BAMFair: Barycenter Aligned Mediation for Fairness across Multiple Sensitive Attributes

Hengyu Yue[1,2], Fan Wang[3], Weiming Liu[3*], Yuwen Liu[1,2], Lianyong Qi[1,2*], Haolong Xiang[4], Xiaolong Xu[4], Xuyun Zhang[5], Shichao Pei[6], Qiang Ni[7]

[1]College of Computer Science and Technology, China University of Petroleum (East China)
[2]Shandong Key Laboratory of Intelligent Oil and Gas Industrial Software
[3]College of Computer Science and Technology, Zhejiang University
[4]School of Software, Nanjing University of Information Science and Technology
[5]School of Computing, Macquarie University
[6]Department of Computer Science, University of Massachusetts Boston
[7]School of Computing and Communications, Lancaster University
{hengyuyue1, yuwenliu97}@gmail.com, {fanwang97, 21831010}@zju.edu.cn,
lianyongqi@upc.edu.cn, haolong.xiang@hdr.mq.edu.au, xlxu@nuist.edu.cn,
xuyun.zhang@mq.edu.au, shichao.pei@umb.edu, q.ni@lancaster.ac.uk

Abstract

Achieving fairness in machine learning models while maintaining high accuracy is an important but complex task, especially when handling multiple sensitive attributes. Traditional fairness methods often struggle to eliminate bias within subgroups divided by sensitive attributes. Several key challenges have been identified in this context: (1) *Multiple sensitive attribute scalability* challenge, where methods fail to ensure fairness as the number of sensitive attributes increases, despite scenarios with multiple sensitive attributes being prevalent in real-world applications; (2) *Multiple objective optimization conflict* challenge, where simultaneously optimizing for accuracy, fairness, and other relevant objectives leads to conflicting gradient updates, causing suboptimal performance. To address these challenges, we propose `BAMFair`, a Barycenter Aligned Mediation framework for fairness across multiple sensitive attributes. It comprises two core modules: a *Global Barycentric Alignment (GBA)* module and a *Nash Fairness Mediator (NFM)* module. Specifically, GBA innovatively introduces a global fair barycenter and minimizes the distances from subgroups divided by sensitive attributes to it, providing a scalable and efficient solution for fairness optimization across multiple sensitive attributes. Subsequently, NFM negotiates an agreement among inconsistent gradient updates between different objectives. Extensive experiments on four real-world datasets validate that `BAMFair` outperforms state-of-the-art methods in scenarios with multiple sensitive attributes.

*Corresponding authors.

1 Introduction

In recent years, machine learning has been increasingly applied to assist decision-making in critical social scenarios including visual recognition [Xie *et al.*, 2026c; Xie *et al.*, 2026a], recommendation systems [Xie *et al.*, 2026b; Liu *et al.*, 2026; Liu *et al.*, 2025b], and justice [Mehrabi *et al.*, 2021], etc. With the rapid advancement of machine learning, concerns about its fairness have intensified [Zhang and Harman, 2021]. In practice, machine learning algorithms often exhibit unfair behaviors related to sensitive attributes such as gender, race, age, and region. For instance, Amazon's recruitment model exhibited gender bias [Smith Brain Trust, 2018], favoring male candidates; in fairness testing via the FairLLM benchmark [Zhang *et al.*, 2023], ChatGPT was found to have fairness flaws across eight sensitive attributes in movie recommendations.

Therefore, algorithmic fairness has become increasingly important to reduce unintentional biases in machine learning. Existing research has proposed various fairness algorithms [Smith Brain Trust, 2018]. Recently, main approach is first to divide data into different subgroups based on predefined sensitive attributes, and then to improve fairness by addressing the imbalance among different subgroups. For instance, Reweighting (RW) [Kamiran and Calders, 2011] assigns different weights to different subgroups to improve fairness; Fair-SMOTE [Chakraborty *et al.*, 2021] achieves fairness by creating synthetic data for minority subgroups to fix the imbalance. Currently, most existing work on fairness focuses on debiasing for a single sensitive attribute. However, it is common for multiple sensitive attributes occurring simultaneously in real-world applications. Consider an example of a loan approval scenario with two sensitive attributes (e.g., gender and race), as shown in Figure 1. We observe that when considering only gender, the approval rate for black people is significantly lower. This will not only exacerbate gender inequality, but also may undermine public trust in the fairness

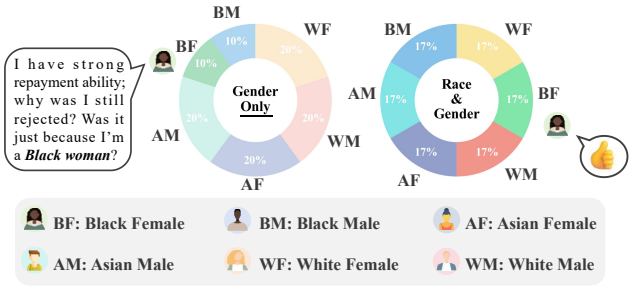


Figure 1: An example of loan approval for different clients. Clients with same repayment ability are divided into six subgroups based on race and gender. A fair loan approval process should consider clients' repayment ability, rather than race and gender. When only gender is considered as a sensitive attribute, black people are discriminated against. Only when both race and gender are considered as sensitive attributes can fairness be achieved for both groups, thereby ensuring the fairness of the loan approval process.

of the financial system [Smith Brain Trust, 2018]. Only by considering both race and gender can we ensure that different subgroups have comparatively fair opportunities in loan approval. Notably, other sensitive attributes (e.g., marital status and educational level) can also influence decision outcomes in practice. Ignoring such attributes may cause severe discrimination in the real world. Thus, it is critical to address fairness across multiple sensitive attributes.

However, existing methods primarily focus on single sensitive attribute and lack exploration of multiple sensitive attributes. Therefore, the first challenge is ***(CH1): Fairness models based on sensitive attribute grouping cannot remain effective when extended to multiple (two or more) sensitive attributes.*** To verify the importance of multiple sensitive attributes, we conducted a preliminary experiment comparing scenarios with different sensitive attributes, as shown in Table 1. When we consider only gender, the fairness of gender improves. However, racial fairness becomes even worse than the baseline with no sensitive attributes. Only by considering both gender and race can the model achieve fairness for both groups. Although some studies have explored fairness with two sensitive attributes, they remain difficult to extend to scenarios with more sensitive attributes. For instance, although Fair-SMOTE can be extended to more sensitive attributes, it requires $C_{2^n}^2$ alignment operations between 2^n subgroups (formed by pairwise combinations of sensitive attributes) in a binary classification task with n sensitive attributes, which leads to a high computational cost.

