Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Interaction-Data-guided Conditional Instrumental Variables for Debiasing Recommender Systems

Zhirong Huang[1,2], **Debo Cheng**[*3,4], **Lin Liu**[4], **Jiuyong Li**[4], **Guangquan Lu**[1,2] and **Shichao Zhang**[*1,2]

[1]Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin, 541004, China

[2]Guangxi Key Lab of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin, 541004, China

[3]School of Computer Science and Technology, Hainan University, Haikou, Hainan, 570228, China

[4]UniSA STEM, University of South Australia, Mawson Lakes, Adelaide, Australia

huangzr@stu.gxnu.edu.cn, chengdb2016@gmail.com,
{zhangsc,lugq}@mailbox.gxnu.edu.cn, {Lin.Liu, Jiuyong.Li}@unisa.edu.au

Abstract

It is often challenging to identify a valid instrumental variable (IV), although the IV methods have been regarded as effective tools of addressing the confounding bias introduced by latent variables. To deal with this issue, an Interaction-Data-guided Conditional IV (IDCIV) debiasing method is proposed for Recommender Systems, called IDCIV-RS. The IDCIV-RS automatically generates the representations of valid CIVs and their corresponding conditioning sets directly from interaction data, significantly reducing the complexity of IV selection while effectively mitigating the confounding bias caused by latent variables in recommender systems. Specifically, the IDCIV-RS leverages a variational autoencoder (VAE) to learn both the CIV representations and their conditioning sets from interaction data, followed by the application of least squares to derive causal representations for click prediction. Extensive experiments on two real-world datasets, Movielens-10M and Douban-Movie, demonstrate that IDCIV-RS successfully learns the representations of valid CIVs, effectively reduces bias, and consequently improves recommendation accuracy.

1 Introduction

With the rapid development of the Internet, the amount of information has exploded, making it increasingly difficult for users to sift through vast amounts of data to find content that aligns with their preferences [Luo *et al.*, 2025; Gao *et al.*, 2021; Zhang, 2021]. Recommender systems have emerged as a critical solution to this problem by analysing user behaviour data to deliver personalised recommendations, thereby enhancing user engagement and satisfaction [Wang *et al.*, 2020]. Recommender systems have become an integral part of many digital platforms, finding extensive applications across various domains, including e-commerce [Shoja and Tabrizi, 2019], streaming media [Gomez-Uribe and Hunt, 2015], and social networks [Liao *et al.*, 2022], significantly improving information retrieval efficiency and user experience.

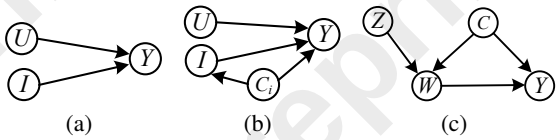

Figure 1: Three causal graphs illustrate the assumptions underlying existing work. The variables are defined as follows: U: user preferences, I: displayed items, Y: user feedback or outcome variables, C_i: confounding factors affecting items, C: confounding factors, Z: instrumental variables, W: treatment variables (i.e., embeddings of user-item pairs). (a) The causal graph represents the traditional recommendation model; (b) the causal graph represents the debiasing method using causality; (c) the causal graph representation of the debiasing method using standard IV.

The performance of recommendation systems is often affected by latent confounders that are neither directly observable nor reflected in historical user-item interactions. For instance, factors such as social influence or peer preferences can significantly shape user behavior but are rarely captured in logged data. This misalignment may cause to recommendation systems misinterpret user intent, ultimately degrading recommendation quality. Effectively addressing latent confounders is thus vital for enhancing both the accuracy and robustness of recommender systems.

User behaviour data is crucial for recommender systems in predicting user preferences. Existing models often assume this data is unbiased and accurately reflects user preferences, meaning the data accurately reflects user preferences [Lan *et al.*, 2024]. Based on this assumption, many methods have

*Corresponding author

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

been proposed, such as Matrix Factorisation (MF) [Koren *et al.*, 2009] and Neural Network-based Collaborative Filtering (NCF) [He *et al.*, 2017]. These methods achieve the goal of predicting user preferences by fitting user behaviour data, as depicted in the causal Directed Acyclic Graph (DAG) [Pearl, 2009; Cheng *et al.*, 2022] shown in Figure 1 (a), which illustrates the causal relationships between user preferences (U) and user feedback (Y), as well as between displayed items (I) and Y. However, in the real-world, user behaviour is inevitably influenced by various unobservable confounding factors, such as item popularity (e.g., frequent recommendation of an item limits users' choices) and user psychology (e.g., choosing to watch a movie for socialising) [Chen *et al.*, 2023]. These factors introduce biases like popularity and conformity bias, leading models to learn false correlations and reducing their ability to accurately predict preferences.

Causal inference [Zhang *et al.*, 2024; Li *et al.*, 2024] has been applied to reduce bias in recommender systems by designing causal DAGs to model data generation, identify biases, and guide model design [Wang *et al.*, 2020; Gao *et al.*, 2024]. For instance, Zhang et al. [Zhang *et al.*, 2021] introduced a causal graph to analyse item popularity's impact and proposed the PDA training paradigm to correct popularity bias. Similarly, Zheng et al. [Zheng *et al.*, 2021] addressed conformity bias with the DICE model, disentangling user interest from conformity. However, these methods rely on the assumption that the real data follows the designed causal graphs, which may not hold in practice [Cai *et al.*, 2024]. Complex real-world confounding makes such assumptions difficult to satisfy, potentially limiting the models' effectiveness.

