FairSMoE: Mitigating Multi-Attribute Fairness Problem with Sparse Mixture-of-Experts

Changdi Yang[1] , **Zheng Zhan**[1,2*†] , **Ci Zhang**[3] , **Yifan Gong**[1] , **Yize Li**[1] , **Zichong Meng**[1] , **Jun Liu**[1] , **Xuan Shen**[1] , **Hao Tang**[4] , **Geng Yuan**[3] , **Pu Zhao**[1‡] , **Xue Lin**[1] and **Yanzhi Wang**[1‡]
[1]Northeastern University
[2]Microsoft Research
[3]University of Georgia
[4]Peking University
{yang.changd, zhan.zhe, gong.yifa, li.yize, meng.zic, liu.jun2, shen.xu, p.zhao, xue.lin, yanz.wang}@northeastern.edu, {cz06540, geng.yuan}@uga.edu, haotang@pku.edu.cn

Abstract

Real-world datasets usually contain multiple attributes, making it essential to ensure fairness across all of them simultaneously. However, different attributes may vary in difficulty, and no existing approaches have effectively addressed this issue. Consequently, an attribute-adaptive strategy is needed to achieve fairness for all attributes. Multi-task Learning (MTL) leverages shared information to optimize multiple tasks concurrently, while Sparsely-Gated Mixture-of-Experts (SMoE) can dynamically allocate computational resources to the most needed tasks. In this work, we formulate multi-attribute fairness issue as an MTL problem and employ SMoE to achieve desirable performance across all attributes simultaneously.

We first analyze the feasibility and find the potentiality by formalizing multi-attribute fairness problem into a MTL problem and mitigating it by using SMoE. However, vanilla SMoE could lead to overutilization problem which causes sub-optimal performance. We then proposed an innovative SMoE framework for multi-attribute fair image classification, which further improves multi-attribute fairness by redesigning the MoE layer and routing policy with fairness consideration. Extensive experiments demonstrated the effectiveness. Taking a DeiT-Small as the backbone, we achieve 77.25% and 86.01% accuracy on the ISIC2019 and CelebA dataset respectively with Multi-attribute Predictive Quality Disparity (PQD) score of 0.801 and 0.787, beating current state-of-the-art methods Muffin, InfoFair and MultiFair.

1 Introduction

As AI democratization advances, machine learning (ML) has been increasingly utilized in a variety of applications, including image or video generation [Wu *et al.*, 2022b; Zhan *et al.*, 2021; Zhan *et al.*, 2024a; Shen *et al.*, 2025b; Li *et al.*, 2025; Shen *et al.*, 2025c], autonomous driving [Li *et al.*, 2023a; Yang *et al.*, 2023c; Zhang *et al.*, 2022; Li *et al.*, 2022; Shen *et al.*, 2025a], and language translation [Zhao *et al.*, 2024; Zhan *et al.*, 2024b; Shen *et al.*, 2024; Shen *et al.*, 2025b; Shen *et al.*, 2025d]. Fairness has emerged as a significant and fundamental concern in these applications. Studies have found unfair ML models exhibiting worse performance toward sensitive attributes, such as race [Nanda *et al.*, 2021; Puyol-Antón *et al.*, 2022], gender [Puyol-Antón *et al.*, 2022] and skin tone [Yang *et al.*, 2023b; Yang *et al.*, 2023a], leading to discrimination and undermining the trustworthiness of ML [Li *et al.*, 2023b; Li *et al.*, 2024] from the public.

Many research efforts are devoted to improving fairness in ML, which include two common categories: (1) Mitigating unbalanced data by generation synthetic data or adopting data augmentation techniques [Sattigeri *et al.*, 2019; Xu *et al.*, 2018]; (2) Revisiting training procedure by utilizing adversarial training [Karkkainen and Joo, 2021; Wang and Deng, 2020], discriminate training [Tao *et al.*, 2022] or training with fair objectives [Karkkainen and Joo, 2021]. Although these methods are effective in improving single-attribute fairness, in practice, an individual may have multiple sensitive attributes, and models optimized for just one attribute can still make unfair predictions.

Recently, a few studies have started to investigate multi-attribute fairness optimization. Data pre-processing [Tian *et al.*, 2024] is considered to ensure statistical parity among multiple sensitive attributes, while other works [Hwang *et al.*, 2020; Deng *et al.*, 2023] extend single-attribute fairness optimization techniques to multi-attribute protections by introducing additional constraints or prediction heads for multiple attributes. However, directly adapting existing methods for multi-attribute protection has several limitations. Firstly, as the number of attributes increases, the computational cost rises accordingly. Consequently, current approaches only consider a small number of sensitive attributes. Secondly, extending these methods may inherit or aggravate their drawbacks. For example, adversarial approaches require elaborate tuning to guarantee training convergence due to the inverse gradient updating [Wang and Deng, 2020].

To tackle these challenges, one solution is to formulate the multi-attribute fairness problem as a multi-task learning (MTL) problem. Multi-attribute fairness naturally decomposes into multiple "tasks," each linked to a sensitive attribute or intersection of attributes. By adopting a MTL framework, the model can simultaneously learn shared representations across attributes while preserving each attribute's unique features. To increase model capacity while maintaining similar computational cost, the Sparse Mixture of Experts (SMoE) is particularly suited because it can dynamically allocate specialized experts to each task [Chen *et al.*, 2023; Chen *et al.*, 2022]. This methodology optimizes fairness for all tasks, improves computational efficiency, and avoids the instability of other architectural approaches.

However, several challenges must be addressed to build an effective MTL framework for multi-attribute fairness. First, conventional MTL setups use a fixed number of experts per task, which can lead to suboptimal performance due to the varying difficulty of optimizing different sensitive attributes. Second, without proper regularization, the router may develop deterministic patterns when dealing with multi-attribute fairness problem. This leads to over-utilization of certain experts, hindering the capacity and reducing the efficiency of SMoE. Moreover, from the perspective of training dynamics, when certain experts are rarely used, the gradients flowing back to update their parameters are somewhat sparse, meaning that these experts are updated very slowly or not at all, which destabilizes the entire training process. Thus, appropriate regularization is required to balance expert utilization and maintain fairness across multiple attributes.

