entropy

MDPI

Article

Gradient-Based Multiple Robust Learning Calibration on Data Missing-Not-at-Random via Bi-Level Optimization

Shuxia Gong [1,*] and Chen Ma [2,*]

[1] Mogo Co., Ltd., Beijing 100000, China
[2] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China
* Correspondence: silencesliver@hotmail.com (S.G.); machen2001@ruc.edu.cn (C.M.)

Abstract: Recommendation systems (RS) have become integral to numerous digital platforms and applications, ranging from e-commerce to content streaming field. A critical problem in RS is that the ratings are missing not at random (MNAR), which is due to the users always giving feedback on items with self-selection. The biased selection of rating data results in inaccurate rating prediction for all user-item pairs. Doubly robust (DR) learning has been studied in many tasks in RS, which is unbiased when either a single imputation or a single propensity model is accurate. In addition, multiple robust (MR) has been proposed with multiple imputation models and propensity models, and is unbiased when there exists a linear combination of these imputation models and propensity models is correct. However, we claim that the imputed errors and propensity scores are miscalibrated in the MR method. In this paper, we propose a gradient-based calibrated multiple robust learning method to enhance the debiasing performance and reliability of the rating prediction model. Specifically, we propose to use bi-level optimization to solve the weights and model coefficients of each propensity and imputation model in MR framework. Moreover, we adopt the differentiable expected calibration error as part of the objective to optimize the model calibration quality directly. Experiments on three real-world datasets show that our method outperforms the state-of-the-art baselines.

Keywords: causal recommendation; multiple robust; calibration; bi-level optimization

check for updates

Academic Editor: Deniz Gençağa

Received: 13 December 2024
Revised: 24 January 2025
Accepted: 10 February 2025
Published: 13 February 2025

Citation: Gong, S.; Ma, C. Gradient-Based Multiple Robust Learning Calibration on Data Missing-Not-at-Random via Bi-Level Optimization. *Entropy* **2025**, *27*, 196. https://doi.org/10.3390/e27020196

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recommendation systems (RS) is an effective tool to address the problem of information overload and has been widely used in e-commerce, social media, and entertainment [1]. RS aims to predict user preferences for items based on collected historical interaction data [2,3]. However, the collected data cannot include all ratings from users to items, and the ratings are missing not at random (MNAR) due to users' self-selection behavior, i.e., users can choose the item to rate freely, which is also known as selection bias problem [4,5]. The MNAR problem indicates that the collected dataset is not representative of the target population of interest (all user-item pairs), and the training distribution differs from the target test distribution. Ignoring such distributional shift will inevitably lead to sub-optimal recommendation performance [2,3,6]. To address the MNAR problem, one line of previous research proposed to use error imputation-based (EIB) methods, which first impute the missing ratings and then train the prediction model based on both observed and imputed ratings [6,7]. Additionally, another category of methods leverages propensity scores, which computes the probability of an event being observed, to reweight the observed ratings and align the distribution of observed data with the target population [3,8]. Furthermore, Doubly Robust (DR) method combines the error imputation

and the inverse propensity re-weighting to achieve double robustness, which means the DR estimator achieves unbiasedness if either the imputed errors or the learned propensities are correct [2,9,10]. Furthermore, the Multiple Robust (MR) method is proposed to mitigate inaccuracies in single-model propensity scores or error imputations found in DR method [11]. By considering multiple candidate propensity and imputation models, MR estimator achieves unbiasedness if any of the propensity models, imputation models, or a linear combination of these models accurately estimate the true propensities or prediction errors.

However, we argue that the imputed errors and estimated propensity scores are miscalibrated in the existing MR method, which cannot reflect the ground-truth likelihood of the correctness of the true error or true propensity. For instance, if we have 100 user-item pairs with estimated propensity scores equal to 0.2, there should be exactly 20 ratings being observed and 80 ratings being unobserved. Although previous study has proposed to adopt calibration experts to calibrate the single propensity model and imputation model in DR estimator [12], this approach cannot be directly extended to the MR estimator, as calibrating each model individually is expensive and unreasonable due to the unbiasedness condition of MR in terms of linear combinations is not considered. Furthermore, the calibration metric previously used in [12] is non-differentiable and cannot be directly optimized.

To fill this gap, we propose the calibrated multiple robust learning (Cali-MR) method to calibrate the linear combinations of multiple imputation models and propensity models using bi-level optimization, which aims to learn an ensemble model that simultaneously possesses strong prediction performance and calibration ability. In this bi-level optimization, we adopt differentiable expected calibration errors to quantify the calibration ability that allows it to be directly optimized. The calibrated linear combination of propensity and imputation models is then used to train the prediction model based on a joint learning algorithm. The contributions of this paper are summarized as follows.

- We propose a novel MR calibration method using bi-level optimization via calibrating the ensemble imputation and propensity models and address the non-differentiable issue by adopting differentiable expected calibration errors.
- We further propose a bi-level calibrated multiple robust learning algorithm to update the calibrated imputation models and the prediction model. To the best of our knowledge, this is the first work to perform calibration for the MR estimator.
- We conduct extensive experiments on three real-world datasets, showing the effectiveness of our method compared to the state-of-the-art debiasing methods.

2. Related Works

2.1. Debiased Recommendation

The rating missing not at random (MNAR) problem results in the distribution of the observed population being different from the target population, hindering the prediction model from learning users' true preferences [4,5,13–15]. There are many methods proposed to address this issue [16–21].

Specifically, methods including the Inverse Propensity Scoring (IPS), EIB, and DR were proposed to mitigate the MNAR problem in RS [6,22]. EIB methods might produce out-of-bound predictions; while the IPS method may suffer from large variance with small propensities [23]. DR methods combine the advantages of both the EIB and IPS methods, guaranteeing unbiasedness if either the error imputation model or propensity model is correctly specified.

There have been quantities of variants of DR methods to improve the debiasing performance, such as Multi-DR [24], BRD-DR [25], SDR [26], TDR [27], CDR [28], N-DR [29], DT-DR [30], UIDR [31], and OME-DR [32]. Besides, Multiple robust (MR) [11] combines

multiple imputation models and propensity models, and is unbiased when there exists a linear combination of them is correct. In addition, Liu et al. [33] use an information bottleneck-based method and Yang et al. [34] and Wang et al. [35] use adversarial learning for debiasing. However, these methods fail to consider model calibration properties.

To mitigate this issue, DCE-DR [12] is proposed to calibrate the propensity and imputation model in the DR method. However, calibrating each imputation model and propensity model in MR is expensive and unreasonable, due to the unbiasedness condition of MR based on linear combinations is not taken into account. In this paper, we propose the Cali-MR method to calibrate multiple propensity and imputation models to further enhance the debiasing performance and reliability of MR method.

2.2. Model Calibration

Calibration means that the probability associated with the predicted class label should reflect its ground truth correctness likelihood [36,37], which plays an important role in building reliable, robust AI systems, especially in safety-critical fields such as medical diagnosis [38,39], self-driving [40,41], and financial decision making [42,43]. Early research demonstrated good calibration performance of simple neural networks on binary classification tasks [44]. However, with the rapid development of deep learning techniques, recent deep and complex neural networks are no longer well-calibrated [36].

Calibration methods can be divided into the following four categories [45]: post-hoc calibration, regularization methods, uncertainty estimation, and hybrid calibration methods. Post-hoc calibration methods aim to calibrate a model after training, including non-parametric calibration histogram binning [46], isotonic regression [47], and parametric methods such as Platt scaling [48]. Regularization methods adopt penalty terms such as the L_2 regularization [36], entropy regularization [49], difference between confidence and accuracy [50], and calibration errors [51,52] to ensure the calibration property. Uncertainty Estimation aims to alleviate model miscalibration by injecting randomness using Bayesian neural networks [53], model ensemble [54], Monte Carlo dropout [55], and Gumbel-softmax [56] based approaches. Hybrid calibration methods combine two or more methods to achieve calibration. For example, Zhang et al. [57] combines ensemble and temperature scaling and Laves et al. [58] adopts monte-carlo dropout with temperature scaling. We conclude the categorization of model calibration methods in Figure 1.

In this paper, we adopt a differentiable expected calibration error as part of the objective to ensure the model calibration. Compared to other techniques for calibration, our approach has the following advantages. First, unlike post-hoc methods that adjust calibration after training [47,48], our method based on calibration error regularization explicitly optimizes the calibration metric during the training process. Secondly, uncertainty estimation methods such as Monte Carlo Dropout typically rely on complex sampling processes [55], resulting in high computational costs, while hybrid calibration methods which combine multiple techniques, are similarly complex and challenging to implement [57,58]. By comparison, our proposed approach is both simple and computationally efficient. Furthermore, unlike other indirect regularization methods, such as L_2 or entropy regularization [36,49], our method directly incorporates differentiable expected calibration error as a loss function, enabling precise optimization of the calibration objective.

