Collaborative Cognitive Diagnosis with Disentangled Representation Learning for Learner Modeling

Weibo Gao[1] Qi Liu[1,2]* Linan Yue[1] Fangzhou Yao[1] Hao Wang[1]
Yin Gu[1] Zheng Zhang[1]

[1] State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
weibogao@mail.ustc.edu.cn; qiliuql@ustc.edu.cn; {lnyue, fangzhouyao}@mail.ustc.edu.cn;
wanghao3@ustc.edu.cn; {gy128, zhangzheng}@mail.ustc.edu.cn

Abstract

Learners sharing similar implicit cognitive states often display comparable observable problem-solving performances. Leveraging collaborative connections among such similar learners proves valuable in comprehending human learning. Motivated by the success of collaborative modeling in various domains, such as recommender systems, we aim to investigate how collaborative signals among learners contribute to the diagnosis of human cognitive states (i.e., knowledge proficiency) in the context of intelligent education. The primary challenges lie in identifying implicit collaborative connections and disentangling the entangled cognitive factors of learners for improved explainability and controllability in learner Cognitive Diagnosis (CD). However, there has been no work on CD capable of simultaneously modeling collaborative and disentangled cognitive states. To address this gap, we present Coral, a Collaborative cognitive diagnosis model with disentangled representation learning. Specifically, Coral first introduces a disentangled state encoder to achieve the initial disentanglement of learners' states. Subsequently, a meticulously designed collaborative representation learning procedure captures collaborative signals. It dynamically constructs a collaborative graph of learners by iteratively searching for optimal neighbors in a context-aware manner. Using the constructed graph, collaborative information is extracted through node representation learning. Finally, a decoding process aligns the initial cognitive states and collaborative states, achieving co-disentanglement with practice performance reconstructions. Extensive experiments demonstrate the superior performance of Coral, showcasing significant improvements over state-of-the-art methods across several real-world datasets. Our code is available at `https://github.com/bigdata-ustc/Coral`.

1 Introduction

It is a common notion that individuals with similar implicit states frequently exhibit similar explicit behaviors. Therefore, establishing interconnections among similar users is crucial for understanding human behaviors. For instance, social connections play a pivotal role in understanding current consumer preferences and predicting future behaviors [42]. Similarly, in the
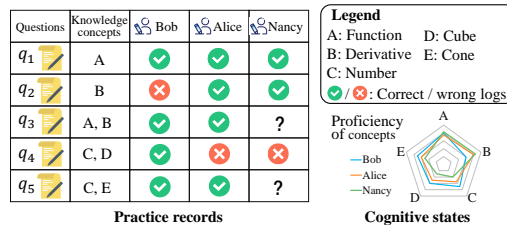

Figure 1: An example of human learning, where learners individually select questions to practice. Each question tests at least one knowledge concept.

---

*Indicates Corresponding Author.

context of intelligent education, a better modeling of like-minded learners with similar learning experiences, is essential for understanding the human learning process [31], analyzing their knowledge proficiency and facilitating personalized tutoring tailored to individual needs [55].

As illustrated in Figure 1, we can infer Nancy is likely to answer the *Cone*-related question q_5 correctly according to the correct practice responses of Bob and Alice, who share similar learning behaviors with Nancy. The underlying psychological assumption is that learners with similar experiences generally possess similar cognitive states — how well the learner masters each knowledge concept, influencing their subsequent responses. To gain a deeper understanding of the human learning process, it is crucial to explicitly diagnose unobservable cognitive states. Existing Cognitive Diagnosis (CD) methods seek to enhance diagnostic accuracy by fully utilizing the inner-learner information (i.e., individual attributions [41] and explicit practice records) and question-side features (e.g., difficulty [16], textual content [26], and educational relations [12, 14]). However, the issue of how similar (a.k.a. collaborative) connections among inter-learners with similar states facilitate understanding of learners' knowledge proficiency remains largely unexplored.

In this study, to efficiently harness collaborative information among similar learners and thereby more accurately diagnose the cognitive states of each individual, we advocate for the incorporation of inter-learner connections into the CD process. However, designing a collaborative CD model in educational scenarios presents two distinct challenges due to the complexity of human learning.

- First, acquiring explicit collaborative connections among learners proves to be a formidable challenge. On the one hand, unlike many well-defined social scenarios (e.g., *Twitter.com*), where user preference similarities are manifested through explicit social actions such as following and liking, the directly available social behaviors among learners in learning environments (e.g., *LeetCode.com*) cannot be used for diagnosis modeling since these social attributes cannot reflect true cognitive-oriented connections. On the other hand, some related studies [28, 13] attempt to design different similarity functions based on practice data to compute cognitive similarities among learners. However, these approaches pose a significant challenge of manually selecting appropriate metrics and corresponding thresholds, introducing additional inductive biases. Although various methods for constructing user relationships have been proposed in other domains [22, 9], these approaches do not consider the domain-specific attributes of students in learning scenarios and cannot be directly applied in educational contexts.

- Second, an ideal collaborative diagnosis procedure requires disentangling and uncovering the mixed explanatory latent factors hidden in the observed learning behaviors. The basic motivation is that learners demonstrate complex and diverse patterns driven by entangled states across both inner- and inter-learner perspectives. For instance, from an inner perspective, Nancy may not master *Cone* since she does not practice *Cone*-related questions. However, based on inter-learner data, one can infer a high probability that she has mastered *Cone*. Most prior attempts can not fulfill this requirement since they learn representations in an entangled way. Although recent models [8, 50] achieve a dimension-level disentanglement of cognitive states, they lack consideration of modeling the influence of collaborative connections, ignoring the complex relations between inner- and inter-learner connections of different individuals. Thereby, it needs to find a suitable way to achieve the co-disentanglement from both the inner- and inter-learner views for cognitive representations with higher interpretability and controllability.

To tackle the above challenges, we propose Coral, a <u>Co</u>llabo<u>ra</u>tive cognitive diagnosis model with disentang<u>l</u>ed representation learning, to reveal learner cognitive states while simultaneously modeling both inner- and inter-learner learning information. Specifically, our approach begins with the disentangled cognitive representation encoding to establish initial disentangled learner states through reconstructing their practice performance from the inner-learner perspective. Next, our focus shifts to effectively learning collaborative cognitive representations from the inter-learner perspective. The most significant point is to find the implicit collaborative relations between learners. To address this challenge, we present a context-aware collaborative graph learning mechanism that automatically explores all K-optimal neighbors for each learner given their basic cognitive states to facilitate the explicit modeling of collaborative connections among learners. Based on the constructed graph, collaborative information can be effectively fused into disentangled learner cognitive states through learning collaborative node representations. Finally, a decoding and reconstruction process is conducted to merge initial states and collaborative states so as to achieve co-disentanglement from both the inner- and inter-learner perspectives. Extensive experiments demonstrate the superior performance of Coral, showing significant improvements over SOTA methods across several datasets.

2

2 Related Work

Cognitive Diagnosis As a fundamental task, cognitive diagnosis (CD) has been well-researched for decades in educational psychology [25, 4, 53]. It aims to profile the implicit cognitive states (i.e., the proficiency of specific knowledge concepts) of learners by exploiting observed practice records (e.g., correct or wrong). Existing research on CD assumes that learners' knowledge proficiency is proportional to their practice performance and thus can be diagnosed through predicting their practice performance [12]. Since the diagnostic results can be applied to many intelligent applications, such as exercise recommendation [18] and learning path suggestions [55], many CD models have been proposed in recent years. The early works from psychology like IRT [16] and MIRT [1] focus on modeling learners' answering process by predicting the probability of a learner answering a question correctly, which utilizes latent factors as the learner's ability. These methods lack interpretability, i.e., they are inability to output explicit multidimensional diagnostic results on each knowledge concept. To achieve better interpretability, later diagnostic models focus on incorporating knowledge concepts of questions to diagnose learners' proficiency on all knowledge concepts [38, 45, 46, 32]. Representative NCDM [40] adopts neural networks to model non-linear interactions instead of handcrafted interaction functions in previous works (e.g., IRT, and MIRT). In summary, existing CD studies enhance diagnostic accuracy by fully utilizing the inner-learner information (i.e., individual attributions and explicit practice records) [41, 50]and question-side features (e.g., difficulty [16, 38], textual content [26], and educational relations [12, 14, 8]). However, to the best of our knowledge, the problem of collaborative diagnostic modeling remains largely unexplored.

Collaborative modeling in Education Collaborative connections among learners in the education context commonly refer to learners with similar explicit practice behaviors, testing scores and implicit knowledge proficiency [28, 54, 44]. However, due to the complexity and implicitness of the human learning process, these relations are commonly not explicitly and directly available. Existing studies [28, 13] in AI Education have attempted to design different similarity functions based on practice data to compute cognitive similarities among learners. However, these methods pose a significant challenge of manually selecting appropriate metrics and corresponding thresholds, introducing additional inductive biases.

Disentangled Representation Learning Disentangled Representation Learning (DRL)[3], which aims to produce robust, controllable, and explainable representations, has become one of the core problems in machine learning. Typical methods include variational method [20], weakly supervised models [21], as well as the recent combination with the diffusion model [6]. DRL has a wide range of applications in user modeling to disentangle attributes. For example, recommendation with several aspects of users' interests [24, 33], fair user representation to disentangle sensitive attributes [10]. In education, DCD [8] attempts to disentangle learners' cognitive representations via variational framework, which motivates us to conduct a further study on collaborative CD setups.

3 Coral

We first introduce the problem setup, followed by details on three core components of Coral: i) Disentangled Cognitive Representation Encoding, ii) Collaborative Representation Learning and iii) Decoding and Reconstruction. Figure 2 shows the framework. The algorithm is listed in Algorithm 1.

3.1 Problem Setup

Our setup considers the human learning dataset D including the practice records between M learners and N questions. The practice records of each learner u is denoted by $\mathbf{x}_u = \{x_{u,i}\}$, where $x_{u,i}$ equals 1 or 0, representing that learner u answered question i correctly or not, respectively. Each question is related to at least one knowledge concept. The association relations between N questions and C knowledge concepts is represented by $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^N$, where $\mathbf{c}_i \in \mathbb{R}^C$ and $c_{i,c}$ equals 1 or 0 denoting that question i is related to concept c or not. The practice records are regarded as the explicit inner-learning information in our context.

