Causality-Based Conformal Imputation Correction with Non-Random Missing Labels

Chunyuan Zheng
Peking University
Beijing, China
cyzheng@stu.pku.edu.cn

Xiang Li
Peking University
Beijing, China
lilixiang222@gmail.com

Hang Pan
University of Science and Technology of China
Hefei, China
hungpaan@mail.ustc.edu.cn

Eric Wang
Peking University
Beijing, China
ho-ward@outlook.com

Haoxuan Li
Peking University
Beijing, China
hxli@stu.pku.edu.cn

Yang Zhang
National University of Singapore
Singapore, Singapore
zyang1580@gmail.com

See-Kiong Ng
National University of Singapore
Singapore, Singapore
seekiong@nus.edu.sg

Xiao-Hua Zhou*
Peking University
Beijing, China
azhou@math.pku.edu.cn

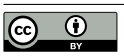
Abstract

Collected data with non-random missing labels poses a widely recognized challenge for unbiased learning. For example, in recommender systems, users are free to choose whether or not to rate an item. To achieve unbiased learning under MNAR data, a variety of methods have been proposed, such as reweighting and imputation. Among them, doubly robust (DR) based methods are widely adopted due to their appealing theoretical guarantees. However, these guarantees rely on strong assumptions that either the propensity or the imputation is accurate for all units (such as user-item pairs), which is very hard to achieve in real-world scenarios. Previous studies show that a small error in imputation can lead to a large bias in DR-based methods. Furthermore, for units with missing labels, we lack an effective method to evaluate the imputation quality. In this work, we propose a model-agnostic framework to assess the accuracy of imputed labels and to correct imputations with large bias based on conformal prediction. Specifically, we leverage conformal prediction to construct a valid prediction set for units with unobserved labels, and revise imputations that fall outside this set. Extensive experiments are conducted on three real-world datasets and one semi-synthetic dataset to show the effectiveness of our proposed method. Our code is available at https://github.com/lixiang-222/conformal-prediction-for-MNAR.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Information systems** → **Data mining**.

*Xiao-Hua Zhou is the corresponding author.



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818158

Keywords

Causality, Non-Random Missing Labels, Conformal Prediction

ACM Reference Format:
Chunyuan Zheng, Xiang Li, Hang Pan, Eric Wang, Haoxuan Li, Yang Zhang, See-Kiong Ng, and Xiao-Hua Zhou. 2026. Causality-Based Conformal Imputation Correction with Non-Random Missing Labels. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818158

1 Introduction

Data missing not at random (MNAR) means that the distribution of collected data differs from that of the target population, which is also known as selection bias [10]. Ignoring selection bias makes machine learning methods difficult to achieve unbiased prediction and reliable evaluation. Taking recommender systems (RS) as an example, which aim to deliver high-quality items to users to enhance personalized experiences [27, 46], and have been widely applied in various domains such as e-commerce [21, 22, 56, 58] and social media [25, 41, 52, 54]. In RS, models are usually trained on collected historical feedback, while users are free to choose which items to interact with or rate and are more likely to provide feedback on items they prefer [32, 33]. Therefore, the observed ratings are not representative of all user-item pairs. Directly training prediction models on such MNAR data can lead to biased risk estimation and sub-optimal prediction performance.

To solve this issue, existing methods can be broadly divided into three types. The error-imputation-based (EIB) methods [1, 36] estimate pseudo-labels for missing label instances and train prediction models on both observed and imputed labels. The inverse propensity score (IPS) [33, 45, 53] based methods adopt a reweighting strategy on observed samples to adjust the observed distribution to the target data (i.e., all user-item pairs) distribution. To combine the strengths of both methods, doubly robust (DR) [31, 47, 53] based methods are proposed, which uses imputation and reweighting

simultaneously. In addition, many enhanced DR methods [4, 8, 44] are further proposed to control the bias, variance, calibration, and balancing of the vanilla DR. Due to the theoretical advantages over IPS, such as double robustness and variance reduction, DR-based methods are widely used in practice.

However, accurate imputation remains challenging under non-random missing labels. Previous methods show that even minor imputation errors can induce significant bias in DR-based estimators [34, 44]. To improve imputation accuracy, we first need to assess whether the imputation is approximately equal to the true rating, but it is difficult for samples with missing ratings. Previous methods, such as CDR [34], attempt to quantify imputation accuracy, but they rely on many assumptions (e.g., the prediction error obeys normal distribution). Moreover, upon finding the samples with poor imputation accuracy, these methods can only discard them rather than revise the imputation results.

In this paper, we take RS with non-random missing labels as an example, aiming to answer the following two questions: (1) how to quantify the accuracy of imputations, especially for user-item pairs with missing ratings, and (2) how to correct the high-bias imputations to train a better prediction model. Specifically, we propose to adopt conformal prediction (CP) [38, 39] to construct a prediction interval for the true value with a pre-defined confidence. However, a direct application of standard CP is invalid, as the observed data and unobserved data follow different distributions. To address this distribution shift, we propose a model-agnostic weighted conformal prediction (WCP) method inspired by [39], which can be adapted to any DR-based method as a plug-in. Specifically, we first split the observed dataset into two parts, the training set and the calibration set. Second, we pretrain a prediction model on the training set, and we compute the prediction error on the calibration set. Based on the prediction error, we construct a weighted prediction interval for user-item pairs with missing ratings. Then, we train a DR-based method, examining whether the imputation falls into the prediction interval to decide weather imputation has a large bias, and revise the imputation by clipping it into the interval or directly dropping during the training process. The main contributions are as follows:

- We propose a model-agnostic weighted conformal prediction framework to assess the accuracy of imputations and correct imputations with large bias without assumptions.
- We propose a DR-based learning algorithm to adapt the weighted conformal prediction with theoretical guarantee.
- We conduct extensive experiments on one semi-synthetic dataset and three real-world datasets to show the effectiveness of the proposed method.

2 Related Work

2.1 Non-Random Missing Labels

Non-random missing labels is a pervasive challenge in the collected data [16, 17, 42, 48]. In RS, it arises from the fact that users selectively interact with items they are interested in [29, 35]. In other words, there is a distribution shift between the training data and the target population (all user-item pairs) [20, 51]. Directly training models on such biased data typically results in suboptimal performance [33, 50]. To mitigate this issue [43, 49, 57, 59], existing debiasing methods can be broadly categorized into three paradigms. Based on the causal inference techniques, the first category is EIB methods [1, 9, 35], which aim to impute the errors of missing data using an imputation model. The second category is IPS [10, 26, 32], which reweights the observed samples by the inverse of propensity scores to reweight the observational distribution to the target one. However, IPS suffers from high variance, especially when propensity scores are small. To combine the strengths of both methods, DR methods [31, 44] have been proposed. DR estimators augment IPS with an error imputation term, achieving unbiasedness if either the propensity model or the imputation model is correctly specified. Recently, DR has been one of the most popular methods and is widely used in practice due to the promising theoretical guarantees. In addition, many enhanced DR methods are proposed to further mitigate bias and imputation misspecification [14, 15, 19], control variance [8], enhance propensity model balancing [18, 20], and enhance model calibration [7, 12]. In addition, many DR-enhanced methods have been proposed when combining a small portion of unbiased data [2, 24, 45].

However, DR-based methods are highly sensitive to these imputations: even a small error in the imputation model can lead to a large bias in the ideal loss estimation. Therefore, correcting the high-bias imputations is important. Although a few studies have attempted to achieve this goal, such as CDR [34], they rely on strong assumptions, such as the normality of prediction error. Additionally, these methods can only discard biased imputation rather than revise the imputations. To this end, we propose a weighted conformal prediction framework to address the inaccurate imputations.

2.2 Conformal Prediction

Conformal prediction (CP) has emerged as a rigorous framework for uncertainty quantification [38, 40]. Unlike traditional parametric methods, CP constructs prediction sets that possess coverage guarantees without imposing restrictive assumptions on the underlying data distribution. To extend the applicability of CP to large-scale datasets, Lei et al. [13] propose split conformal prediction by partitioning the data into disjoint training and calibration sets. In addition, since standard CP relies on the exchangeability assumption, it fails when training and test data follow different distributions. To address this, Tibshirani et al. [39] proved that coverage validity can be restored by reweighting residuals based on the likelihood ratio of the covariate shift. Driven by these advancements, recent literature has a growing trend of applying CP to enhance the model reliability in complex environments [3, 6, 23, 50, 55], validating its potential as a robust tool for trustworthy machine learning.

However, existing works mainly utilize CP for better uncertainty quantification. They do not discuss how to leverage the conformal intervals during the learning process. Our work bridges this gap by proposing a learning algorithm to adapt weighted conformal sets to filter and correct high-bias imputations, leading to the enhancement of the prediction model.