Our contributions are summarized below:

- We target multi-attribute fairness optimization problem and first formalize the problem as a MTL problem. We further analyze the feasibility and potentiality of using SMoE in this problem and tackle the challenge of unbalanced utilization of experts and task difficulties.
- We propose an innovative SMoE framework for multi-attribute fair image classification to improve fairness by redesigning the SMoE layer with fairness considerations. We regulate experts with fairness constraints and dynamically allocate expert numbers for each task.
- Extensive experiments demonstrated our effectiveness. Taking a DeiT-Small as the backbone, FairSMoE achieves 77.25% and 86.01% accuracy on the ISIC2019 and CelebA dataset respectively with Multi-attribute Predictive Quality Disparity (PQD) score of 0.801 and 0.787, beating current state-of-the-art methods such as Muffin and MultiFair. FairSMoE alleviates unbalanced routing and gradient conflict issue.

2 Related Works

Single-Attribute Fairness. Distribution-based methods aim to better represent minority groups or eliminate undesired biases in datasets. For instance, [Derman, 2021; Stafanovičs *et al.*, 2020] propose algorithms that adjust objects in datasets based on predefined rules, while [Yan *et al.*, 2020] discusses sampling techniques to address under-representation of protected groups. However, undersampling strategies are impractical for DNNs as they reduce dataset size making training infeasible.

One-step training methods incorporate fairness into the main training procedure. The works [Gaci *et al.*, 2022; Xu *et al.*, 2019] utilize adversarial frameworks to train models avoiding undesired biases. These methods often require annotations of protected variables, which can be limiting. Additionally, optimization methods have been proposed to enhance fairness during training [Du *et al.*, 2023; Wu *et al.*, 2022a] balancing fairness and accuracy.

Multi-Attribute Fairness. Recent methods addressing multi-attribute fairness primarily focus on data preprocessing and augmentation [Deng *et al.*, 2023; Tian *et al.*, 2024]. For example, [Sheng *et al.*, 2023] proposed a Neural Architecture Search framework to automatically search for fair combinations in multi-attribute models. However, this can cause attribute turbulence among model candidates, leading to performance degradation. [Tian *et al.*, 2024] focus on data augmentation by introducing mix-up procedures to generate synthetic data.

Fairness in Multi-Task Learning. As MTL becomes increasingly prevalent in SOTA models [Ruder, 2017; Zhang and Yang, 2021], understanding the interaction between fairness and MTL is essential. [D'Amour *et al.*, 2020] investigates fairness in multi-task regression models using rank-based non-parametric independence tests. [Zhao and Chen, 2020] proposes MTL enhanced with fairness constraints to jointly learn classifiers leveraging information across sensitive groups.

Mixture of Experts. The initial concept of MoE [Jacobs *et al.*, 1991] involves dividing input space into regions and training specialized experts for each region, with a gating network selecting appropriate experts. Recent advancements leverage SMoE [Jiang *et al.*, 2024; Du *et al.*, 2022] to handle increasing complexity of modern datasets. SMoE [Shazeer *et al.*, 2017] reduces computational overhead by dynamically routing inputs to expert subsets. GShard [Lepikhin *et al.*, 2020] extends MoE models to multilingual settings, scaling to handle over 100 languages simultaneously.

3 Discovery and Analysis

Existing works have shown model fusion success on multi-attribute fairness optimization [Sheng *et al.*, 2023]. However, SMoE under MTL scenarios has crucial factors—routing policy and number of experts per task—that differentiate it from model fusion methods. To investigate optimal SMoE utilization, we address the following questions.

3.1 Do Different Attributes Have the Same Difficulty To Optimize?

We investigated how various sensitive attributes perform under identical model configurations using the ISIC2019 dataset for dermatology disease classification, analyzing three sensitive attributes: age group, gender, and disease site.

Experimental Setting. DeiT-Small [Touvron *et al.*, 2021] model was employed with the last layer replaced by a sparse SMoE layer. Number of experts was set to 4 with auxiliary

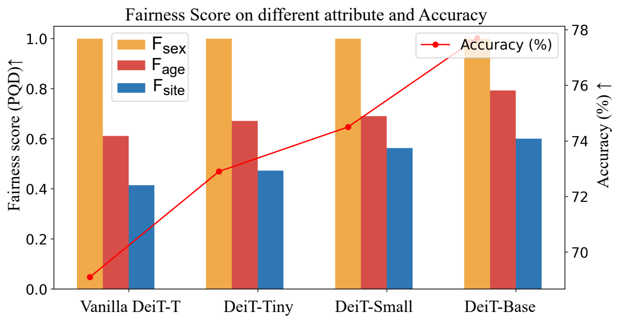

Figure 1: **Results of fairness score on different attributes.** We show that regardless of model type, the different sensitive attribute has variant difficulty to optimize.

loss [Zoph *et al.*, 2022]. Fairness was quantified using Predictive Quality Disparity (PQD) score.

Analysis Results. The results, illustrated in Figure 1, reveal significant disparities in fairness optimization across these attributes. Gender attribute showed high fairness score variance, suggesting minor disparity and balanced outcomes. Conversely, disease site and age group demonstrated low PQD as shown in Figure 2, indicating greater classification outcome disparity and fairness optimization challenges.

Observation. Not all attributes are equally challenging to optimize for fairness. Attributes with higher variability and less balanced representation (disease site and age group) are more difficult to optimize than gender, highlighting the need for attribute-specific strategies.