Besides, we consider the collaborative connections among learners with similar cognitive states, which provide the inter-learner information. We define collaborative connections as a graph structure $G = (V, E)$ which contains a set of nodes (i.e., learners) V and a set of edges E where $(u, v) \in E$ or $(u, v) \in G$ indicates that the existence of a collaborative connection between learner u and v (i.e., u and v have similar latent cognitive states). Notably, the collaborative connections in educational
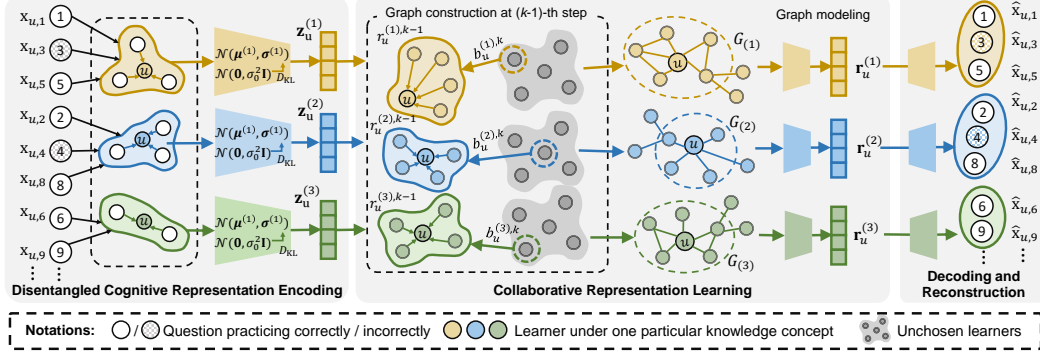
Figure 2: The overall framework of Coral.

scenarios are generally not explicitly or directly available, and it needs to design an adaptive strategy to automatically infer similar learners from observed learning data during the training process.

To achieve cognitive state disentanglement, we initially assign C factorized representations to each learner, i.e., $\mathbf{z}_u = [\mathbf{z}_u^{(1)}; \mathbf{z}_u^{(2)}; \ldots; \mathbf{z}_u^{(C)}] \in \mathbb{R}^{d \times C}$ with Gaussian Mixture initialization since the Gaussian distribution has long been recognized as a proper statistic model for the cognitive states of learners in educational psychology [4]. The component $\mathbf{z}_u^{(c)}$ is expected to capture the learner u's cognitive state over knowledge concept c. We denote Θ as the set of trainable parameters for the proposed model. Based on the above setups, the goal of Coral is to learn co-disentangled representations $\tilde{\mathbf{Z}} = \{\tilde{\mathbf{z}}_u\}_{u=1}^{M}$ for the M learners from both the inner-learner practice perspective and inter-learner collaborative perspective.

3.2 Disentangled Cognitive Representation Encoding

The practice response $\mathbf{x}_u$ of each learner u provides valuable inner-learner insights regarding his/her proficiency since learners' performance on each question is assumed to be proportional to their cognitive proficiency on question-related knowledge concepts [8]. Therefore, we implement an encoder for encoding the disentangled cognitive state $\mathbf{z}_u$ of each learner u by reconstructing their practice responses. For a learner u, we assume that his/her practice performance on candidate questions can be generated from the following distribution:

$$p_\Theta(\mathbf{x}_u) = \mathbb{E}_{p(\mathbf{C})}\left[\int p_\Theta(\mathbf{x}_u \mid \mathbf{z}_u, \mathbf{C}) p_\Theta(\mathbf{z}_u) d\mathbf{z}_u\right], \tag{1}$$

where $p(\mathbf{C}) = p_D(\mathbf{C})$ and $p_\Theta(\mathbf{x}_u \mid \mathbf{z}_u, \mathbf{C})$ is naturally a cognitive diagnosis procedure to predict practice performance. The key point of this task is to learn an optimal encoder $p_\Theta(\mathbf{z}_u)$ via practice records $\mathbf{x}_u$ to encode the cognitive state $\mathbf{z}_u$ of each learner u. To optimize Θ, we introduce a variational distribution $q_\Theta(\mathbf{z}_u \mid \mathbf{x}_u)$ to approximate $p_\Theta(\mathbf{z}_u)$, following the VAE literature [3], through maximizing a lower bound of $\log p_\Theta(\mathbf{x}_u)$ based on the following property.

Property 1. $\max \log p_\Theta(\mathbf{x}_u)$ *is bounded as follows:*

$$\log p_\Theta(\mathbf{x}_u) \geq \mathbb{E}_{p(\mathbf{C})q_\Theta(\mathbf{z}_u \mid \mathbf{X}_u)}[\log p_\Theta(\mathbf{x}_u \mid \mathbf{z}_u)] - \mathbb{E}_{p(\mathbf{C})}[D_{\mathrm{KL}}(q_\Theta(\mathbf{z}_u \mid \mathbf{x}_u) \parallel p_\Theta(\mathbf{z}_u))]. \tag{2}$$

See the Appendix A for the proof.

In Property 1, the first term reconstructs the true practice performance $\mathbf{x}_u$ of learner u and the variational encoder $q_\Theta(\mathbf{z}_u \mid \mathbf{x}_u)$ in the second term approximates the true encoder $p_\Theta(\mathbf{z}_u)$ by minimizing the KL divergence D_{KL}. The variational distribution $q_\Theta(\mathbf{z}_u \mid \mathbf{x}_u)$ and the expectation $\mathbb{E}_{q_\Theta(\mathbf{z}_u \mid \mathbf{X}_u)}$ are intractable, thus we employ the re-parameterization trick [20] for the model optimization.

Furthermore, the diagnosis procedure $p_\Theta(\mathbf{x}_u \mid \mathbf{z}_u, \mathbf{C})$ is achieved by estimating how well a learner u answers question i from both the perspectives of cognitive states and comprehensive abilities. From the perspective of cognitive states, solving question i requires learner u to master all knowledge concepts related to this question. Regarding comprehensive abilities, each learner possesses a latent state reflecting their overall learning ability, which is shared when addressing different questions.

Formally, this process can be described as:

$$p_\Theta\left(\mathbf{x}_u \mid \mathbf{z}_u, \mathbf{C}\right) = \prod_{x_{u,i} \in \mathbf{x}_u} p_\Theta\left(x_{u,i} \mid \mathbf{z}_u, \mathbf{C}\right),$$

$$p_\Theta\left(x_{u,i} \mid \mathbf{z}_u, \mathbf{C}\right) = \sum_{c=1}^{C} c_{i,c} \cdot \phi_\Theta\left(\theta_u \cdot \mathbf{z}_u^{(c)} - \mathbf{h}_i\right), \theta_u = \sum_{c=1}^{C} \psi_\Theta\left(\mathbf{z}_u^{(c)}\right), \tag{3}$$

where $\phi_\Theta(\cdot)$ and $\psi_\Theta(\cdot) : \mathbb{R}^d \to \mathbb{R}^+$ are two shallow neural networks. $\psi_\Theta(\cdot)$ estimates the comprehensive ability of the learner and $\phi_\Theta(\cdot)$ predicts the performance of a learner with a given cognitive state $\mathbf{z}_u^{(c)}$ and a comprehensive ability θ_u over question i in terms of concept c. $\mathbf{h}_i$ is a learnable latent representation for question i. Besides, to ensure psychometric interpretability of prediction, we set the weights of $\phi_\Theta(\cdot)$ are positive values, i.e., $\partial\phi_\Theta(\cdot)/\partial\mathbf{z}_u > 0$, assuming that the probability of correctly answering the question monotonically increases with learners' cognitive state. Please note that we found that the mean operation here can also be replaced with a neural network (i.e., $\phi'_\Theta : \mathbb{R}^C \to \mathbb{R}^+$) with positive weights, formulated as $\phi'_\Theta\left(\mathbf{c}_i \cdot \left(\theta_u \cdot \mathbf{z}_u^{(c)} - \mathbf{h}_i\right)\right)$, as in [40], without affecting prediction performance. Particularly, in contrast to most methods that consider entangled cognitive factors as input, our diagnosis model can better capture learners' proficiency on each knowledge concept by disentangling cognitive states under each concept.

Furthermore, inspired by the outstanding performance of β-TCVAE [7] in disentanglement, we prompt statistical independence among its dimensions to obtain a better trade-off between the reconstruction accuracy and the quality of disentangled representation through $q(\mathbf{z}_u^{(c)}) = \prod_{j=1}^{d} q_\Theta\left(z_{u,j}^{(c)}\right)$ where $q_\Theta(\mathbf{z}_u^{(c)})$ is the aggregated posterior of $\mathbf{z}_u$, i.e., $q_\Theta(\mathbf{z}_u^{(c)}) = \int q_\Theta(\mathbf{z}_u^{(c)} \mid \mathbf{x}_u) p(\mathbf{x}_u) \, d\mathbf{x}_u$ where $p(\mathbf{x}_u) = p_{data}(\mathbf{x}_u)$. This setup is encouraged by the term $D_{\mathrm{KL}}(\cdot)$ in Eq. (2) based on Property 2.

Property 2. *The $D_{\mathrm{KL}}(\cdot)$ in Eq. (2) can be rewritten as:*

$$D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u\right)\right) = I\left(\mathbf{z}_u, \mathbf{x}_u\right) + D_{\mathrm{KL}}\left(q_\Theta(\mathbf{z}_u) \| p_\Theta(\mathbf{z}_u)\right). \tag{4}$$

See Appendix A for the proof. On one hand, $I(\mathbf{z}_u, \mathbf{x}_u)$ maximizes the mutual information (MI) between $\mathbf{z}_u$ and $\mathbf{x}_u$ which obtains the useful information for the diagnosis task as much as possible according to the information bottleneck theory [2]. On the other hand, given a Gaussian distribution $p_\Theta(\mathbf{z}_u^{(c)}) = \prod_{j=1}^{d} p_\Theta\left(z_{u,j}^{(c)}\right)$, the KL divergence term encourages independence among the dimensions of $\mathbf{z}_u^{(c)}$ by preventing each latent variable from deviating too far from specified priors. Compared to prior VAE-based CD models [50, 8], Coral additionally considers ability parameters from psychology [16] to enhance the expressive power of disentangled cognitive states.

Overall, we penalize Eq. (2) by a Lagrange multiplier β resulting in the following objective:

$$\log p_\Theta\left(\mathbf{x}_u\right) \geq \mathbb{E}_{p(\mathbf{C})q_\Theta(\mathbf{z}_u \mid \mathbf{X}_u)}\left[\log p_\Theta\left(\mathbf{x}_u \mid \mathbf{z}_u\right)\right] - \beta \cdot \mathbb{E}_{p(\mathbf{C})}\left[D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u\right)\right)\right]. \tag{5}$$

3.3 Collaborative Representation Learning

Collaborative information among similar learners provides an auxiliary inter-learner insight for cognitive representation learning. However, collaborative connections among learners with similar states are typically not readily accessible. To address this challenge, we design a context-aware graph construction strategy that searches similar neighbors automatically via the initial disentangled cognitive states. Based on the constructed collaborative graph, we can learn collaborative node representations by aggregating collaborative signals to generate collaborative cognitive states.