3 Preliminary

3.1 Debiased Recommendation

We consider a standard rating prediction scenario in RS. Suppose the user set $\mathcal{U} = \{u_1, \ldots, u_m\}$ contains m users and the item set $\mathcal{I} = \{i_1, \ldots, i_n\}$ contains n items. The target population is defined

as all user-item pairs $\mathcal{D} = \mathcal{U} \times \mathcal{I}$. Let $\mathbf{R} \in \mathbb{R}^{m \times n}$ be the ground truth rating matrix, where $r_{u,i}$ represents the rating of user u on item i. For each user-item pair, let $x_{u,i}$ denote the feature vector, and $\hat{r}_{u,i} = f(x_{u,i}; \theta)$ be the prediction model parameterized by θ. Let $O \in \{0,1\}^{m \times n}$ be the binary observation indicator matrix, where $o_{u,i} = 1$ indicates the rating is observed, and $o_{u,i} = 0$ otherwise. If all ratings were observed, the model could be trained by minimizing the ideal loss:

$$\mathcal{L}_{\text{ideal}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} e_{u,i},$$

where $e_{u,i} = \mathcal{L}(\hat{r}_{u,i}, r_{u,i})$ is the prediction error. However, in real-world scenarios, the rating matrix is not fully observed, meaning the ideal loss is inaccessible for those user-item pairs with missing ratings. A naive approach is to use loss on the observed user-item pairs as an estimation of ideal loss, as below:

$$\mathcal{L}_{\text{naive}}(\theta) = \frac{1}{|O|} \sum_{(u,i) \in O} e_{u,i}.$$

However, this naive loss is a biased estimation due to the non-random missing labels, which means that $P_{\text{obs}}(x) \triangleq \mathbb{P}(x|o=1)$ do not equal to the distribution of the missing data $P_{\text{mis}}(x) \triangleq \mathbb{P}(x|o=0)$. To address such bias, the doubly robust (DR) estimator is widely adopted:

$$\mathcal{L}_{\text{DR}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} \left[\hat{e}_{u,i} + \frac{o_{u,i}(e_{u,i} - \hat{e}_{u,i})}{\hat{p}_{u,i}} \right],$$

where $\hat{p}_{u,i} = \pi(x_{u,i}; \psi)$ estimates the propensity score $p_{u,i} := \mathbb{P}(o_{u,i} = 1 | x_{u,i})$, and $\hat{e}_{u,i} = \mathcal{L}(\tilde{r}_{u,i}, \hat{r}_{u,i})$ with $\tilde{r}_{u,i} = h(x_{u,i}; \phi)$ as the imputation model used to estimate the missing rating $r_{u,i}$. The DR-based method has double robustness property, which means that $\mathbb{E}[\mathcal{L}_{\text{DR}}] = \mathcal{L}_{\text{ideal}}$ when either $\hat{p}_{u,i} = p_{u,i}$ or $\tilde{r}_{u,i} = r_{u,i}$ for all user-item pairs.

3.2 Conformal Prediction

The conformal prediction (CP) aims to build a prediction set for each sample to make sure the true rating can fall into this set with a confidence level. Notably, CP does not require any assumption, such as the prediction model structure and the distribution of prediction. In the simple scenario, i.e., we have data $\{(x_{u_k,i_k}, r_{u_k,i_k})\}_{k=1}^{n}$ with each $(x_{u,i}, r_{u,i}) \sim P$, where P is a distribution. Let $\alpha \in (0,1)$ be a nominal error rate. The objective of CP is to construct a prediction set $\hat{C}_n : \mathcal{X} \to \{\text{subsets of } \mathbf{R}\}$, such that for a new i.i.d. test pair $(x_{u',i'}, r_{u',i'}) \sim P$, the true rating is contained within the band with high probability:

$$\mathbb{P}\big(r_{u',i'} \in \hat{C}_n(x_{u',i'})\big) \geq 1 - \alpha.$$

Note that the rank of $r_{u',i'}$ is uniformly distributed in $\{r_{u_k,i_k}\}_{k=1}^{n}$, thus we can found $(1-\alpha)$ quantile in $\{r_{u_k,i_k}\}_{k=1}^{n}$ to construct the set. The validity of CP relies on the exchangeability assumption, i.e., the joint distribution of the data remains invariant under any permutation, to make sure the uniform distribution property.

A more practical issue is that, if we train a prediction model f on $\{(x_{u_k,i_k}, r_{u_k,i_k})\}_{k=1}^{n}$, we want to build a prediction error set for a new i.i.d. test pair $(x_{u',i'}, r_{u',i'})$ to make sure $e_{u',i'}$, such as absolute error $|f(x_{u',i'}) - r_{u',i'}|$, in this set. In this scenario, the rank of $e_{u',i'}$ in $\{e_{u_k,i_k}\}_{k=1}^{n}$ is not uniformly distributed because the prediction error

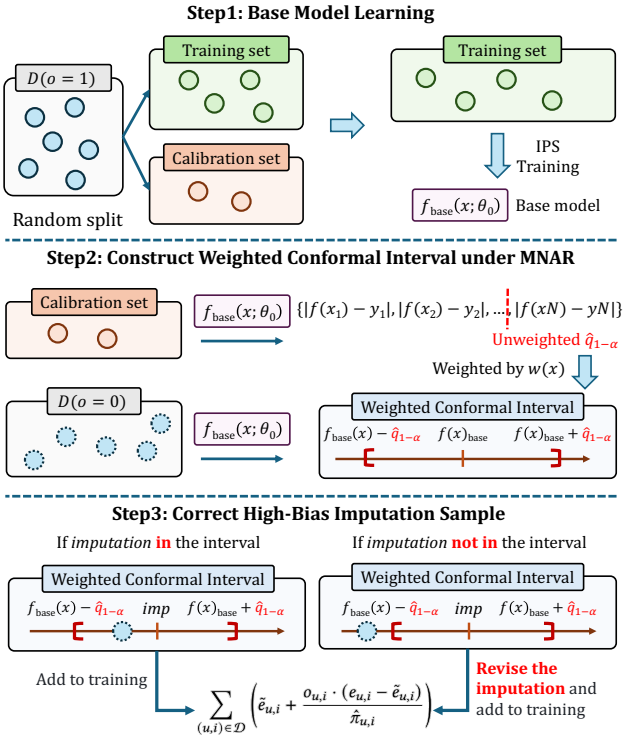


Figure 1: Overview of proposed weighted conformal doubly robust learning method.

Table 1: The quantile of RAIE on ML-100K. * means that statistical significance using paired-t-test with p-value ≤ 0.05.

	RAIE Quantile				
Method	0.1	0.3	0.5	0.7	0.9
DR-JL	1.51%	4.60%	8.23%	13.20%	22.48%
MRDR	2.41%	6.99%	11.32%	16.27%	24.45%
DR-BIAS	1.90%	5.70%	9.65%	14.60%	23.10%
WCP-DR-JL	**0.71%***	**1.93%***	**2.97%***	**4.22%***	**8.04%***
WCP-MRDR	**0.72%***	**1.93%***	**2.96%***	**4.23%***	**11.27%***
WCP-DR-BIAS	**0.72%***	**1.92%***	**2.96%***	**4.24%***	**9.52%***

for test pair (u', i') is more likely larger than the training pair (u, i). To solve this issue, we can first split the training set into D_1 and D_2 to part with $D_1 \cap D_2 = \emptyset$, then we train model f on D_1 and calculate the prediction error $e_{u,i}$ on D_2. Now we can build the prediction set by finding the $1-\alpha$ quantile in $\{e_{u,i}\}_{(u,i) \in D_2}$ due to the rank of $e_{u',i'}$ is uniformly distributed in the $\{e_{u,i}\}_{(u,i) \in D_2}$. However, due to the presence of selection bias, we need to extend the traditional conformal prediction technique, for example, weighted conformal prediction [39], and propose the DR-based learning algorithm to adapt the extended technique.

4 Method

4.1 Motivation

Among the various strategies to mitigate selection bias, the Doubly Robust (DR) estimator has emerged as a dominant paradigm. By enhancing IPS with an error imputation strategy, it provides appealing property such as double robustness and variance reduction. However, the sparsity of user-item interactions and sample selection bias make achieving accurate imputations difficult. Crucially, DR-based methods are highly sensitive to these imputations: even a small error in the imputation model can lead to a large bias in the ideal loss estimation. Without a mechanism to verify the quality of these imputation results, the learned prediction model is likely to be less reliable. Therefore, we aim to answer the following two questions: (1) how to quantify the accuracy of imputations, especially for user-item pairs with missing ratings, and (2) how to correct the high-bias imputations to train a better prediction model.

Case study: To further substantiate the challenge above, we conduct semi-synthetic experiments on the ML-100K dataset. Specifically, we investigate the relative absolute imputation error (RAIE) of exemplar DR-based methods such as DR-JL, MRDR, and DR-BIAS, where RAIE is defined as:

$$\text{RAIE} = \left| \frac{e_{u,i} - \hat{e}_{u,i}}{e_{u,i}} \right|, \quad \forall (u,i) \in \mathcal{D}, \tag{1}$$

where $e_{u,i} = (r_{u,i} - \hat{r}_{u,i})^2$ and $\hat{e}_{u,i} = (\tilde{r}_{u,i} - \hat{r}_{u,i})^2$. According to Table 1, current DR estimators suffer from large imputation bias. For example, the baseline DR-JL estimator reaches an RAIE of 22.48% at the 0.9 quantile. In contrast, our proposed weighted conformal prediction framework significantly reduces RAIE across all quantiles. For example, it reduces the RAIE of DR-JL from 1.51% to 0.71% at the 0.1 quantile, and from 22.48% to 8.04% at the 0.9 quantile. These results demonstrate that our method effectively identifies imputations with large bias and correctly revises them, providing reliable error imputation results essential for the unbiasedness of the DR estimator.

4.2 Method Overview

We provide an overview of our proposed framework for assessing the accuracy of imputed ratings and correcting high-bias imputations, as illustrated in Figure 1. Our method follows a three-stage pipeline: we first train a debiased base prediction model, then we construct weighted conformal prediction intervals for unobserved user-item pairs. Finally, we propose a conformal doubly robust learning algorithm that corrects high-bias imputations by clipping them into these valid intervals or dropping during training.

4.3 Weighted Conformal Prediction under Label Non-Random Missing

As discussed in the preliminary, standard conformal prediction relies on the exchangeability assumption. However, in our scenario, we have a covariate shift between the observed calibration dataset and the unobserved test data. Since $P(x, r|o = 1) \neq P(x, r|o = 0)$, the exchangeability assumption is violated. In addition, we consider $P(x|o = 1) \neq P(x|o = 0)$ and $P(r|x, o = 1) = P(r|x, o = 0)$, which is reasonable in RS, since the intrinsic user preference is not affected by exposure mechanism.