3.2 How Does the Default Routing Policy Behave When Dealt With Multi-Attribute?

Building on observations that different attributes present varying optimization difficulties, we investigated how the default routing policy manages inputs related to both easier and more challenging attributes, and how altering expert numbers and SMoE layer placement affects these dynamics.

Experimental Setting. We explored transformer-based architectures with SMoE layers in various configurations, increasing expert numbers and introducing SMoE layers earlier in the network to observe routing policy distribution across experts.

Analysis Results and Observation. Our findings reveal that, without tailored adjustments, the default routing policy routes similar inputs—regardless of attribute difficulty—to the same experts. This pattern intensifies with increased experts or earlier SMoE layer integration, leading to significant expert overutilization. These results, illustrated in Figure 2, underscore the challenges in maintaining balanced distribution of inputs across experts.

Conclusion and Takeaway. This investigation highlights the critical need for adaptive routing mechanisms within SMoE models for multiple sensitive attributes. Current default routing policies may not sufficiently accommodate input diversity, potentially leading to biased outputs and decreased effectiveness.

4 Problem Formulation

Fairness Metrics. (i) Here we present our definition of fairness used for evaluation. We utilize two mainstream metrics for fairness evaluation: Predictive Quality Disparity (PQD) and Demographic Parity (DP). Details on the definitions of these metrics are available in the Appendix. (ii) *Multi-attribute Fairness* (MF): it measures the overall fairness on multiple attributes. Let $S = \{s_1, s_2, \ldots, s_m\}$ be the set of sensitive attributes in a dataset. The multi-attribute fairness score, MF_Ψ, under the fairness metric Ψ, is defined as:

$$MF_\Psi = \frac{1}{m} \sum_{i=1}^{m} \Psi(s_i) \tag{1}$$

where $\Psi(s_i)$ represents the fairness score for the sensitive attribute s_i, and m is the total number of sensitive attributes.

Problem Formulation. Addressing the multi-attribute fairness problem requires a holistic approach considering the interdependence among all attributes and their impact on the fairness of the model. To this end, we formulate the multi-attribute fairness optimization problem as a multi-task learning (MTL) problem. This approach enables the simultaneous optimization of fairness across various sensitive attributes, alongside the primary task of prediction performance.

In our MTL framework, the tasks are defined as follows:

(i) **Primary Task (Predictive Performance)**: The primary objective is to maximize the overall accuracy of predictions across all groups, defined by:

$$\mathcal{L}_{\text{performance}} = -\sum_{i=1}^{m} \log p(y_i | x_i, \theta) \tag{2}$$

where x_i and y_i represent the features and label of the i-th data point, and θ denotes the parameters of the model.

(ii) **Fairness Tasks**: Each fairness task aims to optimize the fairness score in predictive outcomes related to each sensitive attribute $s_i \in \{s_1, s_2, \ldots, s_m\}$. For each fairness metric Ψ such as PQD and DP, we define a corresponding loss function that penalizes the deviation from fairness:

$$\mathcal{L}_\Psi = \sum_{i=1}^{m} \Psi(s_i) \tag{3}$$

(iii) **Overall Objective**: The overall training objective combines the predictive performance with the fairness tasks, balanced by a set of tunable parameters λ, which regulate the trade-off between accuracy and fairness:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{performance}} + \lambda \mathcal{L}_\Psi \tag{4}$$

By integrating these objectives into a single MTL framework, our approach allows for explicit control over the trade-offs between achieving high predictive performance and ensuring fairness across multiple sensitive attributes.

5 Methodology

5.1 Revisiting Sparse Mixture of Experts

SMoE is proposed to scale up the model capacity while maintaining low per-inference costs. In this work, we insert SMoE layers into a transformer block. The SMoE block consist of n experts $\{E_1, ..., E_n\}$, each of which is a feedforward neural network similar to those in the vision transformer block.

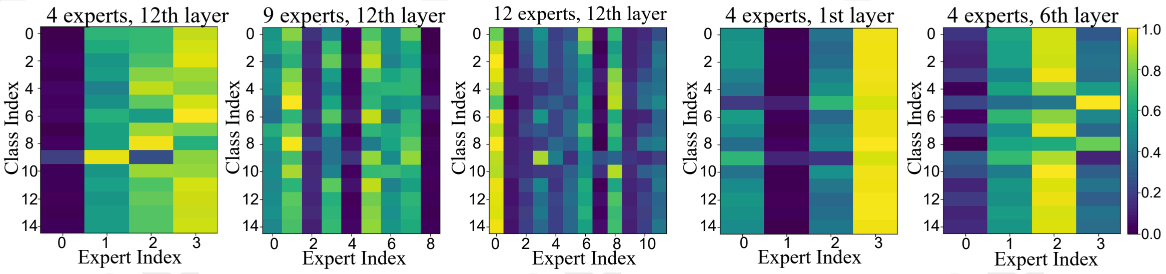

Figure 2: **Distribution of expert activation frequencies** on DeiT-Small with vanilla SMoE layer with different settings. For each heatmap, each row resembles a class while each column indicates different expert. Class from index 0 to 4, 5 to 12 and 13 to 14 belong to three different sensitive attributes (age, site and gener) respectively. Results shows when number of expert increase, utilization rate will decrease and it tends to develop a deterministic pattern for all attributes. We can also observe that the utilization rate for experts will increase when SMoE is put into later layers.

Giving an input embedding x, it is fed into a router network $\mathcal{G}$ and assigned to the most relevant experts. The architecture of the router network is usually one or a few layers of multilayer perceptrons (MLPs). The gating mechanism is defined as following:

$$\mathcal{R}(\boldsymbol{x}) = \texttt{Top-K}(\texttt{softmax}(g(\boldsymbol{x}))) \tag{5}$$

where $g(\cdot)$ are trainable gating networks and $\texttt{Top-K}$ select the largest K values. The final output of an SMoE block is a summarization of features from the activated experts and can be depicted as below:

$$\boldsymbol{y} = \sum_{i=1}^{k} \mathcal{R}(\boldsymbol{x})_i \cdot E_i(\boldsymbol{x}) \tag{6}$$

where $E_i(\boldsymbol{x})$ stands for the feature representations produced from the expert E_i, which is weighted by $\mathcal{R}(\boldsymbol{x})$ to form the final output y.