3.3.1 Context-aware Collaborative Graph Learning

The core goal of constructing the collaborative graph is to find K optimal neighbors for each learner node in V via their initial disentangled cognitive state $\{\mathbf{z}^{(c)}\}_{c=1}^{C}$. For different knowledge concepts, the cognitive connections between the same learner pair are typically different. Thereby, it needs to search C groups of similar neighbors for each learner via each disentangled component $\mathbf{z}^{(c)}$. This means that we would generate C collaborative graphs, i.e., $G = \{G_{(c)}\}_{c=1}^{C}$. Each collaborative

graph $G_{(c)}$ under concept c is expected to characterize the cognitive similarities of learners regarding concept c. Formally, this task is defined as computing $p_\Theta(G \mid V, \mathbf{Z})$ by identifying all the K similar neighbors for each learner covering each concept c. Let $\mathcal{N}_u^{(c)}$ denote the set of K similar neighbors for the learner u, the task can be described as:

$$\max \log p_\Theta(G \mid V, \mathbf{Z}) := \max \sum_{c=1}^{C} \mathbb{E}_{p_\Theta\left(\mathcal{N}_u^{(c)}, \mathbf{z}_u^{(c)}\right)} \left[\log p_\Theta\left(\mathcal{N}_u^{(c)} \mid \mathbf{z}_u^{(c)}\right) \right]$$

$$= \max \sum_{c=1}^{C} I\left(\mathcal{N}^{(c)}; \mathbf{Z}^{(c)}\right) + \sum_{c=1}^{C} \mathbb{E}_{p_\Theta\left(\mathbf{Z}^{(c)}\right)} \left[\log p_\Theta\left(\mathbf{Z}^{(c)}\right) \right] \geq \max \sum_{c=1}^{C} I\left(\mathcal{N}^{(c)}; \mathbf{Z}^{(c)}\right), \quad (6)$$

where $\mathcal{N}^{(c)}$ and $\mathbf{Z}^{(c)} = \{\mathbf{z}_u^{(c)}\}_{u=1}^{M}$ are the neighbor set and feature set of all the learners regarding knowledge concept c, respectively. The number of $\mathcal{N}_u^{(c)}$ equals the combination of arbitrary K neighbors from all M learners for each learner node u under each concept c, i.e., $\left|\mathcal{N}_u^{(c)}\right| = \frac{M!}{K!(M-K)!}$, thus Eq. (6) is computationally expensive especially for larger M and K. To facilitate computation, we transform the Eq. (6) that requires global MI maximization to the task of maximizing MI locally via locally available context information inspired by [22] and derive a lower bound of it as the following Property 3.

Property 3. $\max \log p_\Theta(G \mid V, \mathbf{Z})$ *is bounded as follows:*

$$\max \log p_\Theta(G \mid V, \mathbf{Z}) \geq -\sum_{c=1}^{C} \sum_{u=1}^{M} \sum_{k=1}^{K} \mathcal{L}_u^{(c),k}, \text{ where } \mathcal{L}_u^{(c),k} = -\frac{\exp\left(f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right)\right)}{\sum_{v \in V_u^{(c)}} \exp\left(f_{(c)}\left(v; r_u^{(c),k-1}\right)\right)}.$$
$$(7)$$

See the Appendix A for the proof. The Eq. (7) iteratively searches K neighbors for the learner u under each knowledge concept c from step $k = 1$ to K. $\mathcal{L}_u^{(c),k}$ is the well-known InfoNCE loss function [36]. Let $r_u^{(c),k-1}$ denote the current context at step $(k-1)$ (i.e., the set of $(k-1)$ neighbors selected from step 1 to $(k-1)$). $b_u^{(c),k}$ is the affinity candidate learner in the $(M-k)$ nonneighbor learners. Let $V_u^{(c)}$ denote the current set of nonneighbor learners, and we hence have $b_u^{(c),k} \in V_u^{(c)}$. $f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right)$ is a matching function measuring the similarity between of nonneighbor $b_u^{(c),k}$ and the current context $r_u^{(c),k-1}$, where the higher the scalar score means the higher likelihood of $b_u^{(c),k}$ is a new neighbor.

Furthermore, we have $\mathcal{L}_u^{(c),k} \propto f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right)$. Thus, given the context of $(k-1)$ neighboring learners (i.e., we have found $(k-1)$ neighbors for the learner u) and matching function $f_{(c)}(\cdot)$, our **goal** following the Property 3 is to find a learner $b_u^{(c),k}$ from nonneighbor set $V_u^{(c)}$ that can maximize the matching score $f_{(c)}(\cdot)$ as the k-th neighbor of u. In other words, $p(G \mid V, \mathbf{Z})$ can be optimized through maximizing the matching score $f_{(c)}(\cdot)$ from $k = 1$ to K iteratively. Thereby, at each step k, we sort the scores of the nonneighbor learners and select the learner with the highest score to label as k-th neighbor $b_u^{(c),k}$, i.e., $b_u^{(c),k} \leftarrow \arg\max_v f_{(c)}(v; r_u^{(c),k-1}), v \in V_u^{(c)}$. After obtaining the k-th neighbor $b_u^{(c),k}$, the context $r_u^{(c),k-1}$ is updated to $r_u^{(c),k}$ by absorbing $b_u^{(c),k}$.

The calculation of matching score $f_{(c)}(\cdot)$ usually relies on the node representations (i.e., learner cognitive states). However, the sub-optimal cognitive state learning during the initial training epochs probably results in the matching function exhibiting biases. To enhance the stability of model training, instead of directly aggregating node representations as the context $r_u^{(c),k-1}$ as many graph learning works, we denote it using relative representations w.r.t. the learner u [22]. Without loss of generality, we first establish relative collaborative coordinate systems with learner node u as the origin, and process relationship measurements between node u and each of its neighbors v as $\mathbf{z}_{u,v}^{(c)} = \|\mathbf{z}_u^{(c)} - \mathbf{z}_v^{(c)}\|_2$. Then the context-aware features can be generated by aggregating each node in the context $r_u^{(c),k-1}$, i.e., $\mathbf{rc}_u^{(c),k-1} = \sum_{v \in r_u^{(c),k-1}} \mathbf{z}_{u,v}^{(c)}$. Thereby, let v_k denote $b_u^{(c),k}$ with feature $\mathbf{z}_{v_k}^{(c)}$, we have $f_{(c),v_k} = f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right) = \mathbf{z}_{v_k}^{(c)\mathrm{T}} \cdot \mathbf{rc}_u^{(c),k-1}$.

3.3.2 Collaborative Graph Modeling

After iteratively searching K neighbors under each concept, we can obtain C collaborative graphs regarding each learner, i.e., $\{G_{(c)}\}_{c=1}^{C}$. Then, we consider collaborative modeling as a node representation learning task within each collaborative graph $G_{(c)}$. It relies on a nonlinear kernel function $\varphi_{\Theta}(\cdot)$ to aggregate neighboring information and update each disentangled cognitive state, i.e., $\mathbf{r}_{u}^{(c)} = \varphi_{\Theta}(\mathbf{z}_{u}^{(c)}, \{\mathbf{z}_{v}^{(c)} : (u,v) \in G_{(c)}\})$. Given the disentangled learner cognitive states generated by the variational posterior distribution $q_{\Theta}(\mathbf{z}_{u}|\mathbf{x}_{u})$ from Property 1, $\varphi_{\Theta}(\cdot)$ is naturally expected to contain C channels to extract different concept features from similar learners, though projecting the representation $\mathbf{z}_{u}$ into different subspaces, i.e., $\hat{\mathbf{z}}_{u}^{(c)} = \sigma(\mathbf{W}_{(c)}^{\mathrm{T}}\mathbf{z}_{u}^{(c)} + b_{(c)})/\|\sigma(\mathbf{W}_{(c)}^{\mathrm{T}}\mathbf{z}_{u}^{(c)} + b_{(c)})\|_{2}$, where $\mathbf{W}_{(c)} \in \mathbb{R}^{d}$ and $b_{(c)} \in \mathbb{R}^{d}$ are learnable parameters of channel c and $\sigma(\cdot)$ is a nonlinear activation function (e.g., Sigmoid), and $\|\cdot\|_{2}$ is L_{2} normalization ensuring numerical stability. Then the collaborative learner representation modeling in terms of concept c can be described as:

$$\mathbf{r}_{u}^{(c)} = \frac{1}{|\mathcal{N}_{u}^{(c)}|} \sum_{v \in \mathcal{N}_{u}^{(c)}} s_{u,v}^{(c)} \cdot \hat{\mathbf{z}}_{v}^{(c)}, \; s_{u,v}^{(c)} = \frac{\hat{\mathbf{z}}_{u}^{(c)\mathrm{T}} \cdot \hat{\mathbf{z}}_{v}^{(c)}}{\sum_{j \in \mathcal{N}_{u}^{(c)}} \hat{\mathbf{z}}_{u}^{(c)\mathrm{T}} \cdot \hat{\mathbf{z}}_{j}^{(c)}} + \frac{f_{(c),v}}{\sum_{k=1}^{K} f_{(c),v_{k}}}, \tag{8}$$

where $s_{u,v}^{(c)}$ is the attention weight between u and v, considering both the collaborative aggregation (the first term) commonly used in graph modeling works and the corresponding context-aware attention (the second term) calculated in the iterative graph construction process in Eq. (7). When K is set large in Eq. (7), there is a possibility of introducing non-collaborative noise. In such cases, $s_{u,v}^{(c)}$ can assign lower values to non-collaborative neighbors to mitigate the negative impact of noise, allowing for the adaptive tuning of attention in graph modeling. During training, the channels will remain changing because different subsets of the neighborhood will be searched for dynamically aggregating neighbor information in different iterations.

With Gaussian Mixture initialization from the Disentangled Cognitive Representation Encoding (section 3.2), we derive the theorem on convergence as:

Theorem 1. The Collaborative Representation Learning (section 3.3) procedure is equivalent to an expectation-maximization (EM) algorithm [35] for the mixture model. In particular, it converges to a point estimate of $\{\mathbf{r}_{u}^{(c)}\}_{c=1}^{C}$ that maximizes the marginal likelihood $l\left(\left\{a_{v}^{(c)} : (u,v) \in G_{(c)}\right\}_{c=1}^{C} ; \{\mathbf{r}_{u}^{(c)}\}_{c=1}^{C}\right)$, where $a_{u,v}^{(c)}$ equals 1 or 0 denoting whether learner v is a collaborative neighbor of learner u regarding concept c or not. See the Appendix A for the proof.

3.4 Decoding and Reconstruction

Given the initial disentangled encoding via inner-learner information (section 3.2) and the collaborative representation learning via inter-learner information (section 3.3), this part encourages an alignment between the initial encode $\mathbf{z}_{u}$ and collaborative state $\mathbf{r}_{u}$, formulating a co-disentangled representation as $\tilde{\mathbf{z}}_{u} = \mathbf{z}_{u} + \mathbf{r}_{u}$. This operation is inspired by the residual block [17] to address the second challenge, where $\mathbf{r}_{u}$ can be treated as a disentangled auxiliary information of $\mathbf{z}_{u}$ from collaborative graphs.

