

Adaptive Fair Representation Learning for Personalized Fairness in Recommendations via Information Alignment

Xinyu Zhu*
School of Computer Science
Sichuan University
Chengdu, China
zhuxinyu@stu.scu.edu.cn

Lilin Zhang*
School of Computer Science
Sichuan University
Chengdu, China
zhanglilin@stu.scu.edu.cn

Ning Yang†
School of Computer Science
Sichuan University
Chengdu, China
yangning@scu.edu.cn

ABSTRACT

Personalized fairness in recommendations has been attracting increasing attention from researchers. The existing works often treat a fairness requirement, represented as a collection of sensitive attributes, as a hyper-parameter, and pursue extreme fairness by completely removing information of sensitive attributes from the learned fair embedding, which suffer from two challenges: huge training cost incurred by the explosion of attribute combinations, and the suboptimal trade-off between fairness and accuracy. In this paper, we propose a novel Adaptive Fair Representation Learning (AFRL) model, which achieves a real personalized fairness due to its advantage of training only one model to adaptively serve different fairness requirements during inference phase. Particularly, AFRL treats fairness requirements as inputs and can learn an attribute-specific embedding for each attribute from the unfair user embedding, which endows AFRL with the adaptability during inference phase to determine the non-sensitive attributes under the guidance of the user's unique fairness requirement. To achieve a better trade-off between fairness and accuracy in recommendations, AFRL conducts a novel Information Alignment to exactly preserve discriminative information of non-sensitive attributes and incorporate a debiased collaborative embedding into the fair embedding to capture attribute-independent collaborative signals, without loss of fairness. Finally, the extensive experiments conducted on real datasets together with the sound theoretical analysis demonstrate the superiority of AFRL. The codes and datasets are available on <https://github.com/zhuxinyu2700/AFRL>.

CCS CONCEPTS

• Information systems → Recommender systems;

KEYWORDS

Fair Recommendation, Personalized Fairness, Recommender Systems

*Xinyu Zhu and Lilin Zhang share the first authorship.

†Ning Yang is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0431-4/24/07...\$15.00

<https://doi.org/10.1145/3626772.3657709>

ACM Reference Format:

Xinyu Zhu, Lilin Zhang, and Ning Yang. 2024. Adaptive Fair Representation Learning for Personalized Fairness in Recommendations via Information Alignment. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, July 14–18, 2024, Washington, DC, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3626772.3657709>

1 INTRODUCTION

In the era of exponential information growth, recommender systems play a critical role in alleviating the problem of information overloading [15, 17, 29]. Despite the great success of recommender systems, existing studies have shown that machine learning based recommendation models often suffer from unfair recommendations due to the bias in training data [8, 19, 31, 38]. In recent years, many efforts have been made for fair recommendations from various perspectives [2, 10, 11, 42], among which user-side fairness attracts special attention due to its significance to the improvement of user experience.

A range of approaches have been proposed for user-side fair recommendations, following various technical lines including Pareto optimization [26], adversarial training [24, 36, 37], and disentangled representation learning [13, 30, 33, 40]. The common idea of the existing works is to filter out the information of user sensitive attributes from the learned fair user embeddings, so that recommendations can be made independently of these sensitive attributes. However, the traditional approaches often assume that all users share the common sensitive attributes and consider the user fairness requirement, represented by a collection of user sensitive attributes, as hyper-parameter specified in advance, which limits the model to serving only one fairness requirement. In real world, however, it is the fact that different users are sensitive to different attributes and have *personalized fairness* requirements [24]. For example, some users may want the recommendations to be made for them without consideration of their racial information, while some others want the recommendations to be made without gender bias.

Recently, a few methods have been proposed for personalized fairness in recommendations. For example, to meet diverse fairness requirements, Li *et al.* [24] propose to train a filter for each possible combination of sensitive attributes. Wu *et al.* [37] propose a model PFRec, which builds a set of prompt-based bias eliminators and adapters with customized attribute-specific prompts to learn fair embeddings for different attribute combinations. Creager *et al.* [7] propose FFVAE, a disentangled representation learning model that separates representations into sensitive and non-sensitive subspaces. It addresses personalized fairness by excluding from the

learned fair embeddings the relevant semantic factors corresponding to different sensitive attributes specified by user fairness requirements. However, the problem of personalized fairness in recommendations is still far from being well solved partly due to the following challenges:

(1) **Unacceptable training cost caused by exponential combinations of attributes** Essentially, most existing methods for personalized fairness still treat a fairness requirement as a hyper-parameter, which makes them have to train and store a model for each combination of attributes in a brute force way [24, 37]. Such inflexibility incurs a huge training cost due to the explosion of attribute combinations, which leads to the impracticability of the existing methods.

(2) **Compromise on recommendation accuracy** It is well-recognized by previous works that gains in fairness necessarily come at the cost of losses in accuracy [14, 18]. However, in order to completely erase the sensitive attribute information from the fair embeddings, the existing approaches also remove information that overlaps with non-sensitive attributes [7, 24, 37]. This inevitably causes the loss of the discriminative information of the non-sensitive attributes and results in a suboptimal trade-off between accuracy and fairness. Therefore, it is worth exploring better solutions to achieve personalized fairness with less compromise on recommendation accuracy.

In this paper, to address the above challenges, we propose a novel fair model called Adaptive Fair Representation Learning (AFRL) for personalized fairness recommendation. Particularly, to overcome the challenge of the unacceptable training cost incurred by the explosion of attribute combinations, AFRL treats a user fairness requirement as input, rather than hyper-parameter, by which AFRL can adaptively generate the fair embeddings for different users by fusing the corresponding embeddings of non-sensitive attributes specified by the fairness requirement. The generation of fair embeddings is dynamically controlled by the fairness requirements, which endows AFRL with the adaptability during the inference phase to the exponential combinations of attributes.

To strike a better trade-off between fairness and accuracy, we propose an Information Alignment Module (IAlignM) for AFRL. This module is designed to enable the learned fair embeddings to retain discriminative information from non-sensitive attributes and unbiased collaborative signals from the user’s interaction history. In contrast to existing approaches, where removing sensitive attributes results in the loss of discriminative information from non-sensitive attributes, IAlignM adopts a novel strategy. It learns attribute-specific embeddings from the original user embeddings for each attribute, ensuring the precise alignment of information between the embeddings and their respective attributes. Here *information alignment* refers to preserving and only preserving the discriminative information related to the attributes within the corresponding attribute-specific embeddings. For this purpose, we employ a bilevel optimization approach to ensure that an attribute-specific embedding exactly encodes the information of its corresponding attribute, with a theoretical guarantee rooted in information bottleneck principle. At the same time, IAlignM will further improve the recommendation accuracy by learning a debiased collaborative embedding to explicitly capture into the fair user embeddings the

collaborative signals where bias to any user attribute is filtered out to avoid the loss of fairness.

It is worth mentioning that AFRL is recommendation model-agnostic, which means it can serve as a plugin for an off-the-shelf recommendation model. In this paper, we will test our AFRL for non-sequential and sequential recommendation tasks. In summary, our contributions can be outlined as follows:

- We propose a novel fair model called Adaptive Fair Representation Learning (AFRL) which achieves a real personalized fairness in recommendations due to its advantage of training only one model to adaptively serve different fairness requirements during inference phase.
- To achieve a better trade-off between fairness and accuracy in recommendations, we propose an Information Alignment Module (IAlignM). IAlignM reduces the loss of recommendation accuracy by exactly preserving discriminative information of non-sensitive attributes and incorporating a debiased collaborative embedding into the fair embedding to capture attribute-independent collaborative signals.
- At last, the extensive experiments conducted on real datasets together with the theoretical analysis demonstrate the superiority of AFRL.

2 PRELIMINARIES

2.1 Basic Notations and Definitions

We denote user set by \mathcal{U} and item set by \mathcal{V} . Let $O \in \{0, 1\}^{|\mathcal{U}| \times |\mathcal{V}|}$ be the interaction matrix, where a cell $o_{u,v} = 1$ indicates the observed interaction of user $u \in \mathcal{U}$ with item $v \in \mathcal{V}$, and the u th row $O_{u,*}$ represents the historical user-item interaction set of user u . Let \mathcal{A} be the user attribute matrix of $|\mathcal{U}| \times M$, where M is the number of the attributes of a user, and the cell at u th row and i th column, $a_{u,i}$, is the i th attribute of user u .

