17±2.44	69.77±0.49	**8.52**±3.19	**8.42**±2.77
	Fair-MoE/l14	**67.10**±4.70	**72.80**±2.54	8.79±2.91	13.90±5.86
Language	CLIP/b16	60.10±3.84	67.70±3.13	13.50±3.96	16.40±9.56
	CLIP/l14	59.90±2.01	69.37±1.04	17.27±0.74	20.17±6.09
	FairCLIP/b16	57.97±0.65	68.07±0.57	10.96±4.04	14.25±9.09
	FairCLIP/l14	63.57±1.97	72.40±1.84	8.21±1.99	**11.00**±1.25
	Fair-MoE/b16	63.60±1.85	**73.87**±1.62	**7.48**±4.56	12.30±2.65
	Fair-MoE/l14	**63.80**±1.28	71.37±2.10	15.67±2.99	23.63±14.40

Table 2. Results of ablation study for FO-MoE.

Attr.	Model	ES-AUC	AUC	DPD	EOD
Race	FairCLIP/b16	61.17±1.87	67.47±1.16	10.16±10.05	11.44±11.07
	FairCLIP/l14	67.53±4.26	71.57±2.94	16.01±5.87	17.03±3.74
	FairCLIP/b16 w. FO-MoE	**69.97**±2.60	**72.67**±1.16	**3.19**±2.04	**9.48**±3.74
	FairCLIP/l14 w. FO-MoE	65.53±4.74	67.10±3.85	13.37±8.32	13.24±6.93
Gender	FairCLIP/b16	64.43±1.86	68.47±2.26	2.50±1.47	**4.98**±3.74
	FairCLIP/l14	67.37±1.62	70.80±1.84	**2.11**±1.81	5.24±1.46
	FairCLIP/b16 w. FO-MoE	**67.63**±1.82	**71.37**±2.30	3.14±1.14	7.70±1.66
	FairCLIP/l14 w. FO-MoE	64.33±1.83	68.67±1.87	4.25±3.08	7.22±3.17
Ethnicity	FairCLIP/b16	61.43±1.05	67.33±1.33	10.54±1.52	17.93±4.01
	FairCLIP/l14	64.23±1.11	69.23±0.92	15.37±2.17	15.77±3.17
	FairCLIP/b16 w. FO-MoE	**66.70**±3.30	69.17±2.57	**6.94**±4.40	**9.58**±4.11
	FairCLIP/l14 w. FO-MoE	66.57±3.94	**71.60**±2.71	11.72±2.35	15.97±0.78
Language	FairCLIP/b16	57.97±0.65	68.07±0.57	10.96±4.04	14.25±9.09
	FairCLIP/l14	**63.57**±1.97	72.40±1.84	**8.21**±1.99	11.00±1.25
	FairCLIP/b16 w. FO-MoE	62.47±2.53	**72.53**±1.09	11.97±3.87	23.20±2.12
	FairCLIP/l14 w. FO-MoE	62.33±0.68	65.17±0.90	10.43±0.42	**9.65**±3.05

3 Experiment

3.1 Experimental Setup

Experiments are conducted on the sole fairness-focused medical dataset, Harvard-FairVLMed [20], which comprises 7,000 training samples, 1,000 validation samples, and 2,000 test samples. Each sample in the database includes an SLO fundus image, accompanying clinical notes, labels of image-text pairs and protected attributes such as the patient's race, gender (GEN), ethnicity (ETH), and language (LAN). To fairly compared with the baselines, the training protocol was aligned with that of the FairCLIP [20]. All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU. The Area Under the Curve (AUC) is utilized to measure the model's overall performance. To assess fairness, the Demographic Parity Difference (DPD) and Equal Opportunity Difference (EOD) are used. Additionally, the Equity-Scaled AUC (ES-AUC) is introduced to evaluate the trade-off between performance and fairness.