The decoding process predicts the practice performance of each learner u on candidate questions, given her co-disentangled representation $\tilde{\mathbf{z}}_{u} = \left[\tilde{\mathbf{z}}_{u}^{(1)}, \tilde{\mathbf{z}}_{u}^{(2)}, \ldots, \tilde{\mathbf{z}}_{u}^{(C)}\right]$, i.e., $p_{\Theta}(\hat{\mathbf{x}}_{u}) = \mathbb{E}_{p_{\Theta}(\mathbf{C})}[p_{\Theta}(\hat{\mathbf{x}}_{u} \mid \tilde{\mathbf{z}}_{u}, \mathbf{C})]$, similar to the reconstruction procedure in Eq. (1). Thus, putting Eq. (5) and Eq. (7) together, we have the overall training objective:

$$\arg\min \mathcal{L} = \sum_{u=1}^{M} \Big[\sum_{x_{u,i} \in \mathbf{x}_{u}} \alpha \cdot BCE\left(x_{u,i}, p_{\Theta}\left(x_{u,i} \mid \mathbf{z}_{u}, \mathbf{C}\right)\right) - \beta \cdot D_{\mathrm{KL}}^{u}$$

$$+ \sum_{x_{u,i} \in \mathbf{x}_{u}} BCE\left(x_{u,i}, p_{\Theta}\left(\hat{x}_{u,i}\right)\right) \Big], \tag{9}$$

$$\text{s.t.} \quad \arg\max \sum_{c=1}^{C} \sum_{k=1}^{K} \mathcal{L}_{u}^{(c),k},$$

Datasets	ASSIST	Junyi	NeurIPS2020EC
#students	1,256	1,400	1,000
#questions	16,818	674	919
#knowledge concepts	120	40	30
#concepts per exercise	1.21	1	4.02
#records	199,790	70,797	331,187
#records per student	159,07	50.67	331.19
#correct records / #incorrect records	67.08%	77.20%	53.87%

Table 1: The statistics of three datasets.

where $BCE(\cdot, \cdot)$ is the binary cross entropy loss function between ground-truth practice behaviors $\mathbf{x}_u$ and the reconstructed ones. α and β are hyper-parameters.

By optimizing with minimizing the above loss function Eq. (9), the cognitive state $\tilde{\mathbf{z}}_u$ of each learner u can be jointly refined serving as the diagnostic results. During the testing phase, we evaluate the model performance by matching the difference between the predicted score $p_\Theta(\hat{x}_{u,i})$ and the true score $\mathbf{x}_u$. Specifically, when a proficiency value is required instead of the vector $\tilde{\mathbf{z}}_u^{(c)}$, we can obtain it by averaging each dimension of $\tilde{\mathbf{z}}_u^{(c)}$.

4 Experiments

We empirically evaluate the performances of the proposed Coral model over three real-world datasets and conduct several experiments to prove its effectiveness.

4.1 Experimental Setup

Datasets We conduct experiments on three real-world datasets: ASSIST [11], Junyi [5] and NeurIPS2020EC [43]. The statistics of datasets are listed in Table 1. The details about datasets and preprocessing are depicted in the Appendix C.

Baselines The baselines include the matrix factorization-based model, i.e., PMF [34], the typical latent factor models derived from educational psychology, including IRT [16], MIRT [1], and the neural networks-based models, including NCDM [40], RCD [12], KaNCD [41] and DCD [8].

Evaluation Since cognitive states cannot be directly observed in practice, it is common to indirectly evaluate CDMs through the student performance prediction task on test datasets [4]. To evaluate prediction performance, we adopt ACC and AUC and F1-score as metrics from the perspective of classification, using a threshold of 0.5, and RMSE as metrics from the perspective of regression, following previous work [12, 23].

Settings We set the dimension size d as 20, the layer of graph modeling as 2, and the mini-batch size as 512. In the training stage, we select the learning rate from $\{0.002, 0.005, 0.01, 0.02, 0.05\}$, select α from $\{0.05, 0.1, 0.5, 1\}$ and β from $\{0.25, 0.5, 1\}$, and select neighboring number K from $\{1, 2, 3, 4, 5, 10, 15, 20, 15, 30, 25, 40, 45, 50\}$. All network parameters are initialized with Xavier initialization [15]. Each model is implemented by PyTorch [37] and optimized by Adam optimizer [19]. Specially, for the implementation of baselines, we set the dimensional sizes of each representation in PMF, NCDM, KaNCD, RCD and DCD as the number of knowledge concepts. All experiments are conducted on a Linux server equipped with two 3.00GHz Intel Xeon Gold 5317 CPUs and two Tesla A100 40G GPUs.

4.2 Experimental Results

Prediction Comparison We evaluate prediction performance of Coral against baselines under three setups: normal, sparse, and cold-start scenarios.

Table 2 reports the performance comparison under normal settings for all the models across three datasets on four evaluation metrics. In this setting, we split all the datasets with a 7:1:2 ratio into training sets, validation sets, and test sets. The proposed Coral model significantly outperforms most baselines. This demonstrates two key benefits of Coral. First, the iterative graph construction process effectively generates collaborative connections for modeling. Second, the co-disentangled representation learning successfully discovers disentangled cognitive states for each learner.

We extend our analysis to assess the performance of Coral in sparse scenarios. In order to simulate varied sparse environments, we systematically discard 80%, 60%, 40%, and 20% of the training data from the ASSIST dataset under the normal settings described above. The experimental results

shown in Figure 3 (a) reveal that Coral consistently outperforms baselines across a range of sparse environments. Moreover, our model exhibits robust performance consistently, demonstrating its adaptability and effectiveness in diverse sparse scenarios.

Moreover, we conduct an analysis of Coral's performance in a cold-start environment. To replicate this scenario, we retain solely the cold-start response data for each learner in the test set of Junyi, corresponding to the knowledge concepts they had not previously practiced in the training set. The Figure 3 (b) illustrates the experimental results, highlighting the exceptional performance of the Coral model (with $K = 10$) in a cold-start scenario.

Collaborate Graph Learning We investigate the influence of the generated neighbor number K on diagnostic performance. Figure 3 (c) displays the prediction performances for various values of K on Junyi under the normal scenario. The model performance exhibits improvement as K increases, particularly noticeable when K is small. This observation suggests that the inter-learner information automatically retrieved by Coral contributes positively to the model. However, once K surpasses a threshold, the performance gain becomes less pronounced. This diminishing effect arises because users beyond the threshold (i.e., $K = 10$ in this dataset) may lack significant collaborative relationships, thus limiting the useful clues they can offer. We observe that when K exceeds the threshold, the model's

Dataset	Method	Metric			
		ACC ↑	AUC ↑	F1-score ↑	RMSE ↓
ASSIST	IRT	69.36	69.81	78.14	45.61
	MIRT	71.26	72.59	79.80	44.50
	PMF	71.34	72.27	80.68	48.67
	NCDM	72.27	74.27	79.97	48.67
	KaNCD	**72.43**	**75.38**	80.22	48.67
	RCD	72.04	73.14	80.60	43.74
	DCD	70.33	73.98	79.09	43.94
	Coral	71.53	74.72	**81.16**	**43.66**
Junyi	IRT	79.26	76.46	87.54	38.38
	MIRT	77.74	74.46	86.05	40.29
	PMF	79.65	77.17	88.18	44.10
	NCDM	79.91	78.91	87.73	38.35
	KaNCD	**81.79**	80.93	89.02	36.11
	RCD	81.02	80.22	88.00	37.23
	DCD	79.29	79.55	87.62	37.83
	Coral	81.15	**80.94**	**89.12**	**36.08**
NeurIPS2020EC	IRT	70.11	75.60	71.59	44.68
	MIRT	69.95	75.52	71.24	45.51
	PMF	69.85	75.39	72.62	48.33
	NCDM	71.66	78.57	71.36	43.21
	KaNCD	71.28	77.60	72.50	43.71
	RCD	70.43	77.25	72.64	44.01
	DCD	71.53	75.63	71.13	45.60
	Coral	**71.72**	**78.88**	**72.82**	**43.20**

Table 2: Performance comparison. The best performance is highlighted in **bold**. ↑ (↓) means the higher (lower) score the better (worse) performance, the same as below.

performance remains acceptable, and even the performance improves after K exceeds 30. This indicates that Coral effectively perceives the similarity functions of the scenario and collaborative context. Consequently, it assigns lower similarity scores to non-collaborative neighbors, robustly adjusting the attention weight in graph modeling.

To obtain a more intuitive insight into the iterative graph construction process, we randomly select two learners (called target learners) from the Junyi dataset (with the normal setup) and visualize their neighbor selection process. Initially, we utilize t-SNE [39] to present the aggregated cognitive vector of each learner u, which can be obtained by aggregating each disentangled cognitive component learned by Coral (setting $K = 30$), i.e., $\sum_{c=1}^{C} \tilde{\mathbf{z}}_u^{(c)}$. The embedding of the target learner is highlighted in red, while the nodes representing neighboring learners are color-coded based on the selection steps, with unselected points displayed in gray. The outcomes are illustrated in Figure 4 (a), showcasing how Coral organizes neighbors according to cognitive states and exemplifying a compelling strategy for neighbor selection that takes into account cognitive similarity.

Disentanglement We evaluate the disentanglement level achieved by assessing independence of dimensions within $\mathbf{z}_u$. The independence level $IL(u)$ of each $\mathbf{z}_u$ is quantified as $IL(u) = \frac{1}{C} \sum_{c=1}^{C} \frac{2}{d(d-1)} \sum_{1 \leq i,j \leq d} |z_u^{(c)}[i] - z_u^{(c)}[j]|$, where $z_u^{(c)}[i]$ represents the ith dimension of $\mathbf{z}_u^{(c)}$, following a prior methodology [48, 42, 49]. In Figure 3 (d), we depict $IL = \sum_{u=1}^{M} IL(u)$ and the corresponding model performances at different training epochs on ASSIST (with the normal setup). Notably, Coral (setting $K = 10$) gradually achieves a high degree of disentanglement during the training process, and the model performances generally exhibit a positive correlation with the degree of disentanglement. This observation reveals the effectiveness of the disentanglement process.

Additionally, we visualize the disentangled cognitive component representations (i.e., $\tilde{\mathbf{z}}_u^{(c)}$) of each learner u learned by Coral. We treat each knowledge component in the representation as independent points, with each component colored differently. For visual clarity, we randomly select 200 learners and 5 knowledge concept components for display. The results in Figure 4 (b) uses t-SNE to visualize
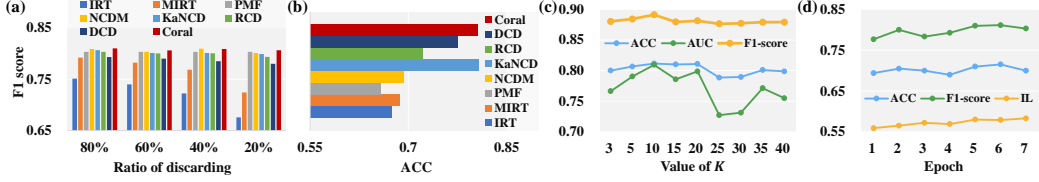
Figure 3: (a) Performance in sparse scenarios. (b) Performance under cold-start scenarios. (c) Performance with different values of K. (d) Disentanglement level and its correlation with performance.
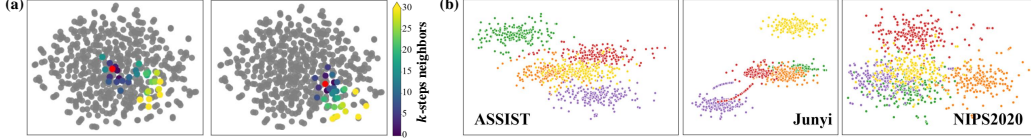

Figure 4: (a) Selected neighbors of the target learner at different steps. (b) t-SNE visualizations of learner representations colored based on knowledge concepts.

learners' cognitive states, with each color representing a distinct category of knowledge-related learner states. This illustrates Coral's ability to achieve well-separated representations.