2.2 Formulation of the Target Problem

Let $y_{u,v} = R(\mathbf{u}, \mathbf{v})$ be a given recommendation model, which infers the probability $y_{u,v}$ that user $u \in \mathcal{U}$ will interact with item $v \in \mathcal{V}$ based on the user embedding $\mathbf{u} \in \mathbb{R}^{d \times 1}$ and the item embedding $\mathbf{v} \in \mathbb{R}^{d \times 1}$, where d is the embedding dimensionality. Let $\mathbf{s}_u \in \{0, 1\}^{Md \times 1}$ representing the fairness requirement of user u , where the i th piece $\mathbf{s}_{u,[1+(i-1)d:i \times d]} = 0$ means the i th attribute is a sensitive attribute of user u , otherwise a non-sensitive attribute. The target problem can be formulated as follow:

Given an unfair recommendation model R , the original user embeddings $\{\mathbf{u}\}$ and item embeddings $\{\mathbf{v}\}$ generated by R , the historical interaction matrix O , and the user attribute matrix \mathcal{A} , we want to train a fair model $FairR$ to generate the fair user embedding \mathbf{u}^* for user u based on u ’s original embedding \mathbf{u} and fairness requirement \mathbf{s}_u , i.e., $\mathbf{u}^* = FairR(\mathbf{u}, \mathbf{s}_u)$.

2.3 Counterfactual Fairness

In this paper, the recommendations made based on \mathbf{u}^* are supposed to be independent of the user’s sensitive attributes specified by \mathbf{s}_u , which is a kind of counterfactual fairness [20]. Intuitively, counterfactual fairness requires the recommendations made in the imaginary counterfactual world, where only the user’s sensitive

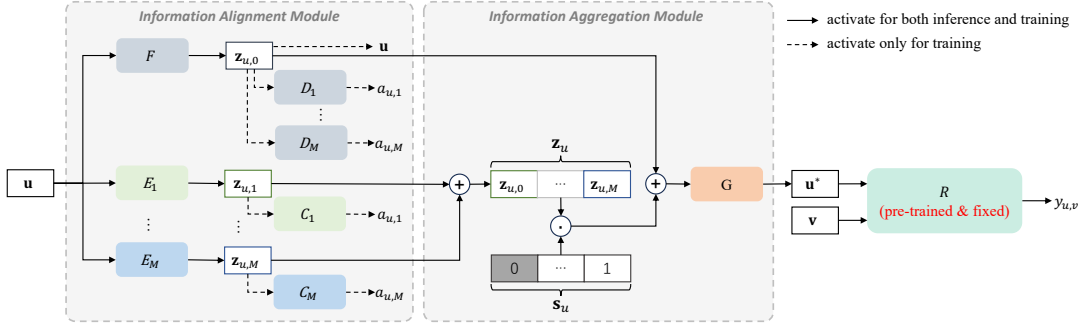


Figure 1: The overview of AFRL. AFRL contains **Information Alignment Module (IAlignM)** and **Information Aggregation Module (IAggM)**. IAlignM generates M attribute-specific embeddings $\{z_{u,i}\}$ ($1 \leq i \leq M$) and a debiased collaborative embedding $z_{u,0}$ from the user embedding u , using attribute encoders $\{E_i\}$ ($1 \leq i \leq M$) and a debiased collaborative signal encoder F , respectively. IAggM aggregates these embeddings with respect to the personalized fairness requirement s_u to produce the final fair user embedding u^* for downstream recommendation.

attributes are modified, should be the same as in the real world. The formal definition of counterfactual fairness in recommender systems is provided as follow:

DEFINITION 1 (COUNTERFACTUALLY FAIR RECOMMENDATION [24]). A recommender model is counterfactually fair if for any possible user with non-sensitive attributes $A_n = a_n$ and sensitive attributes $A_s = a_s$:

$$P(Y_{a_s} | A_n = a_n, A_s = a_s) = P(Y_{a'_s} | A_n = a_n, A_s = a_s)$$

for all Y and for any value a_s attainable by A_s , where Y denotes the generated recommendation results.

3 THE PROPOSED MODEL

3.1 Overview

Fig. 1 gives an overview of the architecture of AFRL. As shown in Fig. 1, AFRL takes as inputs the original user embedding u and the personalized fairness requirement s_u , and produces the fair embedding u^* . We can see that AFRL consists of two modules, the Information Alignment Module (IAlignM) and the Information Aggregation Module (IAggM). IAlignM is responsible for learning M attribute-specific embeddings $\{z_{u,i}\}$ and one debiased collaborative embedding $z_{u,0}$ from the original user embedding u via M **attribute encoders** $\{E_i\}$ ($1 \leq i \leq M$) and a **debiased collaborative signal encoder** F , respectively. In particular, to ensure the **information alignment** between $z_{u,i}$ and the i th attribute $a_{u,i}$ of user u , IAlignM will maximize the mutual information between $z_{u,i}$ and $a_{u,i}$, and minimize the mutual information between $z_{u,i}$ and u , using a bilevel optimization for attribute encoder E_i together with the auxiliary attribute classifier C_i . At the same time, to generate the debiased collaborative embedding $z_{u,0}$, IAlignM will ensure F to remove the correlation between $z_{u,0}$ and the user attributes via an adversarial training together with M attribute discriminators $\{D_i\}$ ($1 \leq i \leq M$), and encode the collaborative signals by maximizing the ability of $z_{u,0}$ to reconstruct the original user embedding u .

Once the attribute-specific embeddings and the debiased collaborative embedding are produced, IAggM will mask the embeddings of sensitive attributes with respect to s_u and fuse the embeddings

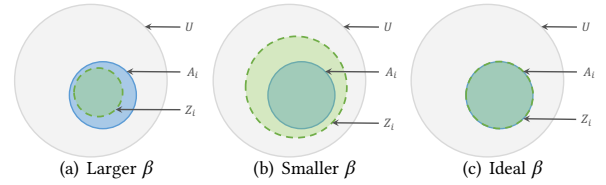


Figure 2: Illustration of the attribute-specific embedding learning for Information Alignment, where (a) and (b) are the cases of information misalignment, while (c) is the optimal case of information alignment.

of non-sensitive attributes with the debiased collaborative embedding using an MLP G to generate the fair embedding u^* . To make u^* legal for the given recommendation model R , the training of IAggM will be supervised by the historical interactions O through R , which locates u^* in the same embedding space as the original user embedding u . Finally, u^* and the item embedding v are fed to R to make a fair prediction of $y_{u,v}$.

3.2 Information Alignment Module

As mentioned before, IAlignM learns attribute-specific embeddings $\{z_{u,1}, \dots, z_{u,M}\}$ and debiased collaborative embedding $z_{u,0}$ from the original user embedding u . Since u is generated by the given recommendation model R , it is reasonable to assume that u has encoded the attribute information and the collaborative signals of user u in an entangling and biased way.

3.2.1 Attribute-Specific Embedding. The attribute-specific embedding $z_{u,i}$ of the i th attribute of user u is extracted by the attribute encoder E_i from the original user embedding u , $1 \leq i \leq M$, i.e.,

$$z_{u,i} = E_i(u). \quad (1)$$

To achieve information alignment, our objective is to make E_i able to encode and only encode the complete information of the i th attribute of users. Since E_i is shared by all users, the output of E_i can be represented by a stochastic variable Z_i with various

values $\{z_{u,i} | u \in \mathcal{U}\}$. Meanwhile, let A_i be the stochastic variable representing the i th column of the user attribute matrix \mathcal{A} , of which the values are $\{a_{u,i} | u \in \mathcal{U}\}$, and let U be the stochastic variable representing different user embeddings $\{\mathbf{u}\}$. Obviously, $Z_i = E_i(U)$. In the light of the information bottleneck principle [34], the optimization objective of E_i is to maximize the mutual information between Z_i and A_i , and simultaneously minimize that between Z_i and U , i.e.,

$$\min_{E_i} -I(Z_i; A_i) + \beta I(Z_i; U), \quad (2)$$

where $I(\cdot, \cdot)$ is mutual information function, β is a balance factor to control the amount of information from U encoded by Z_i .