Furthermore, as the number of sensitive attributes increases, the number of subgroups divided by these attributes grows correspondingly. This makes fairness optimization across subgroups more challenging. For instance, when fairness objectives expand from singular focus on racial fairness to multiple targets encompassing race, gender, and age, conflicts emerge in multi-objective optimization. We thus pose the following challenge ***(CH2): The existing work overlooks to negotiate an agreement among inconsistent gradient updates between accuracy and complex fairness objectives.***

Sensitive Attribute	Acc	M_{EOP}^{Gender}	M_{EOP}^{Race}
None	69.79	0.1252	0.1047
Gender	68.87	0.0118	0.1176
Gender & Race	68.91	0.0123	0.0181

Table 1: Experimental results of fairness optimization on 3 kinds of sensitive attributes in terms of Acc and M_{EOP}[1]. *"None"* represents the baseline without any fairness optimization. Although debiasing for *only gender* improves fairness of gender, racial fairness becomes even *worse than the "None" baseline*. Only by considering *both gender and race* can the model achieve comparatively fair outcomes for both groups simultaneously.

In this paper, we propose a novel fairness framework named `BAMFair`, a Barycenter Aligned Mediation framework for Fairness across Multiple Sensitive Attributes. To address **CH1**, `BAMFair` introduces a global fair barycenter and innovatively transforms the fairness problem into minimizing the distances from individual subgroups to this barycenter. This approach effectively avoids the high complexity of naïve methods that rely on pairwise alignments between all subgroups. By focusing on subgroup-barycenter matching instead, this design significantly reduces the complexity of fairness optimization. We provide a detailed time complexity analysis in Section 4.1. Furthermore, by reducing the number of optimization objectives, the proposed alignment strategy notably alleviates **CH2**. Specifically, GBA changes the number of fairness goals from $O(T^2)$ pairwise matches to only $O(T)$ alignments, where T is the number of subgroups. By significantly reducing the total number of optimization targets, our method makes it easier to negotiate an agreement among inconsistent gradient updates between accuracy and fairness objectives. To further mitigate **CH2**, `BAMFair` incorporates a Nash Fairness Mediator (NFM), which dynamically ensures the accuracy and fairness of the model.

We summarize our main contributions as follows:

- We propose a novel framework `BAMFair`, which realizes the subgroup-barycenter alignment strategy, thus effectively addressing the fairness problem with multiple sensitive attributes.
- The `BAMFair` framework contains two modules: *GBA* and *NFM*. Specifically, GBA innovatively transforms the fairness problem into minimizing the distances from individual subgroups to the barycenter. NFM negotiates an agreement among inconsistent gradient updates between accuracy and fairness objectives.
- Extensive experiments conducted on four real-world datasets demonstrate the effectiveness and scalability of our proposed `BAMFair`, compared with other state-of-the-art methods.

2 Related Work

With the widespread deployment of machine learning, fairness models have gained significant attention. Fairness methods are categorized by their mitigation stage into

[1] M_{EOP} is a fairness metric, where smaller values indicate better fairness. Its formal definitions is provided in Section 3.2.

pre-processing, in-processing, and post-processing. Pre-processing debiases data before training [d'Alessandro *et al.*, 2017], in-processing embeds constraints during training [Dai and Wang, 2021], and post-processing adjusts model predictions [Dai *et al.*, 2024]. While flexible and model-agnostic, the effectiveness of pre-processing and post-processing is limited because the training process itself can introduce new biases [Zhang *et al.*, 2025]. Consequently, many methods focus on adjusting the training process. For instance, [Kamishima *et al.*, 2012] developed a bias remover using discrimination-aware regularization, while [Lokhande *et al.*, 2020; Liu *et al.*, 2019] incorporate fairness metrics directly into loss function.

Despite these advancements, research on multiple sensitive attributes remains limited [Chen *et al.*, 2024]. Although some existing methods such as [Chakraborty *et al.*, 2021] and [Peng *et al.*, 2022] can handle two sensitive attributes, they suffer from high computational complexity and inconsistent gradient between accuracy and complex fairness objectives. A naïve solution to extend current methods to multiple sensitive attributes is pairwise alignment between subgroups, which also leads to a large number of targets and heavy computational costs. In this paper, we propose `BAMFair`, an in-processing method designed for efficient and scalable fairness optimization involving multiple sensitive attributes.

3 Preliminary

3.1 Notations

Consider a binary classification task with dataset $D = \{(x, y, s)\}$, where $x \in \mathbb{R}^d$ is the input attribution, $y \in \{0, 1\}$ is the binary ground truth label, and $s = (s_1, s_2, \ldots, s_k) \in \{0, 1\}$ represent k binary sensitive attributes. We use $\hat{y} \in \{0, 1\}$ to represent the predicted label. Let $\mathcal{G} = (g_1, g_2, \ldots, g_T)$ represent the set of subgroups divided by sensitive attributes, where $T = 2^k$.

3.2 Fairness Evaluation Metrics

We define two fairness **metrics** for a binary label $y \in \{0, 1\}$. To handle multiple sensitive attributes, we evaluate the disparity between any two subgroups $g_r, g_w \in \mathcal{G}$.

Definition 1. *Demographic Parity (DP)* [Dwork *et al.*, 2012]. DP requires the model's predictions to be independent of sensitive attributes. The disparity is quantified by M_{DP}, which measures the maximum difference in acceptance rates between any two subgroups:

$$M_{\text{DP}} = \max_{g_r, g_w \subseteq \mathcal{G}} \left| \mathbb{P}(\hat{y} = 1 \mid g = g_r) - \mathbb{P}(\hat{y} = 1 \mid g = g_w) \right|. \tag{1}$$

where $\mathbb{P}(\cdot)$ denotes the probability measure.