3.2 Comparison with Baselines

To evaluate the performance and fairness of Fair-MoE in medical images, two State-of-the-Art (SoTA) fairness-aware VLMs, i.e., Vanilla and FairCLIP, are chosen as the baselines. Table 1 demonstrates the results of comparing Fair-MoE with CLIP and the SoTA fair medical vision language model Fair-CLIP. For ES-AUC, Fair-MoE outperforms all baselines in all protected attributes. For attribute race, Fair-MoE outperforms baselines 5.00% in ES-AUC. For AUC that measures effectiveness of model, Fair-MoE also outperforms all baselines in all protected attributes. For attribute gender, Fair-MoE achieves 4.91% improvement in AUC. For DPD and EOD that measure the fairness of model, results

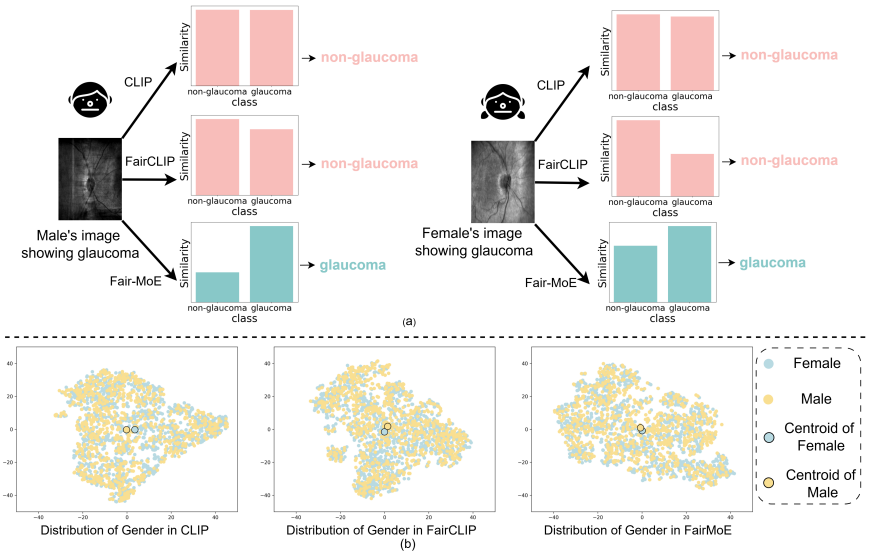

Fig. 3. Case study for Fair-MoE on effectiveness and fairness.

of DPD show that Fair-MoE achieves better fairness than baselines in all attributes. Besides, results of EOD show that Fair-MoE achieves better fairness than baselines in attributes of race, gender, and ethnicity.

The results prove that in addition to achieving a better trade-off between effectiveness and fairness, Fair-MoE can both improve effectiveness and fairness. The parameter counts for CLIP/B16, FairCLIP/B16, and Fair-MoE/B16 are approximately 200M, while those for CLIP/L14, FairCLIP/L14, and Fair-MoE/L14 are approximately 500M. These results demonstrate that Fair-MoE achieves improvements in both accuracy and fairness without a significant increase in model parameter count, maintaining computational efficiency while enhancing performance and fairness.

Case Study: Case study shown in Fig. 3 visualizes how Fair-MoE surpasses FairCLIP and CLIP in both effectiveness and fairness. Fig. 3 (a) illustrates that for glaucoma images from both male and female subjects, CLIP and FairCLIP misclassify them as non-glaucoma, whereas Fair-MoE correctly identifies them, demonstrating its effectiveness. Fig. 3 (b) shows that while the diagnostic feature distributions extracted by CLIP and FairCLIP differ significantly between genders, the distribution obtained by Fair-MoE is more similar across genders, highlighting its fairness.

3.3 Ablation Study

The ablation study for FO-MoE: To assess performance of FO-MoE, FO-MoE is added to fairCLIP. Table 2 demonstrates results of ablation study of FO-

Table 3. Results of ablation study for FOL. The green texts highlights our method.