Latent confounders, which are unobserved variables that simultaneously affect both the treatment (e.g., recommendation process) and the outcomes, present significant challenges for debiasing recommender systems. Instrumental Variables (IVs) are a common solution to this issue [Caner and Hansen, 2004; Pearl, 2009], with a valid IV (e.g., Z in Figure 1(c)) satisfying three key criteria [Pearl, 2009]: (i) relevance to the treatment variable; (ii) an exclusive impact on the outcome through the treatment; and (iii) no shared confounders with the outcome. Several IV-based methods have been proposed for recommendation settings to reduce latent bias without relying on strict causal graph assumptions [Si *et al.*, 2022; Si *et al.*, 2023a; Si *et al.*, 2023b]. For example, IV4Rec [Si *et al.*, 2022] leverages self-collected search data as an IV to effectively address latent confounding. However, verifying the latter two IV conditions from observational data alone is practically infeasible [Brito and Pearl, 2012; Cheng *et al.*, 2023b], making the identification of valid IVs particularly challenging.

In causal inference, researchers have used conditional IVs (CIVs) to tackle the limitations of standard IVs [Pearl, 2009; Brito and Pearl, 2012; Cheng *et al.*, 2024a; Cheng *et al.*, 2023a]. CIVs offer more relaxed application conditions than standard IVs. Recently, Cheng et al. [Cheng *et al.*, 2023b] developed a CIV method (CIV.VAE) based on the variational autoencoder (VAE) [Kingma and Welling, 2013; Schölkopf *et al.*, 2021; Schölkopf, 2022] model, which generates CIVs and their conditional sets from data, significantly relaxing the constraints of standard IVs. However, this method is designed for tabular data, and no attempt has been made to adapt it to interactive data in recommender systems.

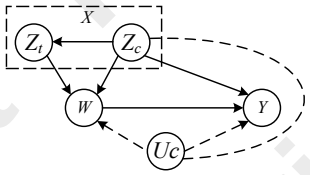

Figure 2: A causal DAG illustrating our proposed IDCIV-RS method for discovering CIVs and their conditional sets from observational data. W, Y, X, and U_c are the treatment, outcome, set of measured pretreatment variables, and latent confounders between W and Y, respectively. Z_t and Z_c denote the representations of the CIVs and their conditional sets learned from X.

To address the challenge of using CIV in interactive data, we first propose a causal DAG, as shown in Figure 2, to represent the causal relationships between observed and latent variables in interaction data within recommender systems. Building on this causal DAG, we develop an interaction-data-guided conditional IV (IDCIV) debiasing method for recommender systems, called IDCIV-RS. Specifically, IDCIV-RS uses the embeddings of user-item pairs (including users with the selected items, i.e., positive samples, and users with their pre-selected items, i.e., negative samples) as the treatment variable W, the user feedback as the outcome Y, and the user interaction data (only positive samples) as the pretreatment variable X. We assume that at least one CIV exists within X, capturing latent information such as user search behaviours that lead to interactions. The assumption is reasonable because in real-world scenarios, users are often influenced by certain external factors that cause their interactions to be not fully consistent with their true preferences. For example, users may interact with items because of certain external incentives (e.g., search recommendations, promotions, etc.) even though these items do not exactly match their true preferences. IDCIV-RS employs a VAE to generate the representations of the CIV and its conditional set from X, denoted as Z_t and Z_c, as shown in Figure 2. The contribution of our work is summarised below:

- We propose a novel causal DAG to represent the causal relationships between observed and latent variables in interaction data within recommender systems.
- We develop an interaction-data-guided conditional IV (IDCIV) debiasing method for recommender systems, called IDCIV-RS, for learning representations of CIVs and its conditional sets under the proposed causal DAG. To the best of our knowledge, this is the first work to generate representations of CIVs and their conditional sets from interaction data for mitigating bias in recommender systems.
- Extensive experiments on two real-world datasets demonstrate that IDCIV-RS achieves optimal debiasing results and recommendation performance compared to state-of-the-art causal debiasing methods.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

2 Related Work

In this section, we review the recommender methods most closely related to our IDCIV-RS, including traditional recommender methods and causal recommender methods.

2.1 Traditional Recommender Methods

Traditional recommender methods, primarily based on Collaborative Filtering (CF), typically assume that user behaviour data is unbiased [Koren *et al.*, 2021]. The mainstream approach is model-based CF, which trains a model on user behaviour data to recommend items that align with user preferences. A classic method is MF [Koren *et al.*, 2009], which decomposes the user-item rating matrix to predict preferences. However, MF assumes that unselected items are incompatible with user preferences, ignoring cases where users may not have encountered those items. To address this, Rendle et al. [Rendle *et al.*, 2012] proposed Bayesian Personalized Ranking (BPR), which assumes users prefer selected items over unselected ones, enhancing preference inference in MF.

With the rise of deep learning [Luo *et al.*, 2024; Zhang *et al.*, 2022], He et al. [He *et al.*, 2017] proposed NCF, which uses multi-layer perceptrons (MLPs) to model non-linear user preferences. To capture richer behavioural signals, Wang et al. [Wang *et al.*, 2019] introduced Neural Graph Collaborative Filtering (NGCF), leveraging Graph Convolutional Networks (GCNs) to embed user-item interactions. He et al. [He *et al.*, 2020] further simplified this with LightGCN, improving both efficiency and accuracy. However, these methods often overlook popularity bias, which can be amplified during training and skew recommendations toward popular items.

2.2 Causal Recommender Methods

To mitigate biases in recommender systems, researchers have increasingly adopted causal inference techniques. Early approaches used the Inverse Propensity Score (IPS) [Wang *et al.*, 2021; Schnabel *et al.*, 2016; Bottou *et al.*, 2013] to reduce bias by assigning an inverse propensity score (e.g., the inverse of item popularity) to user-item interactions during training, balancing the influence of popular and less popular items. However, IPS methods often suffer from high variance and instability. Building on IPS success, researchers have explored causal graph-based methods [Zheng *et al.*, 2021; He *et al.*, 2023; Zhang *et al.*, 2021] that model the generation mechanisms of user behaviour. These methods design specific models to address biases like popularity and conformity bias. Yet, the presence of unobserved confounders in real-world data limits the effectiveness of these causal graph assumptions [Cai *et al.*, 2024].