Fig. 2 illustrates the insight of the attribute-specific embedding learning for information alignment defined by Equation (2), where cycle areas represent the information content of corresponding variables. We can see that the amount of the information encoded in Z_i can be regularized by β . Particularly, as shown in Figs. 2(a) and 2(b), the smaller the β , the more information about U covered by Z_i . Ideally, by regularizing β , the optimal solution E_i to Equation (2) will cause that the area of Z_i exactly overlaps with the area of A_i , as shown in Fig. 2(c). At this ideal time, (1) Z_i covers the information of A_i , which makes the discriminative information of A_i be reserved completely in Z_i ; (2) Z_i only covers the information of A_i , which is favorable to fairness since the correlation between Z_i and other attributes is reduced to minimum. In sharp contrast, the existing works pursue extreme fairness by completely removing the information of sensitive attributes, which will result in information about A_i not being preserved intact if it is a non-sensitive attribute, because the parts of it that overlap with sensitive information are also removed. As we will see later in the experiments, exactly preserving the information of non-sensitive attributes, including the parts overlapping with the information of sensitive attributes, will increase the discriminability of the fair embeddings without loss of fairness, which helps AFRL achieve a better trade-off between fairness and accuracy.

However, the evaluation of mutual information is difficult. To overcome this challenge, we adopt different strategies to approximately maximize $I(Z_i; A_i)$ and minimize $I(Z_i; U)$, respectively.

(1) **The minimization of $I(Z_i; U)$** At first, the definition of $I(Z_i; U)$ is

$$I(Z_i; U) = \mathbb{E}_{(z_{u,i}, \mathbf{u}) \sim p(Z_i, U)} \left[\log \frac{p(z_{u,i} | \mathbf{u})}{p(z_{u,i})} \right]. \quad (3)$$

As previous works did [25, 27], we first assume $p(z_{u,i}) \sim \mathcal{N}(0, \mathbf{I})$. To make Equation (3) tractable, our idea is to view the output of E_i as a sample from a special Gaussian distribution with 0 variance, i.e., $p(z_{u,i} | \mathbf{u}) \sim \mathcal{N}(\mu_u, 0)$. Then the optimal E_i is

$$\argmin_{E_i} I(Z_i; U) = \mathbb{E}_{\mathbf{u} \sim p(U)} \frac{1}{2} \|\mu_u\|_2^2, \quad (4)$$

where $\mu_u = z_{u,i} = E_i(\mathbf{u})$ and the expectation can be evaluated via Monte Carlo method.

(2) **The maximization of $I(Z_i; A_i)$** The definition of $I(Z_i; A_i)$ is

$$I(Z_i; A_i) = \mathbb{E}_{(z_{u,i}, a_{u,i}) \sim p(Z_i, A_i)} \left[\log \frac{p(a_{u,i} | z_{u,i})}{p(a_{u,i})} \right], \quad (5)$$

where $a_{u,i}$ represents the value of the i th attribute of a user u , and $p(a_{u,i})$ and $p(a_{u,i} | z_{u,i})$ are the priori and posterior probabilities of $a_{u,i}$, respectively. Since $p(a_{u,i} | z_{u,i})$ is intractable, we introduce the predicted probability $q(a_{u,i} | z_{u,i}; C_i)$ offered by the classifier C_i :

$$\begin{aligned} I(Z_i; A_i) &= \mathbb{E}_{(z_{u,i}, a_{u,i}) \sim p(Z_i, A_i)} \left[\log \frac{p(a_{u,i} | z_{u,i}) q(a_{u,i} | z_{u,i}; C_i)}{p(a_{u,i}) q(a_{u,i} | z_{u,i}; C_i)} \right] \\ &= \Phi(Z_i, C_i) + \Delta(Z_i, C_i), \end{aligned} \quad (6)$$

where

$$\begin{aligned} \Phi(Z_i, C_i) &= \mathbb{E}_{(z_{u,i}, a_{u,i}) \sim p(Z_i, A_i)} \left[\log \frac{q(a_{u,i} | z_{u,i}; C_i)}{p(a_{u,i})} \right], \\ \Delta(Z_i, C_i) &= \mathbb{E}_{z_{u,i} \sim p(Z_i)} [\text{KL}(p(a_{u,i} | z_{u,i}) \| q(a_{u,i} | z_{u,i}; C_i))]. \end{aligned} \quad (7)$$

To maximize $I(Z_i; A_i)$, we borrow the idea of EM algorithm to iteratively update C_i and Z_i , i.e., at $t + 1$ time step,

$$C_i^{t+1} = \argmin_{C_i^t} \Delta(Z_i^t, C_i^t), \quad (8)$$

$$Z_i^{t+1} = \argmax_{Z_i^t} \Phi(Z_i^t, C_i^{t+1}). \quad (9)$$

By some simple derivations, it is easy to know Equation (8) is equivalent to

$$C_i^{t+1} = \argmin_{C_i} \mathbb{E}_{(z_{u,i}, a_{u,i}) \sim p(Z_i^t, A_i)} [-\log q(a_{u,i} | z_{u,i}; C_i)], \quad (10)$$

and Equation (9) is equivalent to

$$Z_i^{t+1} = \argmin_{Z_i^t} \mathbb{E}_{(z_{u,i}, a_{u,i}) \sim p(Z_i^t, A_i)} [-\log q(a_{u,i} | z_{u,i}; C_i^{t+1})]. \quad (11)$$

By combining Equations (4), (10) and (11), and considering Z_i is updated through E_i since $Z_i = E_i(U)$, we can obtain the following bilevel optimization objective of the attribute encoders:

$$\begin{aligned} \min_{E_i} \min_{C_i} \mathbb{E}_{\mathbf{u} \sim p(U)} \frac{1}{2} \|\mathbf{z}_{u,i}\|_2^2 \\ + \beta \mathbb{E}_{(z_{u,i}, a_{u,i}) \sim p(Z_i, A_i)} [-\log q(a_{u,i} | z_{u,i}; C_i)]. \end{aligned} \quad (12)$$

At the same time, note that the sampling process $(z_{u,i}, a_{u,i}) \sim p(Z_i, A_i)$ is fulfilled by first sampling a user's embedding $\mathbf{u} \sim p(U)$ and then obtaining $(z_{u,i} = E_i(\mathbf{u}), a_{u,i})$. Hence the sampling in the second term in Equation (12) can be equivalently changed to be the same as that in the first term, which leads to the following equivalent tractable bilevel optimization:

$$\min_{E_i} \min_{C_i} \mathbb{E}_{\mathbf{u} \sim p(U)} \left[\frac{1}{2} \|\mathbf{z}_{u,i}\|_2^2 - \beta \log q(a_{u,i} | z_{u,i}; C_i) \right]. \quad (13)$$

3.2.2 Debaised Collaborative Embedding. As mentioned before, the original unfair user embedding \mathbf{u} provided by the given unfair recommendation model R has encoded the collaborative signals biased to user attributes. IAlignM employs a debaised collaborative signal encoder F to generate the debaised collaborative embedding $\mathbf{z}_{u,0}$ to capture from \mathbf{u} the collaborative signals independent of user attributes. For this purpose, we optimize F with two objectives: (1) minimizing the discriminability of $\mathbf{z}_{u,0}$ to user attributes for fairness, and (2) maximizing the information of \mathbf{u} encoded in $\mathbf{z}_{u,0}$ to preserve the collaborative signals as much as possible for recommendation accuracy.

To achieve the objective (1), we introduce a discriminator D_i for i th attribute, $1 \leq i \leq M$, and adjust F to fool D_i with the following adversarial training between F and D_i :

$$\min_F \max_{D_i} -\mathbb{E}_{\mathbf{u} \sim p(U)} \sum_{i=1}^M [-\log q(a_{u,i} | \mathbf{z}_{u,0}; D_i)], \quad (14)$$

where $q(a_{u,i} | \mathbf{z}_{u,0}; D_i)$ is the probability that $\mathbf{z}_{u,0}$ is classified as $a_{u,i}$ by D_i .

