User-side Fairness in Recommender Systems

Weixin CHEN

cswxchen@comp.hkbu.edu.hk

Supervisor: **Prof. Li CHEN**

February 18, 2025

Outline

1. Literature Review

2. Associative and Causal Fairness for Multi-Type Sensitive Attributes

3. Fairness-aware Multimodal Recommendation

4. Controllable Fairness for Recommender Systems

Outline

1 Literature Review

2 Associative and Causal Fairness for Multi-Type Sensitive Attributes

3 Fairness-aware Multimodal Recommendation

4 Controllable Fairness for Recommender Systems

What are recommender systems?

- Learning **user** preference based on their interactions with **items**
- Recommending **items** to **users**

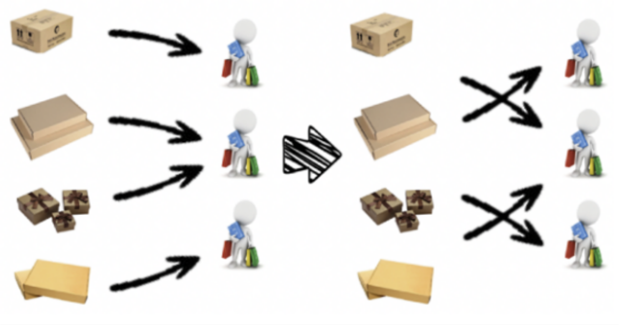
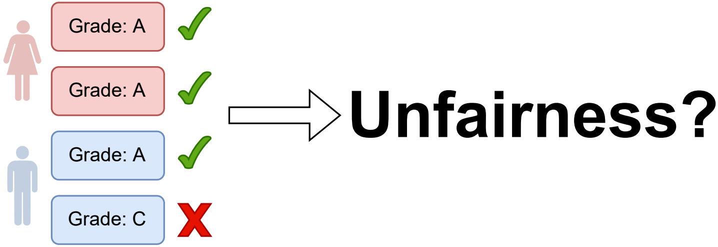


Figure: Paradigm of a recommender system [Li et al., 2021e]

What is user-side fairness in recommender systems?

- User-side fairness in recommender systems could be explained as **to what extent a user's recommendation is unfairly affected by his/her sensitive features?**



Grade: A ✓

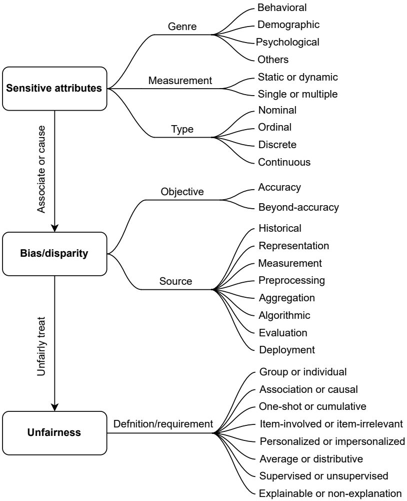
Grade: A ✓

Grade: A ✓

Grade: C ✗

⇒ **Unfairness?**

A simplified model of user-side fairness in recommender systems



Sensitive attributes

Sensitive attributes could be categorized by **attribute genre**:

- Behavioral attribute, e.g., activity level
- Demographic attribute, e.g., age, gender, and occupation
- Psychological attribute, e.g., Big-Five personality and curiosity

, by **measurement**:

- Static or dynamic
- Single or multiple attributes (considered in a research paper)

and by **data type**:

- Nominal/categorical
- Ordinal
- Discrete
- Continuous

Sensitive attributes

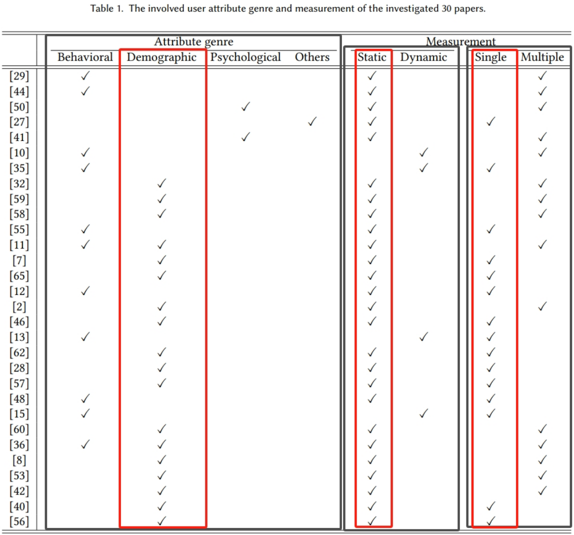


Table 1. The involved user attribute genre and measurement of the investigated 30 papers.

	Attribute genre				Measurement			
	Behavioral	Demographic	Psychological	Others	Static	Dynamic	Single	Multiple
[29]	✓				✓			✓
[44]	✓				✓			✓
[50]			✓		✓			✓
[27]				✓	✓		✓	
[41]			✓		✓			✓
[10]	✓					✓		✓
[35]	✓					✓	✓	
[32]		✓			✓			✓
[59]		✓			✓			✓
[58]		✓			✓			✓
[55]	✓				✓		✓	
[11]	✓	✓			✓			✓
[7]		✓			✓		✓	
[65]		✓			✓		✓	
[12]	✓				✓		✓	
[2]		✓			✓			✓
[46]		✓			✓		✓	
[13]	✓					✓	✓	
[62]		✓			✓		✓	
[28]		✓			✓		✓	
[57]		✓			✓		✓	
[48]	✓				✓		✓	
[15]	✓					✓	✓	
[60]		✓			✓			✓
[36]	✓	✓			✓			✓
[8]		✓			✓			✓
[53]		✓			✓			✓
[42]		✓			✓			✓
[40]		✓			✓		✓	
[56]		✓			✓		✓	

Bias

Bias/disparity on performance, could be categorized by **objective**:

- **Accuracy-oriented objectives.** Most works in recommender systems merely targeted accuracy-oriented objectives. For instance, mean absolute error (MAE) and mean square error (MSE) were used to assess the error of predicted rating in [Liu et al., 2022], and recall and normalized discounted cumulative gain (NDCG) were utilized to evaluate the quality of the ranked recommendation list in [Melchiorre et al., 2020].
- **Beyond-accuracy objectives.** Compared with accuracy objectives, beyond-accuracy objectives have been rarely investigated in recommendation fairness research, even though they were recognized as significantly contributing to the quality of the recommendation list [Kaminskas and Bridge, 2016] and enabling users to explore more than accurate experiences on various items [Herlocker et al., 2004].

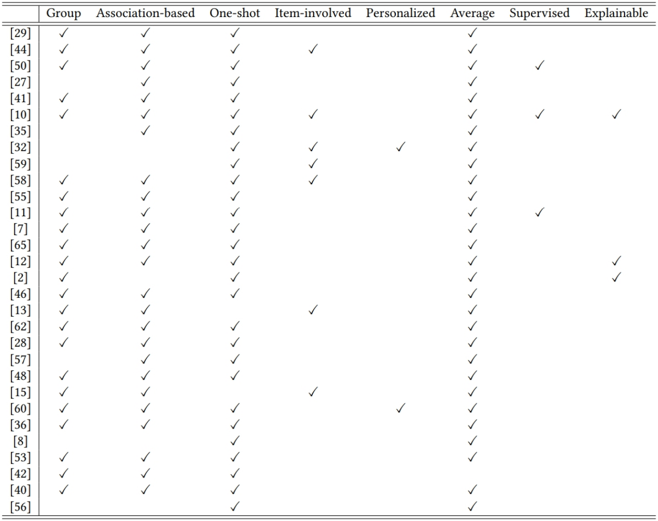
Unfairness

Unfairness definitions/requirements, could be categorized as

- Group fairness vs. Individual fairness
- One-shot fairness vs. Cumulative fairness
- Association-based fairness vs. Causality-based fairness
- Item-involved fairness vs. Item-irrelevant fairness
- Personalized fairness vs. Impersonalized fairness
- Average fairness vs. Distributive fairness
- Supervised (or calibrated) fairness vs. Unsupervised fairness
- Explainable fairness vs. Unexplainable fairness

Fairness deifinitions

Table 2. The fairness definitions of the 30 investigated papers.



	Group	Association-based	One-shot	Item-involved	Personalized	Average	Supervised	Explainable
[29]	✓	✓	✓			✓		
[44]	✓	✓	✓	✓		✓		
[50]	✓	✓	✓			✓	✓	
[27]		✓	✓			✓		
[41]	✓	✓	✓			✓		
[10]	✓	✓	✓	✓		✓	✓	✓
[35]		✓	✓			✓		
[32]			✓	✓	✓	✓		
[59]			✓	✓		✓		
[58]	✓	✓	✓	✓		✓		
[55]	✓	✓	✓			✓		
[11]	✓	✓	✓			✓	✓	
[7]	✓	✓	✓			✓		
[65]	✓	✓	✓			✓		
[12]	✓	✓	✓			✓		✓
[2]	✓		✓			✓		✓
[46]	✓	✓	✓			✓		
[13]	✓	✓		✓		✓		
[62]	✓	✓	✓			✓		
[28]	✓	✓	✓			✓		
[57]		✓	✓			✓		
[48]	✓	✓	✓			✓		
[15]	✓	✓		✓		✓		
[60]	✓	✓	✓		✓	✓		
[36]	✓	✓	✓			✓		
[8]			✓			✓		
[53]	✓	✓	✓			✓		
[42]	✓	✓	✓					
[40]	✓	✓	✓			✓		
[56]			✓			✓		

Fairness methods

Methods for mitigating unfairness:

- **Pre-processing**, i.e., transforming the imbalanced data
- **In-processing**, i.e., modifying the base model
- **Post-processing**, i.e., re-ranking the recommendation list

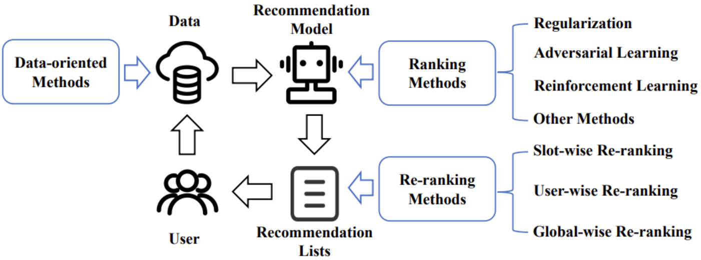


Figure: Taxonomy of fairness methods [Wang et al., 2023]