IVs are commonly used in causal inference to address confounding. Recently, methods leveraging user search data as IVs have emerged, with the IV4Rec framework by Si et al. [Si *et al.*, 2022] being a notable example. IV4Rec uses user search data to decompose user-item representations into causal and non-causal components, addressing some limitations of causal graph-based methods and reducing bias. However, identifying valid user search data as IVs remains challenging. Unlike these approaches, our work focuses on learning the representations of CIVs and their conditional sets, which are less restrictive than standard IVs.

3 The Proposed IDCIV-RS Method

In this section, we first introduce the problem definition, then explain the feasibility and rationality of our method by the causal graph, and then introduce the four main steps of our method. The overall workflow of our proposed IDCIV-RS is visualised in Figure 3.

3.1 Problem Definition

In the recommender system, user behaviour data D usually consists of a user set U and an item set I. D contains two parts, namely: user u and selected items p to form positive sample pairs, and user u and pre-selected items n to form negative sample pairs. The user interaction data X consists of positive sample pairs derived from D. X implicitly contains a wealth of information, including user interactions stemming from search behaviours.

In recommendation models, users and items are usually represented as low-dimensional embedding representations W, and the corresponding user-item pairs can be represented as $W = \{(w_u, w_i) \,|u \in U, i \in I\}$. However, in addition to reflecting user preferences, user behaviour data D contains spurious correlations caused by various latent confounding factors U_c (e.g., item exposure, conformity influence). Although existing methods have mitigated the impact of these confounding factors to some extent through causal inference techniques, they often come with strong assumptions. We aim to address this challenging problem in our work, and our problem definition is described as follows.

Definition 1. *In a recommender system, the latent variables U_c affect the choice made and introduce bias. We assume that at least one CIV exists within X, capturing latent information that leads to interactions. The causal relationships between measured and latent variables are shown in Figure 2. Our goal is to learn the representations of the CIV Z_t and its conditional set Z_c from the user interaction data X to address the confounding biases introduced by U_c.*

3.2 The Proposed Causal DAG

In this work, we proposed a causal DAG $\mathcal{G}$ as shown in Figure 2 to represent the causal relationships between the measured and latent variables. Let $\mathcal{G}_{\underline{W}}$ be the manipulated graph, obtained by deleting all arrows emerging from nodes in W within $\mathcal{G}$. In $\mathcal{G}_{\underline{W}}$, Z_t and W are d-connected when conditioned on Z_c because of the existence of the edge $Z_t \rightarrow W$. However, Z_t and Y are d-separated by Z_c since Z_t, Z_c and U_c form a collider at W, and Z_c blocks the path $Z_t \leftarrow Z_c \rightarrow Y$. Furthermore, the effect of Z_t on Y is mediated solely by W through the causal path $Z_t \rightarrow W \rightarrow Y$. Therefore, Z_t is a CIV, and Z_c is its corresponding conditional set. Note that Z_c may contain information about latent factors U_c due to the complex relationships in interactive data, as indicated by the dashed line between U_c and Z_c [Wu *et al.*, 2022; Cheng *et al.*, 2024b]. Many existing works have shown that confounding factors can affect recommendation outcomes, specifically $Z_c \rightarrow Y$ [Si *et al.*, 2022; Si *et al.*, 2023a; Si *et al.*, 2023b]. Furthermore, since Z_t and Z_c are derived from the user interaction data X, they will affect W, i.e., $Z_t \rightarrow W$ and $Z_c \rightarrow W$ in $\mathcal{G}$.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

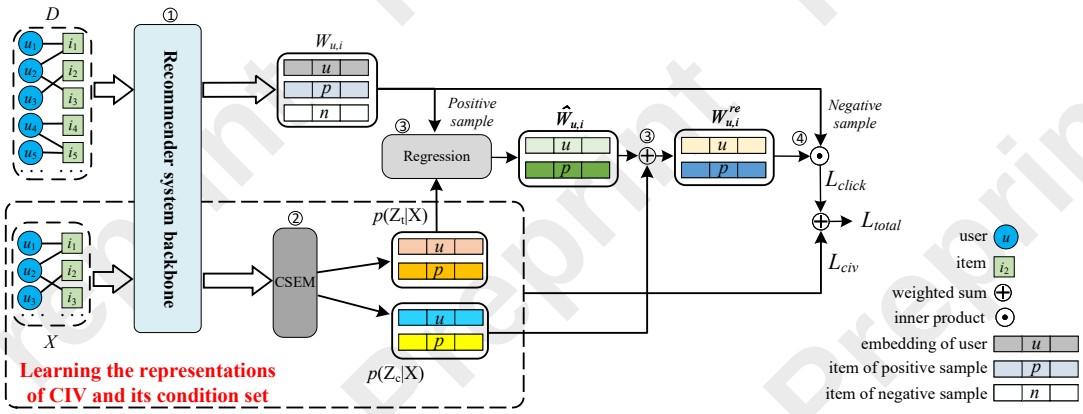

Figure 3: The overall structure of IDCIV-RS consists of four main steps, labelled 1, 2, 3, and 4 in the diagram. First, during feature encoding, IDCIV-RS uses a backbone (MF or LightGCN) to encode the input data into a latent space representation. Second, the CSEM module is used to learn the CIV Z_t and its condition set Z_c from the user-item interaction data X. Third, the representation of CIV is used to decompose the treatment variable $W_{u,i}$, obtain the causal relationship representation $\widehat{W}_{u,i}$, and fuse it with the conditional set representation Z_c to get the reconstructed treatment variable $W^{re}_{u,i}$. Finally, $W^{re}_{u,i}$ is used for the click prediction.