To address this distributional discrepancy, the core idea is to adjust the probability density of each observed data in the original CP. Specifically, we reweight each observed sample based on the likelihood ratio $w(x, r)$, where

$$\begin{aligned} w(x,r) &= \frac{P(x,r|o=0)}{P(x,r|o=1)} = \frac{P(r|x,o=0)P(x|o=0)}{P(r|x,o=1)P(x|o=1)} \\ &= \frac{P(o=0|x)}{P(o=1|x)} \times \frac{P(o=1)}{P(o=0)} \triangleq w(x). \end{aligned}$$

Here, $P(o = 0|x)/P(o = 1|x)$ represents the odds ratio of propensity, the term $P(o = 1)/P(o = 0)$ is a constant scaling factor determined by the global sparsity of the dataset. With the calculated weights, we construct a weighted distribution of the prediction errors. For an observed set $\mathcal{D}_{\text{obs}}$ and a new test point (u', i'), we define the normalized probability weight for each data point as:

$$p^w(x_{u,i}) = \frac{w(x_{u,i})}{\sum_{(k,l)\in\mathcal{D}_{\text{obs}}} w(x_{k,l}) + w(x_{u',i'})}, \ \forall (u,i) \in \mathcal{D}_{\text{obs}} \cup \{(u',i')\}.$$

The corresponding weighted empirical distribution of the prediction error is then given by:

$$\sum_{(u,i)\in\mathcal{D}_{\text{obs}}} p^w(x_{u,i})\delta_{e_{u,i}} + p^w(x_{u',i'})\delta_{\infty},$$

where δ is the Dirac delta function. By using this weighted distribution to determine the quantile, we define the $(1-\alpha)$-th weighted quantile of the prediction error for (u', i') as:

$$\begin{aligned} \hat{q}_{1-\alpha}(x_{u',i'}) = \inf \Bigg\{ q \in \mathbb{R} \cup \{\infty\} : \sum_{(u,i)\in\mathcal{D}_{\text{obs}}} p^w(x_{u,i}) \mathbb{I}(e_{u,i} \le q) \\ + p^w(x_{u',i'}) \mathbb{I}(\infty \le q) \ge 1-\alpha \Bigg\}, \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function. To simplify the notation, we denote $\hat{q}_{1-\alpha}(x_{u',i'})$ as $\hat{q}_{1-\alpha}$. Based on this weighted distribution, the conformal prediction interval is constructed as:

$$\hat{C}_\alpha(x_{u',i'}) = \{ r \in \mathbb{R} : e(u', i') \le \hat{q}_{1-\alpha} \}.$$

The following Theorem shows the validity of the prediction sets.

Theorem 4.1 (Corollary 1 in [39]). *Let $(x_{u,i}, r_{u,i}) \in \mathcal{D}_{obs}$ be n i.i.d. samples drawn from the observed distribution $P(x, r|o = 1)$, and let $(x_{u',i'}, r_{u',i'})$ be a test point drawn from $P(x, r|o = 0)$. Assume the weights are correctly specified as $w(x)$. For any miscoverage level $\alpha \in (0, 1)$, the weighted conformal prediction set $\hat{C}(x_{n+1})$ satisfies:*

$$\mathbb{P}\left(r_{u',i'} \in \hat{C}_\alpha(x_{u',i'}) \right) \ge 1 - \alpha.$$

This theorem shows that the validity of $1 - \alpha$ interval requires accurate weights, which are equivalent to the accurate propensities. One may argue that DR is already unbiased in such a scenario. However, $\mathbb{E}[\mathcal{L}_{\text{DR}}] = \mathcal{L}_{\text{ideal}}$ does not mean that $e_{u,i} = \hat{e}_{u,i}$. Instead, our method can find the accurate $1 - \alpha$ interval of ground-truth rating for each user-item pair, ensuring a more fine-grained refinement.

4.4 Conformal Doubly Robust Learning

After obtaining the prediction sets, we propose to adapt WCP to the DR-based learning algorithm. We use joint learning as an example. Specifically, the algorithm includes three steps as shown below.

Step 1: Base Model Learning. We partition the observed data $\{(x_{u,i}, y_{u,i})\}_{(u,i)\in\mathcal{O}}$ into training ($\mathcal{D}_{\text{tr}}$) and calibration ($\mathcal{D}_{\text{cal}}$) sets.

First, we train a simple base model $f_{\text{base}}(x;\theta_0)$ on $\mathcal{D}_{\text{tr}}$ by minimizing the IPS loss:

$$\min_{\theta_0} \sum_{(u,i)\in\mathcal{D}_{\text{tr}}} \frac{\ell(r_{u,i}, f_{\text{base}}(x_{u,i};\theta_0))}{\hat{p}_{u,i}}.$$

The reason to use IPS loss is that we want to train a debiasing base model and ensure the base model do not train on the unobserved data simultaneously. Once trained, f_{base} is frozen in following steps.

Step 2: Prediction Set Calculation. We leverage $\mathcal{D}_{\text{cal}}$ to construct the validity bounds using the base model. For each $(u,i) \in \mathcal{D}_{\text{cal}}$, we compute the prediction error $e_{u,i}$, such as the absolute residual of the base model:

$$e^{\text{base}}_{u,i} = |r_{u,i} - f_{\text{base}}(x_{u,i})|.$$

Using the importance weights $w(x)$ derived in Section 4.3, we compute the weighted $(1-\alpha)$-th quantile, denoted as $\hat{q}_{1-\alpha}$. This defines a theoretically valid interval $C_\alpha(x_{u',i'})$ for any instance $(x_{u',i'}, r_{u',i'}) \sim P(x,r|o=0)$ and imputation:

$$C_\alpha(x_{u',i'}) = \{r_{u',i'} | e^{\text{base}}(x_{u',i'}) \le \hat{q}_{1-\alpha}\}.$$

For example, if $e^{\text{base}}(x_{u,i})$ is the absolute residual, then

$$C_\alpha(x_{u',i'}) = [f_{\text{base}}(x_{u',i'}) - q_{1-\alpha}, f_{\text{base}}(x_{u',i'}) + q_{1-\alpha}].$$

Step 3: DR-based Prediction Model Training. In this step, we adapt the interval to the DR-based learning algorithm. Using the standard joint learning algorithm as an example, the imputation model is trained using the following imputation loss:

$$\mathcal{L}_e(\phi) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(\hat{e}_{u,i} - e_{u,i})^2}{\hat{p}_{u,i}}.$$

The propensity model is trained using the loss below:

$$\mathcal{L}_p(\psi) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} l(\hat{p}_{u,i}, o_{u,i}),$$

where $l(\cdot,\cdot)$ denote the cross-entropy loss. Notably, we correct the imputations with high bias during the prediction model training. Specifically, we propose to directly drop the imputations outside the interval, or to clip these imputations to the boundary of the interval or the mid-point of the interval, to obtain the modified DR loss, named WCP-DR. Using clip to the boundary as an example, if the imputation $\tilde{r}_{u,i} > f_{\text{base}}(x_{u,i}) + \hat{q}_{1-\alpha}$ the modified imputation $\tilde{r}'_{u,i} = f_{\text{base}}(x_{u,i}) + \hat{q}_{1-\alpha}$. The same thing is applied to those $\tilde{r}_{u,i} < f_{\text{base}}(x_{u,i}) - \hat{q}_{1-\alpha}$. Finally, we optimize the prediction model $f(x;\theta)$ by minimizing the following WCP-DR loss:

$$\mathcal{L}_{\text{WCP-DR}}(\theta) = \sum_{(u,i)\in\mathcal{D}} \left(\tilde{e}_{u,i} + \frac{o_{u,i}\cdot(e_{u,i} - \hat{e}_{u,i})}{\hat{\pi}_{u,i}} \right),$$

where $\tilde{e}_{u,i} = \mathcal{L}(\hat{r}_{u,i}, \tilde{r}'_{u,i})$, and $\tilde{r}'_{u,i} = \tilde{r}_{u,i}$ if $\tilde{r}_{u,i}$ already falls into the interval. Notably, we do not make correction for imputations with $o=1$ due to the user-item pairs with $o=1$ either in D_{tr} or D_{cal}. It is actually the trade-off: if we decrease the size of D_{tr} and D_{cal}, we can evaluate more imputations with $o=1$, but the constructed interval will be less efficient.

For clarity, we give a toy example of how the weighted conformal quantile is computed. Suppose we have four points in D_{cal} and one test point. The residuals for four points are $\{0.05, 0.10, 0.15, 0.20\}$, with normalized weights for five points $\{0.4, 0.3, 0.1, 0.1, 0.1\}$, respectively. For an 80% prediction set, i.e., $\alpha = 0.2$, we compute:

$$\hat{q}_{0.8} = \inf\{q \in \mathbb{R}\cup\{\infty\} : 0.4\mathbb{I}(0.05 \le q) + 0.3\mathbb{I}(0.10 \le q) + 0.1\mathbb{I}(0.15 \le q) + 0.1\mathbb{I}(0.20 \le q) + 0.1\mathbb{I}(\infty \le q) \ge 0.8\}.$$

For any finite q, the last term is zero. Therefore, the cumulative mass reaches 0.8 for the first time at $q = 0.15$, and hence $\hat{q}_{0.8} = 0.15$. The resulting prediction interval is:

$$[f_{\text{base}}(x) - 0.15,\ f_{\text{base}}(x) + 0.15].$$

For example, under boundary clipping, if an imputed rating satisfies $\tilde{r}_{u,i} > f_{\text{base}}(x_{u,i}) + 0.15$, we revise it as:

$$\tilde{r}'_{u,i} = f_{\text{base}}(x_{u,i}) + 0.15.$$

Algorithm 1: Weighted Conformal Prediction Doubly Robust Joint Learning (WCP-DR-JL)

Input: Observed set $\mathcal{D}_{\text{obs}}$, Unobserved set $\mathcal{D}_{\text{mis}}$, Significance level α, learning rate η.
Output: Trained prediction model $f(x;\theta)$

```
Train propensity model π̂(x;ψ) on D_obs;
Randomly partition D_obs into disjoint training set D_tr and calibration set D_cal;
Train base model f_base(x;θ_0) on D_tr via IPS loss and freeze its parameters;
for (u,i) ∈ D_cal do
    Compute base residual: s_{u,i} ← |y_{u,i} − f_base(x_{u,i})|;
    Compute sample importance weight w(x_{u,i});
Initialize imputation model g(x;φ);
while not converged do
    Sample mini-batch B_1 from D_obs ∪ D_mis;
    for (u,i) ∈ B_1 ∩ D_mis do
        Raw imputation: r̃_{u,i} ← g(x;φ);
        Compute interval bounds:
          L ← f_base(x_{u,i}) − q̂_{1−α}(x_{u,i});
          U ← f_base(x_{u,i}) + q̂_{1−α}(x_{u,i});
        Revise imputation if it exceeds [L,U] via boundary clipping;
    Update prediction model: θ ← θ − η∇_θ L_WCP-DR(θ);
    Sample mini-batch B_2 from D_obs;
    Update imputation model parameter φ:
      φ ← φ − η∇_φ L_e(φ), where L_e(φ);
return f(x;θ);
```

4.5 Theoretical Results

In real-world applications of WCP for debiasing, the importance weights $w(x)$ rely on estimated propensity scores $\hat{\pi}(x)$, which may contain error. Thus it is very necessary to investigate the validity of prediction sets under imperfect weight $w(x_{u,i})$. Specifically, we derive a bound that guarantees valid coverage if the propensity estimates are either uniformly bounded (low max-error) or accurate on average (low variance). Specifically, assume both $w(x)$ and

Table 2: Performance comparison with different significance level (α) on ML-100K. We report the δ_{RMSE} and δ_{RMAIE} as the relative reduction by our WCP-based learning algorithm.