5.2 Fairness-Guided Routing (FGR)

The proposed FairSMoE consists of the MoE layer redesigned with fairness consideration. In the training framework, we regulate expert with fairness constraints.

We begin by investigating the routing behavior of a vanilla Sparse Mixture of Experts (SMoE) model trained with a top-2 routing policy using the ISIC2019 dermatology image classification dataset, which includes three sensitive attributes and 16 different classes in total: sex (2), age (5), and general anatomic site (9). Discoveries in section 3 indicate that (i) the routing choices across different attributes are highly similar, and (ii) despite an increase in the number of experts, only a few experts are heavily utilized, leading to a decreased overall utilization rate. These findings suggest that the standard routing paradigm in SMoE is inefficient for multiple sensitive attributes scenarios, potentially leading to under-utilization of model capacity.

To address this inefficiency, we develop a fairness-guided routing policy utilizing disentangled representation learning to mitigate the influence of sensitive attributes on routing decisions. We replace the traditional MLP gating network with a novel structure, as in Figure 3, consisting of the original MLP for gating f_g as a gating branch, a feature extractor network ϕ and an additional sensitive attribute (SA) branch, which consists of a number of m classifiers $\{f_{s_1}, f_{s_2}, ..., f_{s_m}\}$ to predict the attribute class. The input x will first pass the feature extractor ϕ to get a representation $z = \phi(x)$. Then the z will be fed into both the gating branch and SA branch. In the Gating branch, we will get original gating information $e = f_g(z)$. In the SA branch, z will be copied m times and fed into $\{f_{s_1}, f_{s_2}, ..., f_{s_m}\}$ and get class predictions of each attributes $\{p_1, p_2, ..., p_m\}$. After that, we incorporate a confusion loss to further enhance this disentanglement:

$$\mathcal{L}_{\text{confusion}} = -\frac{1}{m}\sum_{i=1}^{m} \log\left(p_i\right) \tag{7}$$

This loss encourages the feature extractor to generate features that are indistinguishable with respect to sensitive attributes, fostering a state of maximum confusion. Output $\{f_{s_1}, f_{s_2}, ..., f_{s_m}\}$ are then fed into a linear projection layer to get the expert choice vector $\{e_1, e_2, ..., e_m\}$ and will add with e. We select the vector from the classifier corresponding to the specific attribute class and then apply a softmax to get the final expert probability distribution.

By fusing losses into the training process of the SMoE model, we ensure that the SA branch of the gating network directs routing decisions based on task relevance and fairness considerations, rather than on detectable sensitive attributes. This approach not only minimizes the influence of protected attributes on routing decisions but also optimizes the utilization of experts by mitigating conventional biases and routing patterns observed in vanilla SMoE models.

This dual-branch structure allows our SMoE model to handle multiple tasks and sensitive attributes more effectively. It aligns with the overarching goals of fairness and improves the model's capacity utilization by diversifying expert deployment across various tasks and contexts.

5.3 Fairness-driven Expert Management (FEM)

Our approach incorporates a comprehensive strategy that includes both dynamic expert allocation and fairness-dependent constraints across experts. These mechanisms are designed to enhance the network's performance and fairness dynamically, adapting expert utilization based on the evolving requirements of the task and fairness objectives.

Attribute-Focused Expert Specialization. In our FairSMoE model, we strategically assign each expert to

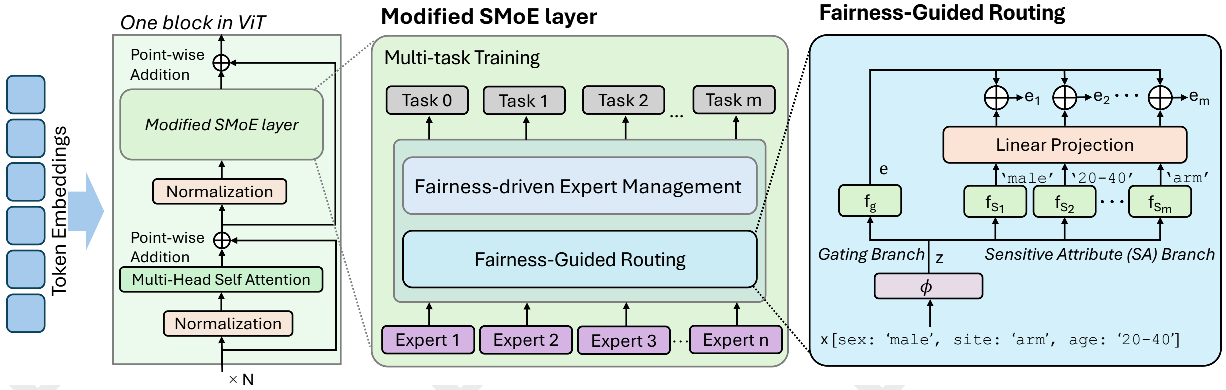

Figure 3: Framework overview. ***(left)*** Overall architecture of FairSMoE. Our proposed method replace the FFN by modified SMoE layer in the last ViT block. ***(middle)*** We re-design the SMoE layer by adopting Fairness-Guided Routing (FGR) mechanism and Fairness-driven Expert Management to enforces attribute-focused constraints and dynamically allocate experts for tasks according to fairness performance. We train the model in a multi-task learning way with 1 primary task and m fairness tasks. ***(right)*** FGR consists of two branches, one is *Gating Branch*, the vanilla gating network. Another is *Sensitive Attribute (SA) Branch*, encoding fairness information into routing consideration.

focus on unique sets of sensitive attributes to foster diverse expertise. Experts are initially assigned attributes based on potential and adjusted dynamically to minimize knowledge overlap and maximize fairness. The specialization is enforced through a tailored loss function:

$$\begin{aligned}\mathcal{L}_{\text{specialization},e} &= \alpha \cdot \sum_{a \in S_e} \text{Loss}_{\text{focus}}(e, a) \\ &\quad - (1-\alpha) \cdot \sum_{a \notin S_e} \text{Loss}_{\text{avoid}}(e, a)\end{aligned} \tag{8}$$

Here, S_e is the set of attributes assigned to expert e, and α controls the balance between focusing on assigned attributes and avoiding non-assigned ones. $\text{Loss}_{\text{focus}}(e, a)$ and $\text{Loss}_{\text{avoid}}(e, a)$ could be a standard loss function such as cross-entropy loss or mean squared error, depending on the task. In this task they are both standard cross-entropy loss.