Pre-processing: balancing data via re-sampling

- “Randomly sampling without replacement the same number of female and male users” [Ekstrand et al., 2018]

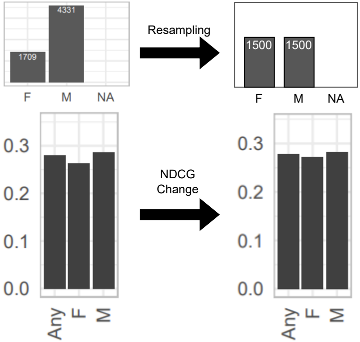


Figure: Gender distribution and corresponding performance before and after re-sampling for ML1M dataset [Ekstrand et al., 2018]

Post-processing: re-ranking

Given a top-*N* recommendation list $\{v_1, v_2, \cdots, v_N \mid u_i\}$ and $s_{i,j}$ that user *i*'s preference score on item *j*, in order to generate a fair top-*K* recommendation list:

$$\begin{array}{ll} \max_{\mathbf{W}_{ij}} & \sum_{i=1}^{n} \sum_{j=1}^{N} \mathbf{W}_{ij}\, s_{i,j} \\ \text{s.t.} & \text{UGF}\left(Z_1, Z_2, \mathbf{W}\right) < \varepsilon \\ & \sum_{j=1}^{N} \mathbf{W}_{ij} = K, \mathbf{W}_{ij} \in \{0, 1\} \end{array} \tag{1}$$

which aims to select *K* items from the given *N* recommended items by maximizing the overall preference score while subject to the ε-fairness requirements.

Fairness definitions

- User-oriented Group Fairness (UGF) [Li et al., 2021a]

$$\mathbb{E}\left[\mathcal{M}(\mathbf{W}) \mid Z = Z_1\right] = \mathbb{E}\left[\mathcal{M}(\mathbf{W}) \mid Z = Z_2\right] \tag{2}$$

- ε-fairness [Li et al., 2021a]

$$UGF\left(Z_1, Z_2, \mathbf{W}\right) = \left| \frac{1}{|Z_1|} \sum_{i \in Z_1} \mathcal{M}\left(\mathbf{W_i}\right) - \frac{1}{|Z_2|} \sum_{i \in Z_2} \mathcal{M}\left(\mathbf{W_i}\right) \right| \leq \varepsilon \tag{3}$$

where $\mathcal{M}(\mathbf{W})$ represents recommendation accuracy and Z represents the belonged user group, e.g., male or female.

In-processing: adversarial learning

To achieve counterfactual fairness requirement, adversarial learning is adapted to train the base model with another objective $\mathbf{r}_u^* \perp \mathbf{Z}_u$:

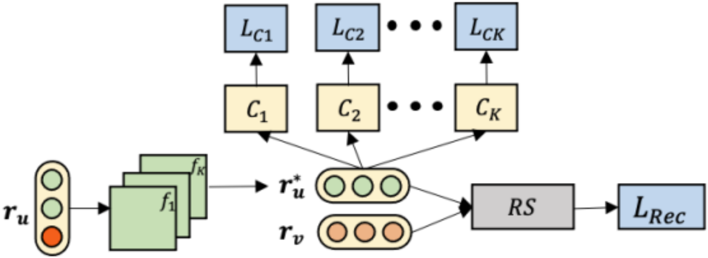


Figure: The architecture of adversarial learning framework [Li et al., 2021c]

Outline

1. Literature Review
2. Associative and Causal Fairness for Multi-Type Sensitive Attributes
3. Fairness-aware Multimodal Recommendation
4. Controllable Fairness for Recommender Systems

Motivation

Current user-side fairness notions in recommendations aim to diminish the association or the causality between users' sensitive attributes and recommended items.

- **Association-based fairness** emphasizes the statistical metric equity across **subpopulations**, for which **regularization** has been proposed as a typical approach to reducing the recommendation imbalance between user groups of different attributes.
- **Causality-based fairness** focuses on independence between sensitive attributes and recommendations for **individual users**, for which **adversarial learning** has been popularly adopted to remove sensitive information from user representation.

Limitations

Little work has investigated the effects of these methods on **balancing** both group-level and individual-level user-side fairness.

- A group-level regularization-based fair recommender may infringe on some individual users with specific sensitive attributes, whose recommendations are negatively affected, because of simply forcing utility parity between sub-populations;
- while an individual-level adversarial-learning-based fair recommender may not be acceptable for the vulnerable group because of a potentially significant performance gap in recommendation accuracy between them and the advantaged group.

Moreover, existing experiments have primarily been performed on **one type** of user attributes, e.g., behavioral attributes or demographic attributes, but neglected other possible sensitive attributes such as psychological attributes.

Definition of association-based fairness

Definition (User-oriented Group Fairness)

A recommender system satisfies user-oriented group fairness (UGF) if it offers equivalent recommendation quality to the user groups with different values on binary sensitive attribute $\boldsymbol{A}$ *[Li et al., 2021b]:*

$$\mathbb{E}\left[\mathcal{M}(L) \mid \boldsymbol{A}=\mathbf{a}_1\right]=\mathbb{E}\left[\mathcal{M}(L) \mid \boldsymbol{A}=\mathbf{a}_2\right] \tag{4}$$

where $\mathbf{a}_1$ *and* $\mathbf{a}_2$ *denote the attainable attribute value of the binary attribute* $\boldsymbol{A}$*, and* $\mathcal{M}$ *represents the recommendation quality of the recommendation list L such as recall and NDCG.*

UGF measurement

User-oriented group fairness can be measured via *UGF* value as proposed in [Li et al., 2021b]:

$$UGF\left(G_{\mathbf{a}_1}, G_{\mathbf{a}_2}, L\right) = \left| \frac{1}{\left|G_{\mathbf{a}_1}\right|} \sum_{i \in G_{\mathbf{a}_1}} \mathcal{M}\left(L_{\mathbf{i}}\right) - \frac{1}{\left|G_{\mathbf{a}_2}\right|} \sum_{i \in G_{\mathbf{a}_2}} \mathcal{M}\left(L_{\mathbf{i}}\right) \right| \qquad (5)$$

where $G_{\mathbf{a}_1}$ (or $G_{\mathbf{a}_2}$) denotes the group of users with sensitive attribute $\mathbf{A} = \mathbf{a}_1$ (or $\mathbf{A} = \mathbf{a}_2$).

UGF can represent the discrepancy level between two user groups in terms of recommendation quality. Therefore, it favors recommendation methods that return results of similar quality for different user groups, while penalizing those that are skewed to the privileged group.

Regularization method

One representative fairness regularization method is **FOCF** [Yao and Huang, 2017] which directly adds the smoothed Huber loss [Huber, 1992] of the targeted unfairness metric as a regularization term $\mathcal{L}_{reg}$ to the learning objective $\mathcal{L}$:

$$\arg\min_{\boldsymbol{\Theta}} \mathcal{L}_{rec} + \lambda_{reg}\mathcal{L}_{reg} \tag{6}$$

where $\boldsymbol{\Theta}$ are all parameters of the training model and λ_{reg} is a parameter to control the utility-fairness trade-off. As *UGF* in Equation (5) is non-differentiable when assembled NDCG as recommendation quality [Bruch et al., 2019], we adopt value unfairness and absolute unfairness as the target unfairness metric *U*.

Regularization method

Value unfairness and absolute unfairness are designed to measure equalized odds in recommendation [Yao and Huang, 2017]:

$$U_{\mathrm{val}} = \frac{1}{n}\sum_{j=1}^{n}\left|\left(\mathrm{E}_{G_{\mathbf{a}_1}}[y]_j - \mathrm{E}_{G_{\mathbf{a}_1}}[r]_j\right) - \left(\mathrm{E}_{G_{\mathbf{a}_2}}[y]_j - \mathrm{E}_{G_{\mathbf{a}_2}}[r]_j\right)\right| \tag{7}$$

$$U_{\mathrm{abs}} = \frac{1}{n}\sum_{j=1}^{n}\left|\left|\mathrm{E}_{G_{\mathbf{a}_1}}[y]_j - \mathrm{E}_{G_{\mathbf{a}_1}}[r]_j\right| - \left|\mathrm{E}_{G_{\mathbf{a}_2}}[y]_j - \mathrm{E}_{G_{\mathbf{a}_2}}[r]_j\right|\right| \tag{8}$$

where $\mathrm{E}_{G_{\mathbf{a}}}[y]_j$ and $\mathrm{E}_{G_{\mathbf{a}}}[r]_j$ respectively denote the average prediction score and the average label value of item j for user group $G_{\mathbf{a}}$ with the sensitive attribute $\mathbf{A} = \mathbf{a}$. Therefore, two variants of FOCF are considered in our experiment, i.e., FOCF w/ U_{val} and FOCF w/ U_{abs}, adopting value unfairness (i.e., $U \leftarrow U_{\mathrm{val}}$) and absolute unfairness (i.e., $U \leftarrow U_{\mathrm{abs}}$) as the optimization targets, respectively.

Definition of causality-based fairness

Definition (Counterfactually fair recommendation)

A recommender system is counterfactually fair if it generates a recommendation list L to any users with insensitive attribute $\boldsymbol{X} = \boldsymbol{x}$ and sensitive attribute $\boldsymbol{A} = \boldsymbol{a}$ as below [Li et al., 2021d]:

$$\Pr\left(L_{\boldsymbol{A}\leftarrow\boldsymbol{a}} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{A} = \boldsymbol{a}\right) = \Pr\left(L_{\boldsymbol{A}\leftarrow\boldsymbol{a}'} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{A} = \boldsymbol{a}\right) \tag{9}$$

for any L and for any value $\boldsymbol{a}'$ attainable by $\boldsymbol{A}$.

Evaluation of counterfactually fair recommendation

- Unlike the statistical measurement of association-based fairness, counterfactual fairness requires experimental evaluation.

- A commonly used evaluation is to train an attacker to infer one of the user's sensitive features (e.g., gender) from the learned user representation for recommendations, given that the user representation is the key connector between sensitive features and recommendation outcome [Li et al., 2021d].