Definition 2. *Equal Opportunity (EOP)* [Hardt *et al.*, 2016]. EOP requires that the true positive rate be consistent across all subgroups. The deviation is measured by M_{EOP}, reflecting the maximum gap in positive prediction probabilities for individuals with positive ground-truth labels:

$$M_{\text{EOP}} = \max_{g_r, g_w \subseteq \mathcal{G}} \left| \mathbb{P}(\hat{y} = 1 \,|\, y = 1,\ g = g_r) - \mathbb{P}(\hat{y} = 1 \,|\, y = 1, g = g_w) \right|. \tag{2}$$

4 Method

In this section, we will introduce the details of our proposed method `BAMFair` for solving **CH1** and **CH2**. The framework of `BAMFair` is shown in Figure 2. `BAMFair` mainly has two modules, i.e., *global barycentric alignment (GBA)* module and *nash fairness mediator (NFM)* module. Specifically, `BAMFair` first obtains embeddings of each subgroup through the backbone network ψ [Liao *et al.*, 2025; Liu *et al.*, 2025a; Xie *et al.*, 2021]. Then, GBA constructs a fair barycenter distribution of all subgroups, and continues to calculate the distance from each subgroup distribution to the barycenter to construct the fairness loss $\mathcal{L}_{\text{fair}}$. Finally, NFM dynamically balances different losses, outputting the optimal parameter update direction.

4.1 Global Barycentric Alignment (GBA)

We first introduce the *Global Barycentric Alignment (GBA)* module in `BAMFair`, which consists of two key steps: *global barycentric construction* and *fairness loss construction*. The global barycentric construction aims to calculate distribution of global fairness barycenter through subgroups. The fairness loss construction aims to quantify the distance between each subgroup and barycenter to form the fairness loss. These two steps align subgroups with the barycenter, thereby enhancing fairness with limited time cost.

Step 1: Global Fair Barycenter Construction

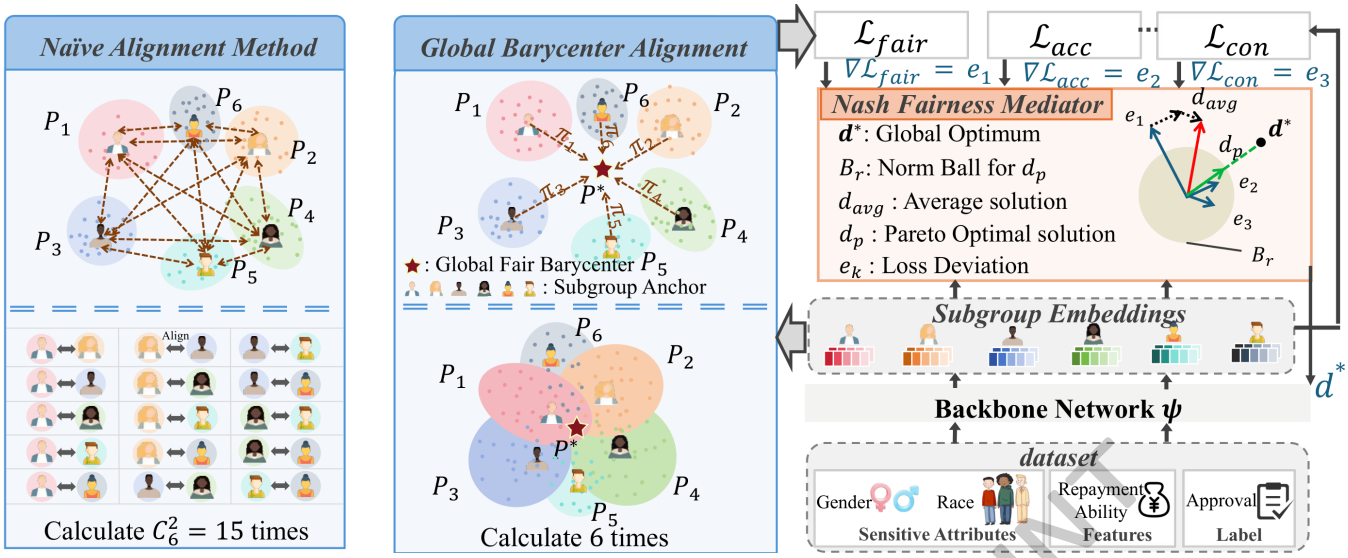
To quantify the prediction distribution of each subgroup, we first discretize the model's output scores into histograms. For subgroup $g_t \in G$, the range of prediction scores is divided into M continuous non-overlapping bins. For the i-th bin of g_t, $v_{t,i}$ represents the center value of the i-th bin in g_t, and $a_{t,i}$ represents the proportion of samples in g_t with scores falling into the i-th bin, satisfying $a_{t,i} \geq 0$ and $\sum_{i=1}^{M} a_{t,i} = 1$. The discretized distribution of g_t is thus $P_t = \{(a_{t,i}, v_{t,i})\}_{i=1}^{M}$. After that, we construct a global fair barycenter P^* as an unbiased reference distribution, which is also represented as a histogram with N bins, i.e., $P^* = \{(b_j, \mu_j)\}$ (for $j = 1$ to N), where μ_j represents the center value of the j-th bin of the barycenter, while b_j represents the probability mass of the j-th bin of the barycenter, satisfying $b_j \geq 0$ and $\sum_{j=1}^{N} b_j = 1$. We define $\pi_{t,ij}$ as the transport plan, representing the mass transferred from the i-th bin of subgroup t to the j-th bin of the barycenter, Then, the optimal transport problem [Wang *et al.*, 2026a] from subgroups to barycenter is formulated as:

$$\begin{aligned} Tp &= \min_{\pi \in \triangle, \mu} \sum_{i,j} \pi_{ij} C_{t,ij} - \varepsilon H(\pi), \\ \text{s.t. } \triangle &= \left\{ \sum_{j} \pi_{t,ij} = a_{t,i}, \sum_{i} \pi_{t,ij} = b_j, \pi_{t,ij} \geq 0 \right\}, \end{aligned} \tag{3}$$

where the distance matrix $C_{t,ij} = \|v_{t,i} - \mu_j\|_2^2$, and $H(\pi) = -\sum_{i,j} \pi_{t,ij} (\log \pi_{t,ij} - 1)$ is the entropy of the transport plan, and ε controls the weight of entropy.

We define potential functions $f_{t,i}, q_j$. The transport plan $\pi_{t,ij}$ can be expressed as:

$$\pi_{t,ij} = \exp\left(\frac{f_{t,i} + q_j - C_{t,ij}}{\varepsilon} \right). \tag{4}$$



(a) Naïve Alignment Method (b) Framework of BAMFair

Figure 2: The subfigure (b) is the overview of `BAMFair`, including two modules: (1) *GBA* constructs a global barycenter from subgroups and only requires $T = 6$ alignments. In contrast, the subfigure (a) shows that naïve pairwise alignment method requires $C_T^2 = 15$ operations for $T = 6$ subgroups. (2) *NFM* negotiates an agreement among inconsistent gradient updates between different objectives, aiming to handle conflicting requirements and achieve the optimal update direction.

