# RDGT: Enhancing Group Cognitive Diagnosis With Relation-Guided Dual-Side Graph Transformer

Xiaoshan Yu, Chuan Qin, *Member, IEEE*, Dazhong Shen, Haiping Ma, Le Zhang, Xingyi Zhang, *Senior Member, IEEE*, Hengshu Zhu, *Senior Member, IEEE*, and Hui Xiong, *Fellow, IEEE*

***Abstract*—Cognitive diagnosis has been widely recognized as a crucial task in the field of computational education, which is capable of learning the knowledge profiles of students and predicting their future exercise performance. Indeed, considerable research efforts have been made in this direction over the past decades. However, most of the existing studies only focus on individual-level diagnostic modeling, while the group-level cognitive diagnosis still lacks an in-depth exploration, which is more compatible with realistic collaborative learning environments. To this end, in this paper, we propose a Relation-guided Dual-side Graph Transformer (RDGT) model for achieving effective group-level cognitive diagnosis. Specifically, we first construct the dual-side relation graphs (i.e., student-side and exercise-side) from the group-student-exercise heterogeneous interaction data for explicitly modeling associations between students and exercises, respectively. In particular, the edge weight between two nodes is defined based on the similarity of corresponding student-exercise interactions. Then, we introduce two relation-guided graph transformers to learn the representations of students and exercises by integrating the whole graph information, including both nodes and edge weights. Meanwhile, the inter-group information has been incorporated into the student-side relation graph to further enhance the representations of students. Along this line, we design a cognitive diagnosis module for learning the groups' proficiency in specific knowledge concepts, which includes an attention-based aggregation strategy to obtain the final group representation and a hybrid loss for optimizing the performance prediction of both group and student. Finally, extensive experiments on 5 real-world datasets clearly demonstrate the effectiveness of our model as well as some interesting findings (e.g., the representative groups and potential collaborations among students).**

***Index Terms*—Computational education, education data mining, cognitive diagnostic models, group-level cognitive diagnosis.**

## I. Introduction

COGNITIVE diagnosis has been recognized as a pivotal task in intelligent education, aimed at determining students' mastery of the corresponding knowledge concepts by exploring their exercise records. It has been applied in various educational scenarios, such as online exercise design [1], [2], computerized adaptive testing [3], and course recommendation [4], [5], leading to more efficient and effective student learning. In past decades, various cognitive diagnostic models (CDMs) have been developed, which can generally be grouped into two categories: psychometric theory-based CDMs [6], [7], [8] and neural network-based CDMs [9], [10], [11]. For instance, item response theory (IRT) [6], multidimensional IRT (MIRT) [7] and deterministic inputs, noisy "and" gate model (DINA) [8] manually design simple student-exercise interaction functions based on psychometric theory to mine ability factors associated with students, while neural cognitive diagnosis (NCD) [9] and relation map driven cognitive diagnosis (RCD) [10] model higher-order student-exercise interactions by incorporating neural networks.

Most of the existing CDMs mainly focus on individual assessments, however, they are not applicable to another classical educational scenario, i.e., collaborative learning, where learners study in groups and influence each other [12]. Similar to individual cognitive diagnosis, group-level cognitive diagnosis (GCD) refers to the assessment of the group's cognitive abilities utilizing the interactive responses of a group of students on given exercises [13]. As illustrated in Fig. 1(a), unlike individual diagnosis, the input of GCD is a group of students and the exercising records of the group where the response results are correct rate and the output is the group's mastery level in specific knowledge concepts.

A major challenge in realizing efficient GCD lies in exploring group representation utilizing sparse group-exercise interaction data. Intuitively, we can incorporate student interaction behavior

Manuscript received 20 June 2023; revised 18 November 2023; accepted 7 January 2024. Date of publication 11 January 2024; date of current version 10 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U21A20512, in part by the National Natural Science Foundation of China under Grant 62107001, and in part by the Anhui Provincial Natural Science Foundation under Grant 2108085QF272. Recommended for acceptance by L. Nie. *(Corresponding authors: Haiping Ma and Chuan Qin.)*

Xiaoshan Yu is with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Artificial Intelligence, Anhui University, Hefei 230093, China, and also with Career Science Lab, BOSS Zhipin, Beijing 100020, China (e-mail: yxsleo@gmail.com).