Adversarial learning

- A typical way to apply adversarial learning is to construct a filter function $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ to filter out the potential information of the sensitive feature from user embedding r_u, through which a filtered user embedding $r_u^* = F(r_u)$ that contains less sensitive information is obtained.
- To simulate the real attacker and train the filter function for the removal ability of sensitive information, a discriminator $D : \mathbb{R}^d \mapsto [0, 1]$ could be assembled to classify the sensitive feature of each user from filtered user embedding r_u^*.
- To this end, a min-max game would be adopted, in which the maximization objective encourages the discriminator to leak sensitive information from filtered user embedding, while the minimization objective stimulates the filter function to deceive the discriminator by obfuscating sensitive feature information in user embedding.

Fairness methods

Group-level association-based:

- **FOCF w/** U_{val} adopting value unfairness (i.e., $U \leftarrow U_{\mathrm{val}}$) as the target unfairness metric
- **FOCF w/** U_{abs} adopting absolute unfairness (i.e., $U \leftarrow U_{\mathrm{abs}}$) as the target unfairness metric

Individual-level causality-based:

- **PCFR** [Li et al., 2021d] is a counterfactual fair recommendation framework, which removes personalized sensitive features from user representation for each user via adversarial learning.
- **FairRec** [Wu et al., 2021a] consists of both a bias-aware model and a bias-free model with orthogonality regularization, for achieving maximized and minimized prediction performance on a user-sensitive feature, respectively.

Base recommendation algorithms

- **PMF** [Salakhutdinov and Mnih, 2007]. Probabilistic Matrix Factorization is a well-known matrix factorization algorithm with probabilistic modeling that assumes the Gaussian distributions of the latent user and item representations.
- **BiasedMF** [Koren et al., 2009]. Biased Matrix Factorization combines the biases of user, item, and global to the rating prediction calculation for matrix factorization.
- **DMF** [Xue et al., 2017]. Deep Matrix Factorization applies a multi-layer perceptron and a non-linear activation function to user and item embeddings.
- **MLP** [Cheng et al., 2016]. MLP is through a deep neural network and non-linear activation functions to match the user and the item by their embedding concatenation.

Datasets (1/3)

- **Insurance.** It is a small dataset for insurance recommendations with 21 product candidates and 29,131 users. There are three **demographic features** of users in this dataset, i.e., *gender*, *occupation*, and *marital status*. We filtered out inactive users with less than three interactions to avoid the cold-start setting.
- **Douban.** Douban is a small dataset originally collected for recommending potential interest groups to users according to their previously joined groups. It contains 1,706 users with their Big-Five personality values respectively about *openness to experience (openness)*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism* as **psychological features**, and 2,396 interest groups as items.

Datasets (2/3)

- **MovieLens.** MovieLens is a widely used large movie dataset containing nearly 1 million interactions involving 6,040 users and 3,706 items. Three **demographic features** of users, i.e., *age*, *gender*, and *occpuation*, are included in this dataset.
- **Taobao-Serendipity.** It is a large dataset consisting of 11,382 users and 834,279 items with more than ten million interactions collected from a Chinese e-commerce platform Mobile Taobao. It contains both **demographic features** (*age* and *gender*) and **psychological features** (Big-Five personality values and *curiosity*).

Datasets (3/3)

Table: Statistics of the preprocessed data for the experiment.

Statistics	**Insurance**	**Douban**
#User	29,131	1,706
#Item	21	2,396
#Interaction	66,339	24,314
Density	10.8441%	0.5948%
Avg. #interactions per user	2.2773	14.2521
Behavioral attributes	activity level	activity level
Demographic attributes	gender, occupation, marital status	N/A
Psychological attributes	N/A	openness, conscientiousness, extraversion, agreeableness, neuroticism

Statistics	**MovieLens**	**Taobao**
#User	6,040	11,382
#Item	3,706	834,279
#Interaction	1,000,209	11,415,471
Sparsity	4.4684%	0.1202%
Avg. #interactions per user	165.5975	1002.9407
Behavioral attributes	activity level	activity level
Demographic attributes	age, gender, occupation	age, gender
Psychological attributes	N/A	openness, conscientiousness, extraversion, agreeableness, neuroticism, curiosity

Selected results (1/3)

Table: Recommendation and fairness performance regarding the behavioral attribute **activity level**.

Method		Insurance			Douban			MovieLens			Taobao		
		NDCG ↑	UGF ↓	AUC ↓	NDCG ↑	UGF ↓	AUC ↓	NDCG ↑	UGF ↓	AUC ↓	NDCG ↑	UGF ↓	AUC ↓
PMF	Original	0.8073	0.0399	0.8352	0.3331	0.0398	0.8676	**0.4768**	0.2512	0.9003	**0.8155**	**0.2476**	0.9275
	+ FOCF w/ U_{val}	0.8105	0.0727	0.8145	0.3312	0.0454	0.8936	0.4482	0.2704	0.9883	0.7941	0.2684	0.9662
	+ FOCF w/ U_{abs}	0.7979	**0.0362**	0.7101	**0.3551**	**0.0265**	0.8686	0.4578	0.2676	0.9264	0.8040	0.2592	0.9438
	+ PCFR	**0.8454**	0.0887	**0.5000**	0.3009	0.0306	0.5130	0.4621	0.2577	0.5671	0.7844	0.2731	**0.6231**
	+ FairRec	0.8399	0.0653	**0.5000**	0.3274	0.0354	**0.5000**	0.3276	**0.2126**	**0.5158**	0.7192	0.3110	0.6376
BiasedMF	Original	**0.8602**	0.0990	0.7661	**0.3604**	0.0151	0.8688	**0.4728**	0.2519	0.8806	**0.8128**	**0.2533**	0.9237
	+ FOCF w/ U_{val}	0.8549	0.1101	0.8008	0.3549	0.0188	0.9237	0.4579	0.2797	0.9901	0.8064	0.2681	0.9544
	+ FOCF w/ U_{abs}	0.8595	0.0990	0.7819	0.3342	0.0561	0.8036	0.4561	0.2754	0.9448	0.8047	0.2631	0.9342
	+ PCFR	0.8285	**0.0779**	**0.5000**	0.3024	0.0145	0.5611	0.4531	0.2504	**0.5285**	0.7788	0.2758	0.6545
	+ FairRec	0.8125	0.1113	**0.5000**	0.3048	**0.0047**	**0.5062**	0.3174	**0.2274**	0.5408	0.7153	0.3115	**0.6099**
DMF	Original	**0.8043**	0.0932	0.5001	0.2433	0.0280	0.5573	0.2634	0.2120	0.8429	0.5603	0.3787	0.8269
	+ FOCF w/ U_{val}	0.7897	**0.0572**	0.6750	0.2177	0.0177	0.7038	0.2467	**0.2044**	0.8228	0.5492	0.3845	0.7656
	+ FOCF w/ U_{abs}	0.7869	0.0650	0.6695	0.2161	0.0127	0.7192	0.2615	0.2181	0.8405	0.5553	0.3860	0.7736
	+ PCFR	0.7880	0.0656	**0.5000**	**0.2503**	0.0289	**0.5206**	0.2550	0.2113	**0.5245**	0.5671	**0.3783**	0.6891
	+ FairRec	0.7713	0.0686	**0.5000**	0.2307	**0.0080**	0.5418	**0.2667**	0.2122	0.5641	**0.5684**	0.3851	**0.5434**
MLP	Original	0.8416	0.0914	0.5307	0.3286	0.0532	0.6597	0.3850	0.2479	0.7639	**0.6940**	0.3556	0.8924
	+ FOCF w/ U_{val}	0.8416	0.0916	0.5173	0.3245	**0.0302**	0.8009	0.3830	0.2648	0.8726	0.6860	0.3783	0.8939
	+ FOCF w/ U_{abs}	0.8416	0.0914	**0.5000**	0.3271	0.0515	0.7032	**0.3935**	0.2533	0.8629	0.6854	0.3723	0.8754
	+ PCFR	0.8419	0.0924	**0.5000**	**0.3299**	0.0573	**0.5410**	0.3873	0.2410	**0.5619**	0.6839	**0.3555**	**0.6458**
	+ FairRec	**0.8539**	**0.0659**	**0.5000**	0.3257	0.0474	0.5554	0.3412	**0.2234**	0.5996	0.6426	0.3788	0.6482

Selected results (2/3)

Table: Recommendation and fairness performance on the MovieLens dataset regarding the demographic attributes **age**, **gender**, and **occupation**.

Method		Age			Gender			Occupation		
		NDCG ↑	UGF ↓	AUC ↓	NDCG ↑	UGF ↓	AUC ↓	NDCG ↑	UGF ↓	AUC ↓
PMF	Original	0.4677	0.0335	0.7144	**0.4674**	0.0604	0.7437	**0.4727**	0.0095	0.5315
	+ FOCF w/ U_{val}	0.4690	0.0399	0.7224	0.4620	0.0678	0.8382	0.4610	0.0092	0.5399
	+ FOCF w/ U_{abs}	**0.4696**	0.0298	0.7123	0.4671	0.0731	0.7693	0.4641	0.0071	0.5288
	+ PCFR	0.4286	**0.0245**	**0.5454**	0.4555	**0.0588**	0.5588	0.4591	**0.0035**	**0.5216**
	+ FairRec	0.3169	0.0429	0.5505	0.3141	0.0825	**0.5382**	0.3233	0.0038	0.5312
BiasedMF	Original	**0.4769**	**0.0313**	0.7208	**0.4765**	0.0590	0.7376	**0.4736**	**0.0049**	0.5344
	+ FOCF w/ U_{val}	0.4654	0.0359	0.7203	0.4610	0.0675	0.8174	0.4624	0.0088	0.5513
	+ FOCF w/ U_{abs}	0.4721	0.0380	0.7141	0.4602	0.0642	0.7614	0.4657	0.0085	0.5372
	+ PCFR	0.4409	0.0321	**0.5550**	0.4528	0.0650	0.5933	0.4436	0.0056	0.5274
	+ FairRec	0.3270	0.0393	0.5569	0.3300	**0.0465**	**0.5766**	0.3399	0.0051	**0.5161**
DMF	Original	0.2580	0.0442	0.6149	0.2665	0.0513	0.6849	0.2615	**0.0020**	0.5321
	+ FOCF w/ U_{val}	**0.2742**	0.0419	0.6012	**0.2747**	**0.0509**	0.6505	**0.2677**	**0.0020**	0.5530
	+ FOCF w/ U_{abs}	0.2668	**0.0368**	0.5993	0.2618	0.0557	0.6455	0.2534	0.0072	0.5577
	+ PCFR	0.2590	0.0396	0.5412	0.2565	0.0556	**0.5699**	0.2529	0.0132	**0.5275**
	+ FairRec	0.2727	0.0443	**0.5267**	0.2673	0.0535	0.5702	0.2558	0.0078	0.5279
MLP	Original	0.3749	**0.0373**	0.6511	0.3743	**0.0508**	0.7037	0.3980	0.0042	0.5247
	+ FOCF w/ U_{val}	0.3973	0.0444	0.6442	**0.3931**	0.0522	0.7122	0.3894	0.0036	0.5367
	+ FOCF w/ U_{abs}	**0.3990**	0.0429	0.6693	0.3906	0.0577	0.7018	**0.3990**	0.0053	0.5458
	+ PCFR	0.3824	0.0404	0.5365	0.3884	0.0594	**0.5566**	0.3885	0.0035	0.5227
	+ FairRec	0.3307	0.0428	**0.5364**	0.3181	0.0601	0.5835	0.3212	**0.0022**	**0.5201**

Selected results (3/3)

Table: Fairness performance on the Taobao-Serendipity dataset regarding the psychological features **openness**, **conscientiousness**, **extraversion**, **agreeableness**, **neuroticism**, and **curiosity**.