Based on the proposed causal DAG, we present IDCIV-RS, an interaction-data-guided CIV debiasing method for recommender systems. IDCIV-RS offers two key advantages over existing approaches: it eliminates the need for domain-specific IV specification and leverages CIVs, which provide a more general framework for mitigating confounding bias from latent factors.

3.3 The Concepts of Treatment Variable, CIV and Its Conditional Set

We define the concepts for the treatment variable, the CIV, and its conditional set based on user behaviour data. The treatment variable $W_{u,i}$ using user-item pairs is defined as:

$$W_{u,i} = \{(w_u, w_i) \,|u \in U, i \in I\}, \tag{1}$$

where w_u and w_i represent the embeddings of user u and item i, respectively. The concepts for the representations of CIV Z_t and its conditional set Z_c from the user interaction data X are as follows:

$$Z_t = \{\left(z_{t_u}, z_{t_p}\right) |u, p \in X\}, \tag{2}$$

$$Z_c = \{\left(z_{c_u}, z_{c_p}\right) |u, p \in X\}, \tag{3}$$

where z_{t_u} and z_{c_u} indicate the representations of the CIV of the user u and its conditional set, respectively, z_{t_p} and z_{c_p} denote the representations of CIV of the corresponding item p and its conditional set, respectively. To estimate the causal effects of users and items on Y (user feedback), it is necessary to construct the CIV and its conditional set for the user u and item p, respectively.

3.4 Learning the Representations of CIV and Its Condition Set

In our IDCIV-RS framework, we employ a VAE structure as the generative model to generate the representations of the CIV Z_t and its conditional set Z_c [Kingma and Welling, 2013; Sohn *et al.*, 2015] from the user interaction data X,

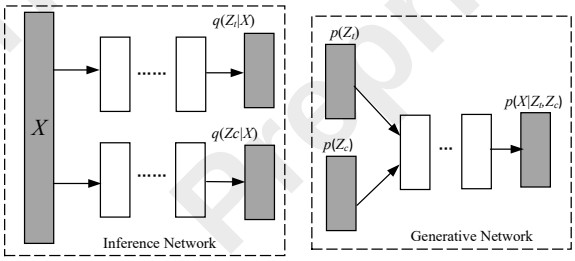

Figure 4: The CSEM components used to learn the latent representation of CIV Z_t and its conditional set Z_c consist of an inference network and a generation network. The grey boxes represent samples drawn from the corresponding distribution, and the white boxes represent the neural network.

referred to as the CIV and conditional Set Extraction Module (CSEM), as shown in Figure 4. We use the inference and generation networks of VAE to approximate the posterior distributions $p\left(Z_t|X\right)$ and $p\left(Z_c|X\right)$ for the two representations Z_t and Z_c.

In the inference network, we use two independent encoders to learn the posterior distributions $q\left(Z_t|X\right)$ and $q\left(Z_c|X\right)$. The variational approximations of their posterior distributions are as follows:

$$q\left(Z_t|X\right) = \prod_{m=1}^{D_{Z_t}} \mathcal{N}\left(\mu = \widehat{\mu}_{Z_{t_m}}, \sigma^2 = \widehat{\sigma}^2_{Z_{t_m}}\right), \tag{4}$$

$$q\left(Z_c|X\right) = \prod_{m=1}^{D_{Z_c}} \mathcal{N}\left(\mu = \widehat{\mu}_{Z_{c_m}}, \sigma^2 = \widehat{\sigma}^2_{Z_{c_m}}\right), \tag{5}$$

where μ and σ are the mean and variance of the Gaussian distribution captured by neural networks. It is worth noting that Z_t and Z_c are composed of pairs of user and item samples, so each $q\left(Z_t|X\right)$ and $q\left(Z_c|X\right)$ has two components: item and user. In the generative network, the prior distribution $p\left(Z_t\right)$

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

follows a Gaussian distribution:

$$p\left(Z_t\right)=\prod_{m=1}^{D_{Z_t}}\mathcal{N}\left(Z_{t_m}|0,1\right). \tag{6}$$

The prior distribution $p\left(Z_c\right)$ is obtained based on the Conditional VAE (CVAE) [Sohn *et al.*, 2015] model. We use Monte Carlo (MC) sampling to condition $p(Z_c)$ on X:

$$Z_c \sim p(Z_c|X). \tag{7}$$

Then, the decoder for X is described as follows:

$$p(X|Z_t,Z_c)=\prod_{m=1}^{D_X} p\left(X_m|Z_t,Z_c\right). \tag{8}$$

For inference, we optimise the parameters by maximising the evidence lower bound (ELBO):

$$\begin{aligned}L_{civ}=&\mathbb{E}_q[\log p\left(X|Z_t,Z_c\right)]-D_{KL}[q\left(Z_t|X\right)||p\left(Z_t\right)]\\&-D_{KL}[q\left(Z_c|X\right)||p\left(Z_c|X\right)].\end{aligned} \tag{9}$$

Note that using CVAE to condition $p(Z_c)$ on X is a critical step for learning the representations of Z_c and Z_t because Z_c and Z_t are independent given X, which ensures that Z_t captures the CIV information, while Z_c capture the confounding information given X in the interactive data.