Chuan Qin is with the Career Science Lab, BOSS Zhipin, Beijing 100020, China, and also with the PBC School of Finance, Tsinghua University, Beijing 100190, China (e-mail: chuanqin0426@gmail.com).

Dazhong Shen is with the Shanghai Artificial Intelligence Laboratory, Shanghai 200240, China (e-mail: dazh.shen@gmail.com).

Haiping Ma is with the Department of Information Materials and Intelligent Sensing Laboratory of Anhui Province, Institutes of Physical Science and Information Technology, Anhui University, Hefei 230093, China (e-mail: hpma@ahu.edu.cn).

Le Zhang is with the Business Intelligence Lab, Baidu Inc, Beijing 100085, China (e-mail: zhangle0202@gmail.com).

Xingyi Zhang is with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230093, China (e-mail: xyzhanghust@gmail.com).

Hengshu Zhu is with the Career Science Lab, BOSS Zhipin, Beijing 100020, China (e-mail: zhuhengshu@gmail.com).

Hui Xiong is with the Hong Kong University of Science and Technology (Guangzhou), Guangzhou 510275, China (e-mail: xionghui@ust.hk).

Digital Object Identifier 10.1109/TKDE.2024.3352640

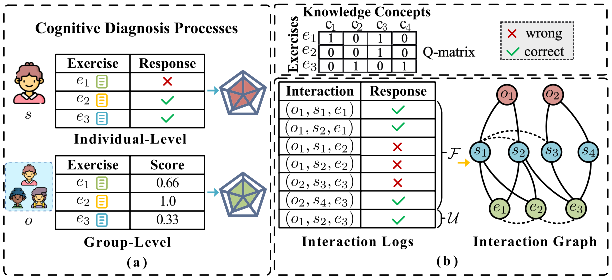


Fig. 1. Illustrative examples of (a) the comparison of group-level cognitive diagnosis and individual cognitive diagnosis; (b) the group-student-exercise heterogeneous construction process. The $Q$-matrix is the correlation matrix between exercises and knowledge concepts, e.g., exercise $e_1$ contains knowledge concepts $c_1$ and $c_3$. The check mark and cross mark indicate that students have answered exercises correctly and incorrectly, respectively. In the interaction graph, solid lines between nodes represent established connections, while dashed lines signify potential relations.

into the group-exercise interaction data, and then aggregate it to form group representations by modeling student representations. As shown in Fig. 1(b), the group interaction $(o_i, e_k, r)$ can be extended into a set of quadruple $(o_i, s_j, e_k, r_{jk})$, where student $s_j$ belong to group $o_i$, $r$ and $r_{jk}$ represent the correct rate and score of $o_i$ and $s_j$ on the exercise $e_k$, respectively. Along this line, one recent approach, named MGCD [13], models student-exercise $(o_i, s_j, e_k, r_{jk})$ and group-exercise $(o_i, e_k, r)$ interactions in a multi-task manner to jointly learn individual and group representations. Despite the impressive effect it achieves, we argue that the intrinsic information between groups, students, and exercises remains unexplored. In this paper, we attempt to comprehensively model the group-student-exercise interaction data, while introducing graph structure to promote information passing between groups and exercises to ameliorate the sparse interaction between them. Moreover, there exists the potential influence among students in a group under the collaborative learning scenario. As shown in Fig. 1(b), students $s_1$ and $s_2$ exhibit similar response behaviors, which may be attributed to their potential collaborative learning behavior. Along this line, we propose a novel relation graph structure to explicitly model such intrinsic associations, which improves both individual and group representation and helps teachers identify the potential collaborations among students in a group.