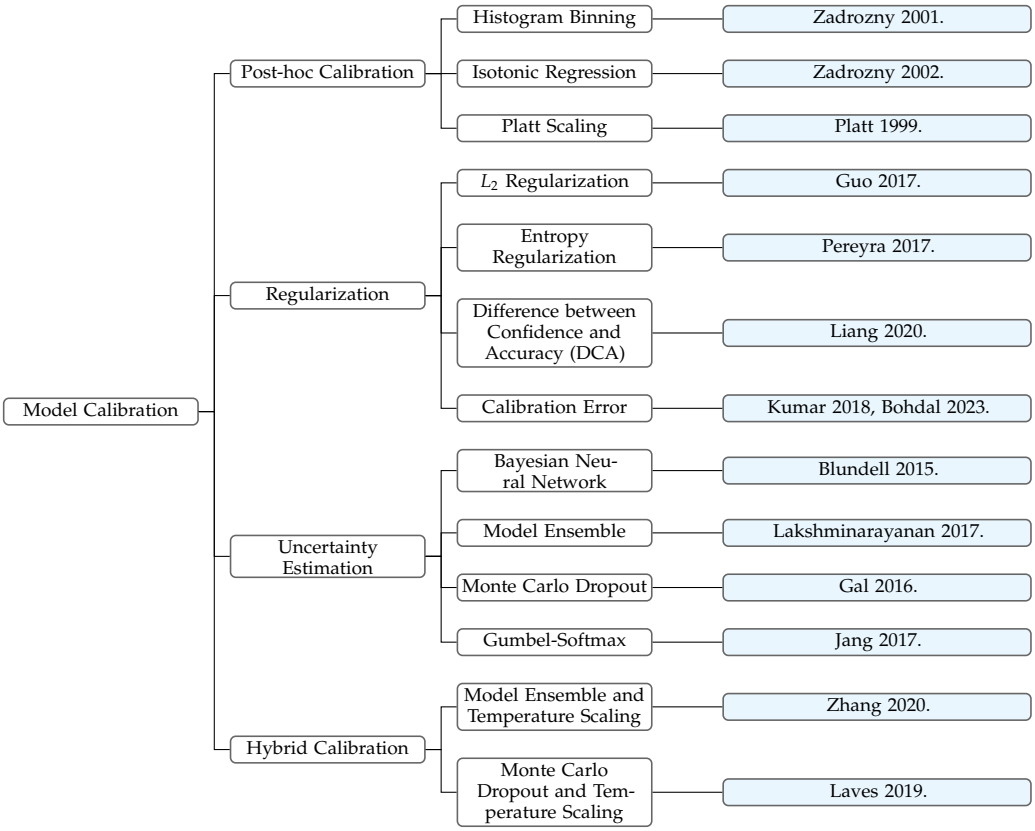


Figure 1. Categorization of model calibration methods: Post-hoc Calibration [46–48], Regularization [36,49–52], Uncertainty Estimation [53–56], and Hybrid Calibration [57,58].

3. Preliminary

3.1. Debiased Recommendation

Let $\mathcal{U} = \{u_1, \cdots, u_m\}$ be the users set, $\mathcal{I} = \{i_1, \cdots, i_n\}$ be the item set, and $\mathcal{D} = \mathcal{U} \times \mathcal{I}$ be the set of all user-item pairs. The rating matrix is denoted as $\mathbf{R} \in \mathbb{R}^{m \times n}$ with $r_{u,i}$ as element. Let $o_{u,i} \in \{0, 1\}$ be the observation indicator, where $o_{u,i} = 1$ indicates the rating $r_{u,i}$ is observed, otherwise is not. Define $x_{u,i}$ be the observed features. We denote the prediction model as $f_\theta(\cdot)$ parameterized by θ and the predicted ratings as $\hat{r}_{u,i} = f_\theta(x_{u,i})$. The goal is to accurately predict $r_{u,i}$ for all user-item pairs, which can be achieved by minimizing the ideal loss

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

where $\mathcal{L}(\cdot, \cdot)$ is the training loss function such as cross-entropy loss. However, in practice, we cannot obtain the complete rating matrix. We denote the set of user-item pairs with observed ratings as $\mathcal{O} = \{(u, i) \mid o_{u,i} = 1\}$. Thus, the naive method optimizes the average loss over the observed user-item pairs

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

Due to the MNAR problem, $\mathbb{E}[\mathcal{L}_{\text{naive}}(\theta)] \neq \mathcal{L}_{\text{ideal}}(\theta)$. Several methods were proposed to unbiasedly estimate the ideal loss, including the EIB, IPS, DR, and their variants. Because EIB and IPS can be regarded as special cases of DR, we only introduce the DR methods here. The loss function of the vanilla DR method is formulated as

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

where $\hat{p}_{u,i} \triangleq \pi(x; \hat{\alpha})$ is the estimation of propensity score $p_{u,i} = \Pr(o_{u,i} = 1 \mid x_{u,i})$, and $\hat{e}_{u,i} = \mathcal{L}(m(x_{u,i}; \beta), \hat{r}_{u,i})$ is the imputed error, while the imputation model is denoted as $m(x_{u,i}; \hat{\beta})$. In addition, the multiple robust (MR) considers J propensity models $\pi_1(x; \hat{\alpha}_1), \ldots, \pi_J(x; \hat{\alpha}_J)$ and K imputation models $m_1(x; \hat{\beta}_1), \ldots, m_K(x; \hat{\beta}_K)$. Let $\hat{p}^j_{u,i} \triangleq \pi_j(x_{u,i}; \hat{\alpha}_j)$, $\hat{m}^k_{u,i} \triangleq m_k(x; \hat{\beta}_k)$, the loss function of MR is shown below:

$$\mathcal{L}_{\mathrm{MR}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \boldsymbol{u}^T(x_{u,i}) \cdot \hat{\boldsymbol{\eta}}(\theta),$$

where $\boldsymbol{u}(x_{u,i}) = \left(1/\hat{p}^1_{u,i}, \cdots, 1/\hat{p}^J_{u,i}, \hat{m}^1_{u,i}, \cdots, \hat{m}^K_{u,i}\right)^T$ and $\hat{\boldsymbol{\eta}}(\theta)$ is the solution by minimizing

$$\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} o_{u,i} \left\{ e_{u,i} - \boldsymbol{u}^T(x_{u,i}) \cdot \boldsymbol{\eta} \right\}^2.$$

The MR estimator is unbiased when there exists a weight $W = (w_1, w_2, \ldots, w_J, 0, 0, \ldots, 0)$ satisfying $W\boldsymbol{u}(x_{u,i}) = 1/p_{u,i}$ or $V = (0, \ldots, 0, v_1, v_2, \ldots, v_K)$ satisfying $V\boldsymbol{u}(x_{u,i}) = e_{u,i}$ for all user-item pairs.

3.2. Calibration

A model is calibrated if its output reflects the ground-truth likelihood of correctness [37]. For the propensity model $\pi(x; \hat{\alpha})$ and the observation indicator o, a formal definition is shown below:

$$\mathbb{E}[o \mid \pi(x; \hat{\alpha}) = \hat{p}] = \hat{p} \quad \forall \hat{p} \in [0, 1].$$

For instance, if we have 100 samples with estimated propensity scores equal to 0.2, there should be exactly 20 samples being observed. Similarly, the formal definition for the calibrated imputation model $m(x; \hat{\beta})$ is formulated below:

$$\mathbb{E}[e \mid m(x; \hat{\beta}) = \hat{e}] = \hat{e} \quad \forall \hat{e} \in \mathbb{R}.$$

To measure the miscalibration of the model, the Expected Calibration Error (ECE) metric is proposed [59]. For a propensity model $\pi(x; \hat{\alpha})$ and imputation model $m(x; \hat{\beta})$, the ECE is defined as follows:

$$\mathrm{ECE}(\hat{\alpha}) = \mathbb{E}_{\hat{p}}[|\mathbb{E}[o \mid \pi(x; \hat{\alpha}) = \hat{p}] - \hat{p}|],$$
$$\mathrm{ECE}(\hat{\beta}) = \mathbb{E}_{\hat{e}}[|\mathbb{E}[e \mid m(x; \hat{\beta}) = \hat{e}] - \hat{e}|].$$

4. Methodology

4.1. Distinctions from Previous Work

Previous studies have proposed to calibrate the single propensity model and imputation model in DR estimator [12], and they propose to use a binning strategy to estimate the ECE metric empirically, for example for the propensity model:

$$\widehat{\mathrm{ECE}}(\hat{p}) = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \frac{\sum_{(u,i)\in B_m} o_{u,i}}{|B_m|} - \frac{\sum_{(u,i)\in B_m} \hat{p}_{u,i}}{|B_m|} \right|,$$

where B_m is the predefined m-th bin and N is the corresponding number of samples in the bin.

However, how to properly calibrate multiple propensity and imputation models for MR estimators remains unexplored. A naive approach is individually calibrating each propensity and imputation model in the MR estimator. Note that this method is computationally expensive and overlooks the robust property of the MR estimator, that is, the MR estimator achieves unbiased if a linear combination of multiple candidate models is accurate.