Periodic performance reviews guide dynamic reassignments, ensuring experts develop deep, relevant expertise without redundancy. The routing logic is also adapted to align with these specializations, optimizing decision-making and enhancing model adaptability and fairness. This integrated approach ensures that experts not only excel in their designated domains but also contribute effectively to the model's overall accuracy and fairness.

Fairness-aware Expert Allocation. Recognizing that different sensitive attributes may require varying levels of optimization complexity, we have implemented a dynamic expert allocation mechanism. This data-driven approach dynamically adjusts the number of allocated experts based on real-time assessments of fairness and performance. Initially, we evaluate the single-attribute fairness score (SF_i) for each task on the validation set. If SF_i remains stable or improves over n iterations, we consider increasing the number of experts assigned to that attribute. Conversely, if adding more experts results in a worse validation loss than previously observed, it indicates potential overfitting or interference among experts, suggesting a reduction in the number of allocated experts.

This method ensures that each expert's capacity is fully utilized, avoiding underutilization or overload, and aligns expert deployment with the fairness needs of each attribute.

Implementation of Expert Management. Our method is outlined in Appendix as Algorithm 2. The combined fairness-dependent constraints and dynamic expert allocation forms the core of our fairness-guided expert management module. This module supports the model in achieving high accuracy while ensuring multi-attribute fairness performance.

Loss Function The overall loss function for the FairSMoE framework is formulated to balance performance, fairness, and expert specialization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{performance}} + \lambda \mathcal{L}_{\Psi} + \sum_{e} \mathcal{L}_{\text{specialization, e}} + \mathcal{L}_{\text{confusion}}$$

Algorithm 1 Fairness-Guided Routing (FGR)

```
Require: Input token $x$, feature extractor $\phi$, sensitive attribute classifiers $\{f_{s_1}, f_{s_2}, ..., f_{s_m}\}$, main task gating network $f_g$, linear projection layer $l$
Ensure: Fair routing decision to an appropriate expert
    $z \leftarrow \phi(x)$   # Extract features from input
    $p_i \leftarrow f_{s_i}(z)$ for $s_i \in S$   # Predict sensitive attributes
    Compute $\mathcal{L}_{\text{confusion}} = -\frac{1}{n}\sum_{i=1}^{n} \log(p_i)$
    $e \leftarrow f_g(z), e_i \leftarrow f_{s_i}(z)$ for $s_i \in S$
    Combine $e$, $e_i$ to update probabilistic distribution $e$   # Incorporate fairness and original gating scores into routing decision
    return Gating score $e$, $\mathcal{L}_{\text{confusion}}$
```

6 Experiments

6.1 Implementation Details

Dataset and Network Backbones. We evaluate our methods on the ISIC 2019 and CelebA datasets, primarily for

Backbone	Method	ISIC2019			CelebA			# Params
		Acc.↑	PQD↑	DP↓	Acc.↑	PQD↑	DP↓	(M)
DeiT-Small	vanilla SMoE	74.49	0.773	3.12	83.26	0.752	3.05	23
	Muffin	72.09	0.778	3.18	82.14	0.763	3.14	45
	MultiFair	75.78	0.794	2.97	84.99	0.774	2.92	22
	FairSMoE	**77.25**	**0.801**	**2.72**	**86.01**	**0.787**	**2.61**	**23**
DeiT-Base	vanilla SMoE	75.77	0.762	3.21	83.90	0.743	3.13	87
	Muffin	73.85	0.774	3.29	82.70	0.753	3.22	173
	MultiFair	76.56	0.784	3.14	85.45	0.764	3.05	86
	FairSMoE	**78.37**	**0.852**	**2.43**	**86.70**	**0.803**	**2.34**	**87**
Swin-Small	vanilla SMoE	76.10	0.784	3.24	84.50	0.763	3.15	51
	Muffin	74.05	0.792	3.31	83.10	0.771	3.24	101
	MultiFair	76.80	0.799	3.12	85.90	0.781	3.05	50
	FairSMoE	**77.84**	**0.814**	**2.73**	**86.30**	**0.793**	**2.64**	**51**
Swin-Base	vanilla SMoE	76.50	0.787	3.23	85.00	0.774	3.14	89
	Muffin	74.30	0.804	3.28	83.50	0.781	3.21	177
	MultiFair	77.10	0.813	3.13	86.20	0.792	3.04	88
	FairSMoE	**78.95**	**0.828**	**2.23**	**87.00**	**0.809**	**2.14**	**89**

Table 1: Results of FairSMoE on ISIC2019 and CelebA dataset using 4 experts. We demonstrate accuracy (Acc. in %) and multi-attribute fairness performance under two different fairness metrics, MF_{PQD} (PQD) and $\text{MF}_{DP}/\times 10^{-3}$ (DP). We select `sex, age, site` and `chubby, goatee, gender` as sensitive attributes for ISIC2019 and CelebA dataset. # indicates activated parameters.

skin lesion analysis and facial attribute recognition tasks, respectively, with more details in Appendix. To demonstrate our generalization, we choose DeiT-Small, DeiT-Base, Swin-Small, and Swin-Base for backbones. As discussed in Discovery and analysis, we follow the observation that applying the SMoE layer in the last transformer block will obtain the best performance.