Meanwhile, by using constraints $\triangle$ imposed on the marginals, we derive:

$$q_{t,j}^{(l+1)} = \varepsilon \log\left(b_{t,j} / \sum_i \exp\left(\frac{f_{t,i}^{(l)} - C_{t,ij}}{\varepsilon}\right)\right). \quad (5)$$

$$f_{t,i}^{(l+1)} = \varepsilon \log\left(a_{t,i} / \sum_j \exp\left(\frac{q_{t,j}^{(l+1)} - C_{t,ij}}{\varepsilon}\right)\right). \quad (6)$$

After updating transport plan, we can fix $\pi_{t,ij}$ and update μ_j:

$$\mu_j = \frac{\sum_t \sum_i \pi_{t,ij} v_{t,i}}{\sum_t \sum_i \pi_{t,ij}}. \quad (7)$$

Upon convergence, we obtain the histogram distribution $P^* = \{b_j, \mu_j\}$ of the global barycenter.

Step 2: Fairness Loss Construction

After getting the barycenter distribution P^*, the fairness loss construction continues to quantify the deviation between each subgroup distribution $P_t = \{a_{t,i}, v_{t,i}\}$ and the barycenter P^* by calculating the Sinkhorn distance [Wang *et al.*, 2024; Wang *et al.*, 2025a]:

$$S(P_t, P^*) = \sum_{i,j} \left[\pi_{t,ij} \cdot \|v_{t,i} - \mu_j\|_2^2 + \varepsilon \pi_{t,ij} (\log \pi_{t,ij} - 1)\right]. \quad (8)$$

Finally, the fairness loss $\mathcal{L}_{\text{fair}}$ is defined as the average of Sinkhorn distances from all subgroups to the barycenter:

$$\mathcal{L}_{\text{fair}} = \frac{1}{T} \sum_{t=1}^{T} S(P_t, P^*). \quad (9)$$

According to the triangle inequality property, the distance between any two arbitrary subgroups g_r and g_w is upper-bounded by their distances to the global barycenter: $S(P_r, P_w) \leq S(P_r, P^*) + S(P_w, P^*)$. Consequently, minimizing $\mathcal{L}_{\text{fair}}$ in Eq.(9) effectively constrains the pairwise disparity across all subgroup combinations. This strategy resolves **CH1** by reducing the complexity from $O(T^2)$ pairwise comparisons to $O(T)$ alignments, providing a scalable fairness guarantee for multiple sensitive attributes.

Complexity Analysis

Aligning subgroups with a global fair barycenter can reduce the computational burden while ensuring fairness across multiple sensitive attributes. For all subgroups, the time complexity of fairness alignment is $O(T \cdot M \cdot N)$, compared to traditional pairwise alignment with time complexity $O(T^2 \cdot M^2)$. This ensures BAMFair maintains efficiency while enhancing fairness for multiple sensitive attributes.

4.2 Nash Fairness Mediator (NFM)

In addition to the fairness loss $\mathcal{L}_{\text{fair}}$, we introduce an accuracy loss $\mathcal{L}_{\text{acc}}$ and a contrastive learning loss $\mathcal{L}_{\text{con}}$. The accuracy loss quantifies the discrepancy between the model's predictions and the ground-truth labels, ensuring the model retains predictive power while optimizing for fairness. For binary classification tasks, it is defined using cross-entropy loss:

$$\mathcal{L}_{\text{acc}} = -\mathbb{E}\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right], \quad (10)$$

where $y \in \{0, 1\}$ is the ground-truth label, $\hat{y} \in [0, 1]$ is the model's predicted probability of positive class, and the expectation is taken over the dataset. The contrastive learning loss $\mathcal{L}_{\text{con}}$ enhances the model's ability to represent features irrelevant to sensitive attributes, reducing reliance on sensitive attributes and improving robustness and fairness of

features[Liao *et al.*, 2024b; Liao *et al.*, 2024a]:

$$\mathcal{L}_{\text{con}} = -\mathbb{E}\left[\log \frac{\exp(\text{sim}(h_m, h_n)/\tau)}{\sum_{s\neq m}\exp(\text{sim}(h_m, h_s)/\tau)}\right], \quad (11)$$

where h_m and h_n form a positive pair, with h_s as negative samples. $\text{sim}(\cdot,\cdot)$ uses cosine similarity to measure feature vector similarity, and the temperature parameter τ regulates the sharpness of the similarity distribution. Positive sample pairs consist of samples that differ in sensitive attributes yet exhibit high similarity in non-sensitive attributes and share consistent labels. This design compels the model to learn the common, discriminative features across such pairs, thereby reducing its over-reliance on sensitive attributes for prediction and enhancing fairness.

Inspired by [Navon *et al.*, 2022], the *nash fairness mediator* module aims to dynamically balance conflicts in multi-objective optimization, specifically coordinating gradient conflicts between the accuracy loss $\mathcal{L}_{\text{acc}}$, fairness loss $\mathcal{L}_{\text{fair}}$, and contrastive learning loss $\mathcal{L}_{\text{con}}$, ensuring the model does not sacrifice accuracy while improving fairness[Wang *et al.*, 2025b; Wang *et al.*, 2026b]. First, compute the gradients of each loss: $\nabla\mathcal{L}_{\text{acc}}$ (accuracy gradient), $\nabla\mathcal{L}_{\text{fair}}$ (fairness gradient), and $\nabla\mathcal{L}_{\text{con}}$ (contrastive learning gradient), and construct the gradient matrix $G = [\nabla\mathcal{L}_{\text{acc}}, \nabla\mathcal{L}_{\text{fair}}, \nabla\mathcal{L}_{\text{con}}]$, denoted as $E = [e_1, e_2, e_3]$.

The Nash bargaining solution aims to find non-negative weights $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ that satisfy the Nash equilibrium condition through a weighted gradient combination:

$$\alpha_k \cdot \frac{\partial}{\partial \alpha_k}\sum_{o=1}^{3}(\alpha_o e_o^\top d) = \alpha_m \cdot \frac{\partial}{\partial \alpha_m}\sum_{r=1}^{3}(\alpha_r e_r^\top d), \quad \forall k, m, \quad (12)$$

where d is the parameter update direction, and $\cdot^\top$ denotes the transpose operation. This condition ensures equal marginal gains for each loss, preventing a single loss from dominating the update direction due to scale advantages.

