Debiased Cognitive Diagnosis: A Contrastive Counterfactual Modeling Method via Variational Autoencoder

Shangshang Yang[1,5], Xuewen Duan[1], Xiaoshan Yu[2], Ziwen Wang[1], Haiping Ma[3,4*], Xingyi Zhang[1]

[1]School of Computer Science and Technology, Anhui University, China
[2]School of Artificial Intelligence, Anhui University, China
[3]Institutes of Physical Science and Information Technology, Anhui University, China
[4]State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology
[5]Anhui Province Key Laboratory of Intelligent Computing and Applications
{yangshang0308, yxsleo, wzw12sir, xyzhanghust}@gmail.com, e24201063@stu.ahu.edu.cn, hpma@ahu.edu.cn

Abstract

Cognitive diagnosis (CD), inferring student knowledge mastery based on historical response records, is crucial for personalized educational services such as adaptive practice and learning path planning. Existing CD models were built based on the assumption that student's response data is integral, overlooking the nonrandom missingness of data caused by student answering exercises selectively. This missingness generally leads to biased and incomplete observations, where confounders, such as selection bias and exposure bias, significantly undermine the accuracy of student knowledge modeling. To address missingness, we propose a Debiased Cognitive Diagnosis (**DBCD**) framework through the perspective of counterfactual modeling to remove exogenous confounders from the response data. Specifically, the proposed DBCD achieves debiasing for CD by applying the idea of contrastive learning to constrain the model's prediction distributions on both factual and counterfactual data. For a student, the factual data is his/her original response records, while the counterfactual data is generated by sampling the same number of exercises from all exercises of each concept through a similarity-based counterfactual sampling strategy. Considering the difficulty of directly removing the exogenous confounders for student, we devise a β-Variational Autoencoder to model their exogenous confounders within the latent representations of knowledge proficiency by leveraging exercise priors and student response patterns. Then, the learned representations are further combined with the vanilla student's ability embedding via a gating mechanism-based fusion for final diagnosis prediction of the model. Extensive experiments on real-world educational datasets demonstrate that the proposed DBCD effectively mitigates confounders and even outperforms existing methods, thereby validating the feasibility and effectiveness of the DBCD framework.

Introduction

Cognitive diagnosis (CD) is a fundamental task in intelligent education, which aims to uncover student's mastery of knowledge concepts from their historical response records (Yu et al. 2024a; Gao et al. 2024), providing theoretical support for downstream personalized educational services (Yang et al. 2023a), such as exercise recommendation and individualized learning path design (Ma et al. 2022b; Gao et al. 2025; Yu et al. 2025b). Up to now, there have been a lot of cognitive diagnosis models (CDMs) proposed based on different techniques (Wang et al. 2024), such as the assumption of educational psychology (Lord 2012), common neural networks (Wang et al. 2020a; Yu et al. 2025a), and graph learning (Gao et al. 2021; Yang et al. 2025).

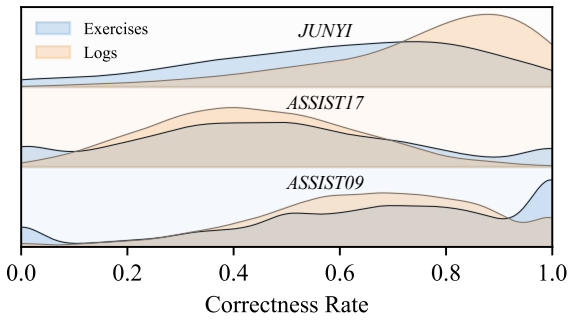

Figure 1: Exercise and response log distributions under different rates of answering correctly on three datasets.

Although existing models have achieved promising results, they generally assume that student's response data is integral and largely overlooked the issue of missing-not-at-random (MNAR) data (Marlin and Zemel 2009). As shown in Figure 1, we conducted a density distribution analysis on three real-world datasets, i.e., JUNYI (Chang et al. 2015), ASSIST17, and ASSIST09 (Feng, Heffernan, and Koedinger 2009), based on the overall exercise correctness rate and found a discrepancy in the distribution between the number of related exercises and the number of student-exercise interactions. Specifically, some students interact with a limited subset of exercises, often biased towards either easy or difficult exercises, making it challenging to accurately assess their knowledge mastery. This nonrandom missingness is raised by two key types of biases: 1) Selection Bias: Student may selectively choose which exercises to attempt based on their current understanding, e.g., prefer-

*Corresponding author.

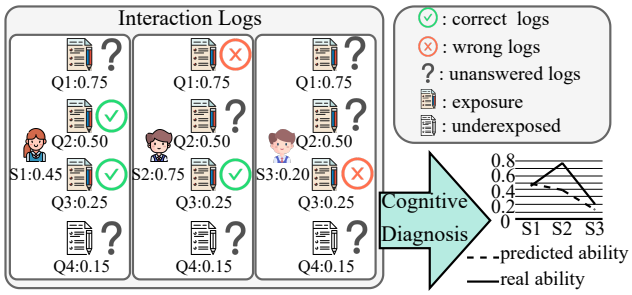

Figure 2: A case of biased CD under MNAR.

ring easier exercises for consolidation or harder exercises for further challenge (Little and Rubin 2019). 2) Exposure Bias: Due to constraints in practical teaching environments (e.g., varying instructor strategies or curriculum pacing), student may not be equally exposed to all exercises related to a given knowledge concept, resulting in incomplete learning signals (Nissen, Donatello, and Van Dusen 2019). In this paper, we refer to both types of biases as confounders, which contribute to the incompleteness of the observed interaction data and hinder accurate estimation of student ability, as shown in Figure 2. Recent progress, such as HeckmanCD (Han et al. 2024), mitigates selection bias using a Heckman-style two-stage estimation, though its effectiveness declines when the selection process is misspecified or data are sparse.

To address the missingness posed by confounders, an intuitive solution is counterfactual modeling (Nabi et al. 2022). However, it is challenging to directly apply the counterfactual modeling to CD due to the following reasons. Firstly, the mechanism of a student selecting one specific exercise for answering is typically unobserved, and it is difficult to explicitly model the response generation process (Xu et al. 2020). Secondly, due to the sparsity of student response logs (Ma et al. 2022a), it is challenging to infer student's reactions under counterfactual scenarios, thus making counterfactual evaluations difficult to validate. Finally, CD aims to infer student's knowledge proficiency, while the prediction of counterfactual modeling requires estimating their potential performance on unobserved exercises, which is an inference task with higher uncertainty and will introduce additional risks to conduct such estimation without accurate knowledge representations (Piech et al. 2015).

To overcome these challenges, we propose a novel debiased cognitive diagnosis framework based on contrastive counterfactual modeling, termed DBCD. Our contributions are as follows: **(1)** A contrastive counterfactual modeling framework is proposed for CD to remove the bias by constraining the model's prediction distributions on both factual and counterfactual data based on the idea of contrastive learning. The student's factual is his/her original response data. To build the counterfactual data for each student, a similarity-based counterfactual sampling strategy is devised to sample the same number of exercises from all exercises of each concept. **(2)** Compared to completely removing the confounders of student, a β-Variational Autoencoder (β-VAE) is employed to model the confounders within the latent representations of knowledge proficiency by leveraging exercise priors and student response patterns. Afterward, a gating mechanism-based fusion is devised to further combine the student's representations learned in β-VAE and the vanilla student's ability embedding for the final diagnosis prediction. **(3)** Experimental results on real-world datasets and six CD models demonstrate that the proposed DBCD framework achieves competitive diagnosis accuracy while exhibiting improved generalization ability and bias mitigation effectiveness.