To achieve the objective (2), we minimize the following reconstruction loss:

$$\min_F \mathbb{E}_{\mathbf{u} \sim p(U)} \|\mathbf{z}_{u,0} - \mathbf{u}\|_2^2. \quad (15)$$

By combining Equations (14) and (15), we obtain the following optimization objective of F :

$$\min_F \max_{D_i} \mathbb{E}_{\mathbf{u} \sim p(U)} \|\mathbf{z}_{u,0} - \mathbf{u}\|_2^2 - \lambda \sum_{i=1}^M [-\log q(a_{u,i} | \mathbf{z}_{u,0}; D_i)], \quad (16)$$

where λ controls the trade-off between recommendation accuracy and fairness.

3.3 Information Aggregation Module

After obtaining attribute-specific embeddings $\{\mathbf{z}_{u,i}\}$ ($1 \leq i \leq M$) and the debiased collaborative embedding $\mathbf{z}_{u,0}$, the Information Aggregation Module (IAgM), which is implemented as an MLP G , combines these embeddings based on the personalized fairness requirement \mathbf{s}_u to produce the final fair embedding \mathbf{u}^* as follow:

$$\mathbf{u}^* = G\left\{\mathbf{z}_{u,0} \oplus [(\mathbf{z}_{u,1} \oplus \dots \oplus \mathbf{z}_{u,M}) \odot \mathbf{s}_u]\right\}, \quad (17)$$

where \oplus represents concatenation, \odot represents element-wise product, and $\mathbf{s}_u \in \{0, 1\}^{M \times 1}$ represents the fairness requirement of user u . The i th part $\mathbf{s}_{u,[1+(i-1)d:i \times d]} = 0$ means the i th attribute is a sensitive attribute of user u , otherwise a non-sensitive attribute.

To ensure \mathbf{u}^* to be compatibility with the original recommendation model R , G will be adjusted with respect to the supervision offered by R as follow:

$$\min_G - \sum_{(u, v^+, v^-)} \ln \sigma(y_{u,v^+} - y_{u,v^-}), \quad (18)$$

where v^+ and v^- are a positive example and a negative example of u , respectively, $\sigma(\cdot)$ is the sigmoid function, and $y_{u,v} = R(\mathbf{u}^*, \mathbf{v})$. Note that Equation (18) is a function of G rather than R , which means it is G that will be adjusted during the optimization of Equation (18) with fixed R . Algorithm 1 gives the training process of AFRL.

3.4 Theoretical Justification

3.4.1 Convergence Analysis. Now we analyze the convergence of the optimization of Equation (13), which is the tractable approximation of Equation (2) to realize the information alignment. At first, it is easy to see that $I(Z_i; U)$ will achieve its minimum as defined in Equation (4). So, we only justify the iterative optimization defined Equations (8), (9) will lead to the maximum of $I(Z_i; A_i)$, which equivalently to prove $I(Z_i^{t+1}; A_i) - I(Z_i^t; A_i) \geq 0$. At first,

Algorithm 1 Training of AFRL

Input: user-item interaction matrix O , user attribute matrix \mathcal{A} , unfair recommendation model R , and the set of original user embeddings U and item embeddings V generated by R ;

Output: F , $\{E_i\}$ ($1 \leq i \leq M$), and G ;

```

1: Randomly initialize  $F$ ,  $\{D_i\}$ ,  $\{E_i\}$ ,  $\{C_i\}$  and  $G$ ,  $1 \leq i \leq M$ ;
2: for 1 to  $T$  do
3:   for each  $\mathbf{u} \in U$  do
4:     obtain  $\mathbf{z}_{u,0} = F(\mathbf{u})$  and  $\{\mathbf{z}_{u,i} = E_i(\mathbf{u})\}$ ;
5:     randomly sample an  $\mathbf{s}_u$  from a uniform distribution over different attribute combinations;
6:     obtain  $\mathbf{u}^*$  by Equation (17);
7:     update  $\{C_i\}$  by inner minimization of Equation (13);
8:     update  $\{E_i\}$  by outer minimization of Equation (13);
9:     update  $\{D_i\}$  by inner maximization of Equation (16);
10:    update  $F$  by outer minimization of Equation (16);
11:    update  $G$  by Equation (18).
12:   end for
13: end for
```

according to Equation (6), we can obtain

$$\begin{aligned} & I(Z_i^{t+1}, A_i) - I(Z_i^t, A_i) \\ &= \underbrace{\Phi(Z_i^{t+1}, C_i) - \Phi(Z_i^t, C_i)}_A + \underbrace{\Delta(Z_i^{t+1}, C_i) - \Delta(Z_i^t, C_i)}_B. \end{aligned} \quad (19)$$

By taking a close look at Equation (6), we can observe that $I(Z_i, A_i)$ is independent of C_i . Based on this observation, we can set C_i to the optimal $C_i^{t+1} = \arg\min_{C_i^t} \Delta(Z_i^t, C_i^t)$ (Equation (8)) without changing $I(Z_i, A_i)$, which gives:

$$\begin{aligned} & I(Z_i^{t+1}, A_i) - I(Z_i^t, A_i) \\ &= \underbrace{\Phi(Z_i^{t+1}, C_i^{t+1}) - \Phi(Z_i^t, C_i^{t+1})}_A + \underbrace{\Delta(Z_i^{t+1}, C_i^{t+1}) - \Delta(Z_i^t, C_i^{t+1})}_B. \end{aligned} \quad (20)$$

Since Δ is a KL-divergence, the optimal C_i^{t+1} makes $\Delta(Z_i^t, C_i^{t+1}) = 0$, and hence $B = \Delta(Z_i^{t+1}, C_i^{t+1}) \geq 0$. At the same time, since Z_i^{t+1} maximizes $\Phi(Z_i^t, C_i^{t+1})$ (see Equation (9)), $\Phi(Z_i^{t+1}, C_i^{t+1}) \geq \Phi(Z_i^t, C_i^{t+1})$, i.e., $A \geq 0$. Thus $I(Z_i^{t+1}, A_i) - I(Z_i^t, A_i) \geq 0$ holds.

3.4.2 Informativeness Analysis. Now we show that AFRL completely extracts the information of a non-sensitive attribute into the fair embedding. Without loss of generality, let A_n , Z_n , U , and U^* be the stochastic variables representing a non-sensitive attribute, the embedding of A_n , the original user embedding, and the fair user embedding, respectively, $1 \leq i \leq M$. Obviously, the information encoded by U can be divided into two parts: $H(U) = I(U; A_n) + H(U|A_n)$, where $I(U; A_n)$ is the part relevant to A_n . We want to prove U^* completely encodes $I(U; A_n)$, which is equal to proving $I(A_n; U|U^*) = 0$.

In terms of the theory of information bottleneck [34], AFRL can be framed as a Markov Chain $A_n \rightarrow U \rightarrow Z_n \rightarrow U^*$. Suppose the attribute encoder E_n is the optimal solution to the bilevel optimization defined in Equation (13), which means $Z_n = E_n(U)$ completely encodes the A_n 's information from U , and equivalently, $I(A_n; U|Z_n) = 0$. Considering $I(A_n; U|Z_n) = I(A_n; U) -$

$I(U; A_n; Z_n) = I(A_n; U) - I(A_n; Z_n)$, we have

$$I(A_n; U) = I(A_n; Z_n). \quad (21)$$

Let Z_s be the stochastic variable representing the embeddings of sensitive attributes, and due to the Markov Chain $Z_s \leftarrow U \rightarrow Z_n$, Z_n and Z_s are conditional independent, i.e., $Z_n \perp Z_s | U$. Meanwhile, the variable of debiased collaborative embedding, Z_0 , is optimized to be independent of any Z_n . Therefore, removing Z_s from and incorporating Z_0 to U^* will not affect the mutual information between Z_n and U^* , $I(Z_n; U^*)$. When E_n and G are both optimal, $I(Z_n; U^*)$ achieves maximum, which results in

$$H(Z_n | U^*) = 0. \quad (22)$$

Based on Equations (21) and (22), we have the following derivations:

$$\begin{aligned} I(A_n; U | U^*) &= I(A_n; U) - I(A_n; U; U^*) \\ &= I(A_n; Z_n) - (I(A_n; U^*) - I(A_n; U^* | U)) \\ &= I(A_n; Z_n) - I(A_n; U^*) \\ &= I(A_n; Z_n) - (I(A_n; U^* | Z_n) + I(A; U^*; Z_n)) \\ &= I(A_n; Z_n) - I(A_n; U^*; Z_n) \\ &= I(A_n; Z_n) - I(Z_n; U^*) + I(Z_n; U^* | A_n) \\ &= (H(Z_n) - H(Z_n | A_n)) - (H(Z_n) - H(Z_n | U^*)) \\ &\quad + (H(Z_n | A_n) - H(Z_n | U^*, A_n)) \\ &= H(Z_n | U^*) - H(Z_n | U^*, A_n) \\ &= -H(Z_n | U^*, A_n) \geq 0. \end{aligned} \quad (23)$$

Note that during the above derivations, $I(A_n; U^* | U)$ and $I(A_n; U^* | Z_n)$ are both equal to 0 because of the Markov Chain $A_n \rightarrow U \rightarrow Z_n \rightarrow U^*$. At the same time, since entropy is non-negative, we also have $-H(Z_n | U^*, A_n) \leq 0$, and thus $H(Z_n | U^*, A_n) = 0$, and consequently $I(A_n; U | U^*) = 0$.