Method		Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
		UGF ↓	AUC ↓	UGF ↓	AUC ↓	UGF ↓	AUC ↓	UGF ↓	AUC ↓	UGF ↓	AUC ↓
PMF	Original	0.0156	0.5810	0.0366	0.5537	0.0034	0.5886	0.0048	0.5871	0.0025	0.5358
	+ FOCF w/ U_{val}	0.0151	0.5871	0.0351	0.5734	0.0035	0.5805	0.0058	0.5925	0.0025	0.5406
	+ FOCF w/ U_{abs}	0.0211	0.5781	0.0351	0.5540	0.0042	0.5854	0.0040	0.5901	**0.0016**	0.5373
	+ PCFR	0.0224	**0.5058**	0.0464	**0.5000**	0.0049	**0.5000**	**0.0023**	0.5708	0.0166	**0.5000**
	+ FairRec	**0.0092**	0.5261	**0.0261**	0.5313	**0.0031**	0.5508	0.0064	**0.5000**	0.0178	0.5634
BiasedMF	Original	0.0234	0.5604	0.0455	0.5883	0.0111	0.5710	0.0027	0.5424	0.0083	0.5615
	+ FOCF w/ U_{val}	0.0226	0.5518	0.0360	0.5694	0.0092	0.5760	**0.0021**	0.5496	0.0065	0.5575
	+ FOCF w/ U_{abs}	0.0293	0.5606	0.0453	0.5841	0.0125	0.5685	0.0033	0.5575	0.0178	0.5732
	+ PCFR	**0.0020**	**0.5000**	0.0385	**0.5000**	0.0157	**0.5000**	0.0166	**0.5000**	**0.0021**	**0.5069**
	+ FairRec	0.0159	0.5407	**0.0307**	0.5212	**0.0075**	0.5427	0.0312	**0.5000**	0.0290	0.5659
DMF	Original	0.0114	0.5493	**0.0038**	0.5502	0.0247	0.5443	**0.0014**	0.5457	0.0194	0.5520
	+ FOCF w/ U_{val}	0.0085	0.5344	0.0068	0.5615	0.0169	0.5861	0.0083	0.5448	0.0195	0.5487
	+ FOCF w/ U_{abs}	0.0194	0.5291	0.0134	0.5630	0.0189	0.5496	0.0041	0.5416	0.0202	0.5553
	+ PCFR	**0.0056**	**0.5000**	0.0137	0.5627	0.0103	**0.5441**	0.0201	0.5600	0.0208	**0.5000**
	+ FairRec	0.0202	**0.5000**	0.0321	**0.5292**	**0.0070**	0.5503	0.0085	**0.5000**	**0.0052**	**0.5000**
MLP	Original	0.0095	0.5958	0.0347	0.5437	**0.0001**	0.5304	**0.0010**	0.5460	0.0114	0.5444
	+ FOCF w/ U_{val}	0.0114	0.5369	**0.0297**	0.5555	0.0009	0.6278	0.0015	0.6128	0.0078	0.5549
	+ FOCF w/ U_{abs}	0.0099	0.5365	**0.0297**	**0.5352**	0.0028	0.5628	0.0017	0.5492	0.0119	0.5644
	+ PCFR	0.0068	**0.5283**	0.0310	0.5398	0.0009	0.5384	0.0054	**0.5342**	**0.0061**	**0.5402**
	+ FairRec	**0.0052**	0.5463	0.0309	0.5420	0.0012	**0.5116**	0.0026	0.5428	0.0097	0.5485

Conclusions (1/2)

- Directly adding fairness objectives to eliminate disparity can enhance association-based fairness performance in certain scenarios, but may lead to more sensitive information leakage during training.
- Adversarially filtering out the sensitive information from the user representation can effectively improve the causality-based fairness performance in most cases, but may aggravate the recommendation bias between the advantaged and disadvantaged groups.
- We also find that adversarial learning can significantly reduce the recommendation disparity in some scenarios, which implies that it owns the potential to achieve a better balance between group-level and individual-level user-side fairness.

Conclusions (2/2)

- The fairness performance regarding the behavioral attributes is the worst in terms of both group-level association and individual-level causality, which might be because users with more interactions would be more accurately modeled
- The unfairness level exhibited by demographic features varies and is highly domain-dependent
- Regarding psychological features, the weak but stable unfairness of causality is hard to eradicate, indicating challenging implications for fairness

Outline

1 Literature Review

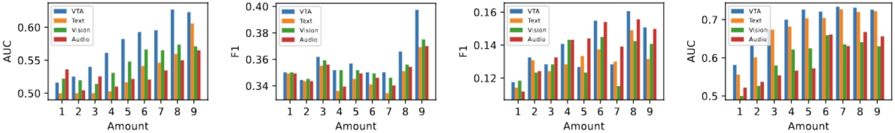
2 Associative and Causal Fairness for Multi-Type Sensitive Attributes

3 Fairness-aware Multimodal Recommendation

4 Controllable Fairness for Recommender Systems

Motivation

Although multimodal recommendations (MMR) exhibit notable improvements in accuracy compared to traditional recommenders, we empirically validate that an increase in **the *quantity* and *variety* of modalities leads to a higher degree of users' sensitive information leakage**, resulting in potential unfairness issues.



(a) Gender on MovieLens (b) Age on MovieLens (c) Occupation on MovieLens (d) Gender on MicroLens

Figure: Sensitive attribute prediction performance given different numbers and types of modalities on two datasets MovieLens and MicroLens.

However, to our knowledge, there is a lack of studies to address the unfairness issue with MMR.

Challenges

One intuitive approach is to incorporate multimodal representations as additional knowledge during the process of learning fair user representations. This approach, however, faces two key challenges:

- The **entanglement** present in multimodal content poses difficulty in eliminating sensitive information while utilizing non-sensitive information to ensure accuracy. It should be important to separate sensitive and non-sensitive information from coupled multimodal content so that fairness concerns can be addressed without significant accuracy loss.
- The **heterogeneity** between items' multimodal representations and user representations hinders leveraging the modality-based knowledge to promote fair user representation learning. Specifically, the multimodal representations of items and the user representations are in highly distinct semantic spaces, posing a considerable obstacle to their interactions.

Solution

To tackle these two challenges for achieving fairness-aware MMR, in this work, we propose a **f**air **m**ulti**m**odal **rec**ommendation approach (referred to **FMMRec**) through **fairness-oriented modal disentanglement** along with **relation-aware fairness learning**.

- For the **fairness-oriented disentanglement** of modal embeddings, we propose to maximize the potential sensitive information of the biased embeddings and minimize that of filtered embeddings, while maintaining sufficient non-sensitive information of filtered embeddings for preserving personalized information.
- For **relation-aware fairness learning**, instead of forcing interactions between multimodal item representations and user representations, FMMRec mines dual user-user relations given the disentangled modal embeddings to learn fair and informative user representations.

Preliminaries (1/2)

Notations for multimodal recommendation problem:

- User set: $\mathcal{U} = \{u_1, u_2, \cdots, u_N\}$
- item set: $\mathcal{V} = \{v_1, v_2, \cdots, v_M\}$
- Embedding of user u: $\boldsymbol{e}_u \in \mathbb{R}^d$
- Embedding of item v: $\boldsymbol{e}_v \in \mathbb{R}^d$
- Historical binary interaction matrix: $\mathcal{R} \in \mathbb{R}^{N \times M}$, wherein each unit $r_{ij} \in \mathcal{R}$ would be filled by 1 if user u_i has interacted with item v_j, otherwise by 0
- Modal embedding of item v: $\boldsymbol{e}_v^m \in \mathbb{R}^{d_m}$, where m denotes a specific modality (e.g., visual modality), d_m is the embedding dimension for the m modality, and $m \in \mathcal{M} = \{\mathrm{v}, \mathrm{t}, \mathrm{a}\}$

Preliminaries (2/2)

Definition (Counterfactually fair recommendation)

A recommender system is counterfactually fair if it generates a recommendation list L to any users with insensitive attribute $\boldsymbol{X} = \boldsymbol{x}$ and sensitive attribute $\boldsymbol{A} = \boldsymbol{a}$ as below [Li et al., 2021d]:

$$\Pr\left(L_{\boldsymbol{A} \leftarrow \boldsymbol{a}} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{A} = \boldsymbol{a}\right) = \Pr\left(L_{\boldsymbol{A} \leftarrow \boldsymbol{a}'} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{A} = \boldsymbol{a}\right) \tag{10}$$

for any L and for any value $\boldsymbol{a}'$ attainable by $\boldsymbol{A}$.