To be specific, we propose a Relation-guided Dual-side Graph Transformer (RDGT) model for delivering a more productive group-level cognitive diagnosis. First, we construct dual-side relation graphs, i.e., student-side and exercise-side, from the group-student-exercise heterogeneous interaction data for explicitly modeling associations between students and exercises, respectively. Moreover, we implement two improved graph transformers to learn the student representations within the group as well as the exercise representations, by introducing relation encoding to better capture the holistic information about the dual-side graphs including node and edge features. Then, a cognitive diagnosis module is designed for learning the groups' mastery on specific knowledge concepts, which includes an attention-based aggregation strategy to obtain the final group representation and a hybrid loss on group and student performance prediction to learn model parameters. Finally, we conduct extensive experiments on real-world datasets that clearly demonstrate the effectiveness of our model and two case studies reveal that our model can be utilized to identify representative groups and potential collaborations among students. In summary, our key contributions are listed as follows:

- We propose a novel group-level cognitive diagnosis model, namely RDGT, which innovatively introduces a dual-side graph structure to explicitly mine potential associations between students within groups, as well as feature correlations between exercises.
- We implement two novel graph transformers by designing the relation-guided encoding, which utilizes the respective edge attributes on the same interaction path between both student and exercise nodes to calculate the association distance, tailored for mining relationships in educational scenarios.
- We design an inter-group information enhancement module to facilitate prospective inter-group information mining by associating top-k most similar out-group students for each student.
- We conduct extensive experiments on five real-world educational datasets, and the results show the effectiveness of the proposed RDGT model. Furthermore, we perform two meaningful case studies, which demonstrate that the representations learned by RDGT can help us understand representative individuals in the group, and observing the learned correlation matrix of students within the group could assist us in identifying inter-student effects.

## II. Related Work

### A. Cognitive Diagnosis

Cognitive diagnosis (CD) is a type of assessment method for characterizing students' proficiency profile based on their interactive behaviors [14], which is originated from educational psychology and subject to the pedagogical assumption that the cognitive state of each student is stable for a short period of time [15]. Existing research has developed numerous effective cognitive diagnostic models which are mainly classified into two categories, i.e., psychometric theory-based and neural network-based, respectively. The first category of models is designed based on psychological theories for portraying student proficiency state by latent factors (e.g., Item Response Theory (IRT) [6], Multidimensional IRT (MIRT) [7], and Deterministic Inputs, Noisy And gate model (DINA) [8]). For example, in DINA, each student is characterized as a binary vector denoting whether the student has mastered the knowledge concepts corresponding to the exercises, and all relevant skills are needed to have the highest positive response probability.

Another category of models focuses on modeling the complex relationship between students, exercises, and knowledge concepts by incorporating neural networks to accurately profile students' mastery attributes (e.g., NeuralCD [9], RCD [10] and HierCDF [16]). Specifically, NeuralCD, as a representative neural CDM, utilizes multidimensional parameters to depict the cognitive states of students and the attributes of exercises at a fine-grained level, and neural networks are introduced to capture complex interactions from heterogeneous data. In addition, RCD mainly learns more effective representations by modeling the interactive and structural relations with the student-question-concept relation map. HierCDF [16] primarily exploits the dependencies of knowledge concepts to assist in modeling the interaction between students and exercises which proves that the introduction of attribute hierarchy can effectively improve the performance of diagnosis.

Existing research on CD mainly focuses on the assessment of individuals, however, they are not applicable to another classical educational scenario, i.e., collaborative learning, where learners study in groups and influence each other [12]. Recently, MGCD [13] has been proposed to conduct a group-level assessment from a multi-task perspective by jointly training student response records and group interaction data. Although the method performs effectively and achieves attractive results for the GCD task, it has yet to deeply explore the underlying relationship between the group-student-exercise. Specifically, MGCD simply considers student-exercise interactions as additional data for assisting in training the main task of group performance prediction, and it ignores the information coupling between group-exercise instances and student-exercise instances. Besides, students' learning behaviors under the group dimension always influence each other, and the mining of this relationship has yet to be reflected in MGCD.