$\alpha^{(0)}$ is initialized to a uniform distribution, with $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$. In each iteration, update the weights α by solving the following optimization problem:

$$\alpha^{(t+1)} = \arg\max_{\alpha} \sum_k \log(\alpha_k(e_k^\top d^{(t)})), s.t. \alpha \geq 0, \sum_k \alpha_k = 1. \quad (13)$$

This objective function is derived from the Nash bargaining solution, where $e_k^\top d^{(t)}$ represents the projection of the gradient e_k onto the current update direction $d^{(t)}$, reflecting the contribution of the loss k to the current update. The constraints ensure that the weights are non-negative and sum to 1, maintaining the property of a convex combination. To solve this optimization problem efficiently, we can use Lagrange multipliers. Construct the Lagrangian function:

$$\mathcal{L}(\alpha, \lambda) = \sum_k \log(\alpha_k(e_k^\top d^{(t)})) + \lambda\left(1 - \sum_k \alpha_k\right). \quad (14)$$

The final solution for the weights can be expressed as:

$$\alpha_k^{(t+1)} = \frac{\frac{1}{e_k^\top d^{(t)}}}{\sum_{q=1}^{3}\frac{1}{e_q^\top d^{(t)}}}. \quad (15)$$

This formula demonstrates that the weight of each loss is inversely proportional to its projection onto the current update direction, balancing the contributions of different gradients.

After obtaining the new weights $\alpha^{(t+1)}$, update the parameter update direction:

$$d^{(t+1)} = \sum_{k=1}^{3}\alpha_k^{(t+1)} e_k. \quad (16)$$

This direction optimizes all losses simultaneously, avoiding bias caused by gradient scale differences.

4.3 Putting Together

Model parameters θ are updated using the coordinated direction d from NFM, with a learning rate η:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot d^{(t)}. \quad (17)$$

By operating on gradients (which encode local optimization directions) rather than raw loss values, the model avoids bias from differing loss magnitudes.

5 Experiment

In this section, we conduct extensive experiments to address five research questions: **RQ1:** How does `BAMFair` perform compared to the state-of-the-art fairness method? **RQ2:** What are the effects of different model components? **RQ3:** How does our method perform in terms of accuracy, fairness and time cost when it is extended to multiple sensitive attributes? **RQ4:** Why do we choose *NFM* over other multi-objective optimization methods?

5.1 Experimental Setup

Datasets

We conduct research using four real-world datasets: Adult [Becker and Kohavi, 1996], Compas [ProPublica, 2016], Credit [Yeh, 2009], and Mep16 [AHRQ, 2016]. For experiments involving two sensitive attributes, we select gender and race from the datasets; for three sensitive attributes, we add age as an additional sensitive attribute. For each experiment trial, we split the data into 70% training, 10% validation, and 20% testing sets. We repeat this procedure 10 times for statistical analysis and report the mean and standard deviation.

Comparison Methods

We compare `BAMFair` with several state-of-the-art fairness improvement methods, covering pre-processing, in-processing, and post-processing methods. Pre-processing includes RW [Kamiran and Calders, 2011], Fair-SMOTE [Chakraborty *et al.*, 2021]. In-processing includes MAAT [Chen *et al.*, 2022], FairGP [Luo *et al.*, 2025]. Post-processing includes ROC [Kamiran *et al.*, 2012], FairMask [Peng *et al.*, 2022], SFM [Hu *et al.*, 2024]. We also compare our method with a traditional model without fairness restriction named Default [Chen *et al.*, 2024].

Evaluation Metrics

The evaluation metrics for the model include Accuracy (Acc)[2], M_{DP}, and M_{EOP}. Smaller values of the fairness metrics M_{DP} and M_{EOP} indicate the model is fairer.

[2]In this paper, we use the percentage of correctly predicted samples as Acc.

Methods	Compas			Adult		
	Acc(%) ↑	M_{DP}(%) ↓	M_{EOP}(%) ↓	Acc(%) ↑	M_{DP}(%) ↓	M_{EOP}(%) ↓
Default	**69.7932 ± 0.5819**	13.5147 ± 1.2436	10.2758 ± 0.9123	**81.7061 ± 0.4728**	9.4873 ± 0.8345	22.3982 ± 1.5267
ROC	66.7853 ± 0.3941	6.0682 ± 0.4159	3.3764 ± 0.3572	77.9938 ± 0.5314	20.7895 ± 1.6723	21.3471 ± 1.2958
RW	67.1689 ± 0.6274	12.4935 ± 0.9561	8.8729 ± 0.7846	79.4325 ± 0.3187	10.1364 ± 0.5932	18.7753 ± 0.7689
MAAT	67.7541 ± 0.4583	21.2867 ± 1.5892	15.0178 ± 1.1349	79.4186 ± 0.6251	17.7248 ± 1.3476	24.0139 ± 1.7824
Fair-SMOTE	68.6827 ± 0.3759	2.7763 ± 0.2984	1.5371 ± 0.2468	78.9542 ± 0.4439	8.2957 ± 0.6518	15.6124 ± 0.9375
FairMask	67.4485 ± 0.5136	12.2479 ± 0.8842	4.8835 ± 0.4921	79.6678 ± 0.3594	7.9263 ± 0.4857	6.7489 ± 0.6234
SFM	68.0274 ± 0.4356	3.1589 ± 0.3742	2.8461 ± 0.2957	79.2845 ± 0.4173	6.7329 ± 0.5486	5.9274 ± 0.4731
FairGP	68.1356 ± 0.4928	2.3287 ± 0.3469	9.7341 ± 0.8573	78.3169 ± 0.5842	10.9138 ± 0.7765	4.1259 ± 0.5182
BAMFair	68.9173 ± 0.4265	**1.7948 ± 0.3692**	**0.4683 ± 0.2157**	81.4092 ± 0.5578	**5.8034 ± 0.7196**	**1.3572 ± 0.3849**