Inspired by this, we propose to calibrate the linear combination of multiple models instead of calibrating each model individually. In addition, note that the previously used $\widehat{\text{ECE}}$ involves assigning each sample to a specific hard bin, making it non-differentiable and thus unsuitable for direct incorporation into the training objective. To address this issue, we employ a soft binning strategy to develop the differentiable expected calibration error metric and leverage it to construct a regularization term that constrains the model's calibration error, which can be used for model training. Next, we will introduce the proposed Cali-MR in detail.

4.2. Differentiable Expected Calibration Error

To address the non-differentiable problem of $\widehat{\text{ECE}}$, we leverage the soft binning strategy [52], using the following differentiable expected calibration error (DECE) that allows directly optimize calibration quality to mitigate the model miscalibration. For example, the DECE for a propensity model $\pi(x;\hat{\alpha})$ is defined as:

$$\text{DECE}(\hat{\alpha}) = \frac{1}{|\mathcal{D}|}\sum_{m=1}^{M}\left|\sum_{(u,i)\in\mathcal{D}} o_m(x_{u,i};\phi)(o_{u,i}-\hat{p}_{u,i})\right|,$$

where $o_m(x_{u,i};\phi) = \mathbb{P}(x_{u,i}\in B_m|\hat{p}_{u,i})$ denotes the probability that how likely it is that $\hat{p}_{u,i}$ belongs to the m-th bin. In practice, the $o_m(x_{u,i};\phi)$ can be logistic regression or any other model [52].

In our Cali-MR, we adopt DECE as a regularization term and develop a gradient-based learning algorithm model training. Specifically, to calibrate the linear combination of multiple models, we formalize the propensity DECE loss $\mathcal{L}^p_{\text{DECE}}$ as follows:

$$\mathcal{L}^p_{\text{DECE}}(\mathbf{w};\phi_p;\alpha_1,\ldots,\alpha_J) = \frac{1}{|\mathcal{D}|}\sum_{m=1}^{M}\left|\sum_{(u,i)\in\mathcal{D}} o_m(x_{u,i};\phi_p)(o_{u,i}-\sum_{j=1}^{J} w_j\hat{p}^j_{u,i})\right|,$$

where $\mathbf{w} = (w_1,\ldots,w_J)$ is a given set of weight coefficients. This loss measures the calibration error of the combination model, and minimizing this loss improves the calibration ability of the current propensity model combination under the current combination coefficients. We use a one-layer neural network with softmax activation function to model the propensity soft binning model $o_m(x_{u,i};\phi_p)$ with parameters ϕ_p, where m is a predefined hyper-parameter.

Similarly, the imputation DECE loss $\mathcal{L}^e_{\text{DECE}}$ under weight coefficients $\mathbf{v} = (v_1,\ldots,v_K)$ is formalized below:

$$\mathcal{L}^e_{\text{DECE}}(\mathbf{v};\phi_m;\beta_1,\ldots,\beta_K) = \frac{1}{|\mathcal{D}|}\sum_{m=1}^{M}\left|\sum_{(u,i)\in\mathcal{D}} o_m(x_{u,i};\phi_m)\left(\frac{o_{u,i}e_{u,i}}{\sum_{j=1}^{J} w_j\hat{p}^j_{u,i}}-\sum_{k=1}^{K} v_k\hat{e}^k_{u,i}\right)\right|,$$

where we similarly model the imputation soft binning model $o_m(x_{u,i};\phi_m)$ with parameters ϕ_m with a one-layer neural network with softmax activation function. Note that the $e_{u,i}$ is

missing for user-item pairs with $o_{u,i} = 0$, we reweight the observed $e_{u,i}$ using the inverse of the linear combination of the multiple propensity models.

Based on the DECE loss $\mathcal{L}^p_{\text{DECE}}$ or $\mathcal{L}^e_{\text{DECE}}$, we can measure the calibration quality of given multiple models and weight coefficients, and further improve the calibration ability of the current combined model using a gradient-based algorithm by minimizing such differentiable loss.

4.3. Calibrated Multiple Robust Learning

Note that in calculating the DECE loss $\mathcal{L}^p_{\text{DECE}}$ or $\mathcal{L}^e_{\text{DECE}}$, the coefficients $\mathbf{w}$ or $\mathbf{v}$ need to be explicitly specified. However, these coefficients are unknown during the training process of the existing MR method. Therefore, we propose using a gradient-based bi-level optimization method to solve for the optimal coefficients, parameters of the soft binning model, and multiple imputation and propensity models. In addition, we alternatively update the prediction model and the calibrated imputation model based on a joint learning algorithm.

For propensity models, the optimization objective can be formalized as follows:

$$\begin{aligned}&(\alpha_1^*, \ldots, \alpha_J^*) = \arg\min_{\alpha_1,\ldots,\alpha_J} \frac{1}{J}\sum_{j=1}^{J} \mathcal{L}_{p_j}(\alpha_j) + \lambda \mathcal{L}^p_{\text{DECE}}(\mathbf{w}^*) \\ &\text{s.t.} \quad \mathbf{w}^*(\alpha_1, \ldots, \alpha_J) = \arg\min_{\mathbf{w}^*} \mathcal{L}_p(\mathbf{w}^*(\alpha_1, \ldots, \alpha_J)),\end{aligned} \tag{1}$$

where $\mathcal{L}^p_{\text{DECE}}(\mathbf{w})$ loss is the calibration constraints defined in Section 4.2, and $\mathcal{L}_{p_j}(\alpha_j)$ is the training loss for a single propensity model $\pi_J(x; \hat{\alpha}_J)$ ensuring the accuracy of each independent propensity model, which is shown below:

$$\mathcal{L}_{p_j}(\alpha_j) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} [-o_{u,i} \cdot \log p^j_{u,i} - (1 - o_{u,i}) \cdot \log(1 - p^j_{u,i})].$$

$\mathcal{L}_p(\mathbf{w}(\alpha_1, \ldots, \alpha_J))$ is the loss for the combination coefficients $\mathbf{w}$, aiming to learn a set of coefficients such that the linear combination $\sum_{j=1}^{J} w_j \hat{p}^j_{u,i}$ can accurately predict observation indicator $o_{u,i}$, which is shown below:

$$\mathcal{L}_p(\mathbf{w}(\alpha_1, \ldots, \alpha_J)) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} [-o_{u,i} \cdot \log(\sum_{j=1}^{J} w_j \hat{p}^j_{u,i}) - (1 - o_{u,i}) \cdot \log(1 - \sum_{j=1}^{J} w_j \hat{p}^j_{u,i})].$$

In this gradient-based bi-level optimization, we aim to train the propensity models such that each model performs well and their linear combination is well calibrated, where the coefficients also ensure the strong prediction performance of the combined model. For practical implementation, we first assumed update coefficients $\mathbf{w}$ through optimizing $\mathcal{L}_p(\mathbf{w})$. Using these coefficients, we calculate the DECE loss $\mathcal{L}^p_{\text{DECE}}(\mathbf{w})$ and combine it with the base prediction loss of each propensity model denoted as $\mathcal{L}_{p_j}(\alpha_j)$ to form the final loss. This final loss is then used to update the propensity models $\alpha_1, \ldots, \alpha_J$. After that, the loss $\mathcal{L}_p(\mathbf{w})$ and $\mathcal{L}^p_{\text{DECE}}(\mathbf{w})$ are used to update the combination coefficients $\mathbf{w}$ and the soft binning model ϕ_p sequentially.

Multiple Imputation Calibration

With the calibrated propensity model $\sum_{j=1}^{J} w_j \hat{p}^j_{u,i}$ obtained from Equation (1) by the bi-level optimization, we can further calibrate the linear combination of multiple imputation models:

$$(\beta_1^*, \ldots, \beta_K^*) = \arg\min_{\beta_1,\ldots,\beta_K} \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{e_k}(\beta_k) + \lambda \mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v}^*)$$
$$\text{s.t.} \quad \mathbf{v}^*(\beta_1, \ldots, \beta_K) = \arg\min_{\mathbf{v}^*} \mathcal{L}_e(v^*(\beta_1, \ldots, \beta_K)), \tag{2}$$

where $\mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v})$ is the calibration constraints shown in Section 4.2. The naive training loss for each imputation model $m_K(x; \hat{\beta}_K)$ is expressed as

$$\mathcal{L}_{e_k}(\beta_k; \theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(e_{u,i} - \hat{e}^k_{u,i})^2}{\sum_{j=1}^{J} w_j \hat{p}^j_{u,i}},$$

and the loss for the combination coefficients $\mathbf{v}$ is

$$\mathcal{L}_e(\mathbf{v}(\beta_1, \ldots, \beta_K)) = \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(e_{u,i} - \sum_{k=1}^{K} v_k \hat{e}^k_{u,i})^2}{|\mathcal{D}| \sum_{j=1}^{J} w_j \hat{p}^j_{u,i}},$$

which aims to learn a set of weight coefficients such that the linear combination of the imputation models $\sum_{k=1}^{K} v_k \hat{e}^k_{u,i}$ can unbiasedly estimate the prediction error $e_{u,i}$.