	WCP-DR-JL		WCP-MRDR		WCP-DR-BIAS	
α	δ_{RMSE}	δ_{RMAIE}	δ_{RMSE}	δ_{RMAIE}	δ_{RMSE}	δ_{RMAIE}
0.1	44.95%	55.39%	49.84%	60.06%	48.30%	57.48%
0.2	48.08%	60.20%	50.80%	63.99%	50.68%	61.91%
0.3	49.48%	63.07%	52.40%	66.31%	51.70%	64.57%
0.4	50.52%	65.10%	53.04%	68.16%	52.38%	66.32%

$\hat{w}(x)$ are normalized, let $w(x)/\hat{w}(x)$ be the density ratio under the observed data distribution, then we have the following theorem.

Theorem 4.2. *Let $\hat{C}_\alpha$ be the prediction set constructed at a nominal significance level α using estimated weights $\hat{w}$. If the two following scenarios hold:*

- *Scenario 1: The uniform relative error is bounded by $\epsilon \in [0, 1)$:*

$$\sup_x \left| \frac{w(x)}{\hat{w}(x)} - 1 \right| \leq \epsilon.$$

- *Scenario 2: The second moment of the ratio is bounded by σ^2:*

$$\mathbb{E}_{P_{\hat{w}}} \left[\left(\frac{w(x)}{\hat{w}(x)} - 1 \right)^2 \right] \leq \sigma^2.$$

The coverage probability for the user-item pair (u', i') with missing ratings is bounded:

$$\mathbb{P}(r_{u',i'} \in \hat{C}_\alpha(x_{u',i'})) \geq 1 - \alpha - \min\left(\alpha\epsilon, \sigma\sqrt{\alpha}\right).$$

Scenario 2 is an average bound instead of a union bound, in which outliers violating the union bound with small density is allowed. Valid coverage lower bound can be derived if either Scenario 1 or Scenario 2 holds. When both scenarios are satisfied, we can take the minimum of the two bounds to obtain a tighter theoretical guarantee. The proof of Theorem 4.2 is provided in Appendix A. In addition, a potential concern is whether clipping the imputation introduces additional bias when the true rating falls outside the conformal interval. To address this concern, we have the following theorem:

Proposition 1. *Taking boundary clipping as an example, when the imputed rating satisfies $\tilde{r} > f_{\text{base}}(x_{u,i}) + \hat{q}_{1-\alpha}$, the correction reduces the bias for estimating r if and only if:*

$$P(r < r_{\text{mid}}) > P(r > r_{\text{mid}}) \cdot \frac{\mathbb{E}[r - r_{\text{mid}} \mid r > r_{\text{mid}}]}{\mathbb{E}[r_{\text{mid}} - r \mid r < r_{\text{mid}}]},$$

where r is the ground-truth rating and $r_{\text{mid}} = (\tilde{r} + f_{\text{base}}(x_{u,i}) + \hat{q}_{1-\alpha})/2$.

Intuitively, this condition holds if either fewer ratings exceed r_{mid}, i.e., $P(r > r_{\text{mid}})$ is small, or the magnitude of $r - r_{\text{mid}}$ for $r > r_{\text{mid}}$ is small.

5 Experiments

5.1 Semi-Synthetic Experiments

5.1.1 Data Processing and Baselines. We utilize the MovieLens-100K dataset to construct our semi-synthetic data. To obtain a complete ground truth matrix $\mathbf{R} \in \mathbb{R}^{N \times M}$ for evaluation, we first apply Matrix Factorization (MF) to complete the sparse user-item interaction matrix. Let $\tilde{r}_{u,i}$ denote the raw completed rating for user u and item i. Subsequently, following [31], we normalize these raw scores to a continuous range of $[0, 1]$ using Min-Max scaling to obtain the ground truth $r_{u,i}$:

$$r_{u,i} = \frac{\tilde{r}_{u,i} - \min(\tilde{R})}{\max(\tilde{R}) - \min(\tilde{R})}.$$

Following previous studies [8, 33, 44], to simulate the selection bias prevalent in real-world systems, we generate the observation indicator matrix O based on a pre-defined propensities:

$$\pi_{u,i} = \begin{cases} p^4_{\text{base}}, & \text{if } r_{u,i} \in [0, 0.4], \\ p^3_{\text{base}}, & \text{if } r_{u,i} \in (0.4, 0.6], \\ p^2_{\text{base}}, & \text{if } r_{u,i} \in (0.6, 0.8], \\ p^1_{\text{base}}, & \text{if } r_{u,i} \in (0.8, 1.0], \end{cases}$$

with p_{base} fixed as 0.5. This propensity assignment simulates the scenario that the users are more likely to rate items they like. Following [8, 44], we obtain the perturbed propensity as $1/\hat{\pi}_{u,i} = (1 - \beta)/\pi + \beta/\bar{o}$, where $\bar{o}$ is the empirical probability of observed $r_{u,i}$, and β is the propensity perturbation hyperparameter, which is 0.5 by default. We adopt the proposed algorithm to three representative DR-based methods: DR-JL [44], MRDR [8], and DR-BIAS [4].

5.1.2 Implementation Details. We randomly split the generated observed data into a training set and a calibration set with an 8:2 ratio. All models were optimized using the Adam optimizer. The base model is trained with a learning rate of 0.001 and a weight decay of 0.0002. The imputation model is trained with a learning rate of 0.02 and a weight decay of 0.002. We evaluate all methods using RMSE on all user-item pairs to evaluate prediction performance:

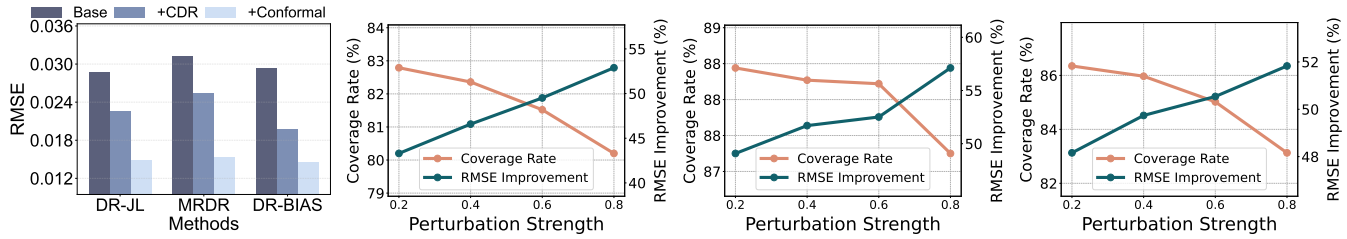
$$RMSE = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} (r_{u,i} - \hat{r}_{u,i})^2}.$$

In addition, we use the mean of RAIE in Equation 1 for imputation evaluation, i.e., RMAIE. We adopt a boundary clipping strategy for imputations that fall outside the interval.

5.1.3 Performance Comparison. Table 2 reports the δ_{RMSE} and δ_{RMAIE} as the relative reduction by our WCP-based learning algorithm with varying significance level. Note that our method reduces the imputation errors by around 55%-70% for all three DR-based methods. In addition, for the prediction model, the improvement is around 45% to 50%. These results validate the effectiveness of our algorithm and model-agnostic property.

5.1.4 Comparison to CDR. As illustrated in Fig. 2 (a), though CDR can also improve the performance of the prediction model, our WCP-based methods consistently outperform the CDR method across all backbones, because we include a fine-grained correction for imputation and can more efficiently find high-bias imputations.

5.1.5 Sensitivity Analysis. As shown in Fig. 2 (b)-(d), we report the RMSE improvement for the prediction model and coverage rate under different perturbation strengths of propensity. Specifically, the coverage rate means that the probability of true ratings falls into a conformal interval, and is robust to the propensity perturbation.



(a) Comparison with CDR (b) WCP-DR-JL (c) WCP-MRDR (d) WCP-DR-BIAS

Figure 2: Performance comparison of our WCP methods and CDR in Figure (a) and performance under varying propensity perturbation strength in Figure (b)-(d).

Table 3: Bias reduction statistics on ML-100K.

Method	Prob. of Bias Decrease ↑	Avg. Bias Decrease ↑	Avg. Bias Increase ↓	Expected Bias Reduction ↑
WCP-DR-JL	99.0 ± 0.2%	0.013 ± 0.001	0.005 ± 0.002	0.012 ± 0.001
WCP-MRDR	99.1 ± 0.1%	0.013 ± 0.001	0.006 ± 0.000	0.013 ± 0.001
WCP-DR-BIAS	99.3 ± 0.6%	0.019 ± 0.013	0.004 ± 0.002	0.018 ± 0.013

Note that we can keep almost the same coverage rate across different scenarios. In addition, the RMSE improvement is increasing as the perturbation strength increases, which further shows that our method is more robust to inaccurate propensities.