Methods	Credit			Mep16		
	Acc(%) ↑	M_{DP}(%) ↓	M_{EOP}(%) ↓	Acc(%) ↑	M_{DP}(%) ↓	M_{EOP}(%) ↓
Default	81.2743 ± 0.3971	2.3685 ± 0.4532	3.0867 ± 0.4195	**86.0883 ± 0.3246**	4.7359 ± 0.6874	3.7642 ± 0.5481
ROC	80.6629 ± 0.2584	1.1573 ± 0.2168	0.6249 ± 0.1435	85.2467 ± 0.4189	9.6932 ± 1.0357	6.4185 ± 0.7692
RW	81.0658 ± 0.4427	1.7639 ± 0.3854	0.6672 ± 0.2589	85.5074 ± 0.2761	4.3482 ± 0.5193	0.3265 ± 0.1874
MAAT	78.2891 ± 0.5163	0.6642 ± 0.3078	1.9053 ± 0.4762	84.0938 ± 0.4652	4.0576 ± 0.7249	11.6983 ± 1.2478
Fair-SMOTE	80.8376 ± 0.3249	2.2581 ± 0.4136	0.5147 ± 0.1953	79.8865 ± 0.6734	5.2048 ± 0.7431	5.3926 ± 0.6859
FairMask	80.5324 ± 0.2895	0.7965 ± 0.2641	1.4483 ± 0.3597	85.4391 ± 0.3376	4.0638 ± 0.5724	0.4691 ± 0.1685
SFM	80.7942 ± 0.3568	1.5274 ± 0.2836	0.8953 ± 0.1672	85.3164 ± 0.3057	4.2185 ± 0.5693	0.5842 ± 0.1736
FairGP	78.1367 ± 0.6348	0.6759 ± 0.2935	1.6084 ± 0.4321	85.3752 ± 0.2987	6.8179 ± 0.8564	0.3489 ± 0.1563
BAMFair	**81.3492 ± 0.4753**	**0.1068 ± 0.1249**	**0.2437 ± 0.1758**	85.5189 ± 0.3842	**3.8057 ± 0.6489**	**0.1643 ± 0.1392**

Table 2: Comparison on utility (Acc) and fairness (M_{DP}, M_{EOP}) in percentage (%). ↑ denotes the larger the better; ↓ denotes the opposite. The best results are bold. The second-best results are underlined.

Implementation Details

For binary classification tasks, we use a logistic regression model to generate classification results. We set the batch size as 128 and adopt Adam optimizer with the learning rate η as 0.002. We set M (number of bins for subgroup distributions) to 100. The temperature parameter τ is set to 0.07. ε (entropy weight in optimal transport) is set to a conventional value of 0.1 [Cuturi, 2013]. In NFM, the initial weights $\alpha^{(0)}$ are uniformly initialized as $(1/3, 1/3, 1/3)$ to ensure equal initial contributions of accuracy, fairness, and contrastive losses. The codes are available at https://github.com/getachAnGe/BAMFair.

5.2 Experimental Results

Performance Comparison (RQ1)

Table 2 presents results across four datasets with race and gender as sensitive attributes. We observe that: **(1) `BAMFair` consistently outperforms all baselines in terms of fairness.** This superiority stems from GBA's effective subgroup alignment (addressing **CH1**) and NFM's optimization of parameter update directions across conflicting losses (addressing **CH2**). **(2) `BAMFair` achieves the highest accuracy among all fairness-optimizing baselines.** Accuracy is slightly lower than Default because we sacrifice marginal performance for fairness. However, this drop remains minimal, and `BAMFair` still provides better accuracy than all other fairness-optimizing baselines. **(3) In-processing methods achieve a better trade-off between accuracy and fairness.** In contrast, pre-processing methods often show unstable performance across datasets, as modifying raw data can lead to the loss of critical feature information. Methods like MAAT, FairGP, and `BAMFair` dynamically integrate fairness constraints into representation learning, making them more robust in complex scenarios.

Module			Metric		
GBA	L_{con}	NFM	Acc(%) ↑	M_{DP}(%) ↓	M_{EOP}(%) ↓
✗	✗	✗	67.7532 ± 0.3918	5.3247 ± 0.4736	4.1958 ± 0.3823
✓	✗	✗	67.8541 ± 0.4583	2.1467 ± 0.2392	4.1478 ± 0.4249
✓	✓	✗	68.3689 ± 0.3574	1.8335 ± 0.1961	0.8129 ± 0.1146
✓	✓	✓	**68.9173 ± 0.4265**	**1.7948 ± 0.3692**	**0.4683 ± 0.2157**

Table 3: Ablation study on Compas. The best results are bold.

To verify if `BAMFair` generates fairer distributions, we visualize model outputs on Compas in Figure 3. While the Default model exhibits clear bias (e.g., higher distance (S) for Black Males), `BAMFair` produces nearly identical distributions across subgroups by minimizing Sinkhorn distances, effectively achieving group fairness.

Ablation Study (RQ2)

To study the effects of individual components, we conduct ablation study on the Compas dataset, with results presented in Table 3. The improvement boost from incorporating GBA proves the effectiveness of aligning subgroups with the barycenter. Adding L_{con} further improves results, validating that contrastive learning strengthens subgroup representation consistency. Finally, the performance improvement through

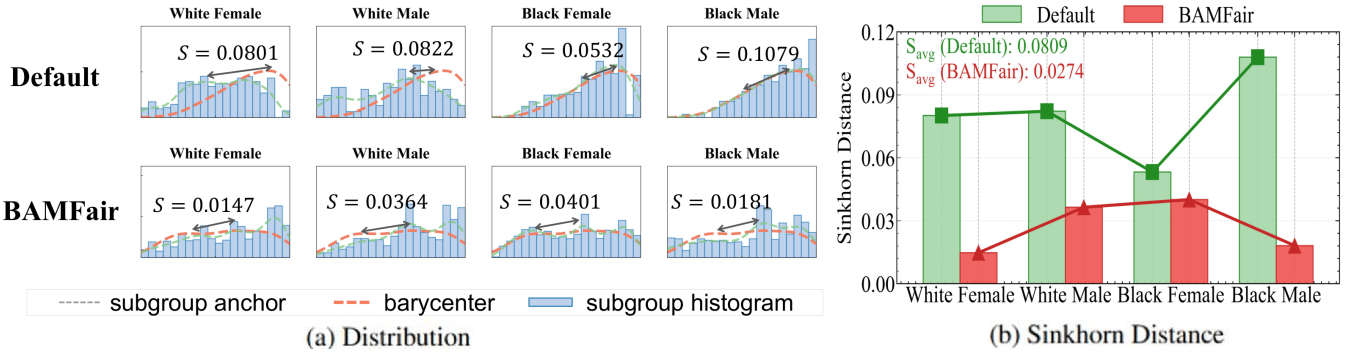


Figure 3: This figure demonstrates the model output distributions of subgroups on Compas with race and gender as sensitive attributes. The results demonstrate that BAMFair successfully generates fairer distributions with smaller Sinkhorn distance.