Similar to updating the propensity models, we first use $\mathcal{L}_e(\mathbf{v})$ to assume update coefficients $\mathbf{v}'$. Based on that, we compute the DECE loss $\mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v}')$, combine it with the base training loss of each imputation model $\mathcal{L}_{e_k}(\beta_k; \theta)$, and use the combined loss to update the imputation models $\beta_1, \ldots, \beta_K$. Then we adopt loss $\mathcal{L}_e(\mathbf{v})$ and $\mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v})$ to update the combination coefficients $\mathbf{v}$ and the soft binning model ϕ_e sequentially. After obtaining the updated imputation models and coefficients, we jointly train the prediction model using the standard multiple robust learning algorithm. Specifically, we use ridge regression to calculate $\hat{\eta}$ in the MR estimator and update the prediction model based on the MR loss $\mathcal{L}_{\mathrm{MR}}$ using different samples. We summarize the above gradient-based bi-level learning algorithm in Algorithm 1.

Algorithm 1: Gradient-Based Bi-level Calibrated Multiple Robust Learning

Input: observed ratings $\mathbf{R}^o$, calibrated propensity model $\sum_{j=1}^{J} w_j \hat{p}^j_{u,i}$, and stabilization parameter λ

while *stopping criteria is not satisfied* **do**
  for *number of steps for training imputation model* **do**
    Sample a batch of user-item pairs $\{(u_l, i_l)\}_{l=1}^{L}$ from observed population $\mathcal{O}$;
    Assumed update coefficients $\mathbf{v}'(\beta_1, \ldots, \beta_K) \leftarrow \mathbf{v} - \alpha_{\mathbf{v}} \nabla_{\mathbf{v}} \mathcal{L}_e(\mathbf{v} \mid \beta_1, \ldots, \beta_K)$
    Update imputation models $(\beta_1, \ldots, \beta_K) \leftarrow \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{e_k}(\beta_k) + \lambda \mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v}'(\beta_1, \ldots, \beta_K))$
    Update coefficients $\mathbf{v} \leftarrow \mathcal{L}_e(\mathbf{v}(\beta_1, \ldots, \beta_K))$
    Update imputation soft binning model $\phi_m \leftarrow \mathcal{L}^e_{\mathrm{DECE}}(\mathbf{v}; \phi_m; \beta_1, \ldots, \beta_K)$
  end
  for *number of steps for training the prediction model* **do**
    Sample a batch of user-item pairs $\mathcal{D}'$ from $\mathcal{D}$;
    Obtain the rated samples in $\mathcal{D}'$ as $\{(u_m, i_m)\}_{m=1}^{M} = \mathcal{O}' \subseteq \mathcal{O}$;
    Update $\eta \leftarrow [\sum_{(u,i)\in\mathcal{O}'} \boldsymbol{u}(x_{u,i}; \alpha, \beta) \cdot \boldsymbol{u}^T(x_{u,i}; \alpha, \beta) + \lambda I]^{-1} [\sum_{(u,i)\in\mathcal{O}'} \boldsymbol{u}(x_{u,i}; \alpha, \beta) \cdot e_{u,i}]$;
    Sample a batch of user-item pairs $\{(u_n, i_n)\}_{n=1}^{N}$ from $\mathcal{D} \setminus \mathcal{D}'$;
    Update prediction model $\theta \leftarrow \theta - \nabla_\theta \mathcal{L}_{\mathrm{MR}}(\theta; \alpha, \beta)$
  end
end

5. Experiments

5.1. Datasets

To evaluate the debiasing performance, we conduct experiments on three benchmark datasets **Coat** (https://www.cs.cornell.edu/~schnabts/mnar/, accessed on 15 January 2025) and **Yahoo! R3** (https://webscope.sandbox.yahoo.com), and **KuaiRec** (https://github.com/chongminggao/KuaiRec, accessed on 15 January 2025) [60], which are widely used in debiased RS with both missing not at random (MNAR) and missing at random (MAR) data. **Coat** dataset consists of 6960 MNAR training samples and 4640 MAR test samples derived from 290 users rating on 300 items. The **Yahoo! R3** dataset includes 311,704 MNAR training samples and 54,000 MAR test samples derived from 15,400 users rating on 1000 items. Both datasets are five-scale, and following previous works [61–63], we binarize the ratings greater than three to 1, and others to 0. The **KuaiRec** dataset is collected from a video-sharing platform and contains 4,676,570 video watching ratios derived from 1411 users evaluating 3327 videos. We binarize the continuous ratios greater than two to 1, otherwise to 0.

5.2. Baselines

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

- **Naive** method [64] naively optimizes the average loss over the observed user-item pairs.
- **IPS** method [3] reweights the observed ratings with the inverse propensity scores.
- **SNIPS** method [65] reweights the observed ratings with self-normalized propensity scores to further reduce the variance.
- **ASIPS** method [22] generates reliable pseudo-ratings to mitigate propensity estimation bias and high variance problem.
- **DR** method [10] combines error imputation and inverse propensity reweighting to construct a doubly robust estimator, where imputed errors are typically set based on label prior knowledge, such as the mean value of the labels.
- **DR-JL** method [2] further proposes modeling error imputation with neural networks and jointly learns the prediction model and imputation model.
- **MRDR** method [66] enhances the **DR-JL** method by explicitly controlling the variance of the DR estimator through imputation model learning.
- **DR-BIAS** method [67] enhances the **DR-JL** method by further reducing the bias of the DR estimator through imputation model learning.
- **DR-MSE** method [67] further combines **MRDR** method and **DR-BIAS** method to achieve bias-variance trade-off and control the generalization error.
- **MR** method [11] adopts multiple candidate propensity and imputation models to mitigate inaccuracies in single-model propensity scores or error imputation in DR methods.
- **TDR** and **TDR-JL** methods [27] correct the imputed errors with targeted learning to reduce the bias and variance simultaneously for existing DR approaches
- **StableDR** method [26] constructs a stabilized DR estimator that has a weaker dependence on extrapolation and is robust to small propensities by learning constrained propensity scores.
- **IPS-V2** and **DR-V2** methods [68] learn the propensity model which can balance some manually selected functions such as the first and second moments of the features.
- **KBIPS** and **KBDR** methods [32] further propose to conduct causal balancing in the reproducing kernel Hilbert space (RKHS) and randomly select some kernel functions to balance for propensity model learning.
- **AKBIPS** and **AKBDR** methods [32] adaptively select the kernel functions which contribute the most to reducing the estimation bias to balance for propensity model learning.

- **DCE-DR** and **DCE-TDR** method [12] propose to calibrate the single propensity model and single imputation model in DR and TDR estimators through Mixture-of-Experts technique.

5.3. Experiment Protocols and Details

We evaluate the prediction performance using three widely adopted evaluation metrics: AUC (Area Under the ROC Curve), NDCG@T (Normalized Discounted Cumulative Gain), and F1@T.

- **AUC** [69] is a performance metric for classifiers that measures the probability of a randomly chosen positive example being ranked higher than a randomly chosen negative one. A higher AUC score reflects better ranking performance in differentiating positive instances from negative ones.
- **NDCG@T** [70] evaluates ranking performance by comparing the Discounted Cumulative Gain (DCG) of the top-T results to the Ideal DCG (IDCG), producing a normalized score between 0 and 1. A higher NDCG@T implies that more relevant items are ranked towards the top.

Let r_i be the relevance of the item at rank i. We first compute the DCG at rank T as well as the IDCG@T by placing the most relevant items in the optimal (ideal) order:

$$\text{DCG@}T = \sum_{i=1}^{T} \frac{2^{r_i} - 1}{\log_2(i+1)} \quad , \quad \text{IDCG@}T = \sum_{i=1}^{T} \frac{2^{r_i^*} - 1}{\log_2(i+1)}.$$

where r_i is the relevance of the i-th item at rank i, and r_i^* denotes the relevance of the i-th item in the ideal ranking. NDCG@T is then defined as:

$$\text{NDCG@}T = \frac{\text{DCG@}T}{\text{IDCG@}T}.$$

- **F1@T** [71] is the harmonic mean of precision and recall computed over the top-T predictions returned by a model. A higher F1@T indicates a better trade-off between precision and recall in the top-T results.

We set $T = 5$ on **Coat** and **Yahoo! R3**, and $T = 20$ on **KuaiRec**. In addition, we tune learning rate in $\{0.01, 0.05\}$ and weight decay in $\{1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 1 \times 10^{-3}, 5 \times 10^{-3}\}$.We use the same hyperparameter search space and follow the results in Li et al. [32].

5.4. Performance Analysis

The experimental results are shown in Table 1, and we find that all the debiasing methods including both IPS-based and DR-based baselines outperform the Naive method, which demonstrates the importance of debiasing. Besides, among all baselines, AKBDR method introduces balancing kernel functions for propensity model training. This method adaptively identifies the most critical kernel functions to balance by fitting prediction errors, and well-trained propensity model effectively eliminates selection bias, thereby demonstrating strong performance on the Coat dataset. On the other hand, TDR introduces a targeted learning technique, which leverages propensity to enhance the error imputation, achieving reductions in both bias and variance. DCE-TDR further improves upon TDR by calibrating propensity model, utilizing the calibrated propensity for targeted learning, and achieving further performance gains. This calibration strategy enables DCE-TDR to deliver competitive results on the Yahoo! R3 and KuaiRec datasets. Furthermore, the proposed Cali-MR method exhibits superior overall performance and significantly outperforms existing methods on **Yahoo! R3** and **KuaiRec** datasets. This shows that calibrating multiple

propensity and imputation models in a Multiple Robust estimator with weight coefficient learned in bi-level optimization further enhances the debiasing performance.