5.1.6 In-depth Analysis. We further verify the theoretical bias reduction property via semi-synthetic experiments on the ML-100K dataset. Experimental results in Table 3 show that our method maintains a bias reduction probability stably above 99% for all DR variants. Meanwhile, the average magnitude of bias reduction is consistently larger than the potential extra bias increment, which guarantees positive expected bias reduction in practical debiasing recommendation tasks.

5.2 Real-World Experiments

5.2.1 Datasets. **Coat**[1] [33], **Yahoo! R3**[2] [28], and **KuaiRec**[3] [5], covering coat, music, and short video recommendation tasks, respectively. Specifically, **Coat** dataset contains 6,960 biased ratings and 4,640 unbiased ratings from 290 users to 300 items, and each user self-selected 24 items to rate and randomly rates 16 items. **Yahoo! R3** dataset contains 311,704 biased ratings and 54,000 unbiased ratings from 15,400 users to 1,000 items, and **KuaiRec** dataset contains 4,676,570 video watching ratio from 1,411 users to 3,327 items. Following [15, 19], we binarize the ratings for **Coat** and **Yahoo! R3** dataset by denoting ratings less than three as negative conversion with label 0, and otherwise as positive conversion with label 1. For **KuaiRec**, we follow the data preprocessing in [12], and we binarize the watch ratio by setting values less than two to 0 and otherwise to 1.

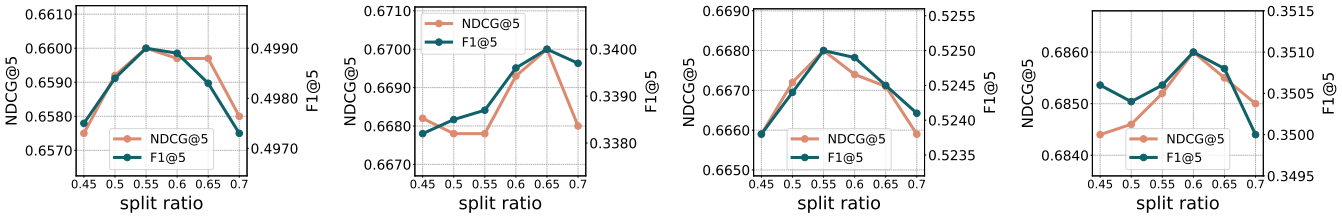
5.2.2 Baselines. We compare the proposed methods with the following baselines: (1) **Naive estimator** [11]: which directly fits the observational data without debiasing; (2) **IPS-based estimators**: IPS [33], SNIPS [37], ASIPS [30], IPS-V2 [18], KBIPS [20], AKBIPS [20]; (3) **DR-based estimators**: DR [31], DR-JL [44], MRDR-JL [8], DR-BIAS [4], DR-MSE [4], MR [14], CDR [34], TDR [15], TDR-JL [15], StableDR [19], DR-V2 [18], KBDR [20], AKBDR [20], DCE-DR [12], DCE-TDR [12], Cali-MR [7]. Following prior studies [12, 15, 18, 30], to ensure fair comparison, both the prediction and imputation models are implemented using Matrix Factorization (MF) as backbone [11] for all baselines. More details are available in Appendix B.

5.2.3 Implementation Details. We implement all methods using PyTorch and employ the Adam optimizer for training. The learning rate is tuned within the range $[0.005, 0.1]$, and the regularization coefficient λ is selected from [1e−5, 5e−3]. We set the batch size to 128 for **Coat** and 2048 for both **Yahoo! R3** and **KuaiRec**. Throughout all real-world experiments, we fix the significance level $\alpha = 0.2$ for main result reporting, and further analyze $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$ in sensitivity experiments. The train-calibration split ratio is set within $[0.55, 0.65]$ for real-world datasets and fixed as 8:2 for the semi-synthetic ML-100K dataset. All propensity models adopt the same network structure and are estimated only once in the initial stage, then reused for subsequent base model training, conformal interval construction, and DR correction learning to maintain theoretical coverage validity. For imputation correction, we mainly adopt boundary clipping and compare it with hard rejection and center clipping in the analysis. To evaluate prediction performance, we adopt three widely used evaluation metrics: AUC, NDCG@K and F1@K. NDCG@K and F1@K are particularly popular in recommender systems as they assess the quality of ranking, which is central to recommendation tasks. We set $K = 5$ for **Coat** and **Yahoo! R3** because there are only 16 / 10 items per user in the test set, and $K = 20$ for **KuaiRec**.

[1]https://www.cs.cornell.edu/~schnabts/mnar
[2]https://www.kaggle.com/datasets/limitiao/yahoor3
[3]https://kuairec.com

Table 4: Performance on AUC, NDCG@K, and F1@K on Coat, Yahoo! R3 and KuaiRec. The best and the second best results are bolded and underlined, where * means statistically significant results (p-value ≤ 0.05) using the paired t-test compared with the best baselines.