Method	Train Dataset			Test Dataset		
	Acc	M_{DP}	M_{EOP}	*Acc*	M_{DP}	M_{EOP}
GBA + PCgrad	68.27%	0.0518	0.0175	67.99%	0.0365	0.0471
GBA + CAgrad	**68.78%**	0.0452	0.0236	67.57%	0.0207	0.0511
GBA + Uncertainty	68.33%	0.0569	0.0342	68.08%	0.0236	0.0327
GBA + MGDA	67.91%	0.0247	0.0212	66.78%	0.0186	0.0089
GBA + NFM	68.45%	**0.0196**	**0.0180**	**68.91%**	**0.0179**	**0.0046**

Table 4: Comparison between different MTL methods on Compas dataset. The best results are bold.

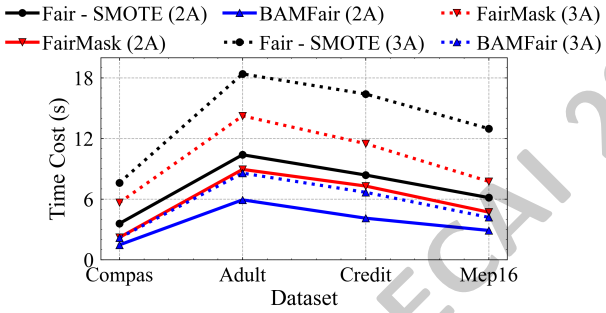


Figure 4: Time cost across datasets (2A for 2 sensitive attributes, 3A for 3 sensitive attributes). BAMFair consistently achieves the lowest time cost, with a significantly smaller increase in runtime as the number of sensitive attributes grows.

augmenting NFM, (i.e., the complete BAMFair framework) achieves the best balance, proving its ability to reach global optimality by reconciling multi-objective conflicts.

Analysis of Scalability (RQ3)

Results with three sensitive attributes highlight the scalability of BAMFair. As shown in Figure 5, BAMFair maintains superior performance while scaling efficiently, making it highly suitable for complex real-world applications. As shown in Figure 4, BAMFair consistently achieves the lowest time cost compared to all scalable methods, regardless of whether two or three sensitive attributes are used. Meanwhile, our method demonstrates a much smaller increase in computational cost as the number of sensitive attributes grows. For

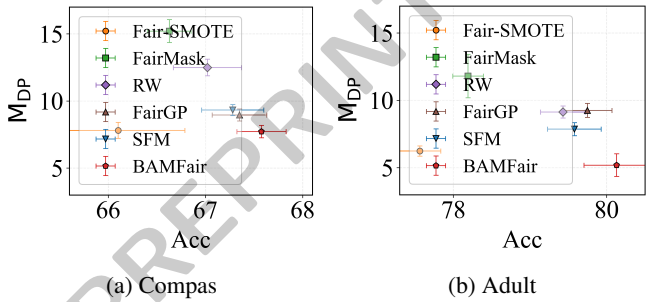


Figure 5: Scalability study on multiple sensitive attributes, where the number of sensitive attributes is extended to three (race, gender, and age). BAMFair outperforms all baselines in terms of both accuracy and fairness.

methods like Fair-SMOTE and FairMask, the time spikes significantly by 141.42% and 156.66% respectively. In contrast, the time cost of BAMFair grows by only 46.88%, which proves the scalability of BAMFair.

Comparison with other MTL (RQ4)

To validate NFM, we compare several main multiple task learning (MTL) methods [Yu *et al.*, 2024] on the Compas dataset. As shown in Table 4, NFM achieves the best performance. Among these methods, MGDA [Sener and Koltun, 2018] biases toward fairness due to its smaller loss scale. Other MTL methods [Yu *et al.*, 2020; Liu *et al.*, 2021], on the other hand, lean towards larger scaled accuracy loss, resulting in poor fairness performance. Although CAgrad gets higher accuracy, its fairness performance is notably worse than that of NFM. Only NFM manages to strike a good balance between fairness and accuracy.

6 Conclusion

To achieve fairness across multiple sensitive attributes, this paper proposes BAMFair. It utilizes GBA to align subgroups with a global barycenter, reducing alignment complexity while enhancing fairness and scalability, and NFM to resolve gradient conflicts. Across four publicly available real-world datasets, BAMFair outperforms state-of-the-art methods in terms of accuracy and fainess.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No.62572486), Natural Science Foundation of Shandong Province (No.ZR2023MF007), European Union under the Horizon Europe programme (grant numbers: 101135930, 101168438), in part by the UKRI through the UK Government's Horizon Europe funding Guarantee (grant numbers: 10119564, 10126241).

References

[AHRQ, 2016] AHRQ. The mep16 dataset (meps hc-192). https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192, 2016.

[Becker and Kohavi, 1996] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

[Chakraborty *et al.*, 2021] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? *Computing Research Repository*, 2021.

[Chen *et al.*, 2022] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. Maat: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *ESEC/FSE 2022*. Association for Computing Machinery, 2022.

[Chen *et al.*, 2024] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairness improvement with multiple protected attributes: How far are we?, 2024.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

[Dai and Wang, 2021] Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. *Computing Research Repository*, 2021.

[Dai *et al.*, 2024] Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, 21(6):1011–1061, 2024.

[d'Alessandro *et al.*, 2017] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134, 2017.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS 2012*, page 214–226. Association for Computing Machinery, 2012.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, page 3323–3331. Curran Associates Inc., 2016.

[Hu *et al.*, 2024] Francois Hu, Philipp Ratz, and Arthur Charpentier. A sequentially fair mechanism for multiple sensitive attributes. *AAAI*, 38(11):12502–12510, 2024.

[Kamiran and Calders, 2011] Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *KAIS*, 33, 2011.

[Kamiran *et al.*, 2012] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *ICDM 2012*, page 924–929. IEEE Computer Society, 2012.

[Kamishima *et al.*, 2012] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML PKDD*, page 35–50. Springer-Verlag, 2012.

[Liao *et al.*, 2024a] Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Fengyuan Yu, Huabin Zhu, Binhui Yao, Tao Wang, Xiaolin Zheng, and Yanchao Tan. Rethinking the representation in federated unsupervised learning with non-iid data. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22841–22850, 2024.

[Liao *et al.*, 2024b] Xinting Liao, Weiming Liu, Pengyang Zhou, Fengyuan Yu, Jiahe Xu, Jun Wang, Wenjie Wang, Chaochao Chen, and Xiaolin Zheng. Foogd: Federated collaboration for both out-of-distribution generalization and detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 132908–132945. Curran Associates, Inc., 2024.

[Liao *et al.*, 2025] Xinting Liao, Weiming Liu, Jiaming Qian, Pengyang Zhou, Jiahe Xu, Wenjie Wang, Chaochao Chen, Xiaolin Zheng, and Tat-Seng Chua. Focoop: enhancing out-of-distribution robustness in federated prompt learning for vision-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

[Liu *et al.*, 2019] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *ICCV*, pages 10051–10060, 2019.