Related Work

Cognitive Diagnosis

Up to now, a wide range of models have been proposed for CD, broadly categorized into two groups: psychometric theory-based models and deep learning-based models. Representatives of the first type include IRT (Lord 2012), MIRT (Chalmers 2012), and DINA (De La Torre 2009). These models typically formulate a student's response to an exercise as a logical function that captures the interaction between student-specific and item-specific characteristics. In contrast, deep learning-based models aim to capture the complex interactions among students, exercises, and knowledge concepts. For instance, NCD (Wang et al. 2020a) and KaNCD (Wang et al. 2022) employ neural networks to model nonlinear interactions between students and exercises. RCD (Gao et al. 2021) and SCD (Wang et al. 2023b) construct multirelational graphs among students, exercises, and knowledge concepts to incorporate structural information for accurate representation learning. HyperCD (Shen et al. 2024) leverages Hypergraph Neural Networks to effectively model the homogeneous influence among students.

Despite their success, existing CD methods rely heavily on historical student–exercise interactions, which are often MNAR. Thus, with incomplete responses, they may fail to accurately infer students' knowledge proficiency.

Debiased Learning

Debiased learning (DeL) aims to remove biases raised from data collection, selection, or labeling mechanisms through statistical or causal approaches, enabling model predictions to align with true data distribution. DeL has been widely used in many areas, e.g., recommendation (Schnabel et al. 2016) and causal inference (Shalit, Johansson, and Sontag 2017), which can be divided into two types regarding adopted strategies: data augmentation-based and representation learning-based approaches. Data augmentation typically adjusts the contribution of different data samples during model training using techniques such as error imputation or data reweighting, and representatives include IPS (Schnabel et al. 2016) and DR (Wang et al. 2019). Here DR achieves unbiased learning by combining inverse propensity weighting with error imputation. The representatives of representation learning-based approaches include AutoDebias (Chen et al. 2021), CVIB (Wang et al. 2020b), and Counter-CLR (Wang et al. 2023a). Here CVIB adopts a variational information bottleneck framework, which enables counterfactual modeling and debiasing in MNAR settings through

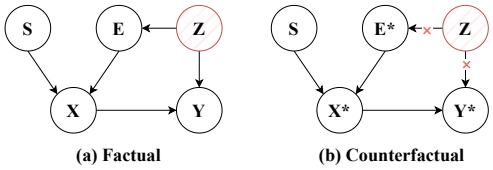

Figure 3: Structural Causal Models: a) Factual: Biased observations under confounding influences; b) Counterfactual: Hypothetical outcomes inferred via causal reasoning.

mutual information decomposition and regularization.

Although DeL methods have achieved notable debiasing results, they are primarily designed for binary relation tasks (e.g., user–item) (Schnabel et al. 2016). In contrast, CD involves complex student–exercise–concept triads, hindering direct application (Ma et al. 2021). Moreover, the lack of explicit confounder modeling often leads to suboptimal objectives and limited debiasing effectiveness.

Problem Statement

For the CD task, there are three types of entities: N students, M exercises, and C knowledge concepts, which can be denoted as $S = \{s_1, s_2, \ldots, s_N\}$,$E = \{e_1, e_2, \ldots, e_M\}$ and $K = \{k_1, k_2, \ldots, k_C\}$, respectively. All historical interaction records are defined as $R = \{(s_i, e_j, r_{ij}) \mid s_i \in S,\ e_j \in E,\ r_{ij} \in \{0, 1\}\}$ where r_{ij} denotes the answer of student s_i on exercise e_j. $r_{ij} = 1$ indicates a correct answer, and $r_{ij} = 0$ indicates an incorrect one. The exercise-concept relation matrix is defined as $Q \in \{q_{jc}\}^{M \times C}$, where $q_{jc} = 1$ if exercise e_j is associated with knowledge concept k_c, and $q_{jc} = 0$ otherwise.

Problem Definition: *Given the student response records R and the relation matrix Q, the goal of debiased CD is to achieve unbiased diagnosis of student's proficiency levels by mitigating the influence of various confounders.*

The Proposed DBCD

Overview of DBCD: A Causal Analysis Perspective

Figure 3 (a) reveals the causal structure underlying the CD process in real world, where some confounders may bias the ability estimation of students. The model involves four endogenous variables: S (students), E (exercises), X (student-exercise interactions), and Y (responses), along with another exogenous confounder Z influencing both exercise exposure and exercise selection. Under such a context, CD is to estimate the causal effect along the path $S, E \rightarrow X \rightarrow Y$. It reflects how to model the student's abilities by predicting responses Y of interactions X. However, there are two backdoor paths for the confounder Z, probably inducing bias:

- $Z \rightarrow E \rightarrow X$ denotes various confounders causing the bias of student-exercise interactions, such as system-driven exposure constraints, student self-selection, and recommendation biases. These confounders cause non-random and selective interactions, thereby distorting the observed causal relationship between X and Y.
- $Z \rightarrow Y$ denotes the direct influence of these confounders on the estimation of student's abilities, such as motivation and fatigue, constituting additional confounding.

The two backdoor paths hinder direct estimation of the intended causal effect. The Backdoor Criterion prescribes adjusting for Z as $P(Y \mid do(X)) = \sum_Z P(Y \mid X, Z)$. However, Z is usually unobserved and hard to infer, making explicit adjustment infeasible.

To this end, we propose a representation-level counterfactual intervention framework (Shalit, Johansson, and Sontag 2017; Wang et al. 2023a), as shown in Figure 3(b), simulating causal intervention without directly observing Z. Specifically, to block the $Z \rightarrow Y$ path, a β-VAE is used to encode student's responses and exercise attributes into a latent representation, which is the student's latent knowledge proficiency $\mathbf{h}_z \in \mathbb{R}^C$. This latent representation is combined with the student's ability embedding $\mathbf{h}_s \in \mathbb{R}^C$ to yield a fused feature $\mathbf{h}_f \in \mathbb{R}^C$, mitigating the direct confounder effect on predictions. To block the $Z \rightarrow E \rightarrow X$ confounder path, we generate counterfactual exercise samples E^* based on exercise similarity. Each student interacts with unbiased alternative exercises, yielding counterfactual interactions X^* and responses Y^* to disentangle confounder effects in exercise selection.

Intuitively, robust predictions may be consistent across factual and counterfactual scenarios. Thus, a contrastive information regularization inspired by CVIB is used to encourage robustness against confounding bias:

$$\mathcal{L}_{fc} = -\frac{1}{N} \sum_{\substack{(s,e,\hat{r}) \in D_f \\ (s,e,\tilde{r}) \in D_{cf}}} \left(\hat{r} \log \tilde{r} + (1 - \hat{r}) \log(1 - \tilde{r})\right), \quad (1)$$

where D_f is the factual data, i.e., R, and D_{cf} denotes the counterfactual data of student-exercise interactions.

Overview. The counterfactual intervention framework for CD, as shown in Figure 4 (a), can be drawn from above causal analysis. Therefore, the implementation of the DBCD can be induced from the intervention as Figure 4 (b). Firstly a VAE pretrained on factual data is used to extract student's latent representation. Then, it will be combined with the ability embedding of student. Next, the fused knowledge proficiency of student will be used for counterfactual modeling of student based on the sampled counterfactual data through contrastive information regularization.