4 EXPERIMENTS

The experiments are designed to address the following research questions:

- **RQ1** How does AFRL perform as compared to the state-of-the-art models?
- **RQ2** How does AFRL make trade-off between fairness and accuracy as compared to the state-of-the-art models?
- **RQ3** How do the attribute-specific embeddings and the debiased collaborative embedding contribute to the performance of AFRL?
- **RQ4** How do the hyper-parameters affect the performance of AFRL?

4.1 Experimental Settings

4.1.1 Datasets. We evaluate AFRL on two datasets:

- **MovieLens-1M¹ (ML-1M)** This dataset consists of 1,000,000 movie ratings provided by 6,040 users with attributes G (Gender), A (Age) and O (Occupation). For simplicity, we convert the ratings into binary values, where ratings not less than 4 are set to 1, while the rest are set to 0.
- **Taobao Display Ad Click² (Taobao)** This dataset comprises more than 26,000,000 interactions between 1,140,000 users and

Table 1: Hyper-parameter setting of AFRL

Method	β is selected from {0.001, 0.01, 0.1, 1, 10, 100}		λ is selected from {0.001, 0.01, 0.1, 1, 10, 100}	
	ML-1M	Taobao	ML-1M	Taobao
AFRL+PMF	0.1	1	0.1	10
AFRL+DeepModel	0.1	1	0.1	1
AFRL+SASRec	1	1	1	10
AFRL+BERT4Rec	1	1	1	10

840,000 advertisements on Taobao website. A user in Taobao is associated with attributes G (Gender), A (Age), and C (Consumption level).

For each dataset, we remove the users with less than 10 instances. We generate interaction sequences for a user with sliding window sizes 100 and 50 on ML-1M and Taobao, respectively. As previous work did [37], for each user, we choose the most recent sequence as a testing instance and the second most recent sequence as a validation instance.

4.1.2 Baselines. We will compare AFRL with following three state-of-the-art models for personalized fairness in recommendations:

- **SM [24]** To meet personalized fairness in recommendations, SM trains a attribute filter for each possible combination of sensitive attributes.
- **PFRec [37]** PFRec builds a set of prompt-based bias eliminators and adapters with predefined attribute-specific prompts to learn fair representations for different attribute combinations.
- **FFVAE [7]** FFVAE disentangles an unfair embedding into latent factors, each of which corresponds to an attribute, and excludes relevant sensitive latent factors based on different fairness requirements of users.

We will check the performance of AFRL and the baseline models in combination with the following four base recommendation models, including two non-sequential models **PMF** [29] and **DeepModel** [6], and two sequential models **SASRec** [17] and **BERT4Rec** [35].

4.1.3 Evaluation Protocols. We evaluate the accuracy of top- N recommendations with two popular metrics Normalized Discounted Cumulative Gain (N@10) and Hit rate (H@10). Following the idea in [17, 35, 37], we sample 1 and 99 negative examples for a positive example during training and testing, respectively. To evaluate the fairness, following the popular method used in the existing works [16, 24], we first train a surrogate classifier for each attribute using the fair embeddings generated by AFRL and the baseline models, respectively, and report the average AUC over all classifiers as the fairness metric. The closer to 0.5 AUC, the better fairness.

4.1.4 Parameter Setting. We set the batch size to 256, the initial learning rate to $5e-5$, and the embedding dimensionality to 64 on both datasets. We employ early stopping as the regularizer for avoiding overfitting, and Adam algorithm as the optimizer. F and $\{E_i\}$ are implemented as a six-layer MLP using ReLU as the activation function. $\{D_i\}$ and $\{C_i\}$ are implemented as a two-layer MLP using ReLU as the activation function. Table 1 gives the settings of the hyper-parameters β and λ in Equations (2) and (16) in different scenarios. For fairness, the hyper-parameters of the baseline models are set to their optimal configuration tuned on validation sets.

¹<https://grouplens.org/datasets/movielens/>

²<https://tianchi.aliyun.com/dataset/56>

Table 2: Fairness (AUC) and Accuracy (N@10, H@10) comparison on ML-1M. A fairness requirement is represented by an attribute combination where each attribute is assumed to be sensitive to users. For instance, G+A+O represents Gender, Age and Occupation are sensitive attributes. The best runs per metric are marked in boldface.

Method	Fairness Requirement																				
	G			A			O			G + A			G + O			A + O			G + A + O		
	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑
PMF	0.7890	0.4157	0.6918	0.7908	0.4157	0.6918	0.6703	0.4157	0.6918	0.7863	0.4157	0.6918	0.7211	0.4157	0.6918	0.7220	0.4157	0.6918	0.7438	0.4157	0.6918
SM+PMF	0.6547	0.3733	0.6492	0.6394	0.3739	0.6496	0.5703	0.3690	0.6467	0.6382	0.3703	0.6471	0.5890	0.3711	0.6498	0.5828	0.3682	0.6440	0.5915	0.3699	0.6464
FFVAE+PMF	0.6433	0.3122	0.5625	0.6843	0.2993	0.5409	0.6036	0.2772	0.5048	0.7261	0.2979	0.5418	0.6553	0.2765	0.5034	0.6258	0.2684	0.4874	0.6394	0.2664	0.4844
AFRL+PMF	0.5310	0.3944	0.6707	0.5238	0.3908	0.6655	0.5112	0.3945	0.6729	0.5274	0.3884	0.6622	0.5211	0.3915	0.6669	0.5175	0.3881	0.6601	0.5220	0.3846	0.6555
DeepModel	0.7656	0.3683	0.6397	0.7365	0.3683	0.6397	0.6542	0.3683	0.6397	0.7589	0.3683	0.6397	0.7113	0.3683	0.6397	0.6876	0.3683	0.6397	0.7202	0.3683	0.6397
SM+DeepModel	0.6529	0.3403	0.6063	0.6319	0.3346	0.6047	0.5709	0.3356	0.6009	0.6495	0.3299	0.5949	0.6169	0.3314	0.5952	0.6024	0.3263	0.5906	0.6158	0.3205	0.5802
FFVAE+DeepModel	0.6289	0.2981	0.5204	0.6491	0.2954	0.5137	0.5000	0.2955	0.5145	0.6258	0.2957	0.5125	0.5647	0.2925	0.5095	0.5000	0.2928	0.509	0.5798	0.2902	0.5029
AFRL+DeepModel	0.5358	0.3504	0.6151	0.5018	0.3474	0.6103	0.5231	0.3525	0.6202	0.5188	0.3430	0.6024	0.5295	0.3471	0.6108	0.5125	0.3438	0.6050	0.5203	0.3377	0.5921
SASRec	0.7554	0.5709	0.8103	0.7259	0.5709	0.8103	0.6143	0.5709	0.8103	0.7284	0.5709	0.8103	0.6879	0.5709	0.8103	0.6634	0.5709	0.8103	0.6883	0.5709	0.8103
SM+SASRec	0.6366	0.4808	0.7474	0.6183	0.4792	0.7447	0.5917	0.4803	0.7454	0.6021	0.4790	0.7427	0.5947	0.4813	0.7486	0.5968	0.4773	0.7378	0.6229	0.4746	0.7382
PFRec+SASRec	0.5179	0.4784	0.7418	0.5159	0.4684	0.7341	0.5288	0.4499	0.7044	0.5146	0.4420	0.7002	0.5179	0.4478	0.7069	0.5073	0.3798	0.5825	0.5173	0.4187	0.6830
FFVAE+SASRec	0.5677	0.4190	0.6585	0.6540	0.4035	0.6376	0.6029	0.3728	0.5995	0.6433	0.4027	0.6383	0.6513	0.3727	0.5966	0.5631	0.3598	0.5785	0.5962	0.3614	0.5829
AFRL+SASRec	0.5543	0.5571	0.7980	0.5346	0.5543	0.7953	0.5200	0.5595	0.7982	0.5445	0.5518	0.7945	0.5372	0.5559	0.7953	0.5273	0.5526	0.7934	0.5463	0.5493	0.7903
BERT4Rec	0.7257	0.5657	0.8103	0.6946	0.5657	0.8103	0.6331	0.5657	0.8103	0.7171	0.5657	0.8103	0.6894	0.5657	0.8103	0.6686	0.5657	0.8103	0.6922	0.5657	0.8103
SM+BERT4Rec	0.5485	0.4764	0.7393	0.5674	0.4731	0.7370	0.5668	0.4744	0.7373	0.6200	0.4660	0.7297	0.6007	0.4702	0.7299	0.6112	0.4587	0.7244	0.6063	0.4606	0.7235
PFRec+BERT4Rec	0.5258	0.4875	0.7400	0.5147	0.4576	0.7113	0.5277	0.4699	0.7183	0.5351	0.4438	0.6963	0.5139	0.4510	0.7037	0.5179	0.3868	0.5936	0.5278	0.4247	0.6765
FFVAE+BERT4Rec	0.6506	0.4229	0.6790	0.5933	0.4090	0.6595	0.5963	0.3528	0.5817	0.6263	0.4080	0.6592	0.6018	0.3549	0.5850	0.5567	0.3431	0.5625	0.5765	0.3434	0.5691
AFRL+BERT4Rec	0.5461	0.5390	0.7906	0.5066	0.5326	0.7830	0.5248	0.5405	0.7903	0.5264	0.5285	0.7798	0.5354	0.5381	0.7897	0.5157	0.5310	0.7817	0.5258	0.5207	0.7770