Explainability We further investigate the interpretability of the cognitive diagnosis outputs based on Coral. We aim to explore whether Coral can provide reasonable predictions for knowledge concepts that learners have not practiced in the training set during actual inference. Firstly, we randomly select a target student u from the Junyi dataset and identify 5 knowledge concepts (denoted as $A \sim E$) that u has not learned in the training data. Subsequently, based on the refined model, we retrieve the top 4 most similar neighboring learners (i.e., $u_1 \sim u_4$) to the target student u. Figure 5 depicts the assessed knowledge concepts ($A \sim E$) and the corresponding mastery levels of selected neighbors using a radar chart. Table 3 presents the diagnostic outputs for the proficiency of u, along with 5 questions related to knowledge $A \sim E$, the predicted scores answering correctly and the actual performances of u. We observe that, despite the cold-start nature of these knowledge concepts for Coral, the model effectively outputs cognitive states that align with the true performance of u by considering the mastery levels of collaborative learners with similar cognitive states. This sufficiently demonstrates the interpretability of Coral's diagnostic results.
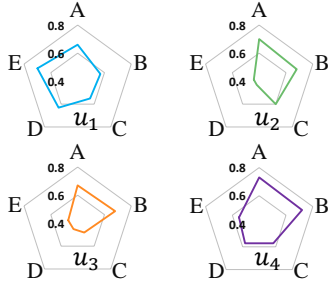

Figure 5: The example of diagnosis output.

Question id	237	213	302	577	620
Knowledge concept	A	B	C	D	E
Proficiency (%)	67.2	68.1	48.3	52.3	52.6
Predicted score (%)	73.4	75.2	47.3	50.8	57.2
True performance	✓	✓	×	×	✓

Table 3: The comparison between the diagnostic results of Coral and the true performance, where ✓ denotes answering correctly and × denotes answering incorrectly.

5 Conclusion

We are pioneering the exploration of collaborative cognitive diagnosis by disentangling the implicit cognitive representations of learners. Extensive experiments demonstrate the superior performance of Coral, showcasing significant improvements over SOTA methods across several real-world datasets. We believe this endeavor marks a crucial step towards collaborative modeling for "AI Education". Furthermore, this work offers valuable insights into conscious-aware learner modeling, under the assumption that human learner proficiency can be effectively represented in a disentangled manner.

Acknowledgments

References

[1] Terry A Ackerman. Multidimensional item response theory models. *Wiley StatsRef: Statistics Reference Online*, 2014.

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2022.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

[4] Haoyang Bi, Enhong Chen, Weidong He, Han Wu, Weihao Zhao, Shijin Wang, and Jinze Wu. Beta-cd: A bayesian meta-learned cognitive diagnosis framework for personalized learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5018–5026, 2023.

[5] Haw-Shiuan Chang, Hwai-Jung Hsu, and Kuan-Ta Chen. Modeling exercise relationships in e-learning: A unified approach. In *EDM*, pages 532–535, 2015.

[6] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. 2023.

[7] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

[8] Xiangzhi Chen, Le Wu, Fei Liu, Lei Chen, Kun Zhang, Richang Hong, and Meng Wang. Disentangling cognitive diagnosis with limited exercise labels. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] Yu Chen, Lingfei Wu, and Mohammed Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. *Advances in neural information processing systems*, 33:19314–19326, 2020.

[10] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.

[11] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, 19:243–266, 2009.

[12] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 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*, pages 501–510, 2021.

[13] Weibo Gao, Qi Liu, Hao Wang, Linan Yue, Haoyang Bi, Yin Gu, Fangzhou Yao, Zheng Zhangm Xin Li, and Yuanjing He. Zero-1-to-3: Domain-level zero-shot cognitive diagnosis via one batch of early-bird students towards three diagnostic objectives. *arXiv preprint arXiv:2312.13434*, 2023.

[14] Weibo Gao, Hao Wang, Qi Liu, Fei Wang, Xin Lin, Linan Yue, Zheng Zhang, Rui Lv, and Shijin Wang. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 983–992, 2023.

[15] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[16] Robert J Harvey and Allen L Hammer. Item response theory. *The Counseling Psychologist*, 27 (3):353–383, 1999.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Yujia Huo, Derek F Wong, Lionel M Ni, Lidia S Chao, and Jing Zhang. Knowledge modeling via contextualized representations for lstm-based personalized exercise recommendation. *Information Sciences*, 523:266–278, 2020.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[21] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[22] Rongfan Li, Ting Zhong, Xinke Jiang, Goce Trajcevski, Jin Wu, and Fan Zhou. Mining spatio-temporal relations via self-paced graph contrastive learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 936–944, 2022.

[23] Rui Li, Liyang He, Qi Liu, Yuze Zhao, Zheng Zhang, Zhenya Huang, Yu Su, and Shijin Wang. Consider: Commonalities and specialties driven multilingual code retrieval framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8679–8687, 2024.

[24] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pages 689–698, 2018.

[25] Qi Liu, Runze Wu, Enhong Chen, Guandong Xu, Yu Su, Zhigang Chen, and Guoping Hu. Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(4):1–26, 2018.

[26] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115, 2019.

[27] Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410, 2023.

[28] Ting Long, Jiarui Qin, Jian Shen, Weinan Zhang, Wei Xia, Ruiming Tang, Xiuqiang He, and Yong Yu. Improving knowledge tracing with collaborative information. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 599–607, 2022.

[29] Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18733–18741, 2024.

[30] Xinwei Long, Jiali Zeng, Fandong Meng, Jie Zhou, and Bowen Zhou. Trust in internal or external knowledge? generative multi-modal entity linking with knowledge retriever. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7559–7569, 2024.

[31] Aleksandra Luszczynska and Ralf Schwarzer. Social cognitive theory. *Fac Health Sci Publ*, pages 225–51, 2015.

[32] Haiping Ma, Changqian Wang, Hengshu Zhu, Shangshang Yang, Xiaoming Zhang, and Xingyi Zhang. Enhancing cognitive diagnosis using un-interacted exercises: A collaboration-aware mixed sampling approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8877–8885, 2024.

[33] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. Learning disentangled representations for recommendation. *Advances in neural information processing systems*, 32, 2019.

[34] Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20, 2007.

[35] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13 (6):47–60, 1996.

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[38] Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. Deep-irt with independent student and item networks. *International Educational Data Mining Society*, 2021.

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[40] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6153–6161, 2020.

[41] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. Neuralcd: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[42] Xin Wang, Zirui Pan, Yuwei Zhou, Hong Chen, Chendi Ge, and Wenwu Zhu. Curriculum co-disentangled representation learning across multiple environments for social recommendation. In *International Conference on Machine Learning*, pages 36174–36192. PMLR, 2023.

[43] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061*, 2020.

[44] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15077–15087, 2024.

[45] Shangshang Yang, Haoyu Wei, Haiping Ma, Ye Tian, Xingyi Zhang, Yunbo Cao, and Yaochu Jin. 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, 2023.

[46] Fangzhou Yao, Qi Liu, Min Hou, Shiwei Tong, Zhenya Huang, Enhong Chen, Jing Sha, and Shijin Wang. Exploiting non-interactive exercises in cognitive diagnosis. *Interaction*, 100(200): 300, 2023.

[47] Fangzhou Yao, Qi Liu, Linan Yue, Weibo Gao, Jiatong Li, Xin Li, and Yuanjing He. Adard: An adaptive response denoising framework for robust learner modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3886–3895, 2024.

[48] Linan Yue, Qi Liu, Yichao Du, Yanqing An, Li Wang, and Enhong Chen. Dare: disentanglement-augmented rationale extraction. *Advances in Neural Information Processing Systems*, 35: 26603–26617, 2022.

[49] Linan Yue, Qi Liu, Ye Liu, Weibo Gao, Fangzhou Yao, and Wenfeng Li. Cooperative classification and rationalization for graph generalization. In *Proceedings of the ACM on Web Conference 2024*, pages 344–352, 2024.

[50] Yunfei Zhang, Chuan Qin, Dazhong Shen, Haiping Ma, Le Zhang, Xingyi Zhang, and Hengshu Zhu. Relicd: A reliable cognitive diagnosis framework with confidence awareness. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 858–867. IEEE, 2023.

[51] Zheng Zhang, Qi Liu, Hao Jiang, Fei Wang, Yan Zhuang, Le Wu, Weibo Gao, and Enhong Chen. Fairlisa: Fair user modeling with limited sensitive attributes information. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[52] Zheng Zhang, Qi Liu, Zirui Hu, Yi Zhan, Zhenya Huang, Weibo Gao, and Qingyang Mao. Enhancing fairness in meta-learned user modeling via adaptive sampling. In *Proceedings of the ACM on Web Conference 2024*, pages 3241–3252, 2024.

[53] Zheng Zhang, Le Wu, Qi Liu, Jiayu Liu, Zhenya Huang, Yu Yin, Yan Zhuang, Weibo Gao, and Enhong Chen. Understanding and improving fairness in cognitive diagnosis. *Science China Information Sciences*, 67(5):152106, 2024.

[54] Hongke Zhao, Songming Zheng, Likang Wu, Bowen Yu, and Jing Wang. Lane: Logic alignment of non-tuning large language models and online recommendation systems for explainable reason generation. *arXiv preprint arXiv:2407.02833*, 2024.

[55] Yan Zhuang, Qi Liu, GuanHao Zhao, Zhenya Huang, Weizhe Huang, Zachary Pardos, Enhong Chen, Jinze Wu, and Xin Li. A bounded ability estimation for computerized adaptive testing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

A Proofs

Property 1. $\max \log p_\Theta\left(\mathbf{x}_u\right)$ *is bounded as follows:*

$$\log p_\Theta\left(\mathbf{x}_u\right) \geq \mathbb{E}_{p(\mathbf{C}) q_\Theta(\mathbf{z}_u | \mathbf{x}_u)}\left[\log p_\Theta\left(\mathbf{x}_u \mid \mathbf{z}_u\right)\right] - \mathbb{E}_{p(\mathbf{C})}\left[D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u\right)\right)\right]. \quad (10)$$

The proof is as follows.