Table 1. Performance on AUC, NDCG@T, and F1@T 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.

	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	$\mathbf{0.745}_{\pm 0.004}$	$0.645_{\pm 0.008}$	$\underline{0.493}_{\pm 0.007}$	$0.692_{\pm 0.002}$	$0.661_{\pm 0.002}$	$0.328_{\pm 0.002}$	$0.782_{\pm 0.003}$	$0.498_{\pm 0.008}$	$0.147_{\pm 0.003}$
DCE-DR	$0.736_{\pm 0.006}$	$0.648_{\pm 0.007}$	$0.489_{\pm 0.005}$	$0.698_{\pm 0.002}$	$0.670_{\pm 0.002}$	$\underline{0.333}_{\pm 0.003}$	$0.795_{\pm 0.004}$	$0.512_{\pm 0.005}$	$0.153_{\pm 0.002}$
DCE-TDR	$0.740_{\pm 0.004}$	$\underline{0.651}_{\pm 0.006}$	$0.489_{\pm 0.007}$	$\underline{0.701}_{\pm 0.002}$	$\underline{0.672}_{\pm 0.002}$	$0.331_{\pm 0.002}$	$\mathbf{0.798}_{\pm 0.005}$	$\underline{0.514}_{\pm 0.006}$	$\underline{0.155}_{\pm 0.002}$
Cali-MR	$\underline{0.741}_{\pm 0.002}$	$\mathbf{0.658}^{*}_{\pm 0.004}$	$\mathbf{0.495}_{\pm 0.004}$	$\mathbf{0.703}^{*}_{\pm 0.002}$	$\mathbf{0.678}^{*}_{\pm 0.002}$	$\mathbf{0.338}^{*}_{\pm 0.004}$	$\mathbf{0.798}_{\pm 0.003}$	$\mathbf{0.521}^{*}_{\pm 0.005}$	$\mathbf{0.158}^{*}_{\pm 0.002}$

5.5. In-Depth Analysis

We explore the difference between calibrating the ensemble model and calibrating individual models separately, with the number of propensity models J and imputation models K taking values from $\{1, 3, 5\}$. The results on the **KuaiRec** and **Yahoo! R3** datasets are shown in Figures 2 and 3, where darker colors in the heatmap represent higher values of the corresponding evaluation metrics, indicating stronger model performance. The proposed method for calibrating the linear combination of multiple candidate models is referred to as 'joint calibration', while the approach of calibrating each individual candidate propensity model and then randomly selecting one calibrated propensity model to calibrate the candidate imputation models is referred to as 'individual calibration'. We find that directly calibrating the ensemble model outperforms calibrating a single model individually, especially on the **Yahoo! R3** dataset. This demonstrates that the proposed Cali-MR method which takes into account the conditions for achieving unbiasedness in the MR estimator achieves better calibration and debiasing performance. On the other hand, examining the impact of the number of candidate models J and K on prediction performance, we find that when $J = 1$ or $K = 1$, i.e., when using only one propensity model or one imputation model, the prediction performance is poor, validating the effectiveness of incorporating more candidate models in the MR estimator. Additionally, the optimal (J, K) combinations for Cali-MR on **Yahoo! R3** and **KuaiRec** datasets are $(5, 5)$ and $(3, 3)$, respectively.

We conduct an ablation study of the proposed Cali-MR on three benchmark datasets, with the experimental results shown in Table 2. It can be observed that the prediction performance declines if either the propensity model or the imputation model is not calibrated, demonstrating the necessity of model calibration. Furthermore, the performance drop is more significant when removing the calibration of the propensity model compared to the imputation model, indicating that propensity models are more miscalibrated in the existing MR estimator. This highlights that the current propensity model learning approach, based on simple fitting to the observed variable indicator o with naive binary cross-entropy loss,

fails to produce well-calibrated models, emphasizing the need to incorporate calibration loss on top of it.

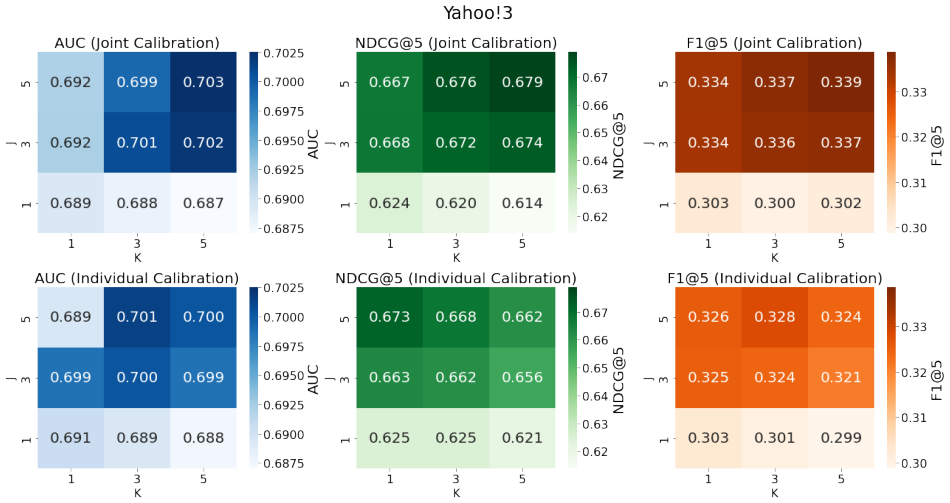


Figure 2. Comparison of joint calibration and individual calibration on the **Yahoo! R3** dataset, with different numbers of candidate propensity and imputation models.

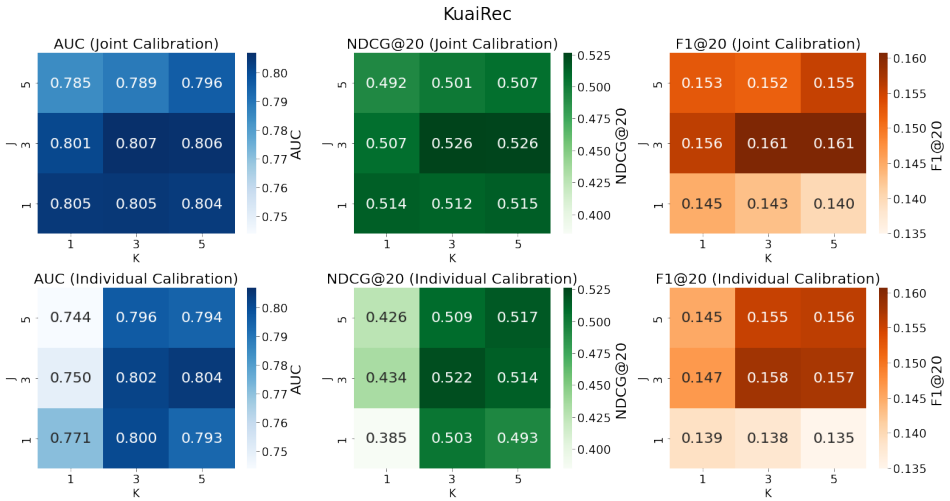


Figure 3. Comparison of joint calibration and individual calibration on the **KuaiRec** dataset, with different numbers of candidate propensity and imputation models.

Table 2. Ablation study of Cali-MR on **Coat**, **Yahoo! R3** and **KuaiRec** datasets.

Method	Coat			Yahoo! R3			KuaiRec		
	AUC	NDCG@5	F1@5	AUC	NDCG@5	F1@5	AUC	NDCG@20	F1@20
Cali-MR	0.741	0.658	0.495	0.703	0.678	0.338	0.798	0.521	0.158
Cali-MR w/o prop	0.730	0.640	0.486	0.700	0.674	0.336	0.787	0.483	0.151
Cali-MR w/o imp	0.736	0.639	0.483	0.703	0.674	0.336	0.793	0.509	0.155
Cali-MR w/o imp & prop	0.727	0.635	0.477	0.698	0.667	0.331	0.783	0.482	0.148

Figures 4 and 5 investigate the impact of varying model calibration hyper-parameter λ in multiple propensity and imputation calibration on prediction performance on **Coat** and **KuaiRec** datasets. We record the results for different values of the hyper-parameter λ and fit a curve to show its variation trend, with a horizontal line representing the baseline method without calibration. We observe that when the hyper-parameter λ falls within an appropriate range, such as $[0.1, 100]$, the proposed method stably outperforms the baseline method, with the best performance achieved around moderate values, such as $\lambda = 10$. Notably, on the **Coat** dataset, when the value of hyperparameter λ is set too

large (e.g., 1000), the model fails to achieve good prediction performance, which indicates the propensity and imputation models overly emphasize calibration at the expense of prediction accuracy, leading to poor performance. This highlights the need to balance the prediction and calibration capabilities of propensity and imputation models during training.