Attr.	Model	ES-AUC	AUC	DPD	EOD
Race	Fair-MoE/b16 w/o FOL	69.97±2.60	72.67±1.16	3.19±2.04	9.48±3.74
	Fair-MoE/l14 w/o FOL	65.53±4.74	67.10±3.85	13.37±8.32	13.24±6.93
	Fair-MoE/b16	69.63±1.21	71.93±0.90	7.25±5.13	7.43±3.04
	Fair-MoE/l14	**72.53**±1.07	**73.93**±0.97	**2.63**±0.65	**4.25**±0.75
Gender	Fair-MoE/b16 w/o FOL	67.63±1.82	71.37±2.30	3.14±1.14	7.70±1.66
	Fair-MoE/l14 w/o FOL	64.33±1.83	68.67±1.87	4.25±3.08	7.22±3.17
	Fair-MoE/b16	68.07±0.96	71.97±1.16	**1.91**±1.02	**3.53**±0.90
	Fair-MoE/l14	**69.97**±3.39	**74.97**±2.90	2.94±1.60	7.33±2.55
Ethnicity	Fair-MoE/b16 w/o FOL	66.70±3.30	69.17±2.57	**6.94**±4.40	9.58±4.11
	Fair-MoE/l14 w/o FOL	66.57±3.94	71.60±2.71	11.72±2.35	15.97±0.78
	Fair-MoE/b16	65.17±2.44	69.77±0.49	8.52±3.19	**8.42**±2.77
	Fair-MoE/l14	**67.10**±4.70	**72.80**±2.54	8.79±2.91	13.9±5.86
Language	Fair-MoE/b16 w/o FOL	62.47±2.53	72.53±1.09	11.97±3.87	23.20±2.12
	Fair-MoE/l14 w/o FOL	62.33±0.68	65.17±0.90	10.43±0.42	**9.65**±3.05
	Fair-MoE/b16	63.60±1.85	**73.87**±1.62	**7.48**±4.56	12.30±2.65
	Fair-MoE/l14	**63.80**±1.28	71.37±2.10	15.67±2.99	23.63±14.4

Table 4. Results of ablation study on different components in Fair-MoE in ES-AUC.

(a) Ablation Study on F_{EI},F_{ET}, F_{FI}, F_{FT}

Model	Race	GEN	ETH	LAN	Model	Race	GEN	ETH	LAN
Fair-MoE/b16	**70.9**	**70.4**	**70.7**	**66.1**	Fair-MoE/l14	**74.0**	**69.5**	**73.4**	**64.1**
Fair-MoE/b16 w/o F_{EI}	62.2	65.4	58.7	60.0	Fair-MoE/l14 w/o F_{EI}	71.4	62.8	64.7	62.5
Fair-MoE/b16 w/o F_{ET}	62.4	62.0	64.4	58.3	Fair-MoE/l14 w/oF_{ET}	64.3	59.2	69.6	62.4
Fair-MoE/b16 w/o F_{FI}	70.4	56.5	61.9	61.7	Fair-MoE/l14 w/o F_{FI}	69.2	63.0	63.4	59.6
Fair-MoE/b16 w/o F_{FT}	60.9	69.8	62.2	48.7	Fair-MoE/l14 w/oF_{FT}	69.3	64.6	70.1	59.7

(b) Ablation study on Patch Embedding-based MoE (EM) and Feature-based MoE (FM)

Model	Race	GEN	ETH	LAN	Model	Race	GEN	ETH	LAN
Fair-MoE/b16	**70.9**	**70.4**	**70.7**	**66.1**	Fair-MoE/l14	**74.0**	**69.5**	**73.4**	**64.1**
Fair-MoE/b16 w/o EM	66.2	68.1	53.5	62.9	Fair-MoE/l14 w/o EM	68.6	66.9	62.0	62.2
Fair-MoE/b16 w/o FM	64.0	66.5	66.3	61.0	Fair-MoE/l14 w/o FM	72.2	65.4	72.1	60.8

(c) Ablation study on MoE modules in Text and Image

Model	Race	GEN	ETH	LAN	Model	Race	GEN	ETH	LAN
Fair-MoE/b16	**70.9**	**70.4**	**70.7**	**66.1**	Fair-MoE/l14	**74.0**	**69.5**	**73.4**	**64.1**
Fair-MoE/b16 w/o Text FO-MoE	66.8	67.2	61.3	63.6	Fair-MoE/l14 w/o Text FO-MoE	72.1	61.3	64.0	63.8
Fair-MoE/b16 w/o Image FO-MoE	69.4	66.8	64.6	54.8	Fair-MoE/l14 w/o Image FO-MoE	66.8	65.3	64.3	58.7