	Coat			Yahoo! R3			KuaiRec		
Method	AUC	NDCG@5	F1@5	AUC	NDCG@5	F1@5	AUC	NDCG@20	F1@20
Naive	$0.703_{\pm 0.006}$	$0.605_{\pm 0.012}$	$0.467_{\pm 0.007}$	$0.673_{\pm 0.001}$	$0.635_{\pm 0.002}$	$0.306_{\pm 0.002}$	$0.753_{\pm 0.001}$	$0.449_{\pm 0.002}$	$0.124_{\pm 0.002}$
IPS	$0.717_{\pm 0.007}$	$0.617_{\pm 0.009}$	$0.473_{\pm 0.008}$	$0.678_{\pm 0.001}$	$0.638_{\pm 0.002}$	$0.318_{\pm 0.002}$	$0.755_{\pm 0.004}$	$0.452_{\pm 0.010}$	$0.131_{\pm 0.004}$
SNIPS	$0.714_{\pm 0.012}$	$0.614_{\pm 0.012}$	$0.474_{\pm 0.009}$	$0.683_{\pm 0.002}$	$0.639_{\pm 0.002}$	$0.316_{\pm 0.002}$	$0.754_{\pm 0.003}$	$0.453_{\pm 0.004}$	$0.126_{\pm 0.003}$
ASIPS	$0.719_{\pm 0.009}$	$0.618_{\pm 0.012}$	$0.476_{\pm 0.009}$	$0.679_{\pm 0.003}$	$0.640_{\pm 0.003}$	$0.319_{\pm 0.003}$	$0.757_{\pm 0.005}$	$0.474_{\pm 0.007}$	$0.130_{\pm 0.005}$
IPS-V2	$0.726_{\pm 0.005}$	$0.627_{\pm 0.009}$	$0.479_{\pm 0.008}$	$0.685_{\pm 0.002}$	$0.646_{\pm 0.003}$	$0.320_{\pm 0.002}$	$0.764_{\pm 0.001}$	$0.476_{\pm 0.003}$	$0.135_{\pm 0.003}$
KBIPS	$0.714_{\pm 0.003}$	$0.618_{\pm 0.010}$	$0.474_{\pm 0.007}$	$0.676_{\pm 0.002}$	$0.642_{\pm 0.003}$	$0.318_{\pm 0.002}$	$0.763_{\pm 0.001}$	$0.463_{\pm 0.007}$	$0.134_{\pm 0.002}$
AKBIPS	$0.732_{\pm 0.004}$	$0.636_{\pm 0.006}$	$0.483_{\pm 0.006}$	$0.689_{\pm 0.001}$	$0.658_{\pm 0.002}$	$0.324_{\pm 0.002}$	$0.766_{\pm 0.003}$	$0.478_{\pm 0.009}$	$0.138_{\pm 0.003}$
DR	$0.718_{\pm 0.008}$	$0.623_{\pm 0.009}$	$0.474_{\pm 0.007}$	$0.684_{\pm 0.002}$	$0.658_{\pm 0.003}$	$0.326_{\pm 0.002}$	$0.755_{\pm 0.008}$	$0.462_{\pm 0.010}$	$0.135_{\pm 0.005}$
DR-JL	$0.723_{\pm 0.005}$	$0.629_{\pm 0.007}$	$0.479_{\pm 0.005}$	$0.685_{\pm 0.002}$	$0.653_{\pm 0.002}$	$0.324_{\pm 0.002}$	$0.766_{\pm 0.002}$	$0.467_{\pm 0.005}$	$0.136_{\pm 0.003}$
MRDR-JL	$0.727_{\pm 0.005}$	$0.627_{\pm 0.008}$	$0.480_{\pm 0.008}$	$0.684_{\pm 0.002}$	$0.652_{\pm 0.003}$	$0.325_{\pm 0.002}$	$0.768_{\pm 0.005}$	$0.473_{\pm 0.007}$	$0.139_{\pm 0.004}$
DR-BIAS	$0.726_{\pm 0.004}$	$0.629_{\pm 0.009}$	$0.482_{\pm 0.007}$	$0.685_{\pm 0.002}$	$0.653_{\pm 0.002}$	$0.325_{\pm 0.003}$	$0.768_{\pm 0.003}$	$0.477_{\pm 0.006}$	$0.137_{\pm 0.004}$
DR-MSE	$0.727_{\pm 0.007}$	$0.631_{\pm 0.008}$	$0.484_{\pm 0.007}$	$0.687_{\pm 0.002}$	$0.657_{\pm 0.003}$	$0.327_{\pm 0.003}$	$0.770_{\pm 0.003}$	$0.480_{\pm 0.006}$	$0.140_{\pm 0.003}$
MR	$0.724_{\pm 0.004}$	$0.636_{\pm 0.006}$	$0.481_{\pm 0.006}$	$0.691_{\pm 0.002}$	$0.647_{\pm 0.002}$	$0.316_{\pm 0.003}$	$0.776_{\pm 0.005}$	$0.483_{\pm 0.006}$	$0.142_{\pm 0.003}$
TDR	$0.714_{\pm 0.006}$	$0.634_{\pm 0.011}$	$0.483_{\pm 0.008}$	$0.688_{\pm 0.003}$	$0.662_{\pm 0.002}$	$0.329_{\pm 0.002}$	$0.772_{\pm 0.003}$	$0.486_{\pm 0.005}$	$0.140_{\pm 0.003}$
TDR-JL	$0.731_{\pm 0.005}$	$0.639_{\pm 0.007}$	$0.484_{\pm 0.007}$	$0.689_{\pm 0.002}$	$0.656_{\pm 0.004}$	$0.327_{\pm 0.003}$	$0.772_{\pm 0.003}$	$0.489_{\pm 0.005}$	$0.142_{\pm 0.003}$
StableDR	$0.735_{\pm 0.005}$	$0.640_{\pm 0.007}$	$0.484_{\pm 0.006}$	$0.688_{\pm 0.002}$	$0.661_{\pm 0.003}$	$0.329_{\pm 0.002}$	$0.773_{\pm 0.001}$	$0.491_{\pm 0.003}$	$0.143_{\pm 0.003}$
DR-V2	$0.734_{\pm 0.007}$	$0.639_{\pm 0.009}$	$0.487_{\pm 0.006}$	$0.690_{\pm 0.002}$	$0.660_{\pm 0.005}$	$0.328_{\pm 0.002}$	$0.773_{\pm 0.003}$	$0.488_{\pm 0.006}$	$0.142_{\pm 0.004}$
KBDR	$0.730_{\pm 0.003}$	$0.631_{\pm 0.005}$	$0.482_{\pm 0.006}$	$0.682_{\pm 0.002}$	$0.648_{\pm 0.003}$	$0.323_{\pm 0.002}$	$0.765_{\pm 0.004}$	$0.460_{\pm 0.006}$	$0.138_{\pm 0.003}$
AKBDR	$0.745_{\pm 0.004}$	$0.645_{\pm 0.008}$	$0.493_{\pm 0.007}$	$0.692_{\pm 0.002}$	$0.661_{\pm 0.002}$	$0.328_{\pm 0.002}$	$0.782_{\pm 0.003}$	$0.498_{\pm 0.008}$	$0.147_{\pm 0.003}$
DCE-DR	$0.736_{\pm 0.006}$	$0.648_{\pm 0.007}$	$0.489_{\pm 0.005}$	$0.698_{\pm 0.002}$	$0.670_{\pm 0.002}$	$0.333_{\pm 0.003}$	$0.795_{\pm 0.004}$	$0.512_{\pm 0.005}$	$0.153_{\pm 0.002}$
DCE-TDR	$0.740_{\pm 0.004}$	$0.651_{\pm 0.006}$	$0.489_{\pm 0.007}$	$0.701_{\pm 0.002}$	$0.672_{\pm 0.002}$	$0.331_{\pm 0.002}$	$0.798_{\pm 0.005}$	$0.514_{\pm 0.006}$	$0.155_{\pm 0.002}$
Cali-MR	$0.741_{\pm 0.002}$	$0.658_{\pm 0.004}$	$0.495_{\pm 0.004}$	$\underline{0.703}_{\pm 0.002}$	$\underline{0.678}_{\pm 0.002}$	$0.338_{\pm 0.004}$	$0.798_{\pm 0.003}$	$\underline{0.521}_{\pm 0.005}$	$0.158_{\pm 0.002}$
CDR-DR-JL	$0.742_{\pm 0.002}$	$0.637_{\pm 0.003}$	$0.481_{\pm 0.004}$	$0.692_{\pm 0.002}$	$0.657_{\pm 0.003}$	$0.327_{\pm 0.004}$	$0.774_{\pm 0.003}$	$0.477_{\pm 0.004}$	$0.138_{\pm 0.005}$
CDR-DR-BIAS	$0.743_{\pm 0.003}$	$0.649_{\pm 0.004}$	$0.483_{\pm 0.005}$	$0.691_{\pm 0.003}$	$0.654_{\pm 0.004}$	$0.330_{\pm 0.005}$	$0.773_{\pm 0.002}$	$0.489_{\pm 0.003}$	$0.139_{\pm 0.004}$
CDR-TDR	$0.743_{\pm 0.004}$	$0.659_{\pm 0.005}$	$0.495_{\pm 0.002}$	$0.692_{\pm 0.004}$	$0.451_{\pm 0.005}$	$0.332_{\pm 0.002}$	$0.803_{\pm 0.004}$	$0.518_{\pm 0.005}$	$0.158_{\pm 0.002}$
WCP-DR-JL (ours)	$0.744_{\pm 0.004}$	$0.660_{\pm 0.005}$	$0.499_{\pm 0.003}$	$0.698_{\pm 0.002}$	$0.670_{\pm 0.004}$	$0.340_{\pm 0.003}$	$\underline{0.812}_{\pm 0.003}$	$0.513_{\pm 0.005}$	$0.160_{\pm 0.004}$
WCP-DR-BIAS (ours)	$\underline{0.750}_{\pm 0.003}$	$\underline{0.664}_{\pm 0.004}$	$\underline{0.518}_{\pm 0.005}$	$0.690_{\pm 0.002}$	$0.659_{\pm 0.003}$	$\underline{0.342}_{\pm 0.004}$	$0.803_{\pm 0.004}$	$\underline{0.521}_{\pm 0.003}$	$\underline{0.162}_{\pm 0.005}$
WCP-DCE-TDR (ours)	$\mathbf{0.763}^{*}_{\pm 0.002}$	$\mathbf{0.668}^{*}_{\pm 0.003}$	$\mathbf{0.525}^{*}_{\pm 0.004}$	$\mathbf{0.719}^{*}_{\pm 0.003}$	$\mathbf{0.686}^{*}_{\pm 0.002}$	$\mathbf{0.351}^{*}_{\pm 0.003}$	$\mathbf{0.819}^{*}_{\pm 0.004}$	$\mathbf{0.525}_{\pm 0.004}$	$\mathbf{0.171}^{*}_{\pm 0.004}$

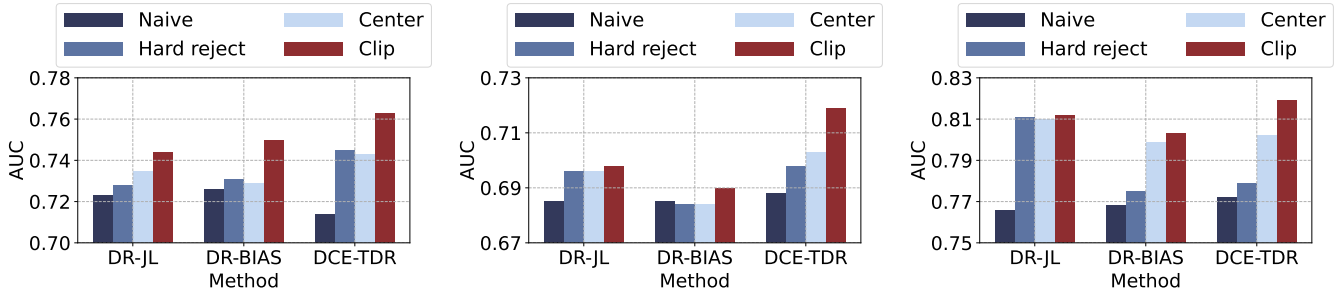


(a) WCP-DR-JL on Coat (b) WCP-DR-JL on Yahoo! R3 (c) WCP-DCE-TDR on Coat (d) WCP-DCE-TDR on Yahoo! R3

Figure 3: Impact of ratio of training set and calibration set on Coat and Yahoo! R3.

5.2.4 Overall Performance. Table 4 presents the performance comparison of various debiasing methods across three real-world datasets. The Naive method, which directly learns from observational data without any debiasing treatment, exhibits relatively poor performance across most metrics, confirming the adverse impact of selection bias. To mitigate this, IPS-based methods utilize inverse propensity weighting to adjust the data distribution, outperforming the Naive baseline. Furthermore, DR-based methods generally achieve better debiasing effects than IPS-based approaches by incorporating error imputation to reduce variance while maintaining unbiasedness. Among the baselines, kernel-based methods and calibration-enhanced methods show the strongest performance.

In comparison, our WCP-enhanced methods (WCP-DR-JL, WCP-DR-BIAS, and WCP-DCE-TDR) consistently and significantly outperform all baselines. By leveraging weighted conformal prediction to construct a valid prediction set and rectifying unreliable



(a) Effect of correction method on Coat (b) Effect of correction method on Yahoo! R3 (c) Effect of correction method on KuaiRec

Figure 4: Effect of different imputation correction methods on Coat, Yahoo! R3, and KuaiRec.

Table 5: Sensitivity analysis of significance level α. The best performance is bolded.

	Coat			Yahoo! R3			KuaiRec		
α	AUC	NDCG@5	F1@5	AUC	NDCG@5	F1@5	AUC	NDCG@20	F1@20
0.1	0.737	0.656	0.494	0.694	0.667	0.338	**0.812**	**0.513**	**0.160**
0.2	0.737	0.654	0.495	0.695	0.668	0.339	0.809	0.509	0.158
0.3	0.740	0.657	0.495	**0.698**	**0.670**	**0.340**	0.807	0.511	0.157
0.4	**0.744**	**0.660**	**0.499**	0.696	0.668	0.338	0.805	0.506	0.158

pseudo-labels, our method substantially enhances the performance of DR-based estimators. Specifically, the results also prove that our methods can easily be adopted to any existing DR-based methods, showing the generality of our method in practice.

5.2.5 Effects of Train-Calibration Split Ratio. We evaluate the impact of the train-calibration split ratio on the performance of WCP-DR-JL and WCP-DCE-TDR on **Coat** and **Yahoo! R3** datasets. As presented in Figure 3, model performance peaks at a moderate split ratio, while deviating from this range (either too small or too large) leads to obvious degradation. This is attributed to the trade-off between the two subsets: an insufficient calibration set undermines the reliability of conformal interval estimation, while an overly small training set impairs the base model's ability to construct a valid prediction set.

5.2.6 Sensitivity Analysis of Significance Level α. The significance level α governs the tightness of conformal intervals and the frequency of imputation correction. We conduct sensitivity analysis for $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$ on three real-world datasets. As shown in Table 5, the optimal α varies slightly across datasets, while all metrics remain stable across the tested range, verifying the hyper-parameter robustness of our method.