Table 3: Fairness and Accuracy comparison on Taobao.

Method	Fairness Requirement																				
	G			A			C			G + A			G + C			A + C			G + A + C		
	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑	AUC↓	N@10↑	H@10↑
PMF	0.7144	0.5329	0.7839	0.6872	0.5329	0.7839	0.6406	0.5329	0.7839	0.7007	0.5329	0.7839	0.6785	0.5329	0.7839	0.6707	0.5329	0.7839	0.6829	0.5329	0.7839
SM+PMF	0.5979	0.4892	0.7420	0.5345	0.5001	0.7546	0.5577	0.4701	0.7270	0.5798	0.4811	0.7364	0.6089	0.4654	0.7271	0.5832	0.4538	0.7169	0.5783	0.4797	0.7313
FFVAE+PMF	0.6867	0.4382	0.6939	0.6212	0.4209	0.6734	0.5360	0.3806	0.6068	0.6325	0.4168	0.6653	0.5767	0.3762	0.6032	0.5456	0.3783	0.6058	0.5722	0.3720	0.5983
AFRL+PMF	0.5659	0.4998	0.7529	0.5443	0.4983	0.7521	0.5073	0.5014	0.7550	0.5551	0.4908	0.7445	0.5366	0.4943	0.7473	0.5258	0.4929	0.7469	0.5391	0.4856	0.7492
DeepModel	0.6582	0.4972	0.7485	0.6016	0.4972	0.7485	0.5652	0.4972	0.7485	0.6301	0.4972	0.7485	0.6095	0.4972	0.7485	0.5835	0.4972	0.7485	0.6082	0.4972	0.7485
SM+DeepModel	0.6057	0.4714	0.7244	0.5541	0.4734	0.7248	0.6066	0.4725	0.7237	0.5800	0.4715	0.7244	0.6010	0.4728	0.7256	0.5803	0.4723	0.7247	0.5845	0.4615	0.7136
FFVAE+DeepModel	0.6221	0.4185	0.6624	0.6392	0.4158	0.6573	0.5175	0.3997	0.6304	0.6133	0.4166	0.6582	0.5441	0.4009	0.6331	0.5337	0.3990	0.6345	0.5447	0.4011	0.6336
AFRL+DeepModel	0.5237	0.4879	0.7444	0.5335	0.4897	0.7389	0.5085	0.4894	0.7380	0.5310	0.4831	0.7303	0.5166	0.4828	0.7292	0.5204	0.4847	0.7359	0.5209	0.4801	0.7273
SASRec	0.7080	0.5399	0.7888	0.6656	0.5399	0.7888	0.5591	0.5399	0.7888	0.6832	0.5399	0.7888	0.6264	0.5399	0.7888	0.6171	0.5399	0.7888	0.6418	0.5399	0.7888
SM+SASRec	0.6305	0.4827	0.7346	0.5868	0.4889	0.7393	0.5762	0.4860	0.7373	0.6121	0.4845	0.7377	0.6042	0.4848	0.7377	0.5826	0.4841	0.7365	0.6011	0.4854	0.7379
PFRec+SASRec	0.6160	0.4957	0.7478	0.6027	0.4848	0.7373	0.5373	0.4945	0.7457	0.6086	0.4808	0.7318	0.5703	0.4860	0.7380	0.5655	0.4851	0.7375	0.5789	0.4794	0.7314
FFVAE+SASRec	0.6506	0.4429	0.6990	0.5933	0.4290	0.6795	0.5500	0.3728	0.6017	0.6000	0.4280	0.6792	0.5700	0.3749	0.6050	0.5567	0.3631	0.5825	0.5765	0.3634	0.5891
AFRL+SASRec	0.5942	0.5266	0.7767	0.5699	0.5263	0.7765	0.5133	0.5271	0.7773	0.5821	0.5211	0.7712	0.5537	0.5216	0.7717	0.5416	0.5212	0.7714	0.5691	0.5161	0.7663
BERT4Rec	0.7119	0.5417	0.7900	0.6752	0.5417	0.7900	0.5863	0.5417	0.7900	0.6895	0.5417	0.7900	0.6331	0.5417	0.7900	0.6282	0.5417	0.7900	0.6543	0.5417	0.7900
SM+BERT4Rec	0.6305	0.4627	0.7246	0.5868	0.4689	0.7293	0.5762	0.4660	0.7273	0.6121	0.4645	0.7277	0.6042	0.4648	0.7277	0.5826	0.4641	0.7265	0.6011	0.4654	0.7279
PFRec+BERT4Rec	0.6020	0.4938	0.7449	0.5810	0.4891	0.7389	0.5451	0.4964	0.7491	0.5803	0.4938	0.7451	0.5724	0.4934	0.7456	0.5544	0.4709	0.7206	0.5737	0.4873	0.7323
FFVAE+BERT4Rec	0.6638	0.4269	0.6760	0.6519	0.4254	0.6746	0.5784	0.3740	0.6045	0.6462	0.4174	0.6644	0.6027	0.3714	0.6015	0.6014	0.3721	0.6042	0.5972	0.3660	0.5945
AFRL+BERT4Rec	0.5879	0.5273	0.7772	0.5799	0.5263	0.7784	0.5082	0.5307	0.7810	0.5839	0.5212	0.7725	0.5480	0.5220	0.7720	0.5440	0.5213	0.7730	0.5586	0.5111	0.7618