MoE. Utilizing FO-MoE achieves higher AUC for all attributes, demonstrating that FO-MoE can achieve higher learning capabilities and advanced effectiveness. For majority of attributes, applying FO-MoE can achieve higher ES-AUC, which indicates that adding FO-MoE achieves a better trade-off between effectiveness and fairness. For attribute race and ethnicity, applying FO-MoE can both improve effectiveness and fairness. This phenomenon proves FO-MoE's ability to filter out bias patch embedding and extract more fair task-relevant information.

The ablation study for FOL: To assess performance of FOL, we remove FOL from Fair-MoE, Table 3 shows how removing FOL from Fair-MoE will affect performance of Fair-MoE. In the case of removing FOL for all four attributes, metrics that measure effectiveness and fairness deteriorate significantly. Removing FOL leads to a drop of 2.56% in AUC for race and 2.34% in ES-AUC for gender. The drop in performance proves that just minimizing the distance be-

tween different attributes' distribution is not enough. Thus, optimizing difference between dispersion of attributes' distribution is indispensable to achieve a leap in both effectiveness and fairness. In addition, optimizing dispersion can improve stability of MoE, letting Fair-MoE better filter out bias patch embedding and utilize its supreme learning capacities to extract fair feature.

The ablation study for detail components: To assess components of FO-MoE and FOL, Table 4 (a) examines the impact of removing F_{EI}, F_{ET}, F_{FI}, and F_{FT}, revealing a degradation of trade-off between fairness and performance across all attributes without each FOL component. Table 4 (b) evaluates Patch Embedding-based MoE and Feature-based MoE, demonstrating that both contribute to a better balance between fairness and performance. Table 4 (c) investigates the application of FO-MoE to the image and text encoders, showing that integrating FO-MoE into both enhances the fairness-performance trade-off.

4 Conclusion

We propose Fair-MoE, a novel method designed to harness fairness in VLMs, enhancing both their effectiveness and fairness in medical diagnosis. Fair-MoE includes two key components: **FO-MoE** and **FOL**. **FO-MoE** is designed to learn unbiased features and filter out biased information. Meanwhile, **FOL** not only optimizes the distance between different protected attributes but also enhances the dispersion among them, guiding the model towards greater fairness and effectiveness. Extensive experiments demonstrate the superiority of Fair-MoE. Detailed ablation studies and visualization provide evidence of the effectiveness of each component within Fair-MoE.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 62476241), the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008), and the Zhejiang University-Angelalign Inc. R&D Center for Intelligent Healthcare.

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

References

1. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
2. Ferrante, R.L.M.E.R.: E addressing fairness in artificial intelligence for medical imaging nat. Commun **13**(1), 4581 (2022)
3. Gai, X., Zhou, C., Liu, J., Feng, Y., Wu, J., Liu, Z.: Medthink: Explaining medical visual question answering via multimodal decision-making rationale. arXiv preprint arXiv:2404.12372 (2024)
4. Germino, J., Moniz, N., Chawla, N.V.: Fairmoe: counterfactually-fair mixture of experts with levels of interpretability. Machine Learning pp. 1–21 (2024)