Baselines. To demonstrate the effectiveness of FairSMoE, we consider three groups of baselines for comparison: (1) transformers with vanilla SMoE layers, (2) Muffin [Sheng *et al.*, 2023] with model fusion, and (3) MultiFair [Tian *et al.*, 2024] with data augmentation.

Training and Evaluation Settings. We applied a batch size of 256 and data augmentation of RandomResizedCrop for all methods on both datasets. Transformers are optimized with AdamW with weight decay of 1×10^{-4}, initial learning rate (LR) of 5×10^{-4}. Training epoch is set to 300 for ISIC2019 and 500 for CelebA. We randomly separate ISIC2019 80:20 for training and test, and randomly select 5% of training set for validation. We set α as 0.6 in Equation (8) and λ as 0.1 in $\mathcal{L}_{\text{total}}$. To evaluate the fairness performance, We use multi-attribute fairness MF_{Ψ} as metric, which is defined in Equation (1). 4 Nvidia A100s are used for training and testing. We are using overall loss adding eq. (4), Equation (7) and Equation (8) together.

6.2 Experiment Results

Comparison with vanilla SMoE and other approaches. We selected two representative vision transformer models and their four variants, DeiT-Small/Base and Swin-Small/Base, to test the effectiveness of FairSMoE. ”Vanilla SMoE” refers to the model where we replace the last layer of each backbone with a vanilla SMoE layer, setting the number of experts to 4, and excluding any auxiliary loss or noisy gating. For Muffin's configuration, we used two backbones without SMoE, following the default fusing method. For MultiFair, we used a single backbone without SMoE, implementing only the data augmentation method proposed in MultiFair. The results are shown in Table 1 and several observations can be drawn: (i) *Performance and Fairness Advancements*: FairSMoE significantly outperforms existing methods like vanilla SMoE, Muffin, and MultiFair in terms of accuracy and fairness. For example, on ISIC2019, FairSMoE enhances accuracy up to 78.95% and improves the PQD metric to 0.828 on Swin-Base, demonstrating its effectiveness in balancing performance with fairness; (ii) *Efficiency in Parameter Utilization*: Despite its enhanced capabilities, FairSMoE maintains efficiency, using no more parameters than the least complex models. Compared with only $25M$ parameters for DeiT-Small, it outperforms models with up to $173M$ parameters, showcasing its ability to achieve optimal performances with a lightweight network. (iii) *Empirical Validation*: The empirical evidence supports the argument that FairSMoE's integrated approach to managing multiple sensitive attributes simultaneously leads to higher performance and fairness. The results validate our design philosophy, which leverages a nuanced understanding of attribute interactions within neural networks to dynamically adjust to varied dataset characteristics and fairness requirements.

Settings	Acc.↑	PQD↑	DP↓
Vanilla SMoE	74.49	0.773	3.12
Ours. w/o FGR	76.08	0.781	2.93
Ours. w/o FEM	75.91	0.783	2.96
Ours. w/ both	77.25	0.801	2.72

Table 2: Ablation studies on FairSMoE of proposed Fairness-Guided Routing (FGR) and Fairness-driven Expert Management (FEM) on ISIC2019 dataset. Backbone is DeiT-Small and we set the number of experts as 4.

Ablation Studies We conducted ablation studies on each component in FairSMoE. We conducted experiments on ISIC2019 dataset with all three attributes, with DeiT-Small as a backbone. As shown in Table 2 and Figure 4, we conducted

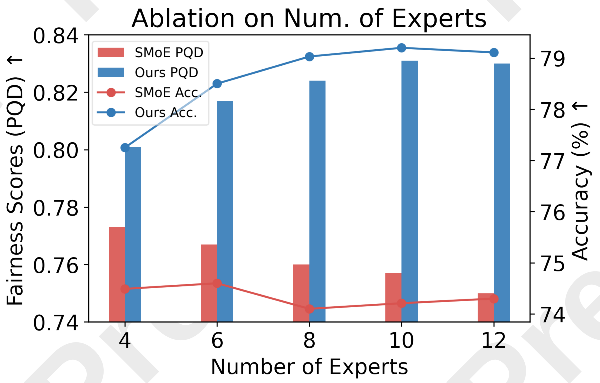

Figure 4: Ablation studies on the number of experts.

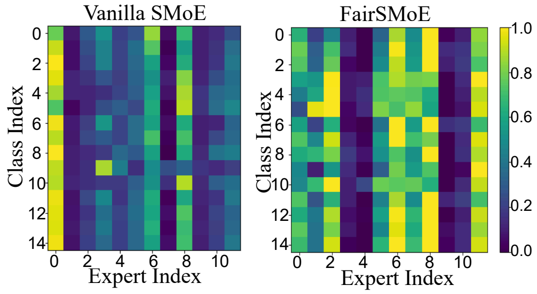

Figure 5: Comparison of expert activation frequencies. We include all three attributes in ISIC2019 and take DeiT-Small as the backbone with SMoE layers on the last layer.