Latent Knowledge Proficiency Modeling & Fusion

Latent Knowledge Proficiency Modeling via β-VAE To mitigate exercise-level bias caused by confounder Z, the idea of multi-exercise aggregation is adopted to infer a latent representation $\mathbf{h}_z$ for a student from his/her historical responses. By doing so, the sensitivity to exercise selection or exercise exposure biases can be reduced, thereby yielding more robust knowledge proficiency representations.

Specifically, we construct the input matrix $\mathbf{X}_i = [\mathbf{r}_i, \mathbf{d}] \in \mathbb{R}^{M \times 2}$ for student s_i. $\mathbf{r}_i = ((r_{i1}, \ldots, r_{ij}, \ldots, r_{iM})) \in \{0, 1\}^M$ is the binary response vector and $\mathbf{d} = (d_1, \ldots, d_j, \ldots, d_M) \in \mathbb{R}^M$ is the exercise difficulty vector. d_j is computed as its empirical correctness rate, serving

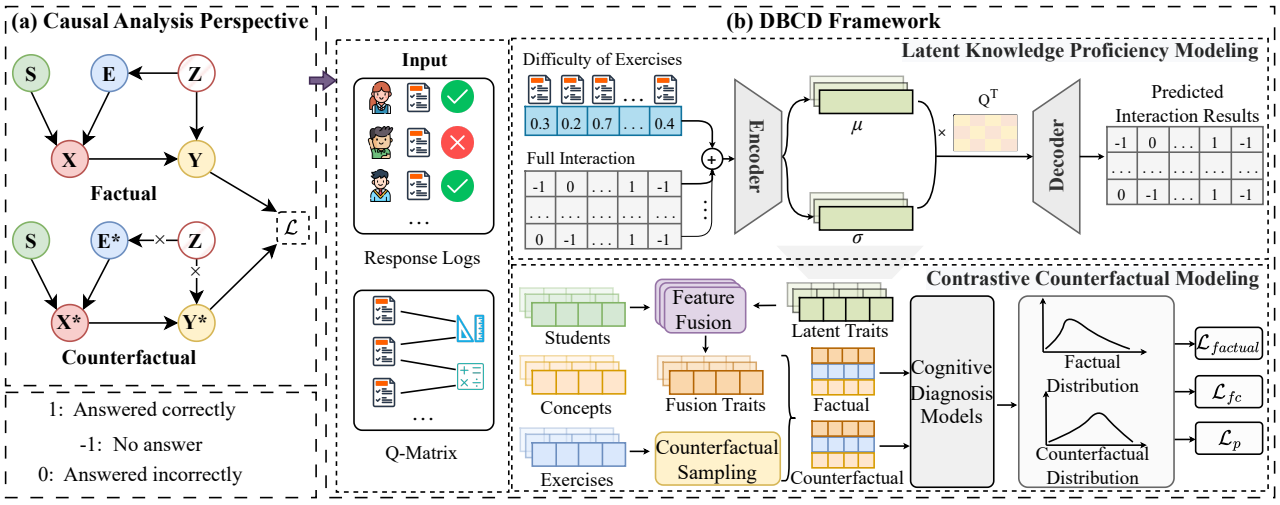

Figure 4: (a) A counterfactual intervention perspective framework based on causal inference; (b) Specific implementation of DBCD induced by causal analysis.

as a bias-insensitive auxiliary input (Chen et al. 2024). Each row $[r_{ij}, d_j] \in \mathbb{R}^{1\times 2}$ is first encoded by the exercise-wise encoder $f_{\text{item}}(\cdot)$ followed by mean pooling to obtain a global representation $\bar{\mathbf{x}}_i \in \mathbb{R}^H$:

$$\bar{\mathbf{x}}_i = \frac{1}{M}\sum_{j=1}^{M} f_{\text{item}}([r_{ij}, d_j]). \tag{2}$$

Then $\bar{\mathbf{x}}_i$ is used to obtain the latent representation $\mathbf{h}_z$ based on a diagonal Gaussian posterior:

$$\begin{aligned} \mathbf{h}_z : q(\mathbf{h}_z \mid \mathbf{X}_i) &= \mathcal{N}(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2)), \\ \boldsymbol{\mu}_i = f_\mu(\bar{\mathbf{x}}_i), \quad \log \boldsymbol{\sigma}_i^2 &= f_{\log\sigma^2}(\bar{\mathbf{x}}_i), \end{aligned} \tag{3}$$

where $f_\mu(\cdot)$ and $f_{\log\sigma^2}(\cdot)$ are learnable encoders. $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ik}, \ldots, \mu_{iC}) \in \mathbb{R}^C$ and $\boldsymbol{\sigma}_i = (\sigma_{ik}, \ldots, \sigma_{ik}, \ldots, \sigma_{iC}) \in \mathbb{R}^C$ denote the mean and standard deviation. To promote knowledge-level alignment, the exercise-concept relation matrix Q is incorporated into $\mathbf{h}_z$ as $\hat{\mathbf{r}}_i = \sigma\left(f_{\text{dec}}(\mathbf{h}_z) \cdot \mathbf{Q}^\top\right)$. Here $f_{\text{dec}}(\cdot)$ is a decoder network and $\sigma(\cdot)$ is a sigmoid activation. By doing so, the prediction of each log only depends on the knowledge concepts associated with each log's exercise, ensuring interpretability. To train the β-VAE stably, the KL divergence loss with free-bits regularization (Chen et al. 2016) and the reconstruction loss are adopted:

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= BCE(\hat{\mathbf{r}}_i, \mathbf{r}_i), \\ \mathcal{L}_{\text{kl}} &= \sum_{k=1}^{C} \max\left(-\tfrac{1}{2}\left(1 + \log\sigma_{ik}^2 - \mu_{ik}^2 - \sigma_{ik}^2\right), \tau\right), \end{aligned} \tag{4}$$

where τ is a lower-bound threshold that enforces minimum information flow for each dimension.

Gating Mechanism-based Knowledge Proficiency Fusion After obtaining the latent representation $\mathbf{h}_z$, a gating mechanism-based fusion is devised to combine the latent representation $\mathbf{h}_z$ with the ability embedding $\mathbf{h}_s$:

$$\mathbf{h}_f = \mathbf{g}_i \odot \mathbf{h}_s + (1 - \mathbf{g}_i) \odot \mathbf{h}_z, \mathbf{g}_i = \sigma([\mathbf{h}_s, \mathbf{h}_z] \times \mathbf{W}_g), \tag{5}$$

where $\mathbf{W}_g \in \mathbb{R}^{2C\times C}$ is a learnable weight matrix, $\odot$ denotes element-wise multiplication, and $[\cdot]$ is the concatenation. With the fused representation $\mathbf{h}_f$, the final prediction of students answering exercises can be made with existing CDMs. This gating mechanism-based fusion enables CDMs to adaptively balance general and concept-specific knowledge signals, enhancing robustness to bias.

Contrastive Counterfactual Modeling for Student's Knowledge Proficiency

To eliminate exercise-level bias caused by confounder Z as much as possible, the contrastive counterfactual modeling is devised, which constrains the model's prediction distributions on both factual and counterfactual data.