3.5 Decomposition of Treatment Variable

After obtaining the representations of CIV (Z_t) and its conditional set (Z_c) from X, we use the CIV (Z_t) to reconstruct the treatment variable (W) and decompose W to derive the causal relationship, specifically, user preference. We apply the least squares (LS) method based on IV4Rec [Si *et al.*, 2022] to decompose W and obtain the representation $\widehat{W}_{u,i}$ that is not affected by U_c:

$$\widehat{W}_{u,i}=f_{pro}\left(Z_t,W_{u,i}\right), \tag{10}$$

where f_{pro} is the projection function that maps Z_t and $W_{u,i}$ into the same space, allowing the unbiased $\widehat{W}_{u,i}$. The function $f_{pro}(\cdot)$ is defined as:

$$f_{pro}\left(Z_t,W_{u,i}\right)=Z_t\cdot\mathcal{W}_{u,i}, \tag{11}$$

where $W_{u,i}$ is the closed-form solution of the LS method, and its calculation formula is as follows:

$$\mathcal{W}_{u,i}=\arg\min_{W_{u,i}}\|Z_t\cdot W_{u,i}-W_{u,i}\|_2^2=Z_t^{\dagger}\cdot W_{u,i}, \tag{12}$$

where $Z_t^{\dagger}$ is the Moore-Penrose pseudo-inverse of Z_t. Thus, the decomposed $W_{u,i}$ captures the causal relationship from Z_t while separating the latent confounding bias introduced by U_c, which reflects user preferences. Additionally, we need to incorporate Z_c to obtain the reconstructed $W_{u,i}^{re}$, as Z_c blocks the confounding bias between Z_t and Y. Therefore, our final reconstructed $W_{u,i}^{re}$ is obtained by:

$$W_{u,i}^{re}=\alpha\widehat{W}_{u,i}+(1-\alpha)Z_c, \tag{13}$$

where α is a hyperparameter used to balance $\widehat{W}_{u,i}$ and Z_c.

3.6 Click Prediction

To improve click prediction, we optimize the reconstructed treatment variables with BPR loss.

$$\begin{aligned}L_{click}=-\sum_{(u,p,n)\in D}&\ln sigma\big(\left\langle w_u^{re},w_p^{re}\right\rangle\\&-\left\langle w_u^{re},w_n\right\rangle\big),\end{aligned} \tag{14}$$

where $\langle\cdot\rangle$ denotes the inner product, and w_n is the negative sample item that user u has not interacted with, selected by the Popularity-based Negative Sampling with Margin (PNSM) strategy [Zheng *et al.*, 2021]. By combining the ELBO and BPR, the final loss function of our IDCIV-RS is:

$$L_{total}=L_{civ}+L_{click}. \tag{15}$$

Therefore, our IDCIV-RS obtain $W_{u,i}^{re}$ for click prediction by learning the representations of CIV Z_t and its condition set Z_c from X and by decomposing W.

4 Experiments

In this section, we conduct experiments on two real-world datasets to validate the recommendation performance and debiasing effectiveness of IDCIV-RS.

4.1 Experimental Settings

Datasets. We use two publicly real-world datasets: the Movielens-10M and the Douban-Movie datasets. Both datasets include user IDs, movie IDs, and ratings (1-5) for movies and are widely used in recommender system debiasing research. We use a dataset setup commonly used in the field of debiased recommender systems; specifically, the training set is biased and the test set is unbiased. We adopt the data pre-processing approach used in previous studies [Zheng *et al.*, 2021].

Dataset	# User	# Item	# Interaction
Movielens-10M	37,962	4,819	1,371,473
Douban-Movie	6,809	1,5012	173,766

Table 1: Statistics of datasets.

Baselines. Causal debiasing methods are typically applied as enhancements to backbone recommendation models. In our experiments, we use MF and LightGCN as the backbone models. We compare our approach against five causality-based debiasing methods:

- **IPS** [Schnabel *et al.*, 2016]: This method assigns weights that are the inverse of an item's popularity, thereby enhancing the impact of less popular items while reducing the influence of more popular ones.
- **IPS-C** [Bottou *et al.*, 2013]: This approach caps the maximum value of IPS weights to reduce variance across the entire weight distribution.
- **CausE** [Bonner and Vasile, 2018]: This method generates two sets of embeddings from the data, which are then aligned using regularization techniques to ensure their similarity.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Dataset		Movielens-10M							
		TopK=20				TopK=50			
Backbone	Method	Recall↑	HR↑	NDCG↑	Imp.↑	Recall↑	HR↑	NDCG↑	Imp.↑
MF	Original	0.1276	0.4397	0.0832	–	0.2332	0.6308	0.1156	–
	IPS	0.1228	0.4210	0.0779	-3.76%	0.2168	0.6016	0.1070	-7.03%
	IPS-C	0.1277	0.4335	0.0809	+0.08%	0.2224	0.6150	0.1102	-4.63%
	CausE	0.1164	0.4144	0.0770	-8.77%	0.2076	0.5940	0.1047	-10.98%
	DICE	0.1626	0.5202	0.1076	+27.42%	0.2854	0.6941	0.1459	+22.38%
	DCCL	0.1503	0.4874	0.0975	+17.79%	0.2636	0.6676	0.1326	+13.04%
	IDCIV-RS-Causal	0.1660	0.5282	0.1108	+30.09%	0.2895	0.7012	0.1495	+24.14%
	IDCIV-RS	**0.1709**	**0.5362**	**0.1148**	**+33.93%**	**0.2973**	**0.7073**	**0.1542**	**+27.49%**
LightGCN	Original	0.1462	0.4831	0.0952	–	0.2631	0.6688	0.1316	–
	IPS	0.1298	0.4438	0.0849	-11.22%	0.2325	0.6196	0.1170	-11.63%
	IPS-C	0.1327	0.4533	0.0871	-9.23%	0.2383	0.6302	0.1201	-9.43%
	CausE	0.1164	0.4099	0.0727	-20.38%	0.2204	0.6080	0.1046	-16.23%
	DICE	0.1810	0.5564	0.1228	+23.80%	0.3109	0.7219	0.1632	+18.17%
	DCCL	0.1462	0.4824	0.0947	0%	0.2644	0.6711	0.1311	+0.49%
	IDCIV-RS-Causal	0.1784	0.5511	0.1205	+22.02%	0.3056	0.7160	0.1602	+16.15%
	IDCIV-RS	**0.1817**	**0.5582**	**0.1241**	**+24.28%**	**0.3119**	**0.7232**	**0.1645**	**+18.55%**

Table 2: The performance of all methods on Movielens-10M. The "original" indicates that only the backbone is used, with no additional causal debiasing methods. The best results are highlighted in bold, and the second-best results are underlined.