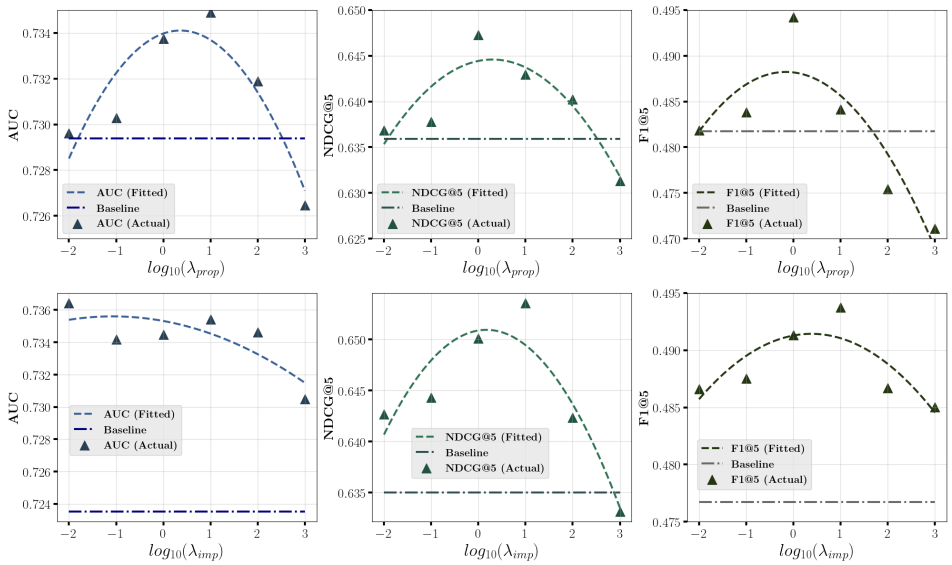


Figure 4. Impact of model calibration hyper-parameter λ_{prop} in multiple propensity calibration and λ_{imp} in multiple imputation calibration on **Coat** dataset.

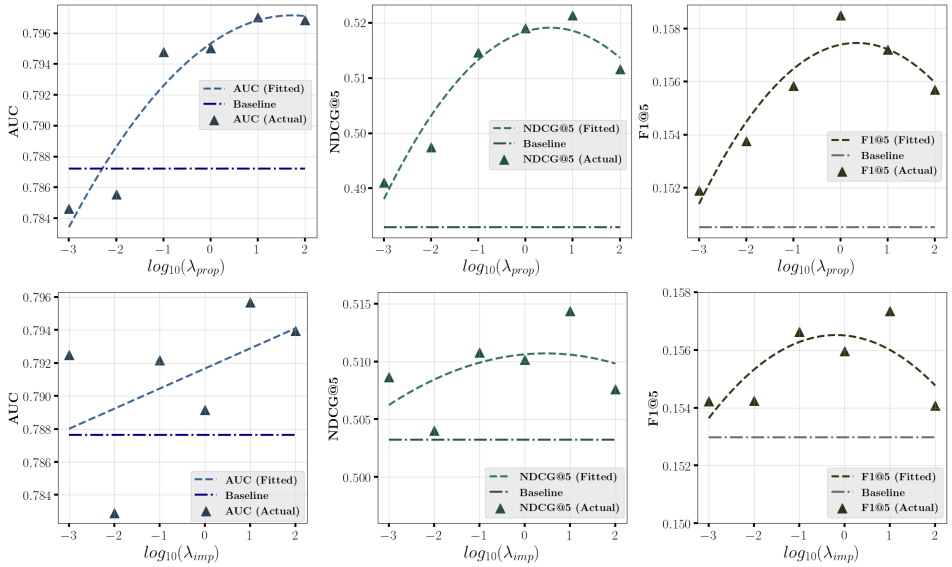


Figure 5. Impact of model calibration hyper-parameter λ_{prop} in multiple propensity calibration and λ_{imp} in multiple imputation calibration on **KuaiRec** dataset.

Figures 6 and 7 explore the effect of varying number of bins M in soft binning strategy on prediction performance on **Coat** and **KuaiRec** datasets. The scatter points represent the results for different values of M, and a fitted curve illustrates the trend of predictive performance as M changes. The results in the figure indicate that the optimal performance is achieved at moderate values of M, such as around 30 for **Coat** dataset and 15 for **KuaiRec** dataset. However, performance declines when M is either too large or too small. A possible reason is that when M is too small, the probability of samples with different prediction values ($\hat{p}$ or $\hat{e}$) being assigned to the same bin increases. Conversely, when M is too large, the probability of samples with similar prediction values ($\hat{p}$ or $\hat{e}$) being assigned to the same

bin decreases. Both situations can lead to inaccurate DECE estimation, thereby reducing the calibration quality of the model.

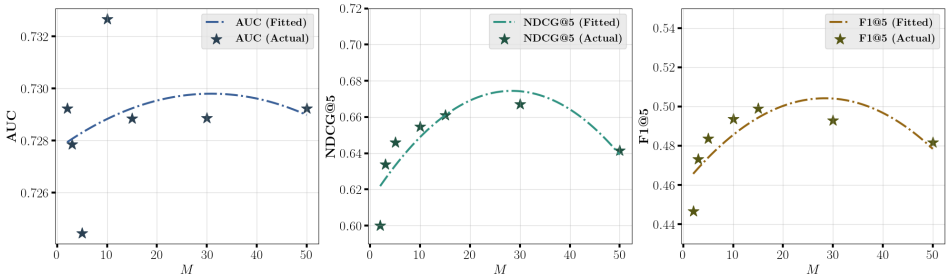


Figure 6. Effect of varying M in soft binning strategy on prediction performance on **Coat** dataset.

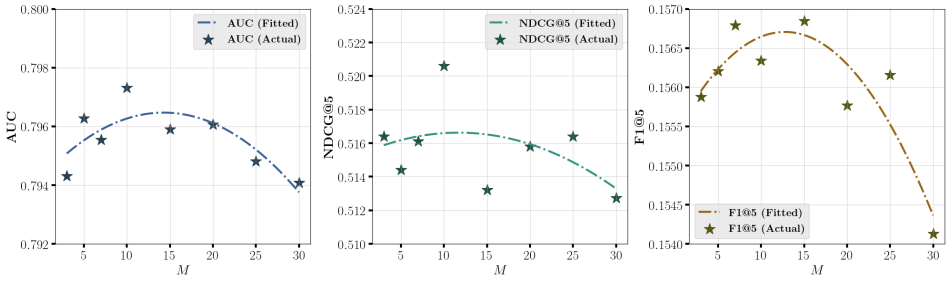


Figure 7. Effect of varying M in soft binning strategy on prediction performance on **KuaiRec** dataset.

The results in Table 3 demonstrate that proposed Cali-MR achieves comparable ECE performance to the current SOTA model, DCE-TDR. However, when evaluating the relative ECE reduction compared to the base models, we observe that DCE-TDR achieves a smaller relative improvement over TDR because TDR itself, leveraging targeted learning, already exhibits a relatively calibrated imputation model. Conversely, the original MR model shows a high ECE, indicating poor calibration, while the proposed Cali-MR significantly reduces the ECE, thereby enhancing calibration quality, which validates the effectiveness of the proposed approach.

Table 3. Comparison of Expected Calibration Error (ECE) on **Coat**, **Yahoo! R3** and **KuaiRec** datasets, where RD refers to the relative decrease value over the corresponding base model.

	Coat		**Yahoo! R3**		**KuaiRec**	
Method	**ECE**	**RD**	**ECE**	**RD**	**ECE**	**RD**
DR-JL	0.1626	-	0.0589	-	0.0999	-
DCE-DR	0.1428	0.0198	0.0554	0.0035	0.0451	0.0548
TDR-JL	0.1468	-	0.0481	-	0.0499	-
DCE-TDR	0.1270	0.0198	0.0476	0.0005	0.0488	0.0011
MR	0.1512	-	0.1594	-	0.2367	-
Cali-MR	0.1239	0.0273	0.0349	0.1245	0.0519	0.1848

We compare the training time and parameter size of different methods on Table 4. The experimental results show that the Naive method, which adopts a single model to fit observed ratings, has a parameter size of 1×. The IPS method introduces an additional propensity model, resulting in a parameter size of 2×, while the DR method further introduces an imputation model, with a parameter size of 3×. In this experiment, the MR method adopts two candidate propensity and imputation models, leading to a parameter size of 5×. As an improvement, our proposed Cali-MR (J = 2, K = 2) adds only a soft binning module, which increases the parameter count marginally compared to the original

MR method, thus still maintaining a parameter size of 5×. In other words, the proposed method achieves significant performance improvements over the existing MR method, while maintaining the same parameter size. Additionally, by comparing the training times for MR and Cali-MR (J = 2, K = 2), we observe that model calibration does not add substantial training time, for example, with times of 127.8 and 144.56 on **Yahoo! R3** dataset, respectively. Furthermore, increasing the number of candidate models (J and K) for Cali-MR does not lead to a significant increase in training time. This is because as the number of candidate models increases, the ensemble model tends to overfit, and due to early stopping strategy, the number of epochs required for convergence varies with the number of candidate models. In summary, compared to the MR method, the proposed Cali-MR method improves model calibration and prediction accuracy with almost the same parameter size and acceptable training time.