Proof.

$$
\begin{aligned}
\log p_\Theta\left(\mathbf{x}_u\right) &= \log \mathbb{E}_{p(\mathbf{C})}\left[p_\Theta\left(\mathbf{x}_u \mid \mathbf{z}_u, \mathbf{C}\right) p_\Theta\left(\mathbf{z}_u\right)\right] \\
&= \mathbb{E}_{p(\mathbf{C})}\left[\log p_\Theta\left(\mathbf{x}_u\right) q_\Theta\left(\mathbf{z}_u | \mathbf{x}_u\right)\right] \\
&= \mathbb{E}_{p(\mathbf{C}) q_\Theta(\mathbf{z}_u | \mathbf{X}_u)}\left[\log p_\Theta\left(\mathbf{x}_u\right)\right] \\
&= \mathbb{E}_{p(\mathbf{C}) q_\Theta(\mathbf{z}_u | \mathbf{X}_u)}\left[\log \frac{p_\Theta\left(\mathbf{x}_u, \mathbf{z}_u\right)}{p_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}\right] \\
&= \mathbb{E}_{p(\mathbf{C}) q_\Theta(\mathbf{z}_u | \mathbf{X}_u)}\left[\log \frac{p_\Theta\left(\mathbf{x}_u, \mathbf{z}_u\right)}{p_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)} \cdot \frac{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}\right] \\
&= \mathbb{E}_{p(\mathbf{C}) q_\Theta(\mathbf{z}_u | \mathbf{X}_u)}\left[\log \frac{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}{p_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}\right] + \mathbb{E}_{p(\mathbf{C}) q_\Theta(\mathbf{z}_u | \mathbf{X}_u)}\left[\log \frac{p_\Theta\left(\mathbf{x}_u, \mathbf{z}_u\right)}{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}\right] \\
&= \mathbb{E}_{p(\mathbf{C})}\left[D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)\right) + \mathbb{E}_{p(\mathbf{C}) q_\Theta(\mathbf{z}_u | \mathbf{X}_u)}\left[\log \frac{p_\Theta\left(\mathbf{x}_u \mid \mathbf{z}_u\right) p_\Theta\left(\mathbf{z}_u\right)}{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}\right]\right. \\
&= \mathbb{E}_{p(\mathbf{C})}\left[D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)\right) - D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u\right)\right)\right. \\
&\quad + \mathbb{E}_{p(\mathbf{C}) q_\Theta(\mathbf{z}_u | \mathbf{X}_u)}\left[\log p_\Theta\left(\mathbf{x}_u \mid \mathbf{z}_u\right)\right] \\
&\geq \mathbb{E}_{p(\mathbf{C})}\left[\mathbb{E}_{q_\Theta(\mathbf{z}_u | \mathbf{X}_u)} \log p_\Theta\left(\mathbf{x}_u \mid \mathbf{z}_u\right) - D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u\right)\right)\right],
\end{aligned}
$$

$$(11)$$

which completes the proof. □

Property 2. *The $D_{\mathrm{KL}}(\cdot)$ in Eq. (2) can be rewritten as:*

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u\right)\right) \\
&= I\left(\mathbf{z}_u, \mathbf{x}_u\right) + D_{\mathrm{KL}}\left(q_\Theta(\mathbf{z}_u) \| p_\Theta(\mathbf{z}_u)\right).
\end{aligned}
\tag{12}
$$

The proof is as follows.

Proof. Given that $p_\Theta\left(\mathbf{x}_u\right) = p_{data}\left(\mathbf{x}_u\right)$ and $q_\Theta\left(\mathbf{z}_u, \mathbf{x}_u\right) = q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) p_\Theta\left(\mathbf{x}_u\right)$, we then have

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \| p_\Theta\left(\mathbf{z}_u\right)\right) \\
&= \mathbb{E}_{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}\left[\log \frac{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) \cdot q_\Theta\left(\mathbf{z}_u\right)}{p_\Theta\left(\mathbf{z}_u\right) \cdot q_\Theta\left(\mathbf{z}_u\right)}\right] \\
&= \mathbb{E}_{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}\left[\log \frac{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}{q_\Theta\left(\mathbf{z}_u\right)}\right] + \mathbb{E}_{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}\left[\frac{q_\Theta\left(\mathbf{z}_u\right)}{p_\Theta\left(\mathbf{z}_u\right)}\right] \\
&= \mathbb{E}_{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) p_{data}\left(\mathbf{x}_u\right)}\left[\log \frac{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}{q_\Theta\left(\mathbf{z}_u\right)}\right] + \mathbb{E}_{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right) p_{data}\left(\mathbf{x}_u\right)}\left[\frac{q_\Theta\left(\mathbf{z}_u\right)}{p_\Theta\left(\mathbf{z}_u\right)}\right] \\
&= \mathbb{E}_{q_\Theta\left(\mathbf{z}_u, \mathbf{x}_u\right)}\left[\log \frac{q_\Theta\left(\mathbf{z}_u \mid \mathbf{x}_u\right)}{q_\Theta\left(\mathbf{z}_u\right)}\right] + \mathbb{E}_{q_\Theta\left(\mathbf{z}_u\right)}\left[\frac{q_\Theta\left(\mathbf{z}_u\right)}{p_\Theta\left(\mathbf{z}_u\right)}\right] \\
&= I\left(\mathbf{z}_u; \mathbf{x}_u\right) + D_{\mathrm{KL}}\left(q_\Theta\left(\mathbf{z}_u\right) \| p_\Theta\left(\mathbf{z}_u\right)\right),
\end{aligned}
\tag{13}
$$

where $I(\mathbf{A}; \mathbf{B})$ calculates mutual information (MI) between $\mathbf{A}$ and $\mathbf{B}$, i.e. $I(\mathbf{A}; \mathbf{B}) = \mathbb{E}_{p(\mathbf{a}, \mathbf{b})}[\log \frac{p(\mathbf{a}|\mathbf{b})}{p(\mathbf{a})}]$. Therefore, the proof is completed. □

Property 3. $\max \log p_\Theta(G \mid V, \mathbf{Z})$ *is bounded as follows:*

$$
\max \log p_\Theta(G \mid V, \mathbf{Z}) \geq -\sum_{c=1}^{C}\sum_{u=1}^{M}\sum_{k=1}^{K} \mathcal{L}_u^{(c),k}
$$
$$
\text{where } \mathcal{L}_u^{(c),k} = -\frac{\exp\left(f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right)\right)}{\sum_{v \in V_u^{(c)}} \exp\left(f_{(c)}\left(v; r_u^{(c),k-1}\right)\right)}.
\tag{14}
$$

The proof is as follows:

Proof. Given the following inequality,

$$
\begin{aligned}
\max \log p_\Theta(G \mid V, \mathbf{Z}) :&= \max \sum_{c=1}^{C} \mathbb{E}_{p_\Theta\left(\mathcal{N}_u^{(c)}, \mathbf{z}_u^{(c)}\right)}\left[\log p_\Theta\left(\mathcal{N}_u^{(c)} \mid \mathbf{z}_u^{(c)}\right)\right] \\
&= \max \sum_{c=1}^{C} I\left(\mathcal{N}^{(c)}; \mathbf{Z}^{(c)}\right) + \sum_{c=1}^{C} \mathbb{E}_{p_\Theta\left(\mathbf{Z}^{(c)}\right)}\left[\log p_\Theta\left(\mathbf{Z}^{(c)}\right)\right] \\
&\geq \max \sum_{c=1}^{C} I\left(\mathcal{N}^{(c)}; \mathbf{Z}^{(c)}\right),
\end{aligned}
\tag{15}
$$

we further derive the term $I\left(\mathcal{N}^{(c)}; \mathbf{Z}^{(c)}\right)$ that search the combination of K similar neighbors from the permutation perspective as follows:

$$
\begin{aligned}
I\left(\mathcal{N}^{(c)}; \mathbf{Z}^{(c)}\right) &= \mathbb{E}_{p_\Theta\left(\mathcal{N}_u^{(c)}, \mathbf{z}_u^{(c)}\right)}\left[\log p_\Theta\left(\mathcal{N}_u^{(c)} \mid \mathbf{z}_u^{(c)}\right)\right] + H\left(\mathcal{N}_u^{(c)}\right) \\
&\geq \mathbb{E}_{p\left(\mathcal{R}_u^{(c)}, \mathbf{z}_u^{(c)}\right)}\left[\log p_\Theta\left(\mathcal{R}_u^{(c)} \mid \mathbf{z}_u^{(c)}\right)\right] + H\left(\mathcal{N}_u^{(c)}\right),
\end{aligned}
\tag{16}
$$

where $H(\mathbf{A}) = -\sum p(\mathbf{a}) \cdot \log p(\mathbf{a})$. $\mathcal{R}_u^{(c)}$ denotes the permutation of neighbors representing routes to $\mathcal{N}_u^{(c)}$ in given orders, where $\left|\mathcal{R}_u^{(c)}\right| = \frac{M!}{(M-K)!}$. However, the search space of $\mathcal{R}_u^{(c)}$ is huge and even prohibitive. Inspired by the equivalent task [22], we next present the Eq. (16) in a heuristic

style by maximizing the MI between the context of selected similar learner neighbors and the next neighbor iteratively.

Specifically, the core goal of Eq. (15) is to find all K neighbors $\mathcal{N}_u^{(c)}$ for each learner u under a specific concept the data at once, from the perspective of maximizing MI globally. However, this task is challenging due to the especially large search spaces of $\mathcal{N}_u^{(c)}$ and $\mathcal{R}_u^{(c)}$. Thereby, we decompose the globally optimal task into an equivalent task in an iterative local optimal process. Concretely, assume that we have found $(k_0 - 1)$ optimal neighbors for the learner u formulating a route $r_u^{(c),k_0-1}$ in a specific order from learner node 1 to $(k_0 - 1)$, we then search the k_0-th neighbor $b_u^{(c),k_0}$ equally from the rest $(M - k_0 + 1)$ learners for the learner u, i.e., $p\left(b_u^{(c),k_0}\right) = \frac{1}{M-k_0+1}$.

Given arbitrary k_0 optimal neighbors for learner u with a specific sub-route $r_u^{(c),k_0}$, we can derive $p_\Theta\left(\mathcal{R}_u^{(c)}, \mathbf{z}_u^{(c)}\right)$ in Eq. (16) as follows:

$$p_\Theta\left(\mathcal{R}_u^{(c)}, \mathbf{z}_u^{(c)}\right) = \mathbb{E}_{p_\Theta\left(\mathcal{R}_u^{(c)}\right)}\left[p_\Theta\left(r_u^{(c),k_0}\right) \prod_{i=k_0+1}^{K} p_\Theta\left(b_u^{(c),i} \mid r_u^{(c),i-1}\right)\right]. \qquad (17)$$

We can derive the log term $\log p_\Theta\left(\mathcal{R}_u^{(c)} \mid \mathbf{z}_u^{(c)}\right)$ in Eq. (16) as follows:

$$\begin{aligned}
\log p_\Theta\left(\mathcal{R}_u^{(c)} \mid \mathbf{z}_u^{(c)}\right) &= \mathbb{E}_{p_\Theta\left(\mathcal{R}_u^{(c)}\right)}\left[\sum_{k=1}^{K} \log p_\Theta\left(b_u^{(c),k} \mid r_u^{(c),k-1}\right)\right] \\
&= \mathbb{E}_{p_\Theta\left(\mathcal{R}_u^{(c)}\right)}\left[\sum_{k=1}^{K} \log \frac{p_\Theta\left(b_u^{(c),k} \mid r_u^{(c),k-1}\right) \cdot p_\Theta\left(b_u^{(c),k}\right)}{p_\Theta\left(b_u^{(c),k}\right)}\right] \\
&= \mathbb{E}_{p_\Theta\left(\mathcal{R}_u^{(c)}\right)}\left[\sum_{k=1}^{K} \log \frac{p_\Theta\left(b_u^{(c),k} \mid r_u^{(c),k-1}\right)}{p_\Theta\left(b_u^{(c),k}\right)} + \sum_{k=1}^{K} \log p_\Theta\left(b_u^{(c),k}\right)\right].
\end{aligned} \qquad (18)$$