Dataset		Douban-Movie							
		TopK=20				TopK=50			
Backbone	Method	Recall↑	HR↑	NDCG↑	Imp.↑	Recall↑	HR↑	NDCG↑	Imp.↑
MF	Original	0.0214	0.0542	0.0128	–	0.0371	0.0933	0.0171	–
	IPS	0.0172	0.0444	0.0099	-19.63%	0.0282	0.0755	0.0130	-23.99%
	IPS-C	0.0166	0.0446	0.0095	-22.43%	0.0271	0.0761	0.0125	-26.95%
	CausE	0.0149	0.0410	0.0074	-30.37%	0.0273	0.0761	0.0108	-26.42%
	DICE	0.0231	0.0615	0.0133	+7.94%	0.0396	0.1012	0.0178	+6.74%
	DCCL	0.0217	0.0595	0.0123	+1.40%	0.0385	0.1040	0.0170	+3.77%
	IDCIV-RS-Causal	0.0278	0.0736	0.0162	+29.91%	0.0462	0.1213	0.0213	+24.53%
	IDCIV-RS	**0.0299**	**0.0777**	**0.0171**	**+39.71%**	**0.0480**	**0.1213**	**0.0220**	**+29.38%**
LightGCN	Original	0.0375	0.0557	0.0118	–	0.0640	0.0908	0.0155	–
	IPS	0.0352	0.0928	0.0208	-6.13%	0.0619	0.1576	0.0281	-3.28%
	IPS-C	0.0368	0.0980	0.0219	-1.87%	0.0643	0.1657	0.0295	-4.69%
	CausE	0.0263	0.0693	0.0143	-29.87%	0.0463	0.1216	0.0199	-27.66%
	DICE	0.0401	0.1088	0.0232	+6.93%	0.0679	0.1755	0.0310	+6.09%
	DCCL	0.0401	0.1046	0.0225	+6.93%	0.0693	0.1732	0.0306	+8.28%
	IDCIV-RS-Causal	0.0407	0.1125	0.0239	+8.53%	0.0704	0.1840	0.0321	+10.00%
	IDCIV-RS	**0.0435**	**0.1178**	**0.0252**	**+16.00%**	**0.0724**	**0.1886**	**0.0332**	**+13.13%**

Table 3: The results of all methods on Douban-Movie. The best is highlighted in bold, and the second-best is underlined.

- **DICE** [Zheng *et al.*, 2021]: This method uses Structural Causal Modeling (SCM) [Pearl, 2009] to define user-item interactions. This approach leverages the collision effect of causal reasoning to enhance training effectiveness.
- **DCCL** [Zhao *et al.*, 2023]: This method uses contrastive learning to address data sparsity and the separation of these components.

We did not compare IDCIV-RS with IV-based methods like IV4Rec [Si *et al.*, 2022], as these require explicit IVs derived from domain knowledge or user search data, which our datasets lack. In contrast, IDCIV-RS learns CIV representations directly from user interactions, avoiding dependence on unavailable or domain-specific data. This confers greater flexibility and applicability in real-world scenarios.

Metrics. We evaluate Top-K recommendation under implicit feedback using Recall, Hit Rate (HR), and NDCG. Results reflect each method's best performance under optimal settings. "Imp." denotes the percentage improvement in Recall over the base model.

4.2 Comparison of Experimental Results

Tables 2 and 3 present the results of IDCIV-RS and all baseline approaches on two real-world datasets. *IDCIV-RS-Causal* is a variant of IDCIV-RS that denotes click prediction using only the unbiased embeddings $\widehat{W}_{u,i}$, which capture the

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

causal relationship.

The analysis of Tables 2 and 3 shows that IDCIV-RS and IDCIV-RS-Causal significantly improve performance metrics compared to the original backbone, with the highest improvement reaching 39.71%, demonstrating statistical significance and the superiority of our approach. Notably, IDCIV-RS consistently outperforms IDCIV-RS-Causal, aligning with the understanding that incorporating appropriate confounders enhances recommendation performance. This confirms that Z_c in IDCIV-RS effectively captures relevant confounders in user interaction data.

Tables 2 and 3 provide several key insights: (1) IPS-based debiasing methods perform poorly due to their reliance on the inverse propensity score, which is sensitive to data distribution. In our experiments, training on a biased dataset and testing on an unbiased one led to distribution mismatches. (2) CausE also underperforms, as it requires an unbiased training dataset to align user-item embeddings. (3) Although DICE and DCCL, which are based on causal graph assumptions, improve performance, they still fall short of optimal results. This is because they target specific biases based on predefined causal graphs, while real-world datasets often contain diverse biases from latent confounders, limiting their effectiveness.

4.3 Evaluation on Debiasing Experiments Ability

We use the *Intersection Over Union* (*IOU*) [Zheng *et al.*, 2021] metric to evaluate the debiasing ability of all methods. A higher IOU reflects more popular items in the recommendations, indicating weaker debiasing performance.