### B. Graph Representation Learning

Research on graph representation learning [17], [18] has received increasing attention in recent years due to the universality of graphs in the real world, e.g., social networks [19], knowledge graph [20], [21], biological networks [22], recommendation systems [23], etc. The goal of graph representation learning is learning the features of nodes or edges and capturing structural information to generate graph representation vectors (aka, the embedding vector) for further support of various graph mining tasks, such as node classification, community detection and link prediction [24], [25], [26], [27], [28], [29]. This is especially important because the quality of the graph representation vectors will directly affect the performance of the downstream tasks. Extensive approaches have been proposed for learning effective graph representations, which are generally categorized into two genres. The one is traditional graph embedding methods, which employ different techniques to capture the information in the graph, including random walks [30], factorization methods [31] and non-GNN based deep learning [32]. Graph neural networks [33], [34], [35] are another category of graph embedding methods that have been proposed recently, where node representations can be effectively explored from rich neighborhood information. For instance, graph convolutional networks (GCN) leverage efficient symmetric-normalized aggregation to approximate the first-order spectral convolutions on graphs [34]. Graph attention networks (GAT) employ a self-attention mechanism to dynamically aggregate node neighbors' information [36]. Furthermore, heterogeneous graph neural networks are studied to address the graph heterogeneity problem in many real-world situations. For example, HAN [37] and HetG [38] consider the multi-level information and the attention mechanism to improve heterogeneous graph learning. HGSL [39] proposes to jointly perform heterogeneous graph structure learning and GNN parameter learning by generating feature relation subgraphs, thus alleviating the noise and incomplete problem.

In addition to the above approaches, Transformers have been introduced in recent years for powerful modeling graphs benefiting from its capability of capturing long-range and global features on the graph [40], [41]. These models have achieved competitive or even superior performance against GNNs in many applications, such as molecule property prediction [41], catalysts discovery [42] and recommendation systems [43]. For example, Graphormer [41] is proposed as a novel graph transformer architecture, which is designed for better modeling graph-structured data by introducing several simple yet effective structural encoding mechanisms. Min et al. proposed a novel Graph-Masked Transformer (GMT) to effectively incorporates different kinds of interactions among the local neighborhood nodes to produce highly representative embeddings [43]. HTGT [44] proposes a heterogeneous temporal graph transformer framework by integrating both spatial and temporal relations while preserving the heterogeneity to learn node representations for malware detection. HGT [45] designs node- and edge-type dependent parameters to characterize the heterogeneous attention, empowering the proposed heterogeneous graph transformer to maintain dedicated representations for different types of nodes and edges. Our RDGT designs effective relation encoding and introduces it into the powerful dual-side graph transformers, tailored for mining relationships among students and modeling the group-level cognitive diagnosis.

## III. Problem Statement

In this section, we introduce the GCD task and notations used in this paper. Specifically, we use bold uppercase and lowercase letters to represent matrices and vectors, respectively. All import notations have been summarized in Table I. Let $\mathcal{O} = \{o_1, o_2, \ldots, o_L\}$ be the set of $L$ groups, $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ be the set of $N$ students, $\mathcal{E} = \{e_1, e_2, \ldots, e_M\}$ be the set of $M$ exercises, and $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ be the set of $K$ knowledge concepts. Each group consists of a certain number of students, e.g., the $i$th group $o_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,|o_i|}\}$, where $s_{i,*} \in \mathcal{S}$ and $|o_i|$ denotes the size of group $g_i$. We collect two kinds of response data without intersection, which are student-exercise interaction records $\mathcal{U}$ and group-exercise interaction records $\mathcal{F}$, respectively. $\mathcal{U}$, which originates from the student's daily practice, is denoted as a set of triple $(s, e, r_{se})$ where $s \in \mathcal{S}$, $e \in \mathcal{E}$ and $r_{se} \in \{0, 1\}$ is the score that student $s$ got on exercise $e$. Meanwhile, $\mathcal{F}$ arises from the group assessment (i.e., all students in the group completed the same batch of exercises), and we denote it as a set of triple $(o, e, y_{oe})$ where $o \in \mathcal{O}$, $e \in \mathcal{E}$ and $y_{oe} \in [0, 1]$ is the correct rate that group $o$ got on $e$. In addition, we define $\mathcal{Q} = \{q_{ij}\}^{M \times K}$ as the $Q$-matrix where $q_{ij} = 1$ if exercise $e_i$ requires knowledge concept $c_j$ and 0 otherwise.

*Problem Definition:* Given the student-exercise response records $\mathcal{U}$, group-exercise response records $\mathcal{F}$ and the $Q$-matrix $\mathcal{Q}$, the goal of the group-level cognitive diagnosis task is to mine groups' proficiency level on specific knowledge concepts.