Overall Architecture

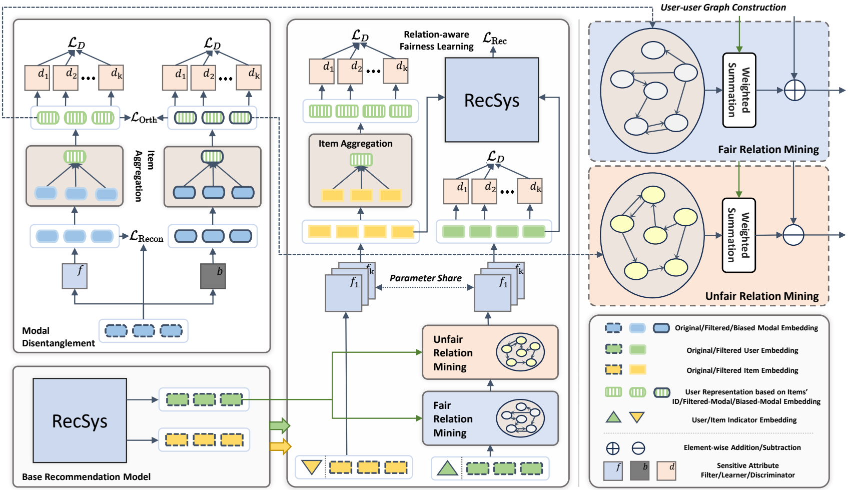
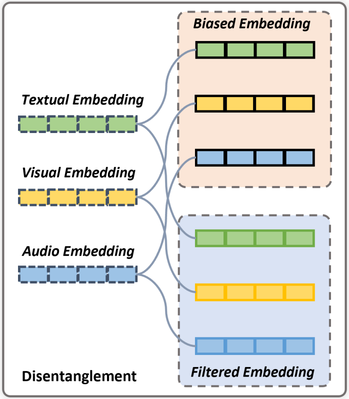

User-user Graph Construction
$\mathcal{L}_D$
d_1 d_2 ... d_k
$\mathcal{L}_{\text{Orth}}$
Item Aggregation
$\mathcal{L}_{\text{Recon}}$
f
b
Modal Disentanglement
RecSys
Base Recommendation Model
Relation-aware Fairness Learning
$\mathcal{L}_{\text{Rec}}$
Item Aggregation
RecSys
f_1
f_k
Parameter Share
Unfair Relation Mining
Fair Relation Mining
Weighted Summation
Fair Relation Mining
Unfair Relation Mining
Original/Filtered/Biased Modal Embedding
Original/Filtered User Embedding
Original/Filtered Item Embedding
User Representation based on Items' ID/Filtered-Modal/Biased-Modal Embedding
User/Item Indicator Embedding
Element-wise Addition/Subtraction
Sensitive Attribute Filter/Learner/Discriminator

Figure: The overall flowchart of our FMMRec.

Fairness-oriented Modal Disentanglement(1/4)

For disentanglement, our objective is to disentangle the original modal embedding $\boldsymbol{e}_v^m \in \mathbb{R}^{d_m}$ into two separate embeddings: $\bar{\boldsymbol{e}}_v^m \in \mathbb{R}^{d_m}$ which contains sensitive information as less as possible, and $\tilde{\boldsymbol{e}}_v^m \in \mathbb{R}^{d_m}$ which contains sensitive information as much as possible, respectively.



Fairness-oriented Modal Disentanglement (2/4)

- To **generate the mentioned two views of modal embedding**, we employ a filter network $f^m : \mathbb{R}^{d_m} \mapsto \mathbb{R}^{d_m}$ and a (biased) learner network $b^m : \mathbb{R}^{d_m} \mapsto \mathbb{R}^{d_m}$ to original modal embedding $\boldsymbol{e}_v^m$:

$$\begin{aligned} \bar{\boldsymbol{e}}_v^m &= f^m(\boldsymbol{e}_v^m), \\ \tilde{\boldsymbol{e}}_v^m &= b^m(\boldsymbol{e}_v^m). \end{aligned} \tag{11}$$

- To **detect how much sensitive information is learned** in the filtered and biased modal embeddings $\bar{\boldsymbol{e}}_v^m$ and $\tilde{\boldsymbol{e}}_v^m$, we adopt two sets of K discriminators D_m^f and D_m^b to infer the ground-truth values of K sensitive attributes of users.

$$\begin{aligned} \bar{\boldsymbol{a}}_{uk}^m = d_{m,k}^f(\bar{\boldsymbol{e}}_u^m),\ \bar{\boldsymbol{e}}_u^m = \frac{\sum_{v \in \mathcal{V}_u} \bar{\boldsymbol{e}}_{\boldsymbol{v}}^{\boldsymbol{m}}}{|\mathcal{V}_u|}, \\ \tilde{\boldsymbol{a}}_{uk}^m = d_{m,k}^b(\tilde{\boldsymbol{e}}_u^m),\ \tilde{\boldsymbol{e}}_u^m = \frac{\sum_{v \in \mathcal{V}_u} \tilde{\boldsymbol{e}}_{\boldsymbol{v}}^{\boldsymbol{m}}}{|\mathcal{V}_u|}, \end{aligned} \tag{12}$$

Fairness-oriented Modal Disentanglement (3/4)

- We employ **two attribute classification losses** for biased and filtered embeddings. Taking the binary attribute as an example:

$$\mathcal{L}_{D^f_m} = \sum_{k=1}^{K} \boldsymbol{a}_{uk} \cdot \log\left(\bar{\boldsymbol{a}}^m_{uk}\right) + (1 - \boldsymbol{a}_{uk}) \cdot \log\left(1 - \bar{\boldsymbol{a}}^m_{uk}\right), \tag{13}$$

$$\mathcal{L}_{D^b_m} = \sum_{k=1}^{K} \boldsymbol{a}_{uk} \cdot \log\left(\tilde{\boldsymbol{a}}^m_{uk}\right) + (1 - \boldsymbol{a}_{uk}) \cdot \log\left(1 - \tilde{\boldsymbol{a}}^m_{uk}\right), \tag{14}$$

- Moreover, we aim to **preserve the non-sensitive representative information of the filtered modal embedding $\bar{\boldsymbol{e}}^m_v$**, making it to be similar to the original modal embedding $\boldsymbol{e}^m_v$ for retaining recommendation performance.

$$\mathcal{L}_{\text{Recon}} = \frac{1}{|\mathcal{V}_u|} \sum_{v \in \mathcal{V}_u} 1 - \frac{\bar{\boldsymbol{e}}^m_v \cdot \boldsymbol{e}^m_v}{\|\bar{\boldsymbol{e}}^m_v\| \|\boldsymbol{e}^m_v\|}. \tag{15}$$

Fairness-oriented Modal Disentanglement (4/4)

- To further filter out the sensitive information in the filtered embedding and elicit sensitive information into the biased embedding, we **push the filtered and biased modal embeddings away from each other in the latent space** by an orthogonality loss:

$$\mathcal{L}_{\text{Orth}} = \frac{1}{|\mathcal{V}_u|} \sum_{v \in \mathcal{V}_u} \max(0, \frac{\bar{\boldsymbol{e}}_v^m \cdot \tilde{\boldsymbol{e}}_v^m}{\|\bar{\boldsymbol{e}}_v^m\| \|\tilde{\boldsymbol{e}}_v^m\|}). \tag{16}$$

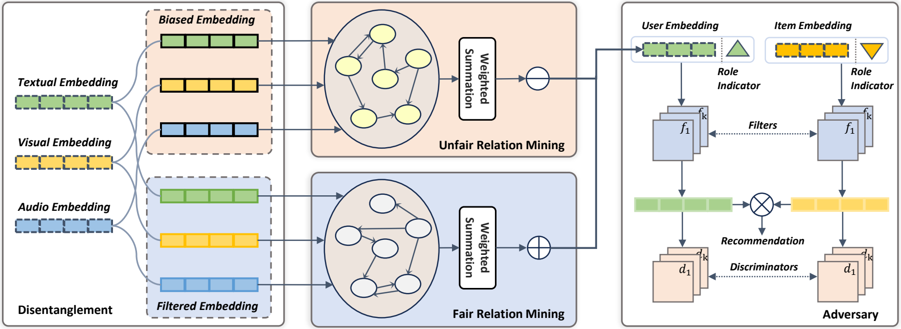
- Similar to adversarial training, we adopt a min-max game for the optimization of the filter network and corresponding discriminator, and optimize the disentanglement learning by **jointly optimizing these four losses with different importance**:

$$\arg\min_{f^m, b^m} \mathcal{L}_{\text{Recon}} + \lambda_{m_0}(\mathcal{L}_{D_m^b} - \mathcal{L}_{D_m^f}) + \lambda_{m_1}\mathcal{L}_{\text{Orth}}, \tag{17}$$

$$\arg\min_{D_m^f, D_m^b} \mathcal{L}_{D_m^f} + \mathcal{L}_{D_m^b}. \tag{18}$$

Relation-aware Fairness Learning (1/4)

We exploit the biased and filtered modal embeddings to mine unfair and fair user-user relations for promoting fairness and expressiveness of user representations, respectively.



Relation-aware Fairness Learning (2/4)

- We adopt the simple and parameter-free cosine similarity to compute **the modality-based user-user "fair" and "unfair" similarity matrix** $\bar{S}^m \in \mathbb{R}^{N\times N}$ and $\tilde{S}^m \in \mathbb{R}^{N\times N}$ by users' filtered/biased aggregated modality-based embedding $\bar{\mathbf{e}}_u^m / \tilde{\mathbf{e}}_u^m$:

$$\begin{aligned} \bar{S}_{uu'}^m &= \frac{\bar{\mathbf{e}}_u^m \cdot \bar{\mathbf{e}}_{u'}^m}{\|\bar{\mathbf{e}}_u^m\| \, \|\bar{\mathbf{e}}_{u'}^m\|} \; \forall u, u' \in \mathcal{U}, \\ \tilde{S}_{uu'}^m &= \frac{\tilde{\mathbf{e}}_u^m \cdot \tilde{\mathbf{e}}_{u'}^m}{\|\tilde{\mathbf{e}}_u^m\| \, \|\tilde{\mathbf{e}}_{u'}^m\|} \; \forall u, u' \in \mathcal{U}. \end{aligned} \tag{19}$$

- **Integrate unimodal adjacency matrices into a multimodal matrix**:

$$\begin{aligned} \bar{S} &= \sum_{m\in\mathcal{M}} \alpha_m \bar{S}^m, \\ \tilde{S} &= \sum_{m\in\mathcal{M}} \alpha_m \tilde{S}^m, \end{aligned} \tag{20}$$

Relation-aware Fairness Learning (3/4)

- **Aggregate the filtered/biased neighbors' representations of each user** u with the constructed modality-based similarities as weights:

$$\begin{aligned} \bar{\boldsymbol{h}}_u &= \sum_{u' \in \mathcal{N}(u;\bar{S})} \bar{S}_{uu'} \boldsymbol{e}_{\boldsymbol{u'}}, \\ \tilde{\boldsymbol{h}}_u &= \sum_{u' \in \mathcal{N}(u;\tilde{S})} \tilde{S}_{uu'} \boldsymbol{e}_{\boldsymbol{u'}}, \end{aligned} \tag{21}$$