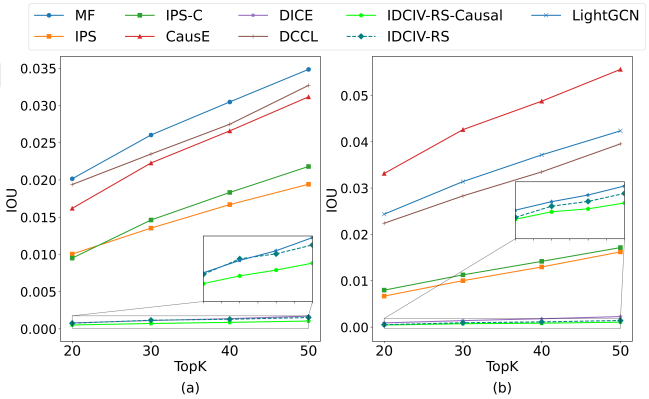

Figure 5: The IOU of recommended items and popular items for all methods on the Douban-movie dataset. (a) IOU of all methods on the MF; (b) IOU of all methods on the LightGCN.

Figure 5 shows the IOU for all methods on the Douban-movie dataset. IDCIV-RS and IDCIV-RS-Causal exhibit the lowest IOU, indicating superior debiasing ability. Notably, the IOU for all baseline methods increases significantly as the number of recommended items grows, suggesting that their debiasing ability diminishes with more recommendations. In contrast, the debiasing ability of our IDCIV-RS and IDCIV-RS-Causal remains relatively stable, demonstrating greater robustness.

Figure 5 also illustrates that the IOU of IDCIV-RS is higher than that of IDCIV-RS-Causal, due to IDCIV-RS incorporating confounding factor information. This suggests that Z_c effectively captures confounding factors in user interaction data, validating our method.

4.4 Ablation Studies

We perform ablation studies to evaluate the effectiveness of each component in IDCIV-RS. To verify the effectiveness of CIV and its condition set, we propose *IDCIV-RS-Con*, which uses only Z_c for click prediction. Figure 6 presents the IOU and Recall of IDCIV-RS and its variants. The results show that IDCIV-RS-Con has the highest IOU and lowest Recall, highlighting the effectiveness of Z_c in capturing confounding factor information. In contrast, IDCIV-RS-Causal exhibits higher Recall but lower IOU than IDCIV-RS-Con, indicating its effectiveness in capturing user preference information and mitigating confounding factors through Z_t. IDCIV-RS, by integrating both user preference and confounding factor information, achieves higher Recall and IOU than IDCIV-RS-Causal, demonstrating the combined effectiveness of Z_t and Z_c.

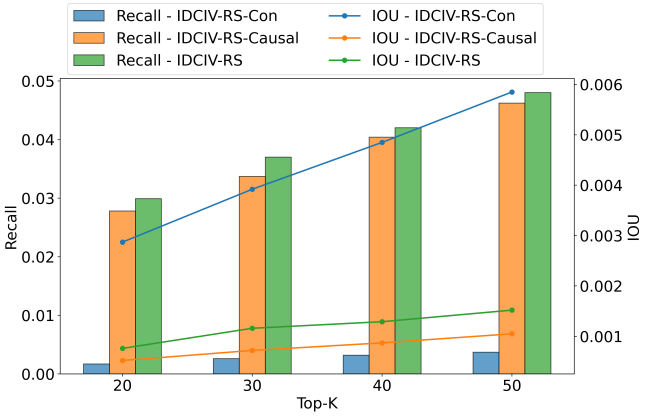

Figure 6: Recall and IOU of MF-based IDCIV-RS and its variants on the Douban-movie dataset. Where the bar represents Recall and the curve represents IOU.

5 Conclusion

In this paper, we propose a data-driven CIV debiasing method called IDCIV-RS. We learn the representations of CIV and its conditional set from user interaction data. The CIV is used to decompose the treatment variable and uncover the causal relationships between variables, while the conditional set captures confounding factors in the user interaction data. Unlike existing IV-based debiasing methods, IDCIV-RS imposes fewer constraints and does not require the selection of specific IVs based on domain knowledge. By integrating confounding factors and the causal relationships of the treatment variable, IDCIV-RS achieves high-quality recommendations and effective debiasing. We conducted extensive experiments on two real-world datasets to validate the effectiveness and superiority of IDCIV-RS in both recommendation and debiasing performance.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Acknowledgments

This work is supported partly by the Project of Guangxi Science and Technology (2025GXNSFFA069015), the National Natural Science Foundation of China (No. 62372119, 62166003), the Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education (No. EBME24-03), the Research Fund of the Guangxi Key Lab of Multi-source Information Mining & Security (MIMS24-M-01), and the Australian Research Council (under grant DP230101122).

References

[Bonner and Vasile, 2018] Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, page 104–112, New York, NY, USA, 2018. Association for Computing Machinery.

[Bottou *et al.*, 2013] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(101):3207–3260, 2013.

[Brito and Pearl, 2012] Carlos Brito and Judea Pearl. Generalized instrumental variables. *arXiv preprint arXiv:1301.0560*, 2012.

[Cai *et al.*, 2024] Miaomiao Cai, Min Hou, Lei Chen, Le Wu, Haoyue Bai, Yong Li, and Meng Wang. Mitigating recommendation biases via group-alignment and global-uniformity in representation learning. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[Caner and Hansen, 2004] Mehmet Caner and Bruce E Hansen. Instrumental variable estimation of a threshold model. *Econometric theory*, 20(5):813–843, 2004.

[Chen *et al.*, 2023] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

[Cheng *et al.*, 2022] Debo Cheng, Jiuyong Li, Lin Liu, Jiji Zhang, Jixue Liu, et al. Ancestral instrument method for causal inference without complete knowledge. *arXiv preprint arXiv:2201.03810*, 2022.

[Cheng *et al.*, 2023a] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Thuc Duy Le, and Jixue Liu. Learning conditional instrumental variable representation for causal effect estimation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 525–540. Springer, 2023.