TABLE I
SUMMARY OF THE PRIMARY NOTATIONS

| Symbols | Description |
|---|---|
| $\mathcal{O}, \mathcal{S}, \mathcal{E}, \mathcal{C}$ | The set of groups, students, exercises, and knowledge concepts, respectively. |
| $o, s, e, c$ | The group, the student, the exercise, and the knowledge concept. |
| $\mathcal{F}, \mathcal{U}, \mathcal{Q}$ | The group-exercise interaction records, the student-exercise interaction records, and the Q-matrix. |
| $r_{se}, y_{oe}$ | The score that student $s$ got on exercise $e$ and the correct rate that group $o$ got on exercise $e$. |
| $\mathcal{H}, \mathcal{H}_{o \leftrightarrow s}, \mathcal{H}_{s \leftrightarrow e}$ | The global group-student-exercise graph, the group-student affiliation graph, and the student-exercise interaction graph. |
| $\mathcal{G}^S = (\mathcal{X}^S, \mathcal{Z}^S)$ | The student-side relation graph. |
| $\mathcal{X}^S, \mathcal{Z}^S$ | The node features and edge weights of $\mathcal{G}^S$. |
| $\mathcal{G}^E = (\mathcal{X}^E, \mathcal{Z}^E)$ | The exercise-side relation graph. |
| $\mathcal{X}^E, \mathcal{Z}^E$ | The node features and the edge weights of $\mathcal{G}^E$. |
| $\mathbf{G}^S, \mathbf{G}^E, \hat{\mathbf{G}}^S$ | The student-side subgraph and the exercise-side subgraph for group $o$, and the subgraph of $\mathcal{G}^S$ excluding $\mathbf{G}^S$. |
| $\mathcal{N}_{i,j}$ | The common heterogeneous neighbors of node $i$ and node $j$. |
| $\mathbf{X}^S, \mathbf{X}^E$ | The node feature matrices of $\mathbf{G}^S$ and $\mathbf{G}^E$, respectively. |
| $\mathbf{Z}^S, \mathbf{Z}^E$ | The edge weight matrices of $\mathbf{G}^S$ and $\mathbf{G}^E$, respectively. |
| $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ | The query matrix, the key matrix, and the value matrix of the self-attention module. |
| $\mathbf{B}^S, \mathbf{B}^E$ | The student-side relation encoding matrix and the exercise relation encoding matrix. |
| $\mathcal{T}_*$ | The neural interaction function. |
| $\mathbf{h}_{diff}, \mathbf{h}_{disc}$ | The exercise factors representing difficulty and discrimination, respectively. |
| $\lvert \cdot \rvert$ | The cardinality of a set. |
| $L, N, M, K$ | The size of the group set $\mathcal{O}$, the student set $\mathcal{S}$, the exercise set $\mathcal{E}$, and the concept set $\mathcal{C}$, respectively. |
| $N_s, N_e$ | The number of nodes of the subgraph $\mathbf{G}^S$ and $\mathbf{G}^E$, respectively. |
| $d_s, d_e, d_h, d$ | The student representation dimension, the exercise representation dimension, the dimension of each self-attention head, and the hidden dimension, respectively. |

## IV. METHODOLOGY

In this section, we first give an overall overview of our proposed model **RDGT** (short for **R**elation-Guided **D**ual-Side **G**raph **T**ransformer). Afterward, we delve into each part of the model with a comprehensive explanation.

*Overview.* Our RDGT model employs a novel relation-guided graph transformer to explore the internal associations of student-side and exercise-side respectively from the heterogeneous interaction graphs of group-student-exercise. It effectively improves representations by adaptively learning coupling information from the dual-side graphs, allowing for accurate diagnosis of groups' proficiency levels on knowledge concepts. As illustrated in Fig. 2, the architecture of RDGT comprises three components, including the dual-side graph construction, the relation encoding-based graph transformer, and the cognitive modeling module. Specifically, we first utilize the group-student-exercise interaction data to build a global interaction heterogeneous graph and exploit the structural information on the global graph to construct student-side and exercise-side relation sub-graphs, respectively. In the relation encoding-based graph transformer, we encode the structural feature on dual-side relation graphs into relation encoding as a guided inductive bias and introduce it into the self-attention layer to capture intrinsic associations in the graph adaptively for more effective learning of representations. Particularly, an attention-based aggregation strategy is applied to combine the learned student representations and the initial group representation to obtain the final group representation. Finally, we construct a cognitive modeling module, which consists of a neural interaction function and multiple neural network layers, to model the complicated interaction between the group and the exercise for accurately predicting the group's performance on a given exercise.