With Eq. (17) and (18), we can derive the first term in Eq. (16) as follows:

$$\begin{aligned}
&\mathbb{E}_{p_\Theta\left(\mathcal{R}_u^{(c)}, \mathbf{z}_u^{(c)}\right)}\left[\log p_\Theta\left(\mathcal{R}_u^{(c)} \mid \mathbf{z}_u^{(c)}\right)\right] \\
&= \sum_{u=1}^{M} \mathbb{E}_{p_\Theta\left(\mathcal{R}_u^{(c)}\right)}\left[p_\Theta\left(r_u^{(c),k_0}\right) \prod_{i=k_0+1}^{K} p_\Theta\left(b_u^{(c),i} \mid r_u^{(c),i-1}\right) \sum_{k=1}^{K} \log \frac{p_\Theta\left(b_u^{(c),k} \mid r_u^{(c),k-1}\right)}{p_\Theta\left(b_u^{(c),k}\right)}\right] \\
&+ \sum_{u=1}^{M} \mathbb{E}_{p_\Theta\left(\mathcal{R}_u^{(c)}\right)}\left[p_\Theta\left(r_u^{(c),k_0}\right) \prod_{i=k_0+1}^{K} p_\Theta\left(b_u^{(c),i} \mid r_u^{(c),i-1}\right) \sum_{k=1}^{K} \log p_\Theta\left(b_u^{(c),k}\right)\right] \\
&\approx \sum_{u=1}^{M} \mathbb{E}_{p_\Theta\left(\mathcal{R}_u^{(c)}\right)}\left\{\sum_{k=1}^{K}\left[I\left(b_u^{(c),k}; r_u^{(c),k-1}\right) \prod_{i=k}^{K} p_\Theta\left(b_u^{(c),i} \mid r_u^{(c),i-1}\right)\right]\right\} + \epsilon(M, K),
\end{aligned} \qquad (19)$$

where $\epsilon(M, K) \geq 0$ is a constant term regarding M and K. Inspired by [36], we have the lower bound of $I\left(b_u^{(c),k}; r_u^{(c),k-1}\right)$ as follows:

$$I\left(b_u^{(c),k}; r_u^{(c),k-1}\right) \geq \log M - \mathcal{L}_u^{(c),k},$$

$$\text{where} \quad \mathcal{L}_u^{(c),k} = -\frac{\exp\left(f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right)\right)}{\sum_{v \in V_u^{(c)}} \exp\left(f_{(c)}\left(v; r_u^{(c),k-1}\right)\right)}. \qquad (20)$$

The Eq. (20) iteratively searches K neighbors for the learner u under each knowledge concept c from step $k = 1$ to K. $\mathcal{L}_u^{(c),k}$ is the well-known InfoNCE loss function [36]. Let $r_u^{(c),k-1}$ denote the

current context at step $(k-1)$ (i.e., the set of $(k-1)$ neighbors selected from step 1 to $(k-1)$). $b_u^{(c),k}$ is the affinity candidate learner in the $(M-k)$ nonneighbor learners. Let $V_u^{(c)}$ denote the current set of nonneighbor learners, and we hence have $b_u^{(c),k} \in V_u^{(c)}$. $f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right)$ is a matching function measuring the similarity between of nonneighbor $b_u^{(c),k}$ and the current context $r_u^{(c),k-1}$, where the higher the scalar score means the higher likelihood of $b_u^{(c),k}$ is a new neighbor.

Furthermore, we have $\mathcal{L}_u^{(c),k} \propto f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right)$. Thus, given the context of $(k-1)$ neighboring learners (i.e., we have found $(k-1)$ neighbors for the learner u) and matching function $f_{(c)}(\cdot)$, our goal following the Eq. (20) is to find a learner $b_u^{(c),k}$ from nonneighbor set $V_u^{(c)}$ that can maximize the matching score $f_{(c)}(\cdot)$ as the k-th neighbor of u. In other words, $p(G \mid V, \mathbf{Z})$ can be optimized through maximizing the matching score $f_{(c)}(\cdot)$ from $k=1$ to K iteratively. Thereby, at each step k, we sort the scores of the nonneighbor learners and select the learner with the highest score to label as k-th neighbor $b_u^{(c),k}$, i.e., $b_u^{(c),k} \leftarrow \arg\max_v f_{(c)}(v; r_u^{(c),k-1}), v \in V_u^{(c)}$. After obtaining the k-th neighbor $b_u^{(c),k}$, the context $r_u^{(c),k-1}$ is updated to $r_u^{(c),k}$ by absorbing $b_u^{(c),k}$.

Based on the Eq. (15), Eq. (16) and Eq. (20), we have

$$\max \log p(G \mid V, \mathbf{Z}) \geq -\sum_{c=1}^{C} \sum_{u=1}^{M} \sum_{k=1}^{K} \mathcal{L}_u^{(c),k}, \tag{21}$$

which completes the proof.

$\square$

Theorem 1. With Gaussian Mixture initialization from the Disentangled Cognitive Representation Encoding (section 3.2), the Collaborative Representation Learning (section 3.3) procedure is equivalent to an expectation-maximization (EM) algorithm [35] for the mixture model. In particular, it converges to a point estimate of $\{\mathbf{r}_u^{(c)}\}_{c=1}^{C}$ that maximizes the marginal likelihood $l\left(\left\{a_v^{(c)} : (u,v) \in G_{(c)}\right\}_{c=1}^{C}; \{\mathbf{r}_u^{(c)}\}_{c=1}^{C}\right)$, where $a_{u,v}^{(c)}$ equals 1 or 0 denoting whether learner v is a collaborative neighbor of learner u regarding concept c or not.

The proof is as follows.

Proof. The collaborative modeling process can be approximatively equivalent to an expectation-maximization (EM) algorithm for the mixture model. Let $A = \{A_{(c)}\}_{c=1}^{C}$ where $A_{(c)} = \{a_v^{(c)} : (u,v) \in G_{(c)}\}$, where $a_{u,v}^{(c)}$ equals 1 or 0 denoting whether learner v is a collaborative neighbor of learner u regarding concept c or not, which is a type unknown factor in EM algorithm. Given $\mathbf{R} = \{\mathbf{r}^{(c)}\}_{c=1}^{C}$ and $\mathbf{Z} = \{\mathbf{z}^{(c)}\}_{c=1}^{C}$, the EM algorithm maximizes the likelihood $l(A; \mathbf{R}) = \sum_A l(A, \mathbf{Z}; \mathbf{R})$. Let $q(A)$ is the distribution over A, we then have

$$\begin{aligned}
\log l(A; \mathbf{R}) &= \sum_A q(A) \cdot l(\mathbf{Z}; \mathbf{R}) \\
&= \sum_A q(A) \cdot \frac{l(A, \mathbf{Z}; \mathbf{R})}{l(A \mid \mathbf{Z}; \mathbf{R})} \\
&= \sum_A q(A) \cdot \frac{l(A, \mathbf{Z}; \mathbf{R})}{q(A)} + \sum_A q(A) \cdot \frac{q(A)}{l(A \mid \mathbf{Z}; \mathbf{R})}.
\end{aligned} \tag{22}$$

Let $L(\mathbf{R}, q(A))$ denote $\sum_A q(A) \cdot \frac{l(A, \mathbf{Z}; \mathbf{R})}{q(A)}$, we can rewrite Eq. (22) as

$$\begin{aligned}
\log l(\mathbf{Z}; \mathbf{R}) &= L(\mathbf{R}, q(A)) + D_{\text{KL}}(q(A) \| l(A \mid \mathbf{Z}; \mathbf{R})) \\
&\leq L(\mathbf{R}, q(A)),
\end{aligned} \tag{23}$$

where $L(\mathbf{R}, q(A))$ is a lower bound of $l(A; \mathbf{R})$ since the KL divergence from $l(A \mid \mathbf{Z}; \mathbf{R})$ towards $q(A)$ is non-negative.

17

At every E-step, the EM algorithm aims to search the optimal $q(A)$ that tightens the lower bound, which $q(A)$ is set to $l(A \mid \mathbf{Z}; \mathbf{R})$ since the KL divergence will become zero. Given that

$$l(A \mid \mathbf{Z}; \mathbf{R}) = \log p(G \mid V, \mathbf{Z}), \tag{24}$$

and Property 3, the optimal $q(A)$ that tightens the lower bound can be achieved through iteratively maximizing the following objective as

$$\max \sum_{c=1}^{C} \sum_{u=1}^{M} \sum_{k=1}^{K} \frac{\exp\left(f_{(c)}\left(b_u^{(c),k}; r_u^{(c),k-1}\right)\right)}{\sum_{v \in V_u^{(c)}} \exp\left(f_{(c)}\left(v; r_u^{(c),k-1}\right)\right)}, \tag{25}$$

which performs the E-step.

After the E-step, an M-step is performed to maximize the lower bound $L(\mathbf{R}, q(A))$ w.r.t. $\mathbf{R}$ with $q(A)$ fixed found in the E-step. Given that

$$\mathbf{r}_u^{(c)} = \frac{1}{|\mathcal{N}_u^{(c)}|} \sum_{v \in \mathcal{N}_u^{(c)}} s_{u,v}^{(c)} \cdot \mathbf{z}_v^{(c)}, \tag{26}$$

we optimize the $\mathbf{r}^{(c)}$ upon $\frac{\partial L(\mathbf{R}, q(A))}{\partial \mathbf{r}^{(c)}}$ to zero, which is actually performing the M-step.

Let $q(A)^k$ and $\mathbf{r}^k$ be the result of the k-th E-step and the k-th M-step, respectively, we then have

$$
\begin{aligned}
l\left(A^k; \mathbf{R}^{k-1}\right) &= l\left(\mathbf{R}^{k-1}, q(A)^k\right) + D_{\text{KL}}\left(q(A)^k \parallel l\left(A^k \mid \mathbf{Z}; \mathbf{R}^{k-1}\right)\right) \\
&= L\left(\mathbf{R}^{k-1}, q(A)^k\right) \\
&\leq L\left(\mathbf{R}^k, q(A)^k\right) \\
&\leq L\left(\mathbf{R}^k, q(A)^k\right) + D_{\text{KL}}\left(q(A)^k \parallel l\left(A \mid \mathbf{Z}; \mathbf{R}^k\right)\right) \\
&= l\left(A^k; \mathbf{R}^k\right).
\end{aligned}
\tag{27}
$$

Hence, this can prove that the likelihood will increase monotonically and be upper-bounded by zero at the same time. Therefore, the algorithm converges, which completes the proof. □

B   Algorithm

To offer a more comprehensive description of Coral's structure, we outline the algorithm (see Algorithm 1) for the entire model, including four functions and a main function.