Similarity-based Counterfactual Sampling Strategy To collect the counterfactual data, a similarity-based counterfactual sampling strategy is devised to build counterfactual data on response logs R. The main idea of this strategy is to generate counterfactual data according to factual interaction pairs (s_i, e_j), where $s_i \in S, e_j \in E$. To generate reasonable counterfactual data $\{(s_i, e_p) | s_i \in S, e_p \in E\}$, exercise e_p in the counterfactual data pairs needs to keep similar structural or semantic properties to e_j in factual data pairs, which enables the model to better capture underlying causal invariances. Not limited to unattempted exercises, the devised strategy can select candidate counterfactual data from both attempted and unattempted exercises by following the structural similarity criterion:

$$e_p \in \{e_l | e_l \in E \setminus \{e_j\} \text{ and } \text{sim}(e_j, e_l) > \delta\}, \tag{6}$$

where δ is a predefined threshold and $\text{sim}(\cdot)$ is a similarity function. To improve the applicability of this strategy, we consider two complementary similarity perspectives.

Structure similarity (via the exercise-concept relation matrix): Suppose e_j and e_p are associated with knowledge

concept sets $\mathcal{K}_j$ and $\mathcal{K}_p$, their structure similarity can be defined by Jaccard coefficients:

$$\text{sim}_{\text{struct}}(e_j, e_p) = |\mathcal{K}_j \cap \mathcal{K}_p| / |\mathcal{K}_j \cup \mathcal{K}_p|. \quad (7)$$

Semantic similarity (via exercise representations): Suppose the embedding of exercises e_j and e_p are $\mathbf{h}_{e_j}, \mathbf{h}_{e_p} \in \mathbb{R}^C$, their semantic similarity are computed by

$$\text{sim}_{\text{embed}}(e_j, e_p) = (\mathbf{h}_{e_j}^\top \mathbf{h}_{e_p}) / (\|\mathbf{h}_{e_j}\| \cdot \|\mathbf{h}_{e_p}\|). \quad (8)$$

These two similarity measures can be used independently or combined via the weighted sum, depending on the characteristics of the datasets. For each factual pair (s_i, e_j), we sample a single counterfactual exercise e_p from the similarity set to form the counterfactual pair (s_i, e_p).

Counterfactual Modeling via Contrastive Information Regularization With the sampled counterfactual data $\{(s_i, e_p) | s_i \in S, e_p \in E\}$, we aim to align the model's prediction distributions on factual and counterfactual data by contrastive information regularization.

The contrastive information regularization is based on an assumption: student's representations are stable and exercise embeddings capture semantic consistency. Therefore, we expect $\hat{r} \approx \tilde{r}$ when e_j and e_p are structurally or semantically similar, where $\hat{r}$ and $\tilde{r}$ denote the predicted response of student s_i on exercise e_j and counterfactual exercise e_p. To achieve the regularization, we apply an information-level contrastive objective to the models' prediction distributions instead of imposing contrastive constraints directly on the ability embeddings of students. Specifically, the loss $\mathcal{L}_{fc}$ in Eq. (1) is adopted as the contrastive objective to minimize the divergence between predictions in factual and counterfactual data, promoting alignment of the model's causal reasoning. To prevent the model from being overly confident in underdetermined scenarios with the contrastive information regularization, another entropy regularization is added:

$$\mathcal{L}_p = \frac{1}{N} \sum \left[\hat{r} \cdot \log \hat{r} + (1 - \hat{r}) \cdot \log(1 - \hat{r}) \right], \quad (9)$$

which encourages the model's prediction distributions to remain uncertain when appropriate.

Training Procedure of DBCD

Training of the proposed DBCD framework has two stages: 1) training the β-VAE on factual data, and 2) training CDMs with the β-VAE on both factual and counterfactual data.

In the first stage, the β-VAE is trained on the factual data R to learns the latent representation $\mathbf{h}_z$ from each student's response vector $\mathbf{r}_i$ and the prior exercise difficulty vector $\mathbf{d}$ by optimizing the reconstruction loss and the KL divergence loss, with the following objective:

$$\mathcal{L}_{\text{vae}} = \mathcal{L}_{rec} + \beta \cdot \mathcal{L}_{kl}, \quad (10)$$

where β is a hyperparameter controlling the weight of the KL divergence. After training, the β-VAE parameters are fixed, and latent representations $\mathbf{h}_z$ are inferred for all students as input features for the subsequent diagnosis model.

In the second stage, the CDMs are trained on the built counterfactual data by using the latent representation $\mathbf{h}_z$ obtained from the trained β-VAE. Specifically, the gating mechanism-based fusion in Eq. (5) is first employed to integrate $\mathbf{h}_z$ with the original student embedding $\mathbf{h}_s$ as the fused representations $\mathbf{h}_f$. Then, the CDMs can complete the models' prediction based on the $\mathbf{h}_f$ along with the embeddings of corresponding exercises and concepts. Finally, the CDMs can be trained by optimizing the overall loss as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{factual}} + \mathcal{L}_{\text{info}}, \ \mathcal{L}_{\text{info}} = \alpha \cdot \mathcal{L}_{fc} + \gamma \cdot \mathcal{L}_p, \quad (11)$$

where $\mathcal{L}_{\text{factual}}$ denotes the model's loss on factual data, i.e., applying the binary cross-entropy loss to r_{ij} and $\hat{r}_{ij}$.

Dataset	Students	Exercises	Concepts	Logs	Q-Density
ASSIST09	2493	17676	123	267423	1.20
ASSIST17	1704	3162	102	390305	1.23
JUNYI	3418	711	39	138538	1.00

Table 1: Dataset statistics.

Experiments

Experimental Settings

Datasets and Preprocessing We evaluated our method on three real-world datasets: ASSIST09, ASSIST17, and JUNYI. Dataset statistics are summarized in Table 1, where Q-density indicates the average number of knowledge concepts per exercise. Following (Saito 2020), we built standardized test sets for controlled evaluation, as existing benchmarks lack predefined splits under missing-at-random (MAR) conditions. All models are evaluated under three settings: 1) Full, the original test set; 2) Random, a randomly sampled subset; and 3) Uniform, a re-sampled subset with uniform missingness. More processing details and experiments can be found in the **appendix**.

Baselines and Evaluation Metrics We compared DBCD with representative baselines, including MIRT, DINA, NCD, KaNCD, KSCD (Ma et al. 2022a), HyperCD, an adapted CVIB for debiasing and HeckmanCD, a debiasing strategy designed to address bias in CD tasks. All methods employed publicly available implementations with hyperparameters tuned as in their original studies. We adopt three standard metrics: accuracy (ACC) (Yang et al. 2024; Yu et al. 2024b), root mean square error (RMSE) (Pei et al. 2018; Yang et al. 2023b), and area under the ROC curve (AUC) (Bradley 1997; Yu et al. 2024c).

Parameter Settings All models were implemented in PyTorch with Xavier initialization (Glorot and Bengio 2010) and optimized using Adam (Kingma and Ba 2014). For MIRT, the latent trait dimension was set equal to the number of concepts C (Yang, Qin, and Yu 2024). The similarity threshold δ varied between 0 and 0.5, and the dimension of $\bar{\mathbf{x}}_i$ was set to 64. Batch size was fixed at 128. During training, the KL weight β was dynamically varied from 0 to 1, the temperature τ ranged from 0.001 to 0.5, and the regularization weight α spanned 0.001 to 2, with γ fixed at 0.01. Our code is available at https://github.com/sherklock/Intelligent-Education/tree/main/DBCD.