### A. Relation-Guided Dual-Side Graph Construction

In contrast to traditional CD which primarily relies on individual response data to diagnose students, GCD involves a notably distinct context where students practice exercises and are assessed in a group setting. In this context, interaction data including responses from both the same student and different students are no longer isolated, rather there are highly intrinsic correlations that should not be ignored during the exploration of student interaction behavior. Given the advantage of graph data structures in effectively modeling complex connections and intrinsic relationships between entities, we build a heterogeneous graph to integrally model the connections between groups, students, and exercises.

The global group-student-exercise heterogeneous graph $\mathcal{H}(\mathcal{O} \cup \mathcal{S} \cup \mathcal{E}, \mathcal{R}_{os} \cup \mathcal{R}_{se})$ is an undirected graph as shown in the left part of Fig. 2(a), which consists of three categories of nodes and two kinds of relation types, where $\mathcal{O}$, $\mathcal{S}$ and $\mathcal{E}$ are the sets of groups, students, and exercises, respectively, $\mathcal{R}_{os}$ refers the set of group-student affiliations, and $\mathcal{R}_{se}$ denotes the set of student-exercise interactions from training data including both student-based interactions from $\mathcal{F}$ and group-based interaction from $\mathcal{U}$. If the group-student relation $link_{o_i \leftrightarrow s_j} = 1$, student $s_j$ belongs to group $o_i$. In addition, the relation $link_{s_i \leftrightarrow e_j} = 1$ indicates that student $s_i$ has performed a response on exercise $e_j$. It is important to note that the connections between students and exercises are undirected (i.e., both receive information about each other's responses) and that we utilize the results of the interactive responses as the corresponding edge characteristics. The heterogeneous graph structure allows for the efficient end-to-end learning of node representations in graph neural networks and its effectiveness is verified in the experiment section (details in Section V-B). Nevertheless, it neglects the connections of the same type of nodes within the group (e.g., student-student and exercise-exercise), making it crucial to actively mine such potential correlations for group-level dimension diagnosis. Therefore, in this paper, we propose a dual-side graph construction strategy to model the internal associations of students and exercises from both sides, respectively, for explicitly mining the intrinsic information on the group dimension. In what follows, we elaborate on the graph construction of each side, respectively.