- With the filtered and biased neighbor representations, we **incorporate the modality-based fair and unfair relations into the user representation** $\boldsymbol{e}_{\boldsymbol{u}}$ for enhancing its fairness and expressiveness:

$$\hat{\boldsymbol{e}}_{\boldsymbol{u}} = \boldsymbol{e}_{\boldsymbol{u}} + \lambda_h (\bar{\boldsymbol{h}}_u - \tilde{\boldsymbol{h}}_u), \tag{22}$$

Relation-aware Fairness Learning (4/4)

- **Generate filtered user representations** by compositional filters:

$$\bar{\boldsymbol{e}}_u = \frac{\sum_{k=1}^{K} f_k([\boldsymbol{r}_u; \hat{\boldsymbol{e}}_u])}{K} \tag{23}$$

- **Predict sensitive attribute** by discriminators:

$$\bar{\boldsymbol{a}}_{uk} = d_k^u(\bar{\boldsymbol{e}}_u) \tag{24}$$

$$\mathcal{L}_{D_u} = \sum_{k=1}^{K} \boldsymbol{a}_{uk} \cdot \log\left(\bar{\boldsymbol{a}}_{uk}\right) + \left(1 - \boldsymbol{a}_{uk}\right) \cdot \log\left(1 - \bar{\boldsymbol{a}}_{uk}\right), \tag{25}$$

- **Adversarial training**:

$$\underset{\Theta \backslash \{\boldsymbol{D_u}, \boldsymbol{D_v}\}}{\arg\min} \; \mathcal{L}_{\text{Rec}} - \lambda_{D_u}\mathcal{L}_{D_u} - \lambda_{D_v}\mathcal{L}_{D_v}, \tag{26}$$

$$\underset{D_u, D_v}{\arg\min} \, \mathcal{L}_{D_u} + \mathcal{L}_{D_v}, \tag{27}$$

Datasets

Table: Statistics of the two datasets used in our experiments, wherein V, T, and A denote the dimensions of visual, textual, and audio modalities, respectively.

Dataset	#Interactions	#Users	#Items	Sparsity	V	T	A
MovieLens	1, 000, 209	6,040	3,706	95.53%	1,000	384	128
MicroLens	123, 368	5,936	12,414	99.83%	1,000	768	128

Baselines

MMR baselines:

- **VBPR** [He and McAuley, 2016] is the first model that incorporates the visual features into recommender systems, treating visual features as another view of item representations;
- **MMGCN** [Wei et al., 2019] learns modality-specific representations of users and items based on the message-passing mechanism of graph neural network (GNN) for each modality, enhanced by a user-item bipartite graph;
- **LATTICE** [Zhang et al., 2021] mines the latent structure for multimodal recommendation to explicitly learn the semantic item-item relationships for each modality, and learn high-order item affinities based on graph convolutional network with the mined modality-based graphs;
- **FREEDOM** [Zhou, 2023] leverages the modality-based item-item graphs following the same approach as LATTICE, but with two notable differences that it freezes the mined graphs during training and incorporates the degree-sensitive edge pruning techniques to effectively reduce noise in the user-item graph;
- **DRAGON** [Zhou et al., 2023] improves dyadic relations in multimodal recommendations by constructing homogeneous graphs and learning dual representations for both users and items.

Fairness-aware baselines:

- **AL** [Wadsworth et al., 2018] first applies adversarial learning to the single user representation;
- **CAL** [Bose and Hamilton, 2019] introduces compositional filters for fair representation learning in multi-attribute scenarios;
- **FairGo** [Wu et al., 2021b] applies compositional filters to both user and item representations, and applies discriminators to explicit user representation and graph-based user representation, for which the first-order representation corresponds to implicit user representation in our experiments.

Overall performance (RQ1)

- **Overall performance**

Table: Experimental results of baselines for accuracy and fairness (*w.r.t.* gender, age, and occupation respectively) performance. **Bold** and underline for best and second-best results of fairness methods, respectively.

Methods	MovieLens								MicroLens			
	Accuracy		Fairness-Gen.		Fairness-Age		Fairness-Occ.		Accuracy		Fairness-Gen.	
	Recall ↑	NDCG ↑	AUC-E ↓	AUC-I ↓	F1-E ↓	F1-I ↓	F1-E ↓	F1-I ↓	Recall ↑	NDCG ↑	AUC-E ↓	AUC-I ↓
VBPR	0.2136	0.2033	0.7338	0.6472	0.4983	0.3882	0.2144	0.1722	0.0652	0.0331	0.6888	0.7533
MMGCN	0.2180	0.2110	0.7314	0.6279	0.4925	0.3858	0.2243	0.1689	0.0465	0.0231	0.7692	0.7813
LATTICE	0.2476	0.2378	0.7397	0.5428	0.5025	0.3725	0.2202	0.1747	0.0745	0.0382	0.7773	0.7781
FREEDOM	0.2423	0.2357	0.7158	0.6306	0.4826	0.3990	0.2012	0.1755	0.0648	0.0339	0.7644	0.7691
DRAGON	0.2387	0.2331	0.7117	0.5803	0.4884	0.3725	0.2169	0.1614	0.0860	0.0432	0.7582	0.7728
AL	0.2163	0.2066	**0.5172**	0.5750	**0.3560**	0.3891	0.1656	0.1780	0.0734	0.0368	0.6387	0.7748
CAL	0.2143	0.2035	0.5340	0.5594	0.3609	0.3717	0.1672	0.1780	0.0733	0.0375	0.6266	0.7786
FairGo	0.2133	0.2003	0.5431	**0.5000**	0.3659	0.3535	0.1639	0.1647	0.0720	0.0356	0.5932	0.5329
FMMRec	**0.2214**	**0.2079**	0.5224	**0.5000**	0.3576	**0.3526**	**0.1573**	**0.1507**	**0.0746**	**0.0379**	**0.5599**	**0.5050**

Overall performance (RQ1)

Observations.

- **Our FMMRec outperforms all baselines in terms of fairness performance** (i.e., lowest AUC/F1 values overall), indicating that it is vital to consider the sensitive information in multimodal representations for improving fairness in multimodal recommendations.
- Compared with the base multimodal recommender baselines (i.e., LATTICE on the MovieLens dataset and DRAGON on the MicroLens dataset), **applying all fairness-aware methods leads to an accuracy drop**. Notably, FMMRec not only delivers superior fairness performance but also outperforms other fairness-aware methods in terms of accuracy.

Disentanglement Performance (RQ2)

Disentanglement Performance

Our expectations:

- more sensitive information to be learned in the biased modal embedding
- And less sensitive information to be leaked in the filtered modal embedding

Than the original modal embedding

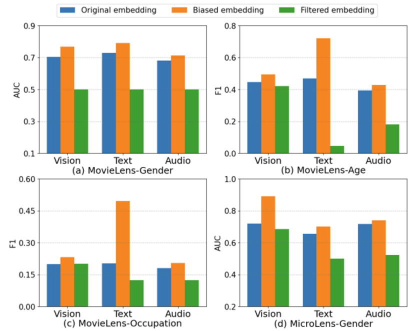


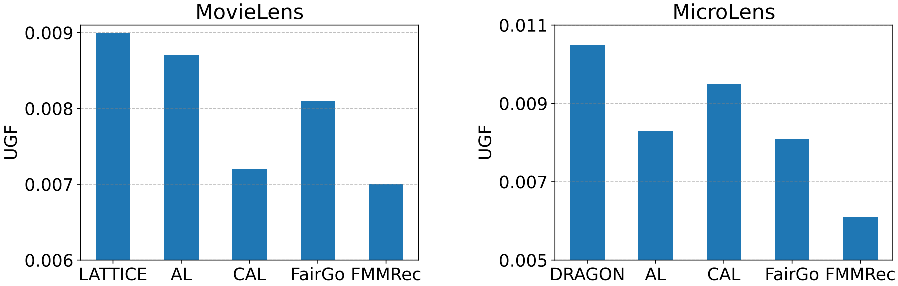
Figure 4: Disentanglement performance (i.e., sensitive attribute prediction accuracy) of each modality on the four dataset-attribute pairs (e.g., MovieLens-Gender denotes the protected attribute gender of the MovieLens dataset).

Ablation Study (RQ3)

Table: Performance of FMMRec with different variants on MovieLens dataset.

Variants	**MovieLens**							
	Accuracy		**Fairness-Gen.**		**Fairness-Age**		**Fairness-Occ.**	
	Recall ↑	NDCG ↑	AUC-E ↓	AUC-I ↓	F1-E ↓	F1-I ↓	F1-E ↓	F1-I ↓
w/o FMD	0.2048	0.1952	0.5902	0.5037	0.3709	**0.3526**	0.1755	0.1531
w/o FRM	0.2194	**0.2082**	0.5587	**0.5000**	0.3584	**0.3526**	0.1623	0.1623
w/o UFRM	0.2162	0.2050	0.5321	**0.5000**	0.3609	**0.3526**	0.1598	0.1689
w/o RI	0.2048	0.1923	0.5225	0.5003	**0.3568**	0.3535	0.1623	0.1623
FMMRec	**0.2214**	0.2079	**0.5224**	**0.5000**	0.3576	**0.3526**	**0.1573**	**0.1507**

Group Fairness Performance (RQ4)



Outline

1 Literature Review

2 Associative and Causal Fairness for Multi-Type Sensitive Attributes

3 Fairness-aware Multimodal Recommendation

4 Controllable Fairness for Recommender Systems

Motivation

- **Limitations of Existing Methods:**
  - *Fixed fairness constraints*: Most fairness-aware methods fix the desired fairness level at training time, lacking flexibility afterward.
  - *Retraining overhead*: Adjusting fairness levels typically demands complete retraining, which is computationally expensive.
- **Controllable Fairness on the Fly:**
  - We introduce a framework (**Cofair**) allowing post-training control of fairness without retraining.
  - Aligns with dynamic demands: diverse stakeholders may request adjustment on fairness requirements over time.