5.2.7 Effects of Imputation Correction Methods. We evaluate the impact of different imputation correction methods, including hard rejection, center clipping (Center), and boundary clipping (Clip), on the performance of DR-JL, DR-BIAS, and DCE-TDR on **Coat**, **Yahoo! R3**, and **KuaiRec** datasets in Figure 4. We observe that hard rejection, Center, and Clip all outperform Naive, with Clip achieving the best results overall. Hard rejection is effective but suffers from inherent bias, while Center and Clip perform more stably. It is attributed to their distinct handling of imputations: Hard rejection sets all imputations to 0 and abandons them directly, while center clipping and boundary clipping revise high-bias imputations to reasonable values within the conformal interval. Boundary clipping further restricts imputation values based on the original prediction, which can maximize the model performance.

6 Conclusion

In this paper, we take RS as an example, addressing the critical challenge of non-random missing labels. While DR estimators provide a theoretical foundation for unbiased learning, their practical performance is always hindered by the difficulty of obtaining accurate imputations. To bridge this gap, we proposed a novel framework that integrates weighted conformal prediction (WCP) into the debiasing pipeline to find and correct high-bias imputations. Our proposed WCP-DR learning algorithm first trains a debiased base prediction model, then constructs weighted conformal prediction intervals for unobserved user-item pairs. Finally, WCP-DR corrects high-bias imputations by clipping them into these valid intervals or dropping during training. In addition, we prove the robustness of the interval coverage guarantee towards inaccurate propensities and the upper bound of the interval to ensure the interval is not too wide. Extensive experiments on one semi-synthetic dataset and three real-world datasets show the effectiveness of the proposed method. Specifically, we show that the convergence rate is also empirically guaranteed in different scenarios, and the real-world performance of our methods is robust to the different training/calibration set ratios.

One potential limitation is that the correction method is straightforward. It is promising to investigate other correction methods to obtain more accurate imputations. In addition, to address the problem of suboptimal weight learning in the weighted conformal method, we can adopt a quantile regression model to obtain a more robust conformal interval.

Acknowledgment

This work is supported by the Beijing Natural Science Foundation (No. L257007), the Beijing Major Science and Technology Project (No. Z251100008425006), and the Science and Technology Innovation Key R&D Program of Chongqing (No. CSTB2024TIAD-STX0040).

References

[1] Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, et al. 2010. Training and Testing Low-Degree Polynomial Data Mappings via Linear SVM. *Journal of Machine Learning Research* 11, 4 (2010).

[2] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. AutoDebias: Learning to Debias for Recommendation. In *International SIGIR Conference on Research and Development in Information Retrieval*. doi:10.1145/3404835.3462919

[3] Zonghao Chen, Ruocheng Guo, Jean-François Ton, and Yang Liu. 2024. Conformal Counterfactual Inference Under Hidden Confounding. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[4] Quanyu Dai, Haoxuan Li, Peng Wu, Zhenhua Dong, Xiao-Hua Zhou, Rui Zhang, Rui Zhang, and Jie Sun. 2022. A Generalized Doubly Robust Learning Framework for Debiasing Post-Click Conversion Rate Prediction. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[5] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A Fully-Observed Dataset and Insights for Evaluating Recommender Systems. In *ACM International Conference on Information and Knowledge Management*.

[6] Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. 2023. Safe Planning in Dynamic Environments Using Conformal Prediction. In *Uncertainty in Artificial Intelligence*.

[7] Shuxia Gong and Chen Ma. 2025. Gradient-Based Multiple Robust Learning Calibration on Data Missing-Not-at-Random via Bi-Level Optimization. *Entropy* 27, 2 (2025).

[8] Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. 2021. Enhanced Doubly Robust Learning for Debiasing Post-Click Conversion Rate Estimation. In *International SIGIR Conference on Research and Development in Information Retrieval*.

[9] José Miguel Hernández-Lobato, Neil Houlsby, and Zoubin Ghahramani. 2014. Probabilistic Matrix Factorization with Non-Random Missing Data. In *International Conference on Machine Learning*.

[10] Guido W Imbens and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge university press.

[11] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009).

[12] Wonbin Kweon and Hwanjo Yu. 2024. Doubly Calibrated Estimator for Recommendation on Data Missing Not at Random. In *International World Wide Web Conference*.

[13] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. 2018. Distribution-Free Predictive Inference for Regression. *J. Amer. Statist. Assoc.* 113, 523 (2018).

[14] Haoxuan Li, Quanyu Dai, Yuru Li, Yan Lyu, Zhenhua Dong, Xiao-Hua Zhou, and Peng Wu. 2023. Multiple Robust Learning for Recommendation. In *AAAI Conference on Artificial Intelligence*.

[15] Haoxuan Li, Yan Lyu, Chunyuan Zheng, and Peng Wu. 2023. TDR-CL: Targeted Doubly Robust Collaborative Learning for Debiased Recommendations. In *International Conference on Learning Representations*.

[16] Haoxuan Li, Kunhan Wu, Chunyuan Zheng, Yanghao Xiao, Hao Wang, Zhi Geng, Fuli Feng, Xiangnan He, and Peng Wu. 2023. Removing Hidden Confounding in Recommendation: A Unified Multi-Task Learning Approach. In *Advances in Neural Information Processing Systems*.

[17] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. 2023. Balancing Unobserved Confounding with a Few Unbiased Ratings in Debiased Recommendations. In *International World Wide Web Conference*.

[18] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, and Peng Cui. 2023. Propensity Matters: Measuring and Enhancing Balancing for Recommendation. In *International Conference on Machine Learning*.

[19] Haoxuan Li, Chunyuan Zheng, and Peng Wu. 2023. StableDR: Stabilized Doubly Robust Learning for Recommendation on Data Missing Not at Random. In *International Conference on Learning Representations*.

[20] Haoxuan Li, Chunyuan Zheng, Yanghao Xiao, Peng Wu, Zhi Geng, Xu Chen, and Peng Cui. 2024. Debiased Collaborative Filtering with Kernel-based Causal Balancing. In *International Conference on Learning Representations*.

[21] Xiang Li, Yanghao Xiao, Chunyuan Zheng, Qian Zou, Bing Cheng, Wei Lin, Haoxuan Li, and Zhouchen Lin. 2026. Optimizing Marketing Subsidies via Counterfactual Learning with Asymmetric Reward Function. In *International SIGIR Conference on Research and Development in Information Retrieval*.

[22] Xiang Li, Zhao-Yu Zhang, Chunyuan Zheng, Qingying Chen, Huiyou Jiang, Haoxuan Li, and Zhouchen Lin. 2026. User Activity Modeling under Inflated Distribution. In *International SIGIR Conference on Research and Development in Information Retrieval*.

[23] Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. 2023. Safe Planning in Dynamic Environments Using Conformal Prediction. *IEEE Robotics and Automation Letters* 8, 8 (2023).

[24] Haochen Liu, Da Tang, Ji Yang, Xiangyu Zhao, Hui Liu, Jiliang Tang, and Youlong Cheng. 2022. Rating Distribution Calibration for Selection Bias Mitigation in Recommendations. In *International World Wide Web Conference*.

[25] Weiming Liu, Chaochao Chen, Xinting Liao, Mengling Hu, Jiajie Su, Yanchao Tan, and Fan Wang. 2024. User Distribution Mapping Modelling with Collaborative Filtering for Cross Domain Recommendation. In *International World Wide Web Conference*.

[26] Jinwei Luo, Dugang Liu, Weike Pan, and Zhong Ming. 2021. Unbiased Recommendation Model Based on Improved Propensity Score Estimation. *Journal of Computer Applications* 3508 (2021).

[27] Yan Lyu, Haoxuan Li, Tianyu Xia, Xiang Li, Xiangnan Feng, Chunyuan Zheng, and Xiao-Hua Zhou. 2026. Hierarchical Denoising Entire Space Multi-Task Model for Post-Click Conversion Rate Prediction with Noisy Labels. In *International SIGIR Conference on Research and Development in Information Retrieval*.

[28] Benjamin M. Marlin, Richard S. Zemel andSam T. Roweis, and Malcolm Slaney. 2007. Collaborative Filtering and the Missing at Random Assumption. In *Conference on Uncertainty in Artificial Intelligence*.

[29] Benjamin M Marlin and Richard S Zemel. 2009. Collaborative Prediction and Ranking with Non-Random Missing Data. In *ACM Conference on Recommender Systems*.

[30] Yuta Saito. 2020. Asymmetric Tri-Training for Debiasing Missing-Not-at-Random Explicit Feedback. In *International SIGIR Conference on Research and Development in Information Retrieval*.

[31] Yuta Saito. 2020. Doubly Robust Estimator for Ranking Metrics with Post-Click Conversions. In *ACM Conference on Recommender Systems*.

[32] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased Recommender Learning From Missing-Not-at-Random Implicit Feedback. In *ACM International Conference on Web Search and Data Mining*.

[33] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *International Conference on Machine Learning*.

[34] Zijie Song, Jiawei Chen, Sheng Zhou, Qihao Shi, Yan Feng, Chun Chen, and Can Wang. 2023. CDR: Conservative Doubly Robust Learning for Debiased Recommendation. In *ACM International Conference on Information and Knowledge Management*.

[35] Harald Steck. 2010. Training and Testing of Recommender Systems on Data Missing Not at Random. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[36] Harald Steck. 2013. Evaluation of Recommendations: Rating-Prediction and Ranking. In *ACM Conference on Recommender Systems*.

[37] Adith Swaminathan and Thorsten Joachims. 2015. The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*.

[38] Ryan Tibshirani. 2023. Conformal Prediction. *UC Berkeley* (2023).

[39] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. Conformal Prediction Under Covariate Shift. *Advances in neural information processing systems* 32 (2019).

[40] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer.

[41] Hao Wang, Zhichao Chen, Zhaoran Liu, Haozhe Li, Degui Yang, Xinggao Liu, and Haoxuan Li. 2024. Entire Space Counterfactual Learning for Reliable Content Recommendations. *IEEE Transactions on Information Forensics and Security* (2024).