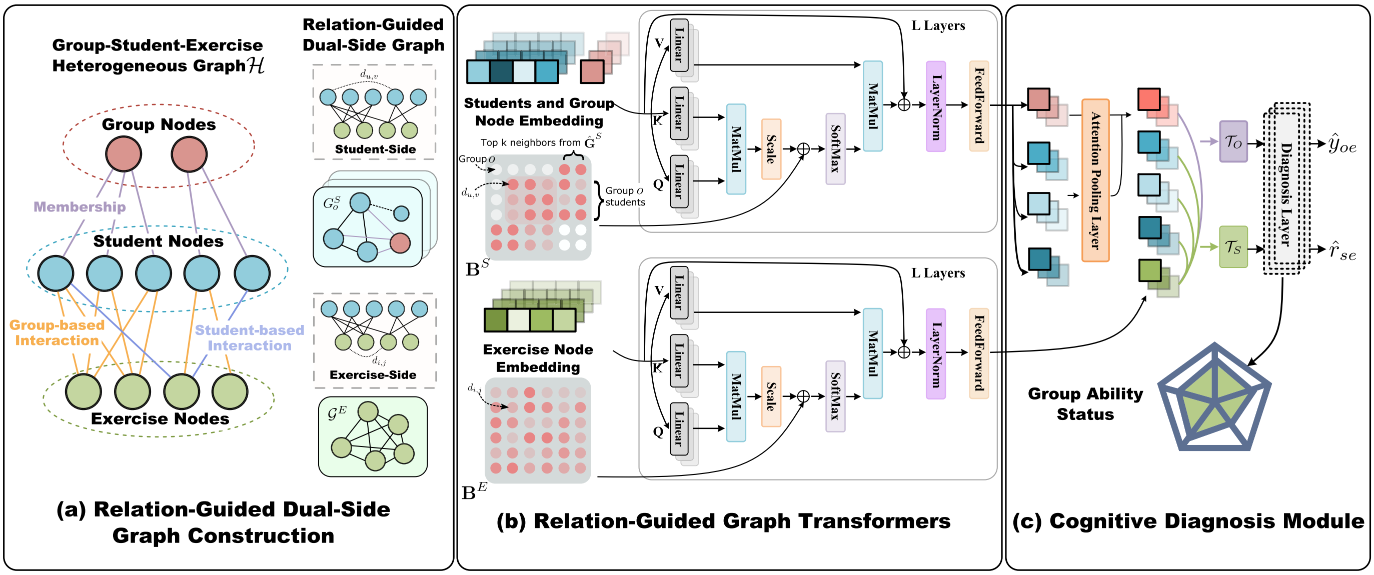


Fig. 2. Overview architecture of RDGT: (a) the construction of relation-guided dual-side graph; (b) the relation-guided graph transformers; and (c) The cognitive diagnosis module.

*1) Student-Side Relation Graph Construction:* In the real world, students in a physically meaningful group (e.g., a class) always have mutual influences on each other during learning activities, for example, students who are deskmates may have very similar study behavior or students with good relationships in the class may enjoy comparable learning abilities. Obviously, it is essential to explore the relationship between individuals within the group during the diagnosis process. Therefore, we propose first to extract the student-exercise interaction graph $\mathcal{H}_{s\leftrightarrow e}$ from the initial heterogeneous graph $\mathcal{H}$. Then, we utilize the interaction features between students and exercises on this graph to mine collaborative information. For any student $u$ and student $v$, we define the association distance $d_{uv}$ as the collaborative similarity between them, as below

$$
\begin{aligned}
d_{u,v} &= sim\left(\mathbf{m}_u, \mathbf{m}_v\right), \\
\mathbf{m}_u &= \left[m_{u,t_1}, m_{u,t_2}, \ldots, m_{u,t_{|\mathcal{N}_{u,v}|}}\right], \\
\mathbf{m}_v &= \left[m_{v,t_1}, m_{v,t_2}, \ldots, m_{v,t_{|\mathcal{N}_{u,v}|}}\right],
\end{aligned} \tag{1}
$$

where $\mathcal{N}_{u,v}$ denotes the common heterogeneous neighbors of node $u$ and node $v$, which should be a set of $|\mathcal{N}_{u,v}|$ exercise nodes, $m_{u,t}$ denotes the feature of the edge $(u, t)$ set as the score that the students $u$ get on the exercise $t$, i.e., $r_{ut}$, and $sim(\cdot)$ denotes the similarity function, such as Cosine similarity, on two vectors $\mathbf{m}_u, \mathbf{m}_v \in \mathbb{R}^{|\mathcal{N}_{v,u}|}$. The proposed method of calculating the association distance is essentially a deep excavation of the similarity in behavioral performance between student pairs. Based on the calculated distance information, we construct the student-side relationship as a complete graph with nodes $\mathcal{S}$ and edges $\mathcal{R}$: $\mathcal{G}^S = (\mathcal{X}^S, \mathcal{Z}^S)$, where $\mathcal{X}^S \in \mathbb{R}^{|\mathcal{S}|\times d_s}$ is student node features and $\mathcal{Z}^S \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denotes edge weights obtained from the association distance. As shown in the blue part of Fig. 2(a), the student relationship graph of each group is a subgraph of $\mathcal{G}^S$, which consists of all student nodes in the group, as well as the edges between each node pair has a weight characterizing the initial proximity.