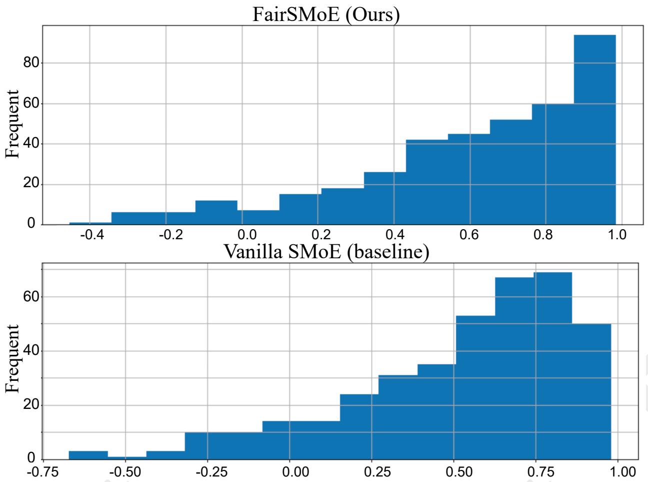

Figure 6: Comparison of gradient conflict on the last layer of DeiT-S

experiments on Fairness-Guided Routing (FGR), Fairness-driven Expert Management (FEM), and the number of experts. Our results demonstrate that (i) the proposed FGR is more effective compared with the vanilla router, demonstrating better accuracy and PQD score compared with the baseline. (ii) when equipped with FEM, the performance is also boosted by 1.42% in accuracy and 0.01 in PQD score, which shows the necessity of selecting the appropriate capacity for each task. (iii) when the number of experts arises, performance on both accuracy and fairness are enhanced. Meanwhile, for vanilla SMoE, the performance on both aspects drop for more experts due to low utilization rate. These promising results demonstrate that FairSMoE has enhanced SMoE with better scalability. We also conducted empirical studies on α in Equation (8). Results are shown in the Appendix.

6.3 In-Depth Discussion of FairSMoE

Given the superiority of our FairSMoE, we further investigate (i) its expert specialization and routing quality, and (ii) the mitigation effects on gradient conflicts from multiple training objectives.

FairSMoE alleviate the unbalanced routing schedule. As mentioned in Discovery and analysis, the default routing mechanism[Shazeer *et al.*, 2017] can not sufficiently accommodate the diversity of multiple sensitive attributes embedded into the input and will tend to develop a deterministic routing pattern and over-utilization of certain experts, which will get worse when the expert number increases. One key advantage of SMoE is to regularize each expert to focus on a set of attributes and improve the overall utilization rate with the increasing expert number. Figure 5 demonstrates the expert activation frequencies on both vanilla SMoE and FairSMoE. We observe that more experts are activated during the routing, which indicates a better utilization rate and leads to better performance on both accuracy and fairness as pointed out in Ablation Studies.

FairSMoE mitigate the issue of gradient conflict among different task. The FairSMoE approach mitigates the issue of gradient conflict among different tasks by employing disentangled representation learning and confusion loss, which ensure that features are insensitive to sensitive attributes. This reduces the interference between tasks related to different attributes, leading to more coherent gradient directions and improved model stability during training. Figure 6 compares the cosine similarity between gradients computed from the `age` and `site` tasks on ISIC2019. The distribution of gradients is more skewed towards 1, indicating higher cosine similarity and reduced gradient conflict.

7 Conclusion

In this paper, we propose FairSMoE, a framework designed to address multi-attribute fairness problem in vision recognition. Our method modifies traditional SMoE models by incorporating fairness-guided routing and dynamic expert management to optimize expert utilization and minimize bias. Through comprehensive testing on multiple datasets with 4 mainstream ViT backbones, we demonstrate notable enhancements in both performance and fairness, also the great generalization and scalability capability on different settings.

Acknowledgments

This work is partially supported by the U.S. National Science Foundation (NSF) CNS-2312158, NSF CMMI-2125326 and Fundamental Research Funds for the Central Universities, Peking University.

Contribution Statement

* Changdi Yang and Zheng Zhan contributed equally. † Work was done during Zheng Zhan's PhD study in Northeastern University. ‡ Pu Zhao and Yanzhi Wang are the corresponding authors.

References

[Chen *et al.*, 2022] Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik Learned-Miller, and Chuang Gan. Mod-squad: Designing mixture of experts as modular multi-task learners, 2022.

[Chen *et al.*, 2023] Tianlong Chen, Xuxi Chen, Xianzhi Du, et al. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *ICCV*, pages 17346–17357, 2023.

[D'Amour *et al.*, 2020] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, and et al. Underspecification presents challenges for credibility in modern machine learning, 2020.

[Deng *et al.*, 2023] Wenlong Deng, Yuan Zhong, Qi Dou, and Xiaoxiao Li. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In *International Conference on Information Processing in Medical Imaging*, 2023.

[Derman, 2021] Ekberjan Derman. Dataset bias mitigation through analysis of cnn training scores. *arXiv preprint arXiv:2106.14829*, 2021.

[Du *et al.*, 2022] Nan Du, Yanping Huang, Andrew M Dai, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*. PMLR, 2022.

[Du *et al.*, 2023] Siyi Du, Ben Hers, Nourhan Bayasi, et al. Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. In *ECCV 2022 Workshops*. Springer, 2023.

[Gaci *et al.*, 2022] Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. Iterative adversarial removal of gender bias in pretrained word embeddings. In *Proceedings of the 37th ACM/SIGAPP Symposium On Applied Computing*, pages 829–836, 2022.

[Hwang *et al.*, 2020] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*, 2020.

[Jacobs *et al.*, 1991] Robert A Jacobs, Michael I Jordan, et al. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[Jiang *et al.*, 2024] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[Karkkainen and Joo, 2021] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[Lepikhin *et al.*, 2020] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, et al. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[Li *et al.*, 2022] Yanyu Li, Pu Zhao, Geng Yuan, Xue Lin, Yanzhi Wang, and Xin Chen. Pruning-as-search: Efficient neural architecture search via channel pruning and structural reparameterization. *arXiv preprint arXiv:2206.01198*, 2022.

[Li *et al.*, 2023a] Yanyu Li, Changdi Yang, Pu Zhao, et al. Towards real-time segmentation on the edge. AAAI'23/IAAI'23/EAAI'23, 2023.

[Li *et al.*, 2023b] Yize Li, Pu Zhao, Xue Lin, Bhavya Kailkhura, and Ryan Goldhahn. Less is more: Data pruning for faster adversarial training. *arXiv preprint arXiv:2302.12366*, 2023.