[42] Hao Wang, Zhichao Chen, Honglei Zhang, Zhengnan Li, Licheng Pan, Haoxuan Li, and Mingming Gong. 2025. Debiased Recommendation via Wasserstein Causal Balancing. *ACM Transactions on Information Systems* (2025).

[43] Jun Wang, Haoxuan Li, Chi Zhang, Dongxu Liang, Enyun Yu, Wenwu Ou, and Wenjia Wang. 2023. CounterCLR: Counterfactual contrastive learning with non-random missing data in recommendation. In *IEEE International Conference on Data Mining*.

[44] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly Robust Joint Learning for Recommendation on Data Missing Not at Random. In *International Conference on Machine Learning*.

[45] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2021. Combating Selection Biases in Recommender Systems with a Few Unbiased Ratings. In *ACM International Conference on Web Search and Data Mining*.

[46] Anpeng Wu, Haoxuan Li, Chunyuan Zheng, Kun Kuang, and Kun Zhang. 2025. Classifying Treatment Responders: Bounds and Algorithms. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[47] Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning Clicks Into Purchases: Revenue Optimization for Product Search in E-Commerce. In *International SIGIR Conference on Research and Development in Information Retrieval*.

[48] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. 2022. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In *International Joint Conferences on Artificial Intelligence*.

[49] Yanghao Xiao, Hao Wang, Xiang Li, Qian Zou, Bing Cheng, Wei Lin, Haoxuan Li, and Zhouchen Lin. 2026. Debiased Recommendation Beyond the Positive Propensity Assumption. In *International SIGIR Conference on Research and Development in Information Retrieval*.

[50] Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. 2024. Doubly Robust Calibration of Prediction Sets Under Covariate Shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86, 4 (2024).
[51] Bowen Yuan, Jui-Yang Hsia, Meng-Yuan Yang, Hong Zhu, Chih-Yao Chang, Zhenhua Dong, and Chih-Jen Lin. 2019. Improving Ad Click Prediction by Considering Non-Displayed Events. In *ACM International Conference on Information and Knowledge Management*.
[52] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Propensity Matters: Measuring and Enhancing Balancing for Recommendation. *ACM computing surveys* 52, 1 (2019).
[53] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. 2020. Large-Scale Causal Approaches to Debiasing Post-Click Conversion Rate Estimation with Multi-Task Learning. In *International World Wide Web Conference*.
[54] Yongfeng Zhang, Xu Chen, et al. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.
[55] Yunchuan Zhang, Sangwoo Park, and Osvaldo Simeone. 2024. Bayesian Optimization with Formal Safety Guarantees via Online Conformal Prediction. *IEEE Journal of Selected Topics in Signal Processing* 19, 1 (2024).
[56] Chunyuan Zheng, Anpeng Wu, Chuan Zhou, Taojun Hu, Qingying Chen, Hongyi Liu, Chenxi Li, Huiyou Jiang, Haoxuan Li, and Zhouchen Lin. 2026. Uplift Modeling with Delayed Feedback: Identifiability and Algorithms. In *AAAI Conference on Artificial Intelligence*.
[57] Chunyuan Zheng, Haocheng Yang, Haoxuan Li, and Mengyue Yang. 2025. Unveiling Extraneous Sampling Bias with Data Missing-Not-At-Random. In *Advances in Neural Information Processing Systems*.
[58] Chuan Zhou, Yaxuan Li, Chunyuan Zheng, Haiteng Zhang, Min Zhang, Haoxuan Li, and Mingming Gong. 2025. A Two-Stage Pretraining-Finetuning Framework for Treatment Effect Estimation with Unmeasured Confounding. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
[59] Chuan Zhou, Lina Yao, Haoxuan Li, and Mingming Gong. 2025. Counterfactual Implicit Feedback Modeling. In *Advances in Neural Information Processing Systems*.

A Proof of Theorem 4.2

Theorem 4.2. Let $\hat{C}_\alpha$ be the prediction set constructed at a nominal significance level α using estimated weights $\hat{w}$. If the two following scenarios hold:

- Scenario 1: The uniform relative error is bounded by $\epsilon \in [0, 1)$:
$$\sup_x \left| \frac{w(x)}{\hat{w}(x)} - 1 \right| \le \epsilon.$$
- Scenario 2: The second moment of the ratio is bounded by σ^2:
$$\mathbb{E}_{P_{\hat{w}}} \left[\left(\frac{w(x)}{\hat{w}(x)} - 1 \right)^2 \right] \le \sigma^2.$$

The coverage probability for the user-item pair (u', i') with missing ratings is bounded:

$$\mathbb{P}(r_{u',i'} \in \hat{C}_\alpha(x_{u',i'})) \ge 1 - \alpha - \min\left(\alpha\epsilon, \sigma\sqrt{\alpha}\right).$$

Proof. Let $E = \{r_{u',i'} \notin \hat{C}_\alpha(x_{u',i'})\}$ denote the miscoverage event. By the validity of weighted conformal prediction under the estimated weights $\hat{w}$, we have

$$P_{\hat{w}}(E) \le \alpha.$$

Since both w and $\hat{w}$ are normalized, the density ratio between the target distribution induced by the true weights and the distribution induced by the estimated weights is

$$\rho(x) = \frac{dP_w}{dP_{\hat{w}}}(x) = \frac{w(x)}{\hat{w}(x)}.$$

Therefore, we have

$$P_w(E) = \mathbb{E}_{P_w}[\mathbb{I}_E] = \mathbb{E}_{P_{\hat{w}}}[\rho(X)\mathbb{I}_E] = \mathbb{E}_{P_{\hat{w}}}[\mathbb{I}_E] + \mathbb{E}_{P_{\hat{w}}}[(\rho(X)-1)\mathbb{I}_E].$$

Since $\mathbb{E}_{P_{\hat{w}}}[\mathbb{I}_E] = P_{\hat{w}}(E) \le \alpha$, it remains to bound

$$\Delta = \mathbb{E}_{P_{\hat{w}}}[(\rho(X)-1)\mathbb{I}_E].$$

Scenario 1. If

$$\sup_x |\rho(x) - 1| \le \epsilon,$$

then

$$|\Delta| \le \mathbb{E}_{P_{\hat{w}}}[|\rho(X)-1|\mathbb{I}_E] \le \epsilon P_{\hat{w}}(E) \le \alpha\epsilon.$$

Thus,

$$P_w(E) \le \alpha + \alpha\epsilon.$$

Scenario 2. By the Cauchy–Schwarz inequality,

$$|\Delta| \le \sqrt{\mathbb{E}_{P_{\hat{w}}}[(\rho(X)-1)^2]}\sqrt{\mathbb{E}_{P_{\hat{w}}}[\mathbb{I}_E^2]}.$$

Since $\mathbb{I}_E^2 = \mathbb{I}_E$ and $P_{\hat{w}}(E) \le \alpha$, the second-moment condition gives

$$|\Delta| \le \sigma\sqrt{P_{\hat{w}}(E)} \le \sigma\sqrt{\alpha}.$$

Thus,

$$P_w(E) \le \alpha + \sigma\sqrt{\alpha}.$$

If both scenarios hold, combining the two bounds yields

$$P_w(E) \le \alpha + \min\left(\alpha\epsilon, \sigma\sqrt{\alpha}\right).$$

Equivalently,

$$\mathbb{P}\left(r_{u',i'} \in \hat{C}_\alpha(x_{u',i'})\right) = 1 - P_w(E) \ge 1 - \alpha - \min\left(\alpha\epsilon, \sigma\sqrt{\alpha}\right).$$

□

B Experimental Baseline

We compare our method with the following baselines for comprehensive evaluations:

- **Naive** method [11] directly minimizes the empirical loss computed on observed user-item interactions without any debiasing treatment.
- **IPS** method [33] applies inverse propensity weighting to adjust the observed ratings, assigning higher weights to under-represented interactions.
- **SNIPS** method [37] employs self-normalized inverse propensity scores for reweighting, which helps stabilize the estimator and reduce variance.
- **ASIPS** method [30] constructs pseudo-labels for unobserved entries to address the high variance issue and propensity estimation errors in IPS-based approaches.
- **DR** method [31] integrates error imputation with inverse propensity weighting, achieving unbiasedness when either component is correctly specified. The imputed errors are often initialized using label statistics.
- **DR-JL** method [44] extends DR by parameterizing the imputation model with neural networks and optimizing both the prediction and imputation models in a joint learning framework.
- **MRDR** method [8] improves upon **DR-JL** by incorporating explicit variance regularization into the imputation model training objective.
- **DR-BIAS** method [4] augments **DR-JL** by introducing bias regularization terms to the imputation model loss function.

- **DR-MSE** method [4] unifies **MRDR** and **DR-BIAS** by jointly minimizing both bias and variance components, thereby controlling the mean squared error.
- **MR** method [14] leverages an ensemble of propensity and imputation models to achieve robustness against misspecification of individual models in the DR framework.
- **TDR** and **TDR-JL** methods [15] employ targeted learning techniques to refine imputed errors, enabling simultaneous reduction of both bias and variance in DR estimators.
- **StableDR** method [19] develops a stabilized DR estimator by learning constrained propensity scores that reduce sensitivity to extrapolation and handle small propensity values more robustly.
- **IPS-V2** and **DR-V2** methods [18] train propensity models that achieve covariate balance by matching pre-specified moment functions of the feature distributions.
- **KBIPS** and **KBDR** methods [20] extend covariate balancing to reproducing kernel Hilbert spaces (RKHS), where randomly sampled kernel functions are used to enforce balance constraints.
- **AKBIPS** and **AKBDR** methods [20] enhance kernel-based balancing by adaptively selecting the most informative kernel functions that maximize bias reduction in propensity estimation.
- **DCE-DR** and **DCE-TDR** methods [12] apply Mixture-of-Experts based calibration to improve the reliability of both propensity and imputation models in DR and TDR frameworks.
- **Cali-MR** method [7] introduces a bi-level optimization approach with differentiable expected calibration error to calibrate multiple propensity and imputation models in the MR estimator.