Dataset & Metric			MIRT			DINA			NCD		
			Base	CVIB	DBCD	Base	CVIB	DBCD	Base	CVIB	DBCD
ASSIST09	Full	ACC% ↑	70.74 ± .19	71.00 ± .11	**72.07 ± .18**	65.75 ± .06	65.20 ± .08	**69.46 ± .04**	71.55 ± .12	71.82 ± .16	**72.55 ± .42**
		AUC% ↑	72.30 ± .33	72.59 ± .01	**74.68 ± .29**	69.01 ± .08	69.89 ± .06	**72.23 ± .22**	72.92 ± .01	73.99 ± .15	**75.28 ± .08**
		RMSE% ↓	44.74 ± .17	**44.61 ± .05**	45.73 ± .14	49.13 ± .04	48.77 ± .01	**46.74 ± .02**	43.93 ± .05	44.25 ± .26	**43.44 ± .23**
	Random	ACC% ↑	70.72 ± .24	71.03 ± .03	**72.13 ± .19**	65.76 ± .09	65.19 ± .12	**69.57 ± .07**	71.70 ± .16	71.97 ± .16	**72.63 ± .40**
		AUC% ↑	72.24 ± .31	72.59 ± .14	**74.75 ± .33**	69.16 ± .14	69.95 ± .11	**72.43 ± .17**	73.19 ± .02	74.21 ± .13	**75.47 ± .05**
		RMSE% ↓	44.74 ± .16	**44.58 ± .05**	45.69 ± .18	49.06 ± .06	48.74 ± .04	**46.64 ± .01**	43.84 ± .06	44.14 ± .26	**43.36 ± .21**
	Uniform	ACC% ↑	71.18 ± .13	71.59 ± .17	**72.82 ± .41**	65.75 ± .10	65.30 ± .14	**70.01 ± .08**	72.17 ± .14	72.37 ± .07	**73.15 ± .36**
		AUC% ↑	72.28 ± .35	72.66 ± .17	**75.40 ± .41**	69.04 ± .11	70.03 ± .15	**72.51 ± .15**	73.02 ± .03	74.02 ± .13	**75.58 ± .02**
		RMSE% ↓	44.41 ± .14	**44.23 ± .03**	44.89 ± .27	48.96 ± .05	48.57 ± .07	**46.31 ± .03**	43.63 ± .05	43.89 ± .30	**43.04 ± .22**
JUNYI	Full	ACC% ↑	77.35 ± .10	77.28 ± .04	**80.07 ± .15**	67.21 ± .26	67.72 ± .39	**76.48 ± .13**	79.19 ± .29	79.37 ± .06	**79.89 ± .07**
		AUC% ↑	75.62 ± .22	74.90 ± .30	**80.46 ± .05**	70.89 ± .05	72.26 ± .31	**76.67 ± .21**	**80.13 ± .03**	79.78 ± .01	79.75 ± .01
		RMSE% ↓	40.89 ± .10	41.10 ± .10	**37.58 ± .05**	45.31 ± .11	45.23 ± .15	**42.57 ± .18**	**37.95 ± .05**	38.26 ± .30	37.96 ± .08
	Random	ACC% ↑	77.54 ± .11	77.57 ± .16	**80.32 ± .12**	67.48 ± .47	68.00 ± .33	**76.71 ± .24**	79.40 ± .23	79.62 ± .13	**80.26 ± .02**
		AUC% ↑	75.77 ± .24	75.22 ± .27	**80.45 ± .20**	71.02 ± .25	72.47 ± .23	**76.49 ± .30**	**80.27 ± .03**	79.97 ± .02	79.86 ± .08
		RMSE% ↓	40.69 ± .07	40.82 ± .13	**37.56 ± .22**	45.08 ± .25	44.96 ± .12	**42.40 ± .18**	37.78 ± .05	38.05 ± .28	**37.76 ± .11**
	Uniform	ACC% ↑	74.13 ± .17	73.96 ± .13	**76.59 ± .24**	64.18 ± .28	64.56 ± .32	**71.42 ± .13**	76.06 ± .32	75.85 ± .23	**76.68 ± .06**
		AUC% ↑	75.59 ± .08	74.91 ± .20	**80.28 ± .13**	70.33 ± .20	71.81 ± .27	**75.63 ± .19**	**79.84 ± .04**	79.45 ± .03	79.32 ± .01
		RMSE% ↓	43.23 ± .08	43.66 ± .05	**40.26 ± .33**	47.89 ± .18	47.83 ± .10	**46.09 ± .22**	**40.31 ± .05**	40.89 ± .42	40.57 ± .08

Table 2: Comparative performance of DBCD, CVIB, and selected baselines on real-world datasets, where the best result in each setting is highlighted in bold.

Dataset & Metric			MIRT-DBCD	
			Experience Difficulty	IRT-based variant
ASSIST09	Full	ACC% ↑	72.07± 0.18	72.19± 0.25
		AUC% ↑	74.68± 0.29	74.62± 0.17
		RMSE% ↓	45.73± 0.14	45.80± 0.08
	Random	ACC% ↑	72.13± 0.19	72.24± 0.26
		AUC% ↑	74.75± 0.33	74.67± 0.21
		RMSE% ↓	45.69± 0.18	45.75± 0.15
	Uniform	ACC% ↑	72.82± 0.41	72.87± 0.30
		AUC% ↑	75.40± 0.41	75.36± 0.48
		RMSE% ↓	44.89± 0.27	44.86± 0.25

Table 3: An ablation verification experiment on the difficulty of the exercises.

Overall Performance

To validate the effectiveness of our proposed DBCD debiasing strategy, we conducted experiments on real-world educational datasets under three distinct test set construction strategies. The performance was compared against six representative cognitive diagnosis models (MIRT, DINA, NCD, KaNCD, KSCD, and HyperCD), the counterfactual-based debiasing baseline CVIB, as well as HeckmanCD which is a debiasing strategy designed to address bias in CD tasks. Detailed results are reported in Table 2, Table 4, and Table 5. The results demonstrate that DBCD consistently outperforms the base models across most settings in terms of ACC, AUC, and RMSE. For simple models such as DINA, it achieves notable improvements, with approximately 6% increase in ACC, 4% increase in AUC, and 2.5% reduction in RMSE. Although the improvements are more modest for complex models such as KaNCD, the method still maintains robust performance. In addition, DBCD surpasses CVIB in nearly all cases, further confirming its effectiveness and suitability for CD tasks. Moreover, while HeckmanCD yields competitive debiasing when its selection model is correctly specified, its performance drops sharply under misspecification or data sparsity. In contrast, DBCD remains stable owing to its representation learning and counterfactual consistency regularization. To conclude, DBCD offers a general and effective framework for debiasing cognitive diagnosis under MNAR settings, achieving consistent and statistically significant gains in predictive accuracy and robustness across diverse model architectures and evaluation setups.