Related Work

Recently, some approaches have begun to address the inflexibility issue in fairness deployment. For instance,

- Some methods [Zhu et al., 2024, Li et al., 2021d] offer more flexibility in terms of selecting **which** attributes to protect, they have limited control over **how much** the fairness criteria (a.k.a. *different fairness levels*).
- Some studies [Song et al., 2019, Cui et al., 2023] provide theoretical fairness guarantees by allowing stakeholders to specify unfairness limits in training.
  However, their constrained optimization frameworks require **retraining** to adjust fairness levels, making them computationally impractical for real-world deployment.

Problem Formulation

- **Notation:**
  - Let $\mathcal{U} = \{1, 2, \ldots, U\}$ be the set of users, and $\mathcal{I} = \{1, 2, \ldots, I\}$ be the set of items.
  - The observed interactions are denoted by $\mathcal{D} = \{(u, i)\}$, where $u \in \mathcal{U}$ and $i \in \mathcal{I}$.
- **Sensitive Attributes:**
  - Each user u has a sensitive feature $a_u \in \{0, 1\}$ (binary for simplicity, but could extend to multi-class).
  - $\mathcal{G}_0 = \{u : a_u = 0\}$, $\mathcal{G}_1 = \{u : a_u = 1\}$.
- **Objective:**
  - Produce user representation $\mathbf{e}_u$ and item representation $\mathbf{e}_i$.
  - Optimize for ranking quality while controlling group fairness *on the fly* (i.e., at inference).

Fairness Metrics

Demographic Parity (DP)

$$\Delta_{\mathrm{DP}} = \left|\mathbb{E}_{u\in\mathcal{G}_1}[G(\mathbf{e}_u)] - \mathbb{E}_{u\in\mathcal{G}_0}[G(\mathbf{e}_u)]\right|$$

Smaller Δ_{DP} is more fair; $G(\cdot)$ is an outcome function (e.g., predicted score $\hat{y}_{u,i}$).

Monotonic Fairness Requirement

We want a sequence of fairness levels $\lambda_1 < \lambda_2 < \cdots < \lambda_T$ such that each user's fairness does not degrade as we move from lower to higher levels.

Model Architecture

- **Shared Representation Layer**: Learns a base user representation capturing overall preference patterns.

$$\mathbf{s}_u = S(\mathbf{e}_u; \theta_s), \tag{28}$$

- **Fairness-Conditioned Adapters**: Each corresponding to a different fairness level λ_t.

$$\mathbf{p}_u^{(t)} = P^{(t)}\left(\mathbf{e}_u; \theta_p^{(t)}\right). \tag{29}$$

- **Output Layer**: Combine shared and fairness-conditioned representation:

$$\mathbf{e}_u^{(t)} = O\Big(\left[\mathbf{s}_u; \mathbf{p}_u^{(t)}\right]; \theta_o\Big). \tag{30}$$

Loss Design for Cofair

Recommendation Loss (BPR)

$$\mathcal{L}_{\text{rec}}(\theta) = -\sum_{(u,i,j)\in\mathcal{D}} \ln\, \sigma\big(\hat{y}_{u,i} - \hat{y}_{u,j}\big).$$

Fairness Loss (Adversarial)

$$\mathcal{L}_{\text{fair}}(\theta) = -\sum_{u\in\mathcal{U}} \mathrm{BCE}\big(D(\mathbf{e}_u^{(t)}),\, a_u\big).$$

User-Level Regularization

$$\mathcal{L}_{\text{reg}} = \sum_{u\in\mathcal{U}} \sum_{t=1}^{T-1} \mathrm{softplus}\Big(\mathcal{L}_{\text{fair}}^{(t)}(u) - \mathcal{L}_{\text{fair}}^{(t+1)}(u)\Big).$$

$\mathcal{L}(\theta) = \sum_{t=1}^{T}\left[\mathcal{L}_{\text{rec}}^{(t)} + \lambda_t \mathcal{L}_{\text{fair}}^{(t)}\right] + \beta\,\mathcal{L}_{\text{reg}}$, with an adaptive update for λ_t.

Experimental Setup: Dataset

Datasets:

- **Movielens-1M [Harper and Konstan, 2015]** is a classic movie-rating dataset with dense user-item interactions. We treat any rating above 0 as a positive interaction, following [Zhao et al., 2023, Islam et al., 2021].
- **Lastfm-360K [Celma Herrada et al., 2009]** is a large music recommendation dataset with play records of users from Last.FM. We performed 20-core filtering and sampled a subset following [Zhao et al., 2023]

We treat **gender** as sensitive attribute, which is the most widely considered sensitive attribute in the fairness literature [Wang et al., 2023, Li et al., 2023, Deldjoo et al., 2024].

Experimental Setup: Evaluation

Evaluation Metrics:

- Ranking Accuracy: Recall@10 and NDCG@10
- Fairness: DP@10 and EOpp@10

$$\begin{gathered}\forall v \in V, f_{G_0}^{v} = \frac{\sum_{u \in G_0} \mathbf{I}_{v \in TopK_u}}{|G_0|}, f_{G_1}^{v} = \frac{\sum_{u \in G_1} \mathbf{I}_{v \in TopK_u}}{|G_1|}, \\ \mathbf{f}_{G_0} = \left[f_{G_0}^{1}, \ldots, f_{G_0}^{v}, \ldots, f_{G_0}^{N}\right], \mathbf{f}_{G_1} = \left[f_{G_1}^{1}, \ldots, f_{G_1}^{v}, \ldots, f_{G_1}^{N}\right],\end{gathered} \tag{31}$$

$$DP@K = JSD\left(\mathbf{f}_{G_0}, \mathbf{f}_{G_1}\right) \tag{32}$$

$$\begin{gathered}\forall v \in V, d_{G_0}^{v} = \frac{\sum_{u \in G_0} \mathbf{I}_{v \in \mathbf{R}_u^t \cap TopK_u}}{|G_0|}, d_{G_1}^{v} = \frac{\sum_{u \in G_1} \mathbf{I}_{v \in \mathbf{R}_u^t \cap TopK_u}}{|G_1|} \\ \mathbf{d}_{G_0} = \left[d_{G_0}^{1}, \ldots, d_{G_0}^{v}, \ldots, d_{G_0}^{N}\right], \mathbf{d}_{G_1} = \left[d_{G_1}^{1}, \ldots, d_{G_1}^{v}, \ldots, d_{G_1}^{N}\right],\end{gathered} \tag{33}$$

$$EOpp@K = JSD\left(\mathbf{d}_{G_0}, \mathbf{d}_{G_1}\right) \tag{34}$$

Experimental Setup: Baselines

Standard collaborative filtering: BPR, LightGCN
Fairness-aware baselines:

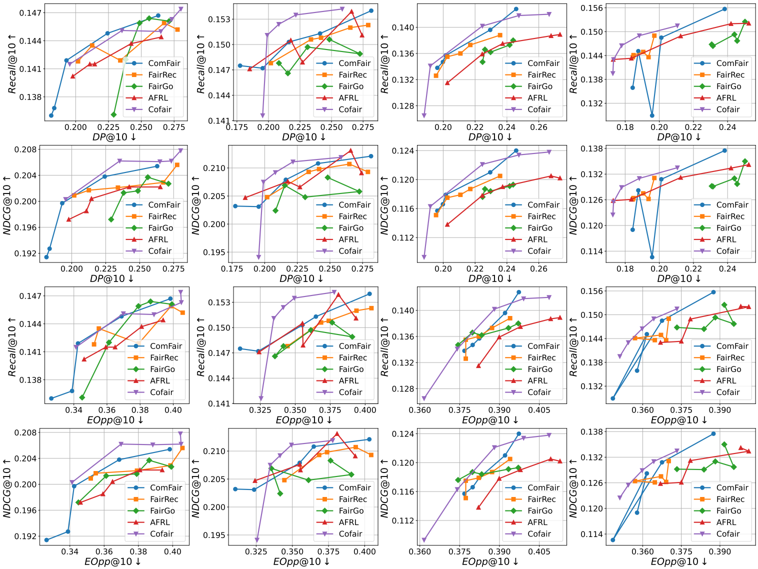
- **ComFair [Bose and Hamilton, 2019]** applies compositional adversarial learning to eliminate sensitive multi-attribute information.
- **FairRec [Wu et al., 2021a]** decomposes adversarial learning to generate a bias-free user representation with minimized sensitive information and a bias-aware user representation.
- **FairGo [Wu et al., 2021b]** applies compositional filters to both user and item representations to user representation and graph-based high-order user representation.
- **AFRL [Zhu et al., 2024]** adaptively learns fair representations by treating fairness requirements as inputs.

Research Questions

We structure our empirical analysis for the following questions:

- **RQ1:** Can Cofair achieve various fairness-accuracy trade-offs without retraining, and how does it compare to baselines?
- **RQ2:** How do the components of our proposed Cofair contribute to the accuracy and fairness performance across levels?
- **RQ3:** How do different hyperparameter configurations influence the performance of our proposed Cofair?
- **RQ4:** Can Cofair enable post-training controllability of fairness for other fairness methods, demonstrating its generalizability?
- **RQ5:** Does Cofair reduce computational overhead?

Overall Performance (RQ1)



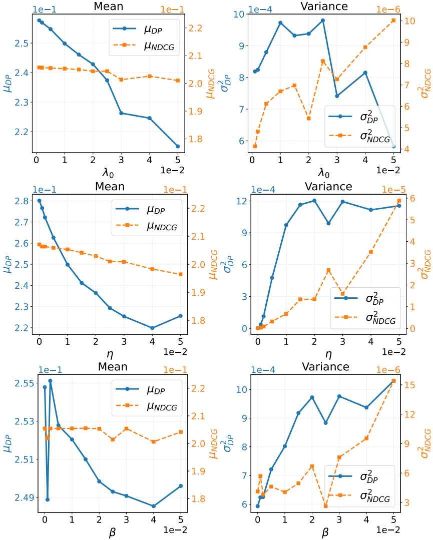
Ablation Study (RQ2)

Table: Ablation study of our proposed method (Cofair) versus its variants without specific components. We report results at five different fairness levels.