*2) Exercise-Side Relation Graph Construction:* Indeed, the correlation among exercises is also necessary for the learning of their characteristics. However, this relationship cannot be adequately understood solely through the attributes of knowledge concept [10], e.g., two exercises that involve similar knowledge concepts but have completely different difficulty properties. Here, we propose to explore the underlying impact between exercises based on their interaction with students. Similar to the student-side graph, we define the association distance between exercise node $i$ and $j$ as

$$
\begin{aligned}
d_{i,j} &= sim\left(\mathbf{n}_i, \mathbf{n}_j\right), \\
\mathbf{n}_i &= \left[n_{i,t_1}, n_{i,t_2}, \ldots, n_{i,t_{|\mathcal{N}_{i,j}|}}\right], \\
\mathbf{n}_j &= \left[n_{j,t_1}, n_{j,t_2}, \ldots, n_{j,t_{|\mathcal{N}_{i,j}|}}\right],
\end{aligned} \tag{2}
$$

where $\mathcal{N}_{i,j}$ denotes the common heterogeneous neighbors of nodes $i$ and $j$, which should be a set of $|\mathcal{N}_{i,j}|$ student nodes, $n_{i,t}$ is the feature of student-exercise edge $(t, i)$ set as the score that the student $t$ get on the exercise $i$, i.e., $r_{ti}$. Afterwards, as shown in the green part of Fig. 2(a), we construct the exercise-side relationship graph $\mathcal{G}^E = (\mathcal{X}^E, \mathcal{Z}^E)$, where $\mathcal{X}^E \in \mathbb{R}^{|\mathcal{E}|\times d_e}$ is exercise node features, and $\mathcal{Z}^E \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ denotes edge weights represented by the association distance.

## *B. Relation-Guided Graph Transformer*

In this part, we propose a novel graph transformer based on relation encoding for effective representation learning, which is described below in terms of both student and exercise, respectively.

*1) Student-Side Graph Transformer:* Mining the fine-grained structure information from the constructed student-side relation graph is crucial for learning student representations, as each node within a group is influenced by others to varying degrees. Encoding graphs with the transformer to uncover deep correlations between nodes is regarded as remarkably powerful due to its ability to capture holistic information from the global receptive field [41]. The iconic transformer architecture [46] comprises multiple transformer layers, each of which is composed of a multi-head self-attention block followed by a position-wise feed-forward network (FFN) layer that features both residual connections and layer normalization operations. Specifically, by sampling all student nodes in the group $o \in \mathcal{O}$ and edge weights among them from $\mathcal{G}^S$, we obtain the student-side subgraph $\mathbf{G}^S$ with corresponding node feature matrix $\mathbf{X}^S \in \mathbb{R}^{N_s \times d_s}$ and edge weight matrix $\mathbf{Z}^S \in \mathbb{R}^{N_s \times N_s}$, where $N_s$ denotes the number of nodes within the subgraph. Then, the multi-head self-attention module at layer $l$ can be defined as,

$$\begin{aligned}
&\mathbf{Q}^S_{l,k} = \mathbf{X}^S_{l-1}\mathbf{W}^{\mathbf{Q}}_{l,k},\ \mathbf{K}^S_{l,k} = \mathbf{X}^S_{l-1}\mathbf{W}^{\mathbf{K}}_{l,k},\ \mathbf{V}^S_{l,k} = \mathbf{X}^S_{l-1}\mathbf{W}^{\mathbf{V}}_{l,k}\\
&\mathbf{A}^S_{l,k} = softmax\left(\frac{\mathbf{Q}^S_{l,k}(\mathbf{K}^S_{l,k})^T}{\sqrt{d_h}}\right),\ \mathbf{X}^S_{l,k} = \mathbf{A}^S_{l,k}\mathbf{V}^S_{l,k},\\
&\mathbf{X}^S_l = concat\left(\mathbf{X}^S_{l,1}, \mathbf{X}^S_{l,2}, \ldots, \mathbf{X}^S_{l,H}\right),
\end{aligned} \tag{3}$$

where $\mathbf{X}^S_{l-1}$ is the node feature matrix at $(l-1)$th layer, $\mathbf{Q}^S_{l,k}, \mathbf{K}^S_{l,k}, \mathbf{V}^S_{l,k} \in \mathbb{R}^{N_s \times d_h}$ are the query matrix, the key matrix, and the value matrix of the $k$th head at the self-attention layer $l$, respectively, $\mathbf{W}^{\mathbf{Q}}_{l,k}, \mathbf{W}^{\mathbf{K}}_{l,k}, \mathbf{W}^{\mathbf{V}}_{l,k} \in \mathbb{R}^{d_s \times d_h}$ are the corresponding trainable parameter matrices, $d_h = \frac