C   Dataset Description and Preprocessing

We conduct experiments on three real-world datasets, i.e., ASSIST [11], Junyi [5] and NeurIPS2020EC [43]. The statistics of these datasets are summarized in Table 1. For all datasets, we preserve the first-time exercise-answering record for the same learner-question pairs to support cognitive diagnosis aligning with common settings used in previous related studies [40]. The detailed information on datasets and preprocessing method are depicted as follows:

- **ASSIST (ASSISTments 2009-2010 "skill builder")** [11] This dataset is an open dataset collected by the ASSISTments online tutoring systems[2], which has become one of the popular benchmark datasets for cognitive diagnosis. We preserve learners with more than 30 practice records for ASSIST to guarantee that each learner has enough data for diagnosis.

- **Junyi** [5] This dataset contains learner online learning logs collected from a Chinese online educational platform called Junyi Academy[3]. Nowadays, Junyi is widely used in the evaluation of online education tasks [47, 8]. We randomly select 1,400 learners with more than 15 practice records from Junyi to guarantee that each learner has enough data for diagnosis.

---

[2]https://sites.google.com/site/assistmentsdata/
[3]https://www.junyiacademy.org/

Algorithm 1 Coral Model

1: **Input:** $\{\mathbf{c}\}_{i=1}^{N}$, practice logs $\{\mathbf{x}_u\}_{u=1}^{M}$;
2: **function** DISENTANGLED_COGNITIVE_REPRESENTATION_ENCODING($\{\mathbf{c}\}_{i=1}^{N}, \mathbf{x}_u$)
3:      // Gaussian Mixture initialization of learner u
4:      $\mathbf{z}_u \leftarrow [\mathbf{z}_u^{(1)}; \mathbf{z}_u^{(2)}; \ldots; \mathbf{z}_u^{(C)}]$
5:      // Calculate ability of learner u
6:      $\theta_u \leftarrow \psi_\Theta(\mathbf{z}_u^{(c)})$, $c = 1, 2, \ldots, C$
7:      // Reconstruct practice performance of learner u
8:      $p_\Theta(x_{u,i} \mid \cdot) \leftarrow c_{i,c} \cdot \phi_\Theta(\theta_u, \mathbf{z}_u^{(c)})$, $c = 1, 2, \ldots, C$
9:      **return** $BCE(x_{u,i}, p_\Theta(x_{u,i} \mid \cdot)), D_{\mathrm{KL}}^u, \mathbf{z}_u$
10: **end function**

11: // Search K neighbors for learner u from all the learners V
12: **function** CONTEXT-AWARE_COLLABORATIVE_GRAPH_LEARNING($V, \mathbf{z}_u^{(c)}$)
13:      **for** $c = 1, 2, \ldots, C$ **do**
14:          // Let be the removing operation
15:          $V_u^{(c)} \leftarrow V \setminus u$
16:          // Initial neighbor set of u, i.e., R_u
17:          $r_u^{(c),k} \leftarrow \{u\}$, where $k = 0$
18:          // Iteratively calculating the cognitive similarity scores between u and each learner in $V_u^{(c)}$, the initial step corresponds to $k = 0$
19:          **for** $k = 1, \ldots, K$ **do**
20:              **for** $v \in V_u^{(c)}$ **do**
21:                  $Score_{u,v} \leftarrow f_{(c)}(v; r_u^{(c),k-1}) // Equation(7)$
22:              **end for**
23:              // Select k-th neighbor for u, denoted as $b_u^{(c),k}$
24:              $b_u^{(c),k} \leftarrow \arg\max_v Score_{u,v}, v \in V_u^{(c)}$
25:              // Update neighbors and non-neighbor learners
26:              $r_u^{(c),k} \leftarrow r_u^{(c),k-1} + \{b_u^{(c),k}\}$
27:              $V_u^{(c)} \leftarrow V_u^{(c)} \setminus b_u^{(c),k}$
28:          **end for**
29:      **end for**
30:      **return** $\{G_{(c)}\}_{c=1}^{C}$
31: **end function**

32: **function** COLLABORATIVE_GRAPH_MODELING($\{G_{(c)}\}_{c=1}^{C}, \mathbf{z}_u$)
33:      **for** $c = 1, 2, \ldots, C$ **do**
34:          $\mathbf{r}_u \leftarrow \varphi(\mathbf{Z}, G)$
35:      **end for**
36:      **return** $\mathbf{r}_u$
37: **end function**

38: **function** DECODING_AND_RECONSTRUCTION($\{G_{(c)}\}_{c=1}^{C}, \mathbf{z}_u, \mathbf{x}_u$)
39:      Calculate $\tilde{\mathbf{z}}_u$
40:      **return** $BCE(x_{u,i}, p_\Theta(\hat{x}_{u,i}))$
41: **end function**

42: **BEGIN MAIN FUNCTION:**
43: **Initialize** $\mathbf{z}_u$, learning rate lr, β, $epoch \leftarrow 0$, $TotalEpoch$
44: **repeat**
45:      **for** $u = 1, 2, \ldots, M$ **do**
46:          $BCE(x_{u,i}, p_\Theta(x_{u,i} \mid \cdot)), D_{\mathrm{KL}}^u, \mathbf{z}_u \leftarrow$ Disentangled_Cognitive_Representation_Encoding($\ldots$) // $\ldots$ denotes omitting parameters
47:          $\{G_{(c)}\}_{c=1}^{C} \leftarrow$ Context-aware_Collaborative_Graph_Learning($\ldots$)
48:          $BCE(x_{u,i}, p_\Theta(\hat{x}_{u,i})) \leftarrow$ Decoding_and_Reconstruction($\ldots$)
49:          Calculate $\arg\min \mathcal{L}$
50:          $\Theta \leftarrow \arg\max_\Theta \mathcal{L}$ by $lr \cdot \nabla_\Theta \mathcal{L}$
51:          $epoch \leftarrow epoch + 1$
52:      **end for**
53: **until** $epoch$ equals $TotalEpoch$

- **NeurIPS2020EC** [43] This dataset is originated from NeurIPS 2020 Education Challenge, which provides learners' practice logs on mathematical questions from Eedi[4]. We randomly select 1,000 learners with more than 15 practice records from NeurIPS2020EC to guarantee that each learner has enough data for diagnosis.

D   Additional Experimental Results

Method	Metric			
	ACC ↑	AUC ↑	F1-score ↑	RMSE ↓
w/o KL	0.693156	0.659914	0.803728	0.452067
w/o collar	0.667321	0.606090	0.786824	0.465231
w/ knn	0.708520	0.721339	0.794198	0.440430
Coral	**0.709710**	**0.721823**	**0.810755**	**0.437818**

Table 4: Ablation study of Coral on ASSIST.

D.1   Ablation Study

We additionally perform ablation studies to assess the impact of key components within Coral. The results in Table 4 depict the performances of Coral (setting $K = 5$) under various conditions: without the KL term for encoding (w/o KL), without the collaborative aggregation during decoding (w/o collar), and replacing the collaborative graph construction procedure using a knn-based methods (w/ knn) used in [13] on the ASSIST dataset. These findings show the effectiveness of each key component in enhancing the overall performance of Coral.

D.2   Efficiency Improvement

We additionally implement three efficiency optimization strategies to further reduce the complexity of Coral. These strategies cannot theoretically guarantee optimal performance, but they can enhance applicability and scalability of Coral through empirical balancing of efficiency and accuracy. We refer them as Coral with n-sample, Coral with m-selections, and Coral with full-kit, as follows:

- Coral with n-sample: During the K iterations of searching for neighbors, randomly sample n subsets from all M learners to replace V in the original approach. This reduces computational efficiency from $M \times K$ to $n \times K$, where $n \ll M$.
- Coral with m-selections: Based on the basic Coral, replace selecting one neighbor per iteration with selecting m neighbors. This decreases computational efficiency from $M \times K$ to $\frac{M \times K}{m}$, where $m < K$.
- Coral with full-kit: A combination of Coral with n-sample and Coral with m-selections, further reducing computational efficiency from $M \times K$ to $\frac{n \times K}{m}$, where $n \ll M$ and $m < K$.

Following the three strategies outlined above, we conduct several experiments on the Junyi dataset, with $K = 40$, to assess prediction performance. The results are summarized in the Figure 6. These experimental results demonstrate Coral's potential to improve computational efficiency while maintaining acceptable performance, as evidenced by the varying levels of accuracy achieved with different optimization strategies.

E   Broader Impact and Limitation

This research delves into modeling human cognitive states within the realm of intelligent education. The proposed Coral model significantly enhances the diagnostic accuracy of implicit learners' knowledge states. This improvement not only provides effective insights for online personalized tutoring services, such as question recommendations but also lays the foundation for further research in this area. Moreover, the automatic construction strategy for collaborative connections among learners
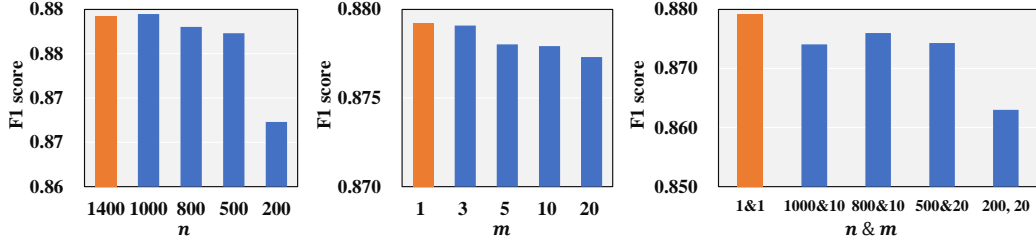
---

[4]https://eedi.com/

Figure 6: The prediction performance of the improved model is illustrated, with the orange bar representing the performance of the original Coral, which achieves the highest F1 score.

offers valuable insights that can contribute to subsequent investigations in this field. Lastly, we anticipate that the proposed techniques can be extended to other domains, including but not limited to user interest modeling and social network modeling. Although our method is effective both theoretically and empirically, it suffers from computational inefficiencies. We have explored preliminary optimization strategies in the Appendix D.2 and will focus on improving computational efficiency in future research. In addition, future research plans to consider issues of fairness [51, 52] and explore the integration of large language models and multi-modal knowledge to enhance interpretability [27, 29, 30]. In essence, our work is dedicated to advancing intelligent education and deepening the understanding of human cognitive proficiency. It cannot cause negative effects. We anticipate its crucial role in fostering progress in both pertinent technologies and societal advancements.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction clearly claim our task (scope), contributions and solutions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide all the proofs in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide our code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide a github repository to publish our work (https://github.com/bigdata-ustc/Coral).

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In the experimental setup (Section 4.1), we provide all the setups.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: In our paper, we provide the average scores. Given the impending submission deadline, we will provide error bars during the rebuttal phase by updating the README file in the anonymous GitHub repository.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide these details in the implementation details (Section 4.1).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: This research conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: These information is discussed in Appendix E.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

   Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

   Answer: [NA]

   Justification: Not Applicable.

   Guidelines:

   - The answer NA means that the paper poses no such risks.
   - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
   - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
   - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

   Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

   Answer: [Yes]

   Justification: We cite the original papers or website links about the dataset and open-source codes.

   Guidelines:

   - The answer NA means that the paper does not use existing assets.
   - The authors should cite the original paper that produced the code package or dataset.
   - The authors should state which version of the asset is used and, if possible, include a URL.
   - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
   - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
   - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
   - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects. Guidelines:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.