Method	**NDCG@10 (*larger* is better)**						
	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	**avg.**	**std.**
w/o SRL	0.2050	0.2044	0.2051	0.2027	0.1945	0.2023	**0.0040**
w/o FCA	0.2044	0.2044	0.2044	0.2044	0.2044	0.2044	0.0000
w/o AWL	0.2073	**0.2069**	**0.2073**	**0.2076**	**0.2063**	**0.2071**	0.0004
w/o URL	**0.2080**	0.2067	0.2064	0.2033	0.2028	0.2054	0.0020
Cofair	0.2015	0.1996	0.1980	0.1987	0.1913	0.1978	0.0035
Improv.	-3.13%	-3.43%	-4.07%	-2.26%	-5.67%	-3.70%	+75.00%

Method	**DP@10 (*smaller* is better)**						
	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	**avg.**	**std.**
w/o SRL	0.2828	0.2716	0.2623	0.2505	0.2403	0.2615	0.0150
w/o FCA	**0.2615**	0.2615	0.2615	0.2615	0.2615	0.2615	0.0000
w/o AWL	0.2802	0.2785	0.2832	0.2803	0.2783	0.2801	0.0018
w/o URL	0.2727	0.2691	0.2652	0.2601	0.2068	0.2548	0.0244
Cofair	0.2707	**0.2508**	**0.2329**	**0.2017**	**0.1891**	**0.2290**	**0.0302**
Improv.	+0.73%	+6.80%	+12.18%	+22.46%	+8.56%	+10.13%	+23.77%

Hyperparameter Analysis (RQ3)



- λ_0 serves as the initial fairness coefficient at the lowest fairness level ($t = 1$)
- η is the learning rate by which fairness coefficients are updated
- β controls the strength of user-level regularization

Framework Study (RQ4)

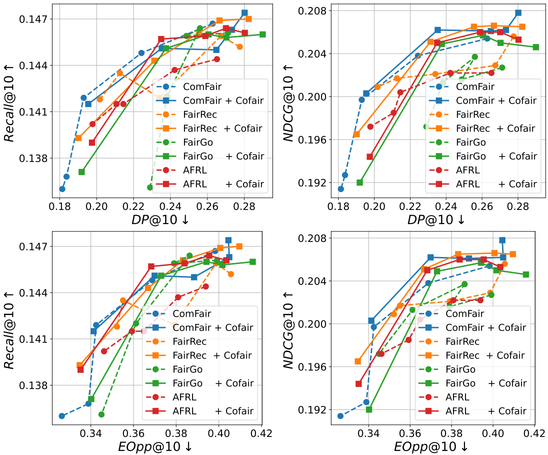


Figure: Comparison of four fairness methods (ComFair, FairRec, FairGo, and AFRL) without and with our *controllable fairness* framework (indicated by ”+ Cofair”).

Efficiency Analysis (RQ5)

Table: Average training time per epoch (in seconds) and the average number of best epochs for fairness methods, for obtaining results for five different fairness levels.

Method		Movielens-1M		Lastfm-360K	
		Time	Epoch	Time	Epoch
BPR	+ ComFair	3.10	78.57	5.15	132.00
	+ FairRec	3.06	44.00	16.36	141.67
	+ FairGo	9.66	114.38	8.68	175.33
	+ AFRL	2.81	57.14	7.56	103.50
	+ Cofair	4.79	7.44	12.24	26.04
LightGCN	+ ComFair	3.64	129.00	23.93	134.50
	+ FairRec	10.42	122.50	27.93	185.56
	+ FairGo	13.26	99.00	22.57	215.67
	+ AFRL	3.53	40.00	11.33	55.00
	+ Cofair	7.55	11.04	27.79	35.58

Future Work

- **Fairness in Cross-domain Recommendation.**
  Transferring knowledge across domains via overlapped users may disadvantage non-overlapped users.
- **LLM-related Fairness.**
  LLMs themselves can embed societal biases, complicating fairness.
- **Fairness with Diversity & Explanations.**
  Providing users with diverse recommendations is beneficial for preventing “filter bubbles” and enhancing user satisfaction.
- **Consumer-Provider Fairness and Beyond.**
  Fair recommendation is multi-sided, thus beyond user-side fairness, future work should incorporate item providers’ interests.

List of Publications

1. WEIXIN CHEN, LI CHEN, YONGXIN NI, YUHAN ZHAO, *Causality-Inspired Fair Representation Learning for Multimodal Recommendation*, **ACM Transactions on Information Systems**, 2025.
   [*under 2nd-round review (1st-round decision: major revision)*]

2. WEIXIN CHEN, LI CHEN, YUHAN ZHAO, *Investigating User-Side Fairness in Outcome and Process for Multi-Type Sensitive Attributes in Recommendations*, **ACM Transactions on Recommender Systems**, 2025.
   [*under 2nd-round review (1st-round decision: major revision)*]

3. WEIXIN CHEN, LI CHEN, YUHAN ZHAO, *Fairness on the Fly: Controllable Fairness for Recommender Systems without Retraining*, **ACM SIGIR Conference on Research and Development in Information Retrieval**, 2025.
   [*under review*]

QA

Thanks!

Any Questions?

Bose, A. and Hamilton, W. (2019).
Compositional fairness constraints for graph embeddings.
In *ICML*.

Bruch, S., Zoghi, M., Bendersky, M., and Najork, M. (2019).
Revisiting approximate metric optimization in the age of deep neural networks.
In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1241–1244.

Celma Herrada, Ò. et al. (2009).
Music recommendation and discovery in the long tail.
Universitat Pompeu Fabra.

Cheng, H., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., and Shah, H. (2016).
Wide & deep learning for recommender systems.
In *DLRS@RecSys*.

Cui, Y., Chen, M., Zheng, K., Chen, L., and Zhou, X. (2023).
Controllable universal fair representation learning.
In *Proceedings of the ACM Web Conference 2023 (WWW'23)*, pages 949–959.

Deldjoo, Y., Jannach, D., Bellogin, A., Difonzo, A., and Zanzonelli, D. (2024).
Fairness in recommender systems: research landscape and future directions.
User Modeling and User-Adapted Interaction (UMUAI'24), 34(1):59–108.

Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., and Pera, M. S. (2018).
All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness.
In *Conference on fairness, accountability and transparency*, pages 172–186. PMLR.

Harper, F. M. and Konstan, J. A. (2015).
The movielens datasets: History and context.
Acm Transactions on Interactive Intelligent Systems, 5(4):1–19.

He, R. and McAuley, J. J. (2016).
VBPR: visual bayesian personalized ranking from implicit feedback.
In *AAAI*.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004).
Evaluating collaborative filtering recommender systems.
ACM Transactions on Information Systems (TOIS), 22(1):5–53.

Huber, P. J. (1992).
Robust estimation of a location parameter.
Breakthroughs in statistics: Methodology and distribution.

Islam, R., Keya, K. N., Zeng, Z., Pan, S., and Foulds, J. (2021).
Debiasing career recommendations with neural fair collaborative filtering.
In *Proceedings of the Web Conference 2021 (WWW'21)*, pages 3779–3790.

Kaminskas, M. and Bridge, D. (2016).
Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems.
ACM Transactions on Interactive Intelligent Systems (TiiS), 7(1):1–42.

Koren, Y., Bell, R. M., and Volinsky, C. (2009).
Matrix factorization techniques for recommender systems.
Computer.

Li, Y., Chen, H., Fu, Z., Ge, Y., and Zhang, Y. (2021a).
User-oriented fairness in recommendation.
In *Proceedings of the Web Conference 2021*, pages 624–632.

Li, Y., Chen, H., Fu, Z., Ge, Y., and Zhang, Y. (2021b).
User-oriented fairness in recommendation.
In *WWW*.

Li, Y., Chen, H., Xu, S., Ge, Y., Tan, J., Liu, S., and Zhang, Y. (2023).

Fairness in recommendation: Foundations, methods and applications.
ACM Transactions on Intelligent Systems and Technology, 14(5):95:1–95:48.

Li, Y., Chen, H., Xu, S., Ge, Y., and Zhang, Y. (2021c).
Towards personalized fairness based on causal notion.
In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1054–1063.

Li, Y., Chen, H., Xu, S., Ge, Y., and Zhang, Y. (2021d).
Towards personalized fairness based on causal notion.
In *SIGIR*.

Li, Y., Ge, Y., and Zhang, Y. (2021e).
Tutorial on fairness of machine learning in recommender systems.
In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2654–2657.

Liu, H., Tang, D., Yang, J., Zhao, X., Liu, H., Tang, J., and Cheng, Y. (2022).
Rating distribution calibration for selection bias mitigation in recommendations.
In *Proceedings of the ACM Web Conference 2022*, pages 2048–2057.

Melchiorre, A. B., Zangerle, E., and Schedl, M. (2020).
Personality bias of music recommendation algorithms.
In *Fourteenth ACM conference on recommender systems*, pages 533–538.

Salakhutdinov, R. and Mnih, A. (2007).
Probabilistic matrix factorization.
In *NeurIPS*.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. (2019).
Learning controllable fair representations.
In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS'19)*, pages 2164–2173.

Wadsworth, C., Vera, F., and Piech, C. (2018).

Achieving fairness through adversarial learning: An application to recidivism prediction.
In *FAT/ML*.

Wang, Y., Ma, W., Zhang*, M., Liu, Y., and Ma, S. (2023).
A survey on the fairness of recommender systems.
ACM Transactions on Information Systems.

Wei, Y., Wang, X., Nie, L., He, X., Hong, R., and Chua, T. (2019).
MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video.
In *MM*.

Wu, C., Wu, F., Wang, X., Huang, Y., and Xie, X. (2021a).
Fairness-aware news recommendation with decomposed adversarial learning.
In *AAAI*.

Wu, L., Chen, L., Shao, P., Hong, R., Wang, X., and Wang, M. (2021b).
Learning fair representations for recommendation: A graph-based perspective.
In *WWW*.

Xue, H., Dai, X., Zhang, J., Huang, S., and Chen, J. (2017).
Deep matrix factorization models for recommender systems.
In *IJCAI*.

Yao, S. and Huang, B. (2017).
Beyond parity: Fairness objectives for collaborative filtering.
In *NeurIPS*.

Zhang, J., Zhu, Y., Liu, Q., Wu, S., Wang, S., and Wang, L. (2021).
Mining latent structures for multimedia recommendation.
In *MM*.

Zhao, C., Wu, L., Shao, P., Zhang, K., Hong, R., and Wang, M. (2023).
Fair representation learning for recommendation: A mutual information perspective.
In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI'23)*, pages 4911–4919.

Zhou, H., Zhou, X., and Shen, Z. (2023).
Enhancing dyadic relations with homogeneous graphs for multimodal recommendation.
In *ECAI*.

Zhou, X. (2023).
Mining latent structures for multimedia recommendation.
In *MM*.

Zhu, X., Zhang, L., and Yang, N. (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)*, pages 427–436.