Table 4. Comparison of training time (seconds) and parameter size (Naive method denoted as 1×) on **Coat**, **Yahoo! R3** and **KuaiRec** datasets.

Method	Coat		Yahoo! R3		KuaiRec	
	Time	Params	Time	Params	Times	Params
Naive	4.04	1×	26.07	1×	11.67	1×
IPS	6.54	2×	32.84	2×	15.14	2×
DR	17.41	3×	43.18	3×	31.20	3×
DR-JL	18.81	3×	166.21	3×	110.01	3×
TDR-JL	21.13	3×	128.88	3×	101.09	3×
MR (J = 2, K = 2)	13.55	5×	127.8	5×	114.78	5×
Cali-DR (J = 1, K = 1)	21.03	3×	132.98	3×	128.31	3×
Cali-MR (J = 2, K = 2)	23.28	5×	144.56	5×	124.58	5×
Cali-MR (J = 1, K = 5)	21.57	5×	139.80	5×	130.34	5×
Cali-MR (J = 1, K = 5)	21.43	7×	136.77	7×	141.46	7×
Cali-MR (J = 3, K = 3)	19.55	5×	127.80	5×	106.66	7×
Cali-MR (J = 3, K = 3)	25.43	7×	151.97	7×	127.47	7×
Cali-MR (J = 3, K = 5)	23.59	9×	166.42	9×	163.02	9×
Cali-MR (J = 5, K = 3)	23.71	9×	143.76	9×	103.60	9×
Cali-MR (J = 5, K = 3)	29.35	9×	132.74	9×	128.03	9×
Cali-MR (J = 5, K = 5)	25.65	11×	153.79	11×	166.73	11×

6. Conclusions

In this paper, we explore how to properly calibrate multiple propensity and imputation models in a Multiple Robust (MR) estimator. First, we argue that calibrating each candidate model individually is too costly and unreasonable due to the unbiasedness condition of MR in terms of linear combinations is not considered. Based on this, we propose using a gradient-based bi-level optimization method to calibrate the linear combination of multiple candidate models. Specifically, in the bi-level optimization, we first assumed to update the combination coefficients to obtain the best-performing combination coefficients under the current candidate model parameters. Then, based on these coefficients, we update the candidate model parameters to ensure that each model maintains good prediction performance, while also ensuring that the combination model achieves strong calibration ability, using the differentiable expected calibration error metric that allows it to be directly optimized. Experimental results on three real-world datasets demonstrate that the proposed multiple propensity and imputation calibration method further enhances the prediction performance.

Regarding the broader impact, this is the first work to address the calibration of MR models, by introducing the differentiable expected calibration error (DECE) to directly

optimize the objective, rather than relying on post-hoc adjustments in previous work. This provides a novel perspective on model calibration within the debiasing recommendation field. The main limitation of this work lies in the use of bi-level optimization to update parameters of multiple candidate models and the optimal combination coefficients within the MR method. This results in longer training times for the proposed Cali-MR method compared to the original MR method. A potential future work is to develop an alternative optimization algorithm to reduce the computational cost. Furthermore, in future works, we plan to investigate whether other calibration metrics, aside from ECE, might be more suitable for debiasing recommendation scenarios.

Author Contributions: Conceptualization, S.G.; Methodology, S.G.; Validation, C.M.; Investigation, S.G. and C.M.; Writing – original draft, S.G.; Writing – review & editing, C.M.; Visualization, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Our code is available at https://github.com/Yilu114/Cali-MR (accessed on 15 January 2025). Data were obtained from GitHub and are available https://github.com/RyanWangZf/CVIB-Rec with the permission of GitHub (accessed on 15 January 2025).

Conflicts of Interest: Author Shuxia Gong is employed by Mogo Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ricci, F.; Rokach, L.; Shapira, B. Introduction to recommender systems handbook. In *Recommender Systems Handbook*; Springer: Boston, MA, USA, 2010; pp. 1–35.
2. Wang, X.; Zhang, R.; Sun, Y.; Qi, J. Doubly Robust Joint Learning for Recommendation on Data Missing Not at Random. In Proceedings of the ICML, Long Beach, CA, USA, 9–15 June 2019.
3. Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; Joachims, T. Recommendations as Treatments: Debiasing Learning and Evaluation. In Proceedings of the ICML, New York, NY, USA, 19–24 June 2016.
4. Chen, J.; Dong, H.; Wang, X.; Feng, F.; Wang, M.; He, X. Bias and Debias in Recommender System: A Survey and Future Directions. *Acm Trans. Inf. Syst.* **2022**, *41*, 1–39. [CrossRef]
5. Wu, P.; Li, H.; Deng, Y.; Hu, W.; Dai, Q.; Dong, Z.; Sun, J.; Zhang, R.; Zhou, X.H. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In Proceedings of the IJCAI, Vienna, Austria, 23–29 July 2022.
6. Steck, H. Training and testing of recommender systems on data missing not at random. In Proceedings of the KDD, Washington, DC, USA, 25–28 July 2010.
7. Chang, Y.W.; Hsieh, C.J.; Chang, K.W.; Ringgaard, M.; Lin, C.J. Training and testing low-degree polynomial data mappings via linear SVM. *J. Mach. Learn. Res.* **2010**, *11*, 1471–1490.
8. Saito, Y.; Yaginuma, S.; Nishino, Y.; Sakata, H.; Nakata, K. Unbiased recommender learning from missing-not-at-random implicit feedback. In Proceedings of the WSDM, Houston, TX, USA, 3–7 February 2020.
9. Morgan, S.L.; Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2015.
10. Saito, Y. Doubly robust estimator for ranking metrics with post-click conversions. In Proceedings of the RecSys, Virtual Event, 22–26 September 2020.
11. Li, H.; Dai, Q.; Li, Y.; Lyu, Y.; Dong, Z.; Zhou, X.H.; Wu, P. Multiple Robust Learning for Recommendation. In Proceedings of the AAAI, Washington, DC, USA, 7–14 February 2023.
12. Kweon, W.; Yu, H. Doubly Calibrated Estimator for Recommendation on Data Missing Not At Random. In Proceedings of the WWW, Singapore, 13–17 May 2024.
13. Luo, H.; Zhuang, F.; Xie, R.; Zhu, H.; Wang, D.; An, Z.; Xu, Y. A survey on causal inference for recommendation. *Innovation* **2024**, *5*, 100590. [CrossRef]