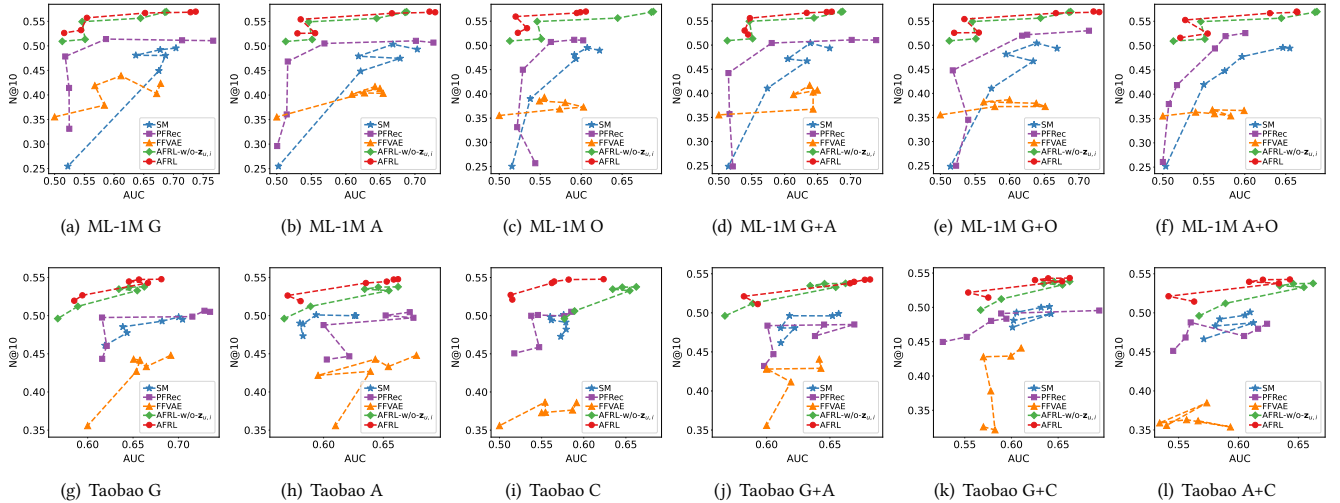


Figure 3: Pareto front curves under different fairness requirements with SASRec [17] as base model. In each curve, the points from left to right correspond to $\lambda \in \{100, 10, 1, 0.1, 0.01, 0.001\}$, respectively.

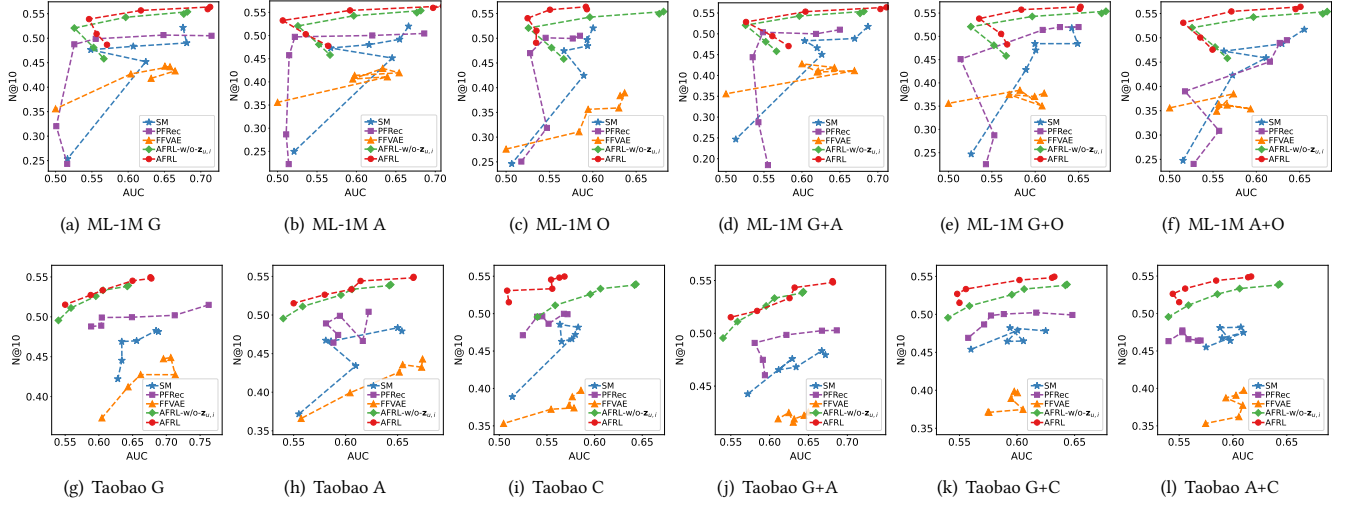


Figure 4: Pareto front curves under different fairness requirements with BERT4Rec [35] as base model. In each curve, the points from left to right correspond to $\lambda \in \{100, 10, 1, 0.1, 0.01, 0.001\}$, respectively.

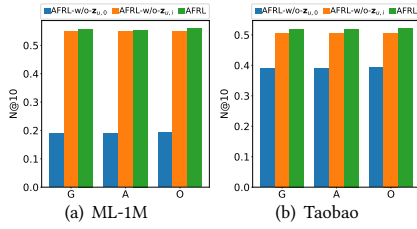


Figure 5: Contributions to accuracy by $z_{u,0}$ and $z_{u,i}$.

4.2 Fairness and Accuracy (RQ1)

Tables 2 and 3 show the performances of AFRL and the baseline models on ML-1M and Taobao, respectively. We can see that on both datasets, compared with the AUC of original model, the AUC after applying any fair model decreases, which means all the fair models can improve the fairness of the original recommendation models. However, the AUC after applying AFRL is closest to 0.5 for most fairness requirements, by which we can conclude that compared with the baseline fair models, AFRL can achieve the best personalized fairness of a recommendation model, due to its capability of learning the attribute-specific embeddings of attributes based on information alignment.

For accuracy, we can observe that the applying of AFRL and the baseline fair models will reduce the accuracy of the original recommendation models in terms of NDCG and Hit rate, which is reasonable because of the loss of information in the fair embeddings learned by the fair models. However, we see that in most cases, AFRL outperforms the baseline fair models, which indicates that AFRL can achieve less accuracy loss, due to its ability to capture more discriminative information from non-sensitive attributes and collaborative signals in the fair representations learned by it.

4.3 Trade-off (RQ2)

AFRL uses λ in Equation (16) to control the trade-off between fairness and accuracy, similar to the baseline fair models. Figures 3 and 4 show the pareto front curves of AFRL and the baseline methods over different λ and different fairness requirements on both datasets, with SASRec [17] and BERT4Rec [35] as the base recommendation models, respectively.

We can first see that in most cases, the point of the curve of AFRL at a specific λ basically locates in the uppermost-left position compared to all other points, i.e., there are no points of the curves of the baseline models in its upper-left region. This result confirms the advantage of AFRL to achieve a better trade-off between fairness and accuracy than the baseline fair models. At the same time, we also note that as λ is decreasing, the points of the curve of AFRL move to upper-right region, i.e., the accuracy improves while the fairness degrades. This is because in Equation (16), the smaller λ will cause more information from original user embedding to be preserved in the fair embedding, which is favorable to its discriminability but unfavorable to its fairness.

4.4 Ablation Study (RQ3)

Now we investigate the contributions of the attribute-specific embeddings $z_{u,i}$ ($1 \leq i \leq M$) and the debiased collaborative embedding $z_{u,0}$. For this purpose, we compare AFRL with its two variants, AFRL-w/o- $z_{u,i}$ where the embeddings $z_{u,i}$ of non-sensitive attributes are removed from the fair user embedding u^* , and AFRL-w/o- $z_{u,0}$ where the debiased collaborative embedding $z_{u,0}$ is removed from u^* . The results are shown in Figures 3, 4 and 5.