5. Giovanola, B., Tiribelli, S.: Beyond bias and discrimination: redefining the ai ethics principle of fairness in healthcare machine-learning algorithms. AI & society **38**(2), 549–563 (2023)
6. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in chest x-ray disease detection models. EBioMedicine **89** (2023)
7. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
8. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3**(1), 79–87 (1991). https://doi.org/10.1162/neco.1991.3.1.79
9. Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z.: Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. arXiv preprint arXiv:2404.10237 (2024)
10. Kelly, C., Hu, L., Hu, J., Tian, Y., Yang, D., Yang, B., Yang, C., Li, Z., Huang, Z., Zou, Y.: Visiongpt-3d: A generalized multimodal agent for enhanced 3d vision understanding. arXiv preprint arXiv:2403.09530 (2024)
11. Kelly, C., Hu, L., Yang, B., Tian, Y., Yang, D., Yang, C., Huang, Z., Li, Z., Hu, J., Zou, Y.: Visiongpt: Vision-language understanding agent using generalized multimodal framework. arXiv preprint arXiv:2403.09027 (2024)
12. Khan, M.O., Afzal, M.M., Mirza, S., Fang, Y.: How fair are medical imaging foundation models? In: Machine Learning for Health (ML4H). pp. 217–231. PMLR (2023)
13. Liu, J., Hu, T., Du, J., Zhang, R., Zhou, J.T., Liu, Z.: Kpl: Training-free medical knowledge mining of vision-language models. arXiv preprint arXiv:2501.11231 (2025)
14. Liu, J., Hu, T., Xiong, H., Du, J., Feng, Y., Wu, J., Zhou, J., Liu, Z.: Vpl: Visual proxy learning framework for zero-shot medical image diagnosis. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 9978–9992 (2024)
15. Liu, J., Hu, T., Zhang, Y., Feng, Y., Hao, J., Lv, J., Liu, Z.: Parameter-efficient transfer learning for medical visual question answering. IEEE Transactions on Emerging Topics in Computational Intelligence (2023)
16. Liu, J., Hu, T., Zhang, Y., Gai, X., Feng, Y., Liu, Z.: A chatgpt aided explainable framework for zero-shot medical image diagnosis. arXiv preprint arXiv:2307.01981 (2023)
17. Liu, J., Wang, Y., Du, J., Zhou, J., Liu, Z.: Medcot: Medical chain of thought via hierarchical expert. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 17371–17389 (2024)
18. Lou, Y., Xue, F., Zheng, Z., You, Y.: Cross-token modeling with conditional computation. arXiv preprint arXiv:2109.02008 (2021)
19. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics **23**(6), bbac409 (2022)
20. 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)
21. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)

22. Parraga, O., More, M.D., Oliveira, C.M., Gavenski, N.S., Kupssinskü, L.S., Medronha, A., Moura, L.V., Simões, G.S., Barros, R.C.: Fairness in deep learning: A survey on vision and language research. ACM Computing Surveys (2023)
23. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
24. Pratt, B., Wild, V., Barasa, E., Kamuya, D., Gilson, L., Hendl, T., Molyneux, S.: Justice: a key consideration in health policy and systems research ethics. BMJ Global Health **5**(4), e001942 (2020)
25. Qin, Z., Yi, H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. arXiv preprint arXiv:2209.15517 (2022)
26. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems **34**, 8583–8595 (2021)
27. Sharma, S., Henderson, J., Ghosh, J.: Feamoe: fair, explainable and adaptive mixture of experts. arXiv preprint arXiv:2210.04995 (2022)
28. Sikstrom, L., Maslej, M.M., Hui, K., Findlay, Z., Buchman, D.Z., Hill, S.L.: Conceptualising fairness: three pillars for medical algorithms and health equity. BMJ health & care informatics **29**(1) (2022)
29. Stanley, E.A., Wilms, M., Mouches, P., Forkert, N.D.: Fairness-related performance and explainability effects in deep learning models for brain image analysis. Journal of Medical Imaging **9**(6), 061102–061102 (2022)
30. Tian, B., Du, R., Shen, Y.: Fairvit: Fair vision transformer via adaptive masking. arXiv preprint arXiv:2407.14799 (2024)
31. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
32. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.H., Kiraly, A., Kazemzadeh, S., Melamed, Z., et al.: Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. arXiv preprint arXiv:2308.01317 (2023)
33. 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)
34. Yang, D., Hu, L., Tian, Y., Li, Z., Kelly, C., Yang, B., Yang, C., Zou, Y.: Worldgpt: a sora-inspired video ai agent as rich world models from text and image inputs. arXiv preprint arXiv:2403.07944 (2024)
35. Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., Fedus, W.: St-moe: Designing stable and transferable sparse expert models. arXiv preprint arXiv:2202.08906 (2022)