Ablation Study

To evaluate the effectiveness of each core component in our proposed DBCD framework, we conducted a comprehensive ablation study involving the following two variants: 1) w/o β-VAE: this variant removed the latent representation inference and fusion module; 2) w/o CCM: this variant removed the contrastive counterfactual modeling module. The detailed experimental results are illustrated in Figure 5. As shown, removing the β-VAE yields the most significant performance drop, marked by increased RMSE and reduced ACC and AUC, indicating its pivotal role in stabilizing training. Excluding the CCM module results in consistent but milder declines, reflecting its contribution to discriminative precision. The full model achieves the best overall performance. In summary, the β-VAE ensures robust representation learning under MNAR conditions, while CCM

Dataset & Metric			KaNCD			KSCD			HyperCD		
			Base	CVIB	DBCD	Base	CVIB	DBCD	Base	CVIB	DBCD
ASSIST09	Full	ACC% ↑	72.39 ± .74	72.88 ± .17	**72.90 ± .03**	72.54 ± .07	72.54 ± .10	**72.74 ± .08**	72.39 ± .07	72.59 ± .07	**72.84 ± .07**
		AUC% ↑	75.95 ± .03	75.83 ± .06	**75.99 ± .05**	75.58 ± .04	75.54 ± .05	**75.84 ± .04**	75.31 ± .07	75.56 ± .03	**75.99 ± .03**
		RMSE% ↓	42.86 ± .36	42.74 ± .01	**42.68 ± .03**	**42.82 ± .05**	42.84 ± .01	42.93 ± .10	42.92 ± .04	43.10 ± .20	**42.78 ± .04**
	Random	ACC% ↑	72.50 ± .67	**72.89 ± .12**	72.87 ± .06	72.54 ± .05	72.53 ± .14	**72.81 ± .12**	72.40 ± .10	72.62 ± .13	**72.90 ± .06**
		AUC% ↑	**76.17 ± .07**	75.97 ± .04	76.12 ± .06	75.71 ± .04	75.66 ± .04	**75.98 ± .04**	75.42 ± .05	75.66 ± .06	**76.12 ± .05**
		RMSE% ↓	42.76 ± .34	42.67 ± .04	**42.64 ± .03**	**42.76 ± .06**	42.79 ± .01	42.86 ± .09	42.88 ± .04	43.05 ± .19	**42.71 ± .03**
	Uniform	ACC% ↑	72.81 ± .79	73.29 ± .18	**73.45 ± .07**	73.05 ± .10	73.10 ± .12	**73.25 ± .10**	72.96 ± .26	73.00 ± .16	**73.35 ± .07**
		AUC% ↑	76.17 ± .06	76.00 ± .07	**76.25 ± .04**	75.80 ± .05	75.71 ± .08	**76.01 ± .09**	75.48 ± .09	75.67 ± .08	**76.17 ± .03**
		RMSE% ↓	42.56 ± .36	42.47 ± .02	**42.38 ± .02**	**42.52 ± .05**	42.57 ± .02	42.69 ± .15	42.61 ± .05	42.91 ± .27	**42.50 ± .06**
JUNYI	Full	ACC% ↑	79.95 ± .13	79.70 ± .13	**80.35 ± .09**	80.80 ± .09	80.99 ± .05	**81.06 ± .08**	80.75 ± .48	80.85 ± .04	**80.93 ± .02**
		AUC% ↑	80.25 ± .17	79.77 ± .06	**81.00 ± .01**	**81.74 ± .04**	81.30 ± .03	81.56 ± .06	81.90 ± .04	81.51 ± .06	**81.93 ± .06**
		RMSE% ↓	37.61 ± .01	37.82 ± .03	**37.26 ± .01**	36.82 ± .04	36.93 ± .09	**36.82 ± .02**	36.85 ± .23	36.96 ± .09	**36.55 ± .12**
	Random	ACC% ↑	80.16 ± .15	79.95 ± .16	**80.64 ± .14**	80.98 ± .14	81.16 ± .09	**81.21 ± .11**	80.90 ± .54	81.04 ± .06	**81.13 ± .10**
		AUC% ↑	80.39 ± .14	79.93 ± .09	**81.09 ± .03**	**81.70 ± .02**	81.30 ± .04	81.61 ± .03	81.91 ± .05	81.51 ± .05	**81.95 ± .05**
		RMSE% ↓	37.39 ± .04	37.61 ± .04	**37.07 ± .01**	36.70 ± .04	36.76 ± .09	**36.65 ± .03**	36.71 ± .23	36.81 ± .06	**36.58 ± .11**
	Uniform	ACC% ↑	76.74 ± .16	76.40 ± .10	**77.19 ± .03**	77.87 ± .06	77.86 ± .05	**77.98 ± .07**	77.77 ± .44	77.68 ± .10	**77.83 ± .12**
		AUC% ↑	80.14 ± .14	79.37 ± .11	**80.91 ± .08**	**81.64 ± .04**	81.01 ± .04	81.42 ± .04	81.75 ± .03	81.29 ± .08	**81.90 ± .02**
		RMSE% ↓	39.99 ± .04	40.33 ± .13	**39.60 ± .04**	**39.03 ± .01**	39.39 ± .15	39.15 ± .06	39.11 ± .16	39.42 ± .19	**39.02 ± .20**

Table 4: Comparative performance of DBCD, CVIB, and selected baselines on real-world datasets, where the best result in each setting is highlighted in bold.

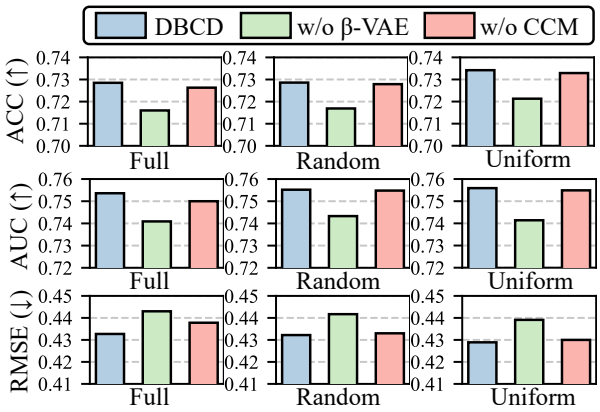

Figure 5: Ablation study of DBCD conducted on ASSIST09.

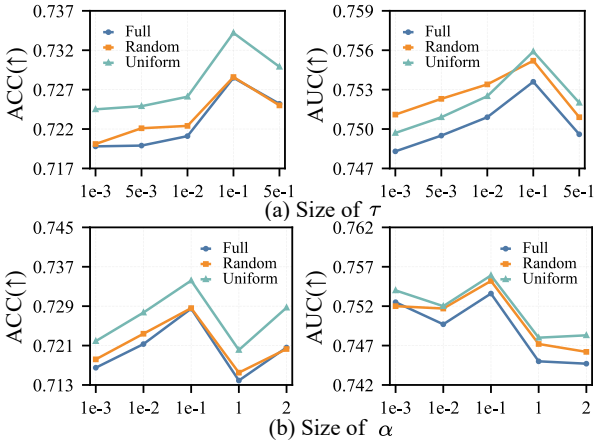

Figure 6: Hyperparameter sensitivity analysis conducted on the ASSIST09 dataset.

enhances predictive accuracy. Together, they effectively improve DBCD's generalization and bias mitigation.

Reasonableness Analysis of Empirical Accuracy

To further validate the rationality of our current use of empirical accuracy as a feature of exercise difficulty in the VAE module, we conducted an ablation validation experiment on the ASSIST09 dataset based on the MIRT model. Specifically, we replaced the empirical accuracy rate with the item difficulty estimated by a 3-parameter IRT model. The results are shown in Table 3. The results show that the performance of our current implementation is statistically indistinguishable from that of the IRT-based variant. These results further validate that our approach remains robust and unbiased with respect to the construction of the item difficulty feature.