[Cheng *et al.*, 2023b] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. Causal inference with conditional instruments using deep generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7122–7130, 2023.

[Cheng *et al.*, 2024a] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, Wentao Gao, and Thuc Duy Le. Instrumental variable estimation for causal inference in longitudinal data with time-dependent latent confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11480–11488, 2024.

[Cheng *et al.*, 2024b] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. Conditional instrumental variable regression with representation learning for causal inference. In *The Twelfth International Conference on Learning Representations*, pages 1–17, 2024.

[Gao *et al.*, 2021] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. Advances and challenges in conversational recommender systems: A survey. *AI open*, 2:100–126, 2021.

[Gao *et al.*, 2024] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems*, 42(4):1–32, 2024.

[Gomez-Uribe and Hunt, 2015] Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):1–19, 2015.

[He *et al.*, 2017] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

[He *et al.*, 2020] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.

[He *et al.*, 2023] Xiangnan He, Yang Zhang, Fuli Feng, Chonggang Song, Lingling Yi, Guohui Ling, and Yongdong Zhang. Addressing confounding feature issue for causal recommendation. *ACM Transactions on Information Systems*, 41(3):1–23, 2023.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Koren *et al.*, 2009] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[Koren *et al.*, 2021] Yehuda Koren, Steffen Rendle, and Robert Bell. Advances in collaborative filtering. *Recommender systems handbook*, pages 91–142, 2021.

[Lan *et al.*, 2024] Wei Lan, Guoxian Zhou, Qingfeng Chen, Wenguang Wang, Shirui Pan, Yi Pan, and Shichao Zhang. Contrastive clustering learning for multi-behavior recommendation. *ACM Transactions on Information Systems*, 43(1):1–23, 2024.

[Li *et al.*, 2024] Chengyu Li, Debo Cheng, Guixian Zhang, and Shichao Zhang. Contrastive learning for fair graph

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

representations via counterfactual graph augmentation. *Knowledge-Based Systems*, 305:112635, 2024.

[Liao *et al.*, 2022] Jie Liao, Wei Zhou, Fengji Luo, Junhao Wen, Min Gao, Xiuhua Li, and Jun Zeng. Sociallgn: Light graph convolution network for social recommendation. *Information Sciences*, 589:595–607, 2022.

[Luo *et al.*, 2024] Renqiang Luo, Huafei Huang, Shuo Yu, Zhuoyang Han, Estrid He, Xiuzhen Zhang, and Feng Xia. Fugnn: Harmonizing fairness and utility in graph neural networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2072–2081, 2024.

[Luo *et al.*, 2025] Renqiang Luo, Huafei Huang, Ivan Lee, Chengpei Xu, Jianzhong Qi, and Feng Xia. Fairgp: A scalable and fair graph transformer using graph partitioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12319–12327, 2025.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Rendle *et al.*, 2012] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[Schnabel *et al.*, 2016] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1670–1679, New York, New York, USA, 20–22 Jun 2016. PMLR.

[Schölkopf *et al.*, 2021] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[Schölkopf, 2022] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804. 2022.

[Shoja and Tabrizi, 2019] Babak Maleki Shoja and Nasseh Tabrizi. Customer reviews analysis with deep neural networks for e-commerce recommender systems. *IEEE access*, 7:119121–119130, 2019.

[Si *et al.*, 2022] Zihua Si, Xueran Han, Xiao Zhang, Jun Xu, Yue Yin, Yang Song, and Ji-Rong Wen. A model-agnostic causal learning framework for recommendation using search data. In *Proceedings of the ACM Web Conference 2022*, pages 224–233, 2022.

[Si *et al.*, 2023a] Zihua Si, Zhongxiang Sun, Xiao Zhang, Jun Xu, Yang Song, Xiaoxue Zang, and Ji-Rong Wen. Enhancing recommendation with search data in a causal learning manner. *ACM Transactions on Information Systems*, 41(4):1–31, 2023.

[Si *et al.*, 2023b] Zihua Si, Zhongxiang Sun, Xiao Zhang, Jun Xu, Xiaoxue Zang, Yang Song, Kun Gai, and Ji-Rong Wen. When search meets recommendation: Learning disentangled search representation for recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1313–1323, 2023.

[Sohn *et al.*, 2015] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[Wang *et al.*, 2019] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174, 2019.

[Wang *et al.*, 2020] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 426–431, 2020.

[Wang *et al.*, 2021] Nan Wang, Zhen Qin, Xuanhui Wang, and Hongning Wang. Non-clicks mean irrelevant? propensity ratio scoring as a correction. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 481–489, 2021.

[Wu *et al.*, 2022] Anpeng Wu, Kun Kuang, Bo Li, and Fei Wu. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, pages 24056–24075. PMLR, 2022.

[Zhang *et al.*, 2021] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 11–20, 2021.

[Zhang *et al.*, 2022] Shichao Zhang, Jiaye Li, Wenzhen Zhang, and Yongsong Qin. Hyper-class representation of data. *Neurocomputing*, 503:200–218, 2022.

[Zhang *et al.*, 2024] Guixian Zhang, Debo Cheng, Guan Yuan, and Shichao Zhang. Learning fair representations via rebalancing graph structure. *Information Processing & Management*, 61(1):103570, 2024.

[Zhang, 2021] Shichao Zhang. Challenges in knn classification. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4663–4675, 2021.

[Zhao *et al.*, 2023] Weiqi Zhao, Dian Tang, Xin Chen, Dawei Lv, Daoli Ou, Biao Li, Peng Jiang, and Kun Gai. Disentangled causal embedding with contrastive learning for recommender system. In *Companion Proceedings of the ACM Web Conference 2023*, page 406–410, New York, NY, USA, 2023. Association for Computing Machinery.

[Zheng *et al.*, 2021] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*, pages 2980–2991, 2021.