[Liu *et al.*, 2021] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *NIPS*, volume 34, pages 18878–18890. Curran Associates, 2021.

[Liu *et al.*, 2025a] Yuwen Liu, Lianyong Qi, Xingyuan Mao, Weiming Liu, Shichao Pei, Fan Wang, Xuyun Zhang, Amin Beheshti, and Xiaokang Zhou. Variational graph auto-encoder driven graph enhancement for sequential recommendation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 3144–3152, 2025.

[Liu *et al.*, 2025b] Yuwen Liu, Lianyong Qi, Xingyuan Mao, Weiming Liu, Fan Wang, Xiaolong Xu, Xuyun Zhang, Wanchun Dou, Xiaokang Zhou, and Amin Beheshti. Hyperbolic variational graph auto-encoder for next poi recommendation. In *Proceedings of the ACM on Web Conference 2025*, pages 3267–3275, 2025.

[Liu *et al.*, 2026] Yuwen Liu, Lianyong Qi, Xingyuan Mao, Weiming Liu, Xuhui Fan, Qiang Ni, Xuyun Zhang, Yang Zhang, Yuan Tian, and Amin Beheshti. Hyperbolic-enhanced mixture-of-experts mamba for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 15403–15411, 2026.

[Lokhande *et al.*, 2020] Vishnu Suresh Lokhande, Aditya Kumar Akash, Sathya N. Ravi, and Vikas Singh. Fairalm: Augmented lagrangian method for training fair models with little regret. In *ECCV*, page 365–381. Springer-Verlag, 2020.

[Luo *et al.*, 2025] Renqiang Luo, Huafei Huang, Ivan Lee, Chengpei Xu, Jianzhong Qi, and Feng Xia. Fairgp: A scalable and fair graph transformer using graph partitioning. In *AAAI*, pages 12319–12327. AAAI, 2025.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 2021.

[Navon *et al.*, 2022] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game, 2022.

[Peng *et al.*, 2022] Kewen Peng, Joymallya Chakraborty, and Tim Menzies. Fairmask: Better fairness via model-based rebalancing of protected attributes, 2022.

[ProPublica, 2016] ProPublica. The compas dataset. https://github.com/propublica/compas-analysis, 2016.

[Sener and Koltun, 2018] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NIPS*, page 525–536. Curran Associates Inc., 2018.

[Smith Brain Trust, 2018] Smith Brain Trust. The problem with amazon's ai recruiter. https://www.rhsmith.umd.edu/research/problem-amazons-ai-recruiter, 2018. Accessed: 2018-10-31.

[Wang *et al.*, 2024] Fan Wang, Chaochao Chen, Weiming Liu, Tianhao Fan, Xinting Liao, Yanchao Tan, Lianyong Qi, and Xiaolin Zheng. Ce-rcfr: Robust counterfactual regression for consensus-enabled treatment effect estimation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3013–3023, 2024.

[Wang *et al.*, 2025a] Fan Wang, Lianyong Qi, Weiming Liu, Bowen Yu, Jintao Chen, and Yanwei Xu. Inter- and intra-similarity preserved counterfactual incentive effect estimation for recommendation systems. *ACM Trans. Inf. Syst.*, 43(6), 2025.

[Wang *et al.*, 2025b] Xiuyuan Wang, Chaochao Chen, Weiming Liu, Xinting Liao, Fan Wang, and Xiaolin Zheng. Efficient source-free unlearning via energy-guided data synthesis and discrimination-aware multitask optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[Wang *et al.*, 2026a] Fan Wang, Hengyu Yue, Bowen Yu, Weiming Liu, Zongxin Yang, Xuyun Zhang, Xiaolin Zheng, Chaochao Chen, and Shuiguang Deng. Spiked-cfr: Causal representation learning from llms via wasserstein projection pursuit. In *ICML*, 2026.

[Wang *et al.*, 2026b] Xiuyuan Wang, Weiming Liu, Hongyu Cai, Xin Gao, Fan Wang, Chaochao Chen, and Xiaolin Zheng. Preference-calibrated optimization with score-level distribution alignment for text-to-image diffusion model unlearning. In *ICML*, 2026.

[Xie *et al.*, 2021] Jianye Xie, Haotong Sun, Junsheng Zhou, Weiguang Qu, and Xinyu Dai. Event detection as graph parsing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1630–1640, 2021.

[Xie *et al.*, 2026a] Jianye Xie, Chunhua Hu, Lianyong Qi, Fan Wang, Xiaolong Xu, Haolong Xiang, Xuyun Zhang, Shichao Pei, Amin Beheshti, Wanchun Dou, et al. Plikd: Prompt learning with instance-aware knowledge distillation for web-scale semantic image classification. In *Proceedings of the ACM Web Conference 2026*, pages 3788–3799, 2026.

[Xie *et al.*, 2026b] Jianye Xie, Lianyong Qi, Weiming Liu, Anqi Wang, Xiaolong Xu, Haolong Xiang, Xuyun Zhang, Wenwen Gong, Yang Zhang, Amin Beheshti, et al. Joint similar user exploration and informative behavior guidance for multi-modal new item recommendation. In *Proceedings of the ACM Web Conference 2026*, pages 6218–6229, 2026.

[Xie *et al.*, 2026c] Jianye Xie, Lianyong Qi, Fan Wang, Anqi Wang, Wenjuan Gong, Danxin Wang, Wanchun Dou, Yang Cao, Shichao Pei, and Xiaokang Zhou. Retrieval-driven reasoning for deliberative visual classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11060–11068, 2026.

[Yeh, 2009] I-Cheng Yeh. Default of credit card clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

[Yu *et al.*, 2020] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NIPS*, volume 33, pages 5824–5836. Curran Associates, Inc., 2020.

[Yu *et al.*, 2024] Jun Yu, Yutong Dai, and Xiaokang Liu. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras. *CoRR*, 2024.

[Zhang and Harman, 2021] Jie M. Zhang and Mark Harman. "ignorance and prejudice" in software fairness. In *ICSE*, page 1436–1447, 2021.

[Zhang *et al.*, 2023] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *RecSys 2023*, pages 993–999. ACM, sep 2023.

[Zhang *et al.*, 2025] Wenbin Zhang, Shuigeng Zhou, Toby Walsh, and Jeremy C. Weiss. Fairness amidst non-iid graph data: A literature review, 2025.