Hyperparameter Analysis

To assess the impact of KL regularization and contrastive alignment on model performance, we conducted experiments on the ASSIST09 dataset with varied hyperparameter settings. The results are shown in Figure 6. As can be seen, setting $\tau = 0.1$ and $\alpha = 0.1$ yields optimal or near-optimal ACC and AUC across all test splits. Moderate KL regularization and contrastive weighting improve model discrimination and robustness. To sum up, our regularization and contrastive objectives effectively enhance DBCD's predic-

Dataset & Metric			MIRT		
			Base	HeckmanCD	DBCD
ASSIST09	Full	ACC% ↑	70.74± 0.19	70.97± 0.30	**72.07± 0.18**
		AUC% ↑	72.30± 0.33	71.11± 0.61	**74.68± 0.29**
		RMSE% ↓	44.74± 0.17	**44.52± 0.06**	45.73± 0.14
	Random	ACC% ↑	70.72± 0.24	71.06± 0.23	**72.13± 0.19**
		AUC% ↑	72.24± 0.31	71.25± 0.48	**74.75± 0.33**
		RMSE% ↓	44.74± 0.16	**44.46± 0.10**	45.69± 0.18
	Uniform	ACC% ↑	71.18± 0.13	71.12± 0.27	**72.82± 0.41**
		AUC% ↑	72.28± 0.35	70.55± 0.62	**75.40± 0.41**
		RMSE% ↓	**44.41± 0.14**	44.47± 0.13	44.89± 0.27
ASSIST17	Full	ACC% ↑	67.98± 0.09	69.20± 0.17	**69.42± 0.14**
		AUC% ↑	73.58± 0.04	75.15± 0.26	**75.80± 0.25**
		RMSE% ↓	46.53± 0.04	44.83± 0.09	**44.72± 0.13**
	Random	ACC% ↑	67.87± 0.04	69.10± 0.25	**69.31± 0.20**
		AUC% ↑	73.47± 0.09	75.02± 0.32	**75.71± 0.28**
		RMSE% ↓	46.59± 0.05	44.89± 0.13	**44.75± 0.11**
	Uniform	ACC% ↑	68.09± 0.13	68.87± 0.27	**69.42± 0.22**
		AUC% ↑	73.86± 0.09	74.92± 0.37	**75.96± 0.24**
		RMSE% ↓	46.31± 0.04	44.98± 0.15	**44.71± 0.15**
JUNYI	Full	ACC% ↑	77.35± 0.10	78.69± 0.27	**80.07± 0.15**
		AUC% ↑	75.62± 0.22	78.80± 0.64	**80.46± 0.05**
		RMSE% ↓	40.89± 0.10	38.90± 0.35	**37.58± 0.05**
	Random	ACC% ↑	77.54± 0.11	78.79± 0.32	**80.32± 0.12**
		AUC% ↑	75.77± 0.24	78.74± 0.56	**80.45± 0.20**
		RMSE% ↓	40.69± 0.07	38.78± 0.36	**37.56± 0.22**
	Uniform	ACC% ↑	74.13± 0.17	75.43± 0.41	**76.59± 0.24**
		AUC% ↑	75.59± 0.08	78.19± 0.70	**80.28± 0.13**
		RMSE% ↓	43.23± 0.08	41.42± 0.43	**40.26± 0.33**

Table 5: Comparative performance of Base, HeckmanCD, and DBCD, with the best result in each setting highlighted.

tive accuracy and robustness by balancing latent information preservation and enforcing consistency.

Visualization Analysis

To examine the physical meaning of the gating mechanism, we visualize concept-level gating values $\mathbf{g}_i$ on the Junyi dataset. Figure 7 shows the distribution of averaged gating values across concepts. Some concepts exhibit notably lower values, indicating stronger reliance on latent representations than on student embeddings. This suggests that the model can adaptively identify concepts with sparse or biased data and automatically strengthen the contribution of latent representations through gate adjustment, thereby achieving a concept-level debiasing effect.

Computational Complexity Analysis of DBCD

To assess the computational requirements of our framework, we analyze its two-stage training procedure. The first stage trains the VAE on factual data. VAE has to process N students. For each student, it has encoding cost $M\nu + \xi$ and decoding cost η_{full}, ν and ξ mean encoding per (student, item) and aggregation. Thus, VAE's complexity is $N * (M\nu + \xi + \eta_{\text{full}})$. The second stage performs counterfactual generation and CD model training. The counterfactual generation's complexity is $O(M(\bar{q}^2 + \bar{s}) + |R|k)$, having three parts: $QQ^\top$ computation costs $O(M\bar{q}^2)$ (suppose each item contains average $\bar{q}$ concepts), item pool construction costs $O(M\bar{s})$ ($\bar{s}$ denotes average retained neighbors), and sampling k counterfactuals per interaction costs $O(|R|k)$. The CD model is then trained on both factual and counterfactual samples, increasing its training cost by a factor of k relative to using factual data only.

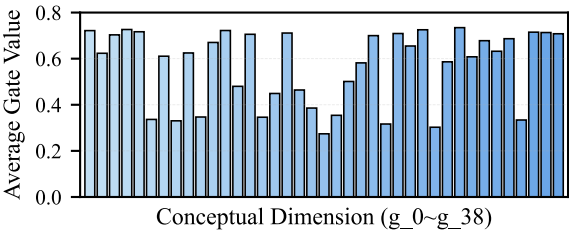

Figure 7: Visual analysis of $\mathbf{g}_i$ on the JUNYI dataset.

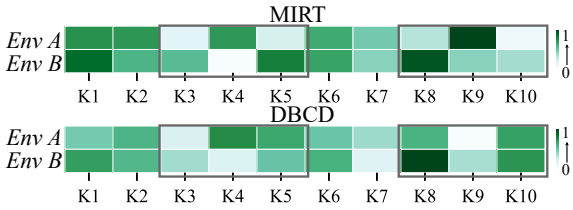

Figure 8: A case study on the MIRT model and the proposed DBCD method conducted on the ASSIST09 dataset.

Case Study

To validate our debiasing strategy, we conducted a case study on the ASSIST09 dataset with two training settings: 1) *Env A* with MNAR data influenced by confounders; 2) *Env B* approximating MAR as an unbiased baseline (Saito 2020). Both the original MIRT and DBCD-enhanced MIRT models were trained accordingly. Results are shown in Figure 8. As shown, the original MIRT exhibits notable bias in *Env A*, especially on concepts K3–K5 and K8–K10, while DBCD substantially reduces this bias, producing estimates closer to the unbiased *Env B*. In summary, our strategy effectively alleviates MNAR bias, improving the accuracy of knowledge state estimation in cognitive diagnosis.

Conclusion

We propose DBCD, a debiasing framework for CD under MNAR conditions, addressing bias from selective student responses. It integrates: (1) latent knowledge proficiency modeling via a VAE fused with ability embeddings through a gating mechanism for robust prediction; and (2) contrastive counterfactual modeling that regularizes predictions between factual and counterfactual samples to reduce bias. Experiments on real-world datasets demonstrate the superiority of DBCD over competitive baselines, validating its effectiveness in mitigating MNAR-induced bias.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.62302010, No.62107001), in part by China Postdoctoral Science Foundation (No.2023M740015), in part by the Postdoctoral Fellowship Program (Grade B) of China Postdoctoral Science Foundation (No.GZB20240002), in part by the Anhui Provincial Natural Science Foundation (No.2508085MF160), and in part by the Anhui Province Key Laboratory of Intelligent Computing and Applications (No. AFZNJS2024KF01).

References

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7): 1145–1159.

Chalmers, R. P. 2012. mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, 48: 1–29.

Chang, H.-S.; Hsu, H.-J.; Chen, K.-T.; et al. 2015. Modeling exercise relationships in E-learning: A unified approach. In *EDM*, 532–535.

Chen, J.; Dong, H.; Qiu, Y.; He, X.; Xin, X.; Chen, L.; Lin, G.; and Yang, K. 2021. AutoDebias: Learning to debias for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 21–30.

Chen, X.; Feng, S.; Yang, M.; Zhao, K.; Xu, R.; Cui, C.; and Chen, M. 2024. Modeling question difficulty for unbiased cognitive diagnosis: A causal perspective. *Knowledge-Based Systems*, 294: 111750.