At first, from Figures 3 and 4 we can see that basically at the same λ , AFRL and AFRL-w/o- $z_{u,i}$ can achieve almost the same fairness, but AFRL has higher accuracy. This indicates that due to the information alignment, the fair embedding learned by AFRL through $z_{u,i}$ preserves more discriminability compared to the fair embedding learned by AFRL-w/o- $z_{u,i}$, without sacrificing fairness. At the

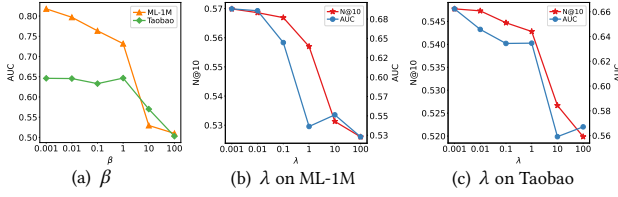


Figure 6: Tuning of hyper-parameters.

same time, Fig. 5 shows that AFRL achieves the higher accuracy than both AFRL-w/o- $\mathbf{z}_{u,0}$ and AFRL-w/o- $\mathbf{z}_{u,i}$, which indicates both $\mathbf{z}_{u,0}$ and $\mathbf{z}_{u,i}$ contribute to the accuracy because of the discriminative information encoded by them. However, we also note that the accuracy of AFRL-w/o- $\mathbf{z}_{u,i}$ is higher than that of AFRL-w/o- $\mathbf{z}_{u,0}$, which suggests the debiased collaborative signals encoded by $\mathbf{z}_{u,0}$ contribute more to the discriminability of the fair embedding.

4.5 Hyper-parameter Tuning (RQ4)

To determine the appropriate value of β , we train a classifier for each attribute A_i using the attribute embeddings Z_i and utilize AUC to evaluate the amount of information of A_i encoded by Z_i . Fig. 6(a) shows the average AUC curves over all surrogate classifiers on both datasets. We can see that basically the smaller β , the higher AUC, and the more information of A_i encoded, which is consistent with our analysis of Fig. 2 in Section 3.2.1. We also note that after β reaches 1, AUC begins to drop sharply. Since our objective is information alignment, i.e., making the information encoded by Z_i exactly cover that of attribute A_i rather than maximize or minimize it, we heuristically choose $\beta = 1$ to approximate the ideal case shown in Fig. 2(c).

As λ controls the trade-off between accuracy and fairness, we need to consider both accuracy and fairness to determine its optimal value. As mentioned in Section 4.1.3, we train a set of surrogate classifiers employing the final fair user embedding \mathbf{u}^* as input to predict the sensitive attribute value, with AUC as the metric for fairness performance. Accuracy is evaluated by NDCG. Fig. 6(b) and Fig. 6(c) show the impact of λ on recommendation accuracy and fairness on ML-1M and Taobao, respectively. We see that the larger λ , the more the fairness (the smaller the AUC), and the less the accuracy (the smaller the NDCG). We choose $\lambda = 1$ for the ML-1M dataset and $\lambda = 10$ for Taobao, because they are a balance point beyond which the accuracy will drastically drop but without significant rise of the fairness.

5 RELATED WORK

The fairness challenge in recommender systems is complex due to the involvement of multiple stakeholders, generally categorized into item-side fairness and user-side fairness. Item-side fairness seeks that different types of items have an equitable opportunity to be recommended, by minimizing prediction errors for various item categories [32] or allocating exposure based on item relevance [3]. User-side fairness aims to make fair recommendations to different users [1, 9, 12], or ensure that similar users receive similar treatments [10, 16, 22].

A range of approaches have been proposed for user-side fairness, following various technical lines including Pareto optimization [2, 5, 26, 39, 41], adversarial training [4, 21, 23, 36], and disentangled representation learning [13, 30, 33, 40]. The common idea of them is to filter out the information of user sensitive attributes from the learned fair embeddings, so that recommendations can be made without bias to sensitive attributes. At the same time, some works apply reinforcement learning to fair recommendations, which usually impose fairness constraints to the cumulative return function for long-term and dynamic fairness [11, 28].

The traditional methods often assume all users share the same sensitive attributes, which is not true in real world where different users have personalized fairness requirements. To address this issue, recently, some models have been proposed for personalized fairness in recommendations. For example, Li *et al.* [24] propose to train a filter for each possible combination of sensitive attributes, while Wu *et al.* [37] propose a model PFRec to build a set of prompt-based bias eliminators and adapters with customized attribute-specific prompts to learn fair embeddings for different attribute combinations. However, they treat a fairness requirement as a hyper-parameter, and train a model for each combination of attributes, which makes them implement the personalized fairness in a brute force way with unacceptable training cost. Creager *et al.* [7] propose the FFVAE model, which disentangles an unfair embedding into latent factors, each of which corresponds to an attribute, and excludes relevant sensitive latent factors based on different fairness requirements of users. Nevertheless, the existing methods often pursue an extreme fairness by completely removing the information of sensitive attributes from the fair embeddings, which makes them suffer from a suboptimal trade-off between accuracy and fairness.

6 CONCLUSION

In this paper, we propose a novel model, called Adaptive Fair Representation Learning (AFRL), for personalized fairness in recommendations. AFRL can overcome the challenge of the explosion of attribute combinations and achieve a better trade-off between accuracy and fairness. At first, AFRL treats fairness requirements as inputs and can learn an attribute-specific embedding for each attribute from the unfair user embedding, which endows AFRL with the adaptability during inference phase to dynamically determine the non-sensitive attributes that should be encoded into the fair embedding under the guidance of the user's unique fairness requirement. With the novel information alignment offered by the proposed Information Alignment Module (IAlignM), AFRL can reduce the loss of recommendation accuracy, without loss of fairness, by exactly preserving discriminative information of non-sensitive attributes and incorporating a debiased collaborative embedding into the fair embedding to capture attribute-independent collaborative signals. At last, the extensive experiments conducted on real datasets together with the sound theoretical analysis demonstrate the superiority of AFRL.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China under grant 61972270.

REFERENCES

- [1] McKane Andrus and Sarah Villeneuve. 2022. Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *FAT**.
- [2] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *SIGKDD*.
- [3] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*.
- [4] Avishek Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. In *ICML*.
- [5] Robin Burke, Nasim Sonboli, and Aldo Ordóñez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *FAT**.
- [6] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishikesh Aradhya, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *DLRS*.
- [7] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*. PMLR.
- [8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *ITCS*.
- [9] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *FAT**.
- [10] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *SIGIR*.
- [11] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. 2021. Towards long-term fairness in recommendation. In *WSDM*.
- [12] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. 2022. Explainable fairness in recommendation. In *SIGIR*.
- [13] Sixue Gong, Xiaoming Liu, and Anil K Jain. 2020. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*.
- [14] Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. 2021. Controllable guarantees for fair outcomes via contrastive information estimation. In *AAAI*.
- [15] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*.
- [16] Wenyue Hua, Yingqiang Ge, Shuyuan Xu, Jianchao Ji, and Yongfeng Zhang. 2023. UP5: Unbiased Foundation Model for Fairness-aware Recommendation. *arXiv preprint arXiv:2305.12090* (2023).
- [17] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*.
- [18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *FAT**.
- [19] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
- [20] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).
- [21] Roger Zhe Li, Julián Urbano, and Alan Hanjalic. 2021. Leave no user behind: Towards improving the utility of recommender systems for non-mainstream users. In *WSDM*.
- [22] Xuran Li, Peng Wu, and Jing Su. 2023. Accurate fairness: Improving individual fairness without trading accuracy. In *AAAI*.
- [23] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *WWW*.
- [24] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards personalized fairness based on causal notion. In *SIGIR*.
- [25] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *WWW*.
- [26] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Recsys*.
- [27] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2021. Mitigating confounding bias in recommendation via information bottleneck. In *Recsys*.
- [28] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. 2020. Balancing between accuracy and fairness for interactive recommendation with reinforcement learning. In *PAKDD*.
- [29] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. In *NeurIPS*.
- [30] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2021. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *AAAI*.
- [31] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.
- [32] Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. 2019. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *WSDM*.
- [33] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. 2020. Fairness by learning orthogonal disentangled representations. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX* 16. Springer.
- [34] Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810* (2017).
- [35] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM*.
- [36] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware news recommendation with decomposed adversarial learning. In *AAAI*.
- [37] Yiqing Wu, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Ao Xiang, Xu Zhang, Leyu Lin, and Qing He. 2022. Selective fairness in recommendation via prompts. In *SIGIR*.
- [38] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Recsys*.
- [39] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *NeurIPS*.
- [40] Chen Zhao, Le Wu, Pengyang Shao, Kun Zhang, Richang Hong, and Meng Wang. 2023. Fair representation learning for recommendation: a mutual information perspective. In *AAAI*.
- [41] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *CIKM*.
- [42] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. 2021. Fairness among New Items in Cold Start Recommender Systems. In *SIGIR*.