14. Li, M.; Sui, H. Causal Recommendation via Machine Unlearning with a Few Unbiased Data. In Proceedings of the AAAI Workshop on Artificial Intelligence with Causal Techniques, Philadelphia, PA, USA, 25 February–4 March 2025.
15. Wang, W.; Zhang, Y.; Li, H.; Wu, P.; Feng, F.; He, X. Causal Recommendation: Progresses and Future Directions. In Proceedings of the SIGIR, Taipei, Taiwan, 23–27 July 2023.
16. Saito, Y.; Nomura, M. Towards Resolving Propensity Contradiction in Offline Recommender Learning. In Proceedings of the IJCAI, Vienna, Austria, 23–29 July 2022.
17. Wang, H.; Yang, W.; Yang, L.; Wu, A.; Xu, L.; Ren, J.; Wu, F.; Kuang, K. Estimating Individualized Causal Effect with Confounded Instruments. In Proceedings of the KDD, Washington, DC, USA, 14–18 August 2022.
18. Zou, H.; Wang, H.; Xu, R.; Li, B.; Pei, J.; Jian, Y.J.; Cui, P. Factual Observation Based Heterogeneity Learning for Counterfactual Prediction. In Proceedings of the CCLR, Tübingen, Germany, 11–14 April 2023.
19. Wang, H.; Kuang, K.; Lan, L.; Wang, Z.; Huang, W.; Wu, F.; Yang, W. Out-of-distribution generalization with causal feature separation. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 1758–1772. [CrossRef]
20. Wang, H.; Kuang, K.; Chi, H.; Yang, L.; Geng, M.; Huang, W.; Yang, W. Treatment effect estimation with adjustment feature selection. In Proceedings of the KDD, Long Beach, CA, USA, 6–10 August 2023.
21. Wu, A.; Kuang, K.; Xiong, R.; Li, B.; Wu, F. Stable estimation of heterogeneous treatment effects. In Proceedings of the ICML, Honolulu, HI, USA, 23–29 July 2023.
22. Saito, Y. Asymmetric Tri-training for Debiasing Missing-Not-At-Random Explicit Feedback. In Proceedings of the SIGIR, Xi'an, China, 25–30 July 2020.
23. Wang, H.; Chang, T.W.; Liu, T.; Huang, J.; Chen, Z.; Yu, C.; Li, R.; Chu, W. Escm2: Entire space counterfactual multi-task model for post-click conversion rate estimation. In Proceedings of the SIGIR, Madrid, Spain, 11–15 July 2022.
24. Zhang, W.; Bao, W.; Liu, X.Y.; Yang, K.; Lin, Q.; Wen, H.; Ramezani, R. Large-scale Causal Approaches to Debiasing Post-click Conversion Rate Estimation with Multi-task Learning. In Proceedings of the WWW, Taipei, Taiwan, 20–24 April 2020.
25. Ding, S.; Wu, P.; Feng, F.; He, X.; Wang, Y.; Liao, Y.; Zhang, Y. Addressing Unmeasured Confounder for Recommendation with Sensitivity Analysis. In Proceedings of the KDD, Washington, DC, USA, 14–18 August 2022.
26. Li, H.; Zheng, C.; Wu, P. StableDR: Stabilized Doubly Robust Learning for Recommendation on Data Missing Not at Random. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023.
27. Li, H.; Lyu, Y.; Zheng, C.; Wu, P. TDR-CL: Targeted Doubly Robust Collaborative Learning for Debiased Recommendations. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023.
28. Song, Z.; Chen, J.; Zhou, S.; Shi, Q.; Feng, Y.; Chen, C.; Wang, C. CDR: Conservative Doubly Robust Learning for Debiased Recommendation. In Proceedings of the CIKM, Birmingham, UK, 21–25 October 2023.
29. Li, H.; Zheng, C.; Ding, S.; Feng, F.; He, X.; Geng, Z.; Wu, P. Be Aware of the Neighborhood Effect: Modeling Selection Bias under Interference for Recommendation. In Proceedings of the ICLR, Vienna, Austria, 7–11 May 2024.
30. Zhang, H.; Wang, S.; Li, H.; Zheng, C.; Chen, X.; Liu, L.; Luo, S.; Wu, P. Uncovering the Propensity Identification Problem in Debiased Recommendations. In Proceedings of the ICDE, Utrecht, The Netherlands, 13–17 May 2024.
31. Li, H.; Zheng, C.; Wang, S.; Wu, K.; Wang, E.; Wu, P.; Geng, Z.; Chen, X.; Zhou, X.H. Relaxing the Accurate Imputation Assumption in Doubly Robust Learning for Debiased Collaborative Filtering. In Proceedings of the ICML, Vienna, Austria, 21–27 July 2024.
32. Li, H.; Zheng, C.; Xiao, Y.; Wu, P.; Geng, Z.; Chen, X.; Cui, P. Debiased collaborative filtering with kernel-based causal balancing. In Proceedings of the ICLR, Vienna, Austria, 7–11 May 2024.
33. Liu, D.; Cheng, P.; Zhu, H.; Dong, Z.; He, X.; Pan, W.; Ming, Z. Debiased representation learning in recommendation via information bottleneck. *ACM Trans. Recomm. Syst.* **2023**, *1*, 1–27. [CrossRef]
34. Yang, M.; Dai, Q.; Dong, Z.; Chen, X.; He, X.; Wang, J. Top-n recommendation with counterfactual user preference simulation. In Proceedings of the CIKM, Virtual Event, 1–5 November 2021.
35. Wang, J.; Li, H.; Zhang, C.; Liang, D.; Yu, E.; Ou, W.; Wang, W. Counterclr: Counterfactual contrastive learning with non-random missing data in recommendation. In Proceedings of the ICDM, Shanghai, China, 1–4 December 2023.
36. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the ICML, Sydney, NSW, Australia, 6–11 August 2017.
37. Kull, M.; Silva Filho, T.; Flach, P. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Proceedings of the AISTATS, Fort Lauderdale, FL, USA, 20–22 April 2017.
38. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the KDD, Sydney, NSW, Australia, 10–13 August 2015.
39. Huang, Y.; Li, W.; Macheret, F.; Gabriel, R.A.; Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 621–633. [CrossRef]
40. Bojarski, M. End to end learning for self-driving cars. *arXiv* **2016**, arXiv:1604.07316.
41. Chen, Z.; Huang, X. End-to-end learning for lane keeping of self-driving cars. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 1856–1860.

42. Büchel, P.; Kratochwil, M.; Nagl, M.; Rösch, D. Deep calibration of financial models: Turning theory into practice. *Rev. Deriv. Res.* **2022**, *25*, 109–136. [CrossRef]
43. Biagini, F.; Gonon, L.; Walter, N. Approximation rates for deep calibration of (rough) stochastic volatility models. *SIAM J. Financ. Math.* **2024**, *15*, 734–784. [CrossRef]
44. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the ICML, Bonn, Germany, 7–11 August 2005.
45. Wang, C. Calibration in deep learning: A survey of the state-of-the-art. *arXiv* **2023**, arXiv:2308.01222.
46. Zadrozny, B.; Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In Proceedings of the ICML, Williamstown, MA, USA, 28 June–1 July 2001.
47. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the KDD, Edmonton, AB, Canada, 23–26 July 2002.
48. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **1999**, *10*, 61–74.
49. Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; Hinton, G. Regularizing neural networks by penalizing confident output distributions. *arXiv* **2017**, arXiv:1701.06548.
50. Liang, G.; Zhang, Y.; Wang, X.; Jacobs, N. Improved trainable calibration method for neural networks on medical imaging classification. *arXiv* **2020**, arXiv:2009.04057.
51. Kumar, A.; Sarawagi, S.; Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In Proceedings of the ICML, Stockholm, Sweden, 10–15 July 2018.
52. Bohdal, O.; Yang, Y.; Hospedales, T. Meta-Calibration: Learning of Model Calibration Using Differentiable Expected Calibration Error. *Trans. Mach. Learn. Res.* **2023**. Available online: https://openreview.net/forum?id=R2hUure38l (accessed on 15 January 2025).
53. Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight uncertainty in neural network. In Proceedings of the ICML, Lille, France, 6–11 July 2015.
54. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Proceedings of the NeurIPS, Long Beach, CA, USA, 4–9 December 2017.
55. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the ICML, New York City, NY, USA, 19–24 June 2016.
56. Jang, E.; Gu, S.; Poole, B. Categorical Reparametrization with Gumble-Softmax. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
57. Zhang, J.; Kailkhura, B.; Han, T.Y.J. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In Proceedings of the ICML, Virtual Event, 13–18 July 2020.
58. Laves, M.H.; Ihler, S.; Kortmann, K.P.; Ortmaier, T. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv* **2019**, arXiv:1909.13550.
59. Naeini, M.P.; Cooper, G.; Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In Proceedings of the AAAI, Austin, TX, USA, 25–30 January 2015.
60. Gao, C.; Li, S.; Lei, W.; Chen, J.; Li, B.; Jiang, P.; He, X.; Mao, J.; Chua, T.S. KuaiRec: A Fully-observed Dataset and Insights for Evaluating Recommender Systems. In Proceedings of the CIKM, Atlanta, GA, USA, 17–21 October 2022.
61. Chen, J.; Dong, H.; Qiu, Y.; He, X.; Xin, X.; Chen, L.; Lin, G.; Yang, K. AutoDebias: Learning to Debias for Recommendation. In Proceedings of the SIGIR, Online, 11–15 July 2021.
62. Li, H.; Wu, K.; Zheng, C.; Xiao, Y.; Wang, H.; Geng, Z.; Feng, F.; He, X.; Wu, P. Removing Hidden Confounding in Recommendation: A Unified Multi-Task Learning Approach. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 10–15 December 2024.
63. Li, H.; Xiao, Y.; Zheng, C.; Wu, P. Balancing Unobserved Confounding with a Few Unbiased Ratings in Debiased Recommendations. In Proceedings of the WWW, Austin, TX, USA, 30 April–4 May 2023.
64. Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *42*, 30–37. [CrossRef]
65. Swaminathan, A.; Joachims, T. The Self-Normalized Estimator for Counterfactual Learning. In Proceedings of the NeurIPS, Montreal, QC, Canada, 7–12 December 2015.
66. Guo, S.; Zou, L.; Liu, Y.; Ye, W.; Cheng, S.; Wang, S.; Chen, H.; Yin, D.; Chang, Y. Enhanced Doubly Robust Learning for Debiasing Post-Click Conversion Rate Estimation. In Proceedings of the SIGIR, Online, 11–15 July 2021.
67. Dai, Q.; Li, H.; Wu, P.; Dong, Z.; Zhou, X.H.; Zhang, R.; Zhang, R.; Sun, J. A generalized doubly robust learning framework for debiasing post-click conversion rate prediction. In Proceedings of the KDD, Washington, DC, USA, 14–18 August 2022.
68. Li, H.; Xiao, Y.; Zheng, C.; Wu, P.; Cui, P. Propensity Matters: Measuring and Enhancing Balancing for Recommendation. In Proceedings of the ICML, Honolulu, HI, USA, 23–29 July 2023.
69. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]

70. Järvelin, K.; Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst. (TOIS)* **2002**, *20*, 422–446. [CrossRef]
71. Lipton, Z.C.; Elkan, C.; Naryanaswamy, B. Optimal thresholding of classifiers to maximize F1 measure. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, 15–19 September 2014; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2014, pp. 225–239.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.