[Li *et al.*, 2024] Yize Li, Pu Zhao, Ruyi Ding, Tong Zhou, Yunsi Fei, Xiaolin Xu, and Xue Lin. Neural architecture search for adversarial robustness via learnable pruning. In *Frontiers in High Performance Computing*, 2024.

[Li *et al.*, 2025] Yize Li, Yihua Zhang, Sijia Liu, and Xue Lin. Pruning then reweighting: Towards data-efficient training of diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[Nanda *et al.*, 2021] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021.

[Puyol-Antón *et al.*, 2022] Esther Puyol-Antón, Bram Ruijsink, Jorge Mariscal Harana, Stefan K Piechnik, and et al. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. *Frontiers in cardiovascular medicine*, 9:859310, 2022.

[Ruder, 2017] Sebastian Ruder. An overview of multi-task learning in deep neural networks, 2017.

[Sattigeri *et al.*, 2019] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

[Shazeer *et al.*, 2017] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[Shen *et al.*, 2024] Xuan Shen, Pu Zhao, Yifan Gong, Zhenglun Kong, Zheng Zhan, Yushu Wu, Ming Lin, Chao

Wu, Xue Lin, and Yanzhi Wang. Search for efficient large language models. In *NeurIPS*, 2024.

[Shen *et al.*, 2025a] Xuan Shen, Weize Ma, Jing Liu, et al. Quartdepth: Post-training quantization for real-time depth estimation on the edge. In *CVPR*, 2025.

[Shen *et al.*, 2025b] Xuan Shen, Zhao Song, Yufa Zhou, et al. Lazydit: Lazy learning for the acceleration of diffusion transformers. In *AAAI*, 2025.

[Shen *et al.*, 2025c] Xuan Shen, Zhao Song, Yufa Zhou, et al. Numerical pruning for efficient autoregressive models. In *AAAI*, 2025.

[Shen *et al.*, 2025d] Xuan Shen, Hangyu Zheng, Yifan Gong, et al. Sparse learning for state space models on mobile. In *ICLR*, 2025.

[Sheng *et al.*, 2023] Yi Sheng, Junhuan Yang, Lei Yang, et al. Muffin: A framework toward multi-dimension ai fairness by uniting off-the-shelf models. In *DAC*. IEEE, 2023.

[Stafanovičs *et al.*, 2020] Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. Mitigating gender bias in machine translation with target gender annotations. *arXiv preprint arXiv:2010.06203*, 2020.

[Tao *et al.*, 2022] Guanhong Tao, Weisong Sun, Tingxu Han, et al. Ruler: discriminative and iterative adversarial training for deep neural network fairness. In *ACM joint european software engineering conference and symposium on the foundations of software engineering*, 2022.

[Tian *et al.*, 2024] Huan Tian, Bo Liu, Tianqing Zhu, Wanlei Zhou, and S Yu Philip. Multifair: Model fairness with multiple sensitive attributes. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, et al. Training data-efficient image transformers & distillation through attention, 2021.

[Wang and Deng, 2020] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.

[Wu *et al.*, 2022a] Yawen Wu, Dewen Zeng, Xiaowei Xu, et al. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention*. Springer, 2022.

[Wu *et al.*, 2022b] Yushu Wu, Yifan Gong, Pu Zhao, et al. Compiler-aware neural architecture search for on-mobile real-time super-resolution. In *ECCV*, pages 92–111. Springer, 2022.

[Xu *et al.*, 2018] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*, pages 570–575. IEEE, 2018.

[Xu *et al.*, 2019] Depeng Xu, Yongkai Wu, Shuhan Yuan, et al. Achieving causal fairness through generative adversarial networks. In *IJCAI*, 2019.

[Yan *et al.*, 2020] Shen Yan, Di Huang, and Mohammad Soleymani. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 361–369, 2020.

[Yang *et al.*, 2023a] Changdi Yang, Yi Sheng, Peiyan Dong, et al. Fast and fair medical ai on the edge through neural architecture search for hybrid vision models. In *ICCAD*. IEEE, 2023.

[Yang *et al.*, 2023b] Changdi Yang, Yi Sheng, Peiyan Dong, et al. Late breaking results: Fast fair medical applications? hybrid vision models achieve the fairness on the edge. In *DAC*. IEEE, 2023.

[Yang *et al.*, 2023c] Changdi Yang, Pu Zhao, Yanyu Li, et al. Pruning parameterization with bi-level optimization for efficient semantic segmentation on the edge. In *CVPR*, 2023.

[Zhan *et al.*, 2021] Zheng Zhan, Yifan Gong, Pu Zhao, Geng Yuan, et al. Achieving on-mobile real-time super-resolution with neural architecture and pruning search. In *ICCV*, pages 4821–4831, 2021.

[Zhan *et al.*, 2024a] Zheng Zhan, Yushu Wu, Yifan Gong, et al. Fast and memory-efficient video diffusion using streamlined inference. In *NeurIPS*, 2024.

[Zhan *et al.*, 2024b] Zheng Zhan, Yushu Wu, Zhenglun Kong, et al. Rethinking token reduction for state space models. In *EMNLP*, pages 1686–1697, Miami, Florida, USA, nov 2024. ACL.

[Zhang and Yang, 2021] Yu Zhang and Qiang Yang. A survey on multi-task learning, 2021.

[Zhang *et al.*, 2022] Yihua Zhang, Yuguang Yao, Parikshit Ram, et al. Advancing model pruning via bi-level optimization. *NeurIPS*, 2022.

[Zhao and Chen, 2020] Chen Zhao and Feng Chen. Rank-based multi-task learning for fair regression, 2020.

[Zhao *et al.*, 2024] Pu Zhao, Fei Sun, Xuan Shen, et al. Pruning foundation models for high accuracy without retraining. In *Findings of EMNLP 2024*. ACL, 2024.

[Zoph *et al.*, 2022] Barret Zoph, Irwan Bello, et al. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.