Chen, X.; Kingma, D. P.; Salimans, T.; Duan, Y.; Dhariwal, P.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*.

De La Torre, J. 2009. DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, 34(1): 115–130.

Feng, M.; Heffernan, N.; and Koedinger, K. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, 19(3): 243–266.

Gao, W.; Liu, Q.; Huang, Z.; Yin, Y.; Bi, H.; Wang, M.-C.; Ma, J.; Wang, S.; and Su, Y. 2021. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 501–510.

Gao, W.; Liu, Q.; Yue, L.; Yao, F.; Lv, R.; Zhang, Z.; Wang, H.; and Huang, Z. 2025. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23923–23932.

Gao, W.; Liu, Q.; Yue, L.; Yao, F.; Wang, H.; Gu, Y.; and Zhang, Z. 2024. Collaborative Cognitive Diagnosis with Disentangled Representation Learning for Learner Modeling. *arXiv preprint arXiv:2411.02066*.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Han, D.; Liu, Q.; Lei, S.; Tong, S.; and Huang, W. 2024. HeckmanCD: Exploiting Selection Bias in Cognitive Diagnosis. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 768–777.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Little, R. J.; and Rubin, D. B. 2019. *Statistical analysis with missing data*. John Wiley & Sons.

Lord, F. M. 2012. *Applications of item response theory to practical testing problems*. Routledge.

Ma, H.; Li, M.; Wu, L.; Zhang, H.; Cao, Y.; Zhang, X.; and Zhao, X. 2022a. Knowledge-sensed cognitive diagnosis for intelligent education platforms. In *Proceedings of the 31st ACM international conference on information & knowledge management*, 1451–1460.

Ma, H.; Zhu, J.; Yang, S.; Liu, Q.; Zhang, H.; Zhang, X.; Cao, Y.; and Zhao, X. 2022b. A prerequisite attention model for knowledge proficiency diagnosis of students. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4304–4308.

Ma, J.; Guo, R.; Chen, C.; Zhang, A.; and Li, J. 2021. Deconfounding with networked observational data in a dynamic environment. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 166–174.

Marlin, B. M.; and Zemel, R. S. 2009. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*, 5–12.

Nabi, R.; Bhattacharya, R.; Shpitser, I.; and Robins, J. 2022. Causal and counterfactual views of missing data models. *arXiv preprint arXiv:2210.05558*.

Nissen, J.; Donatello, R.; and Van Dusen, B. 2019. Missing data and bias in physics education research: A case for using multiple imputation. *Physical Review Physics Education Research*, 15(2): 020106.

Pei, H.; Yang, B.; Liu, J.; and Dong, L. 2018. Group sparse bayesian learning for active surveillance on epidemic dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Saito, Y. 2020. Asymmetric tri-training for debiasing missing-not-at-random explicit feedback. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 309–318.

Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; and Joachims, T. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, 1670–1679. PMLR.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, 3076–3085. PMLR.

Shen, J.; Qian, H.; Liu, S.; Zhang, W.; Jiang, B.; and Zhou, A. 2024. Capturing Homogeneous Influence among Students: Hypergraph Cognitive Diagnosis for Intelligent Education Systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2628–2639.

Wang, F.; Gao, W.; Liu, Q.; Li, J.; Zhao, G.; Zhang, Z.; Huang, Z.; Zhu, M.; Wang, S.; Tong, W.; et al. 2024. A survey of models for cognitive diagnosis: New developments and future directions. *arXiv preprint arXiv:2407.05458*.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Chen, Y.; Yin, Y.; Huang, Z.; and Wang, S. 2020a. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6153–6161.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Yin, Y.; Wang, S.; and Su, Y. 2022. NeuralCD: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8): 8312–8327.

Wang, J.; Li, H.; Zhang, C.; Liang, D.; Yu, E.; Ou, W.; and Wang, W. 2023a. Counterclr: Counterfactual contrastive learning with non-random missing data in recommendation. In *2023 IEEE International Conference on Data Mining (ICDM)*, 1355–1360. IEEE.

Wang, S.; Zeng, Z.; Yang, X.; and Zhang, X. 2023b. Self-supervised graph learning for long-tailed cognitive diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 110–118.

Wang, X.; Zhang, R.; Sun, Y.; and Qi, J. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, 6638–6647. PMLR.

Wang, Z.; Chen, X.; Wen, R.; Huang, S.-L.; Kuruoglu, E.; and Zheng, Y. 2020b. Information theoretic counterfactual learning from missing-not-at-random feedback. *Advances in Neural Information Processing Systems*, 33: 1854–1864.

Xu, D.; Ruan, C.; Korpeoglu, E.; Kumar, S.; and Achan, K. 2020. Adversarial counterfactual learning and evaluation for recommender system. *Advances in Neural Information Processing Systems*, 33: 13515–13526.

Yang, S.; Chen, M.; Wang, Z.; Yu, X.; Zhang, P.; Ma, H.; and Zhang, X. 2025. DisenGCD: A Meta Multigraph-assisted Disentangled Graph Learning Framework for Cognitive Diagnosis. *Advances in Neural Information Processing Systems*, 37: 91532–91559.

Yang, S.; Ma, H.; Bi, Y.; Tian, Y.; Zhang, L.; Jin, Y.; and Zhang, X. 2024. An evolutionary multi-objective neural architecture search approach to advancing cognitive diagnosis in intelligent education. *IEEE Transactions on Evolutionary Computation*.

Yang, S.; Qin, L.; and Yu, X. 2024. Endowing interpretability for neural cognitive diagnosis by efficient kolmogorov-arnold networks. *arXiv preprint arXiv:2405.14399*.

Yang, S.; Wei, H.; Ma, H.; Tian, Y.; Zhang, X.; Cao, Y.; and Jin, Y. 2023a. Cognitive diagnosis-based personalized exercise group assembly via a multi-objective evolutionary algorithm. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(3): 829–844.

Yang, S.; Zhen, C.; Tian, Y.; Ma, H.; Liu, Y.; Zhang, P.; and Zhang, X. 2023b. Evolutionary multi-objective neural architecture search for generalized cognitive diagnosis models. In *2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, 1–10. IEEE.

Yu, X.; Qin, C.; Shen, D.; Ma, H.; Zhang, L.; Zhang, X.; Zhu, H.; and Xiong, H. 2024a. Rdgt: enhancing group cognitive diagnosis with relation-guided dual-side graph transformer. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 3429–3442.

Yu, X.; Qin, C.; Shen, D.; Yang, S.; Ma, H.; Zhu, H.; and Zhang, X. 2024b. Rigl: A unified reciprocal approach for tracing the independent and group learning processes. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4047–4058.

Yu, X.; Qin, C.; Zhang, Q.; Zhu, C.; Ma, H.; Zhang, X.; and Zhu, H. 2024c. DISCO: A Hierarchical Disentangled Cognitive Diagnosis Framework for Interpretable Job Recommendation. In *IEEE International Conference on Data Mining (ICDM) 2024*.

Yu, X.; Yang, S.; Li, J.; Wang, Z.; Qin, C.; Ma, H.; and Zhang, X. 2025a. Rethinking Learner Modeling: A Feedback-Centric Cognitive Disentanglement Perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 3657–3668.

Yu, X.; Yang, S.; Wang, Z.; Song, S.; Ma, H.; Cao, Z.; and Zhang, X. 2025b. LIGHT: Enhancing Learning Path Recommendation via Knowledge Topology-Aware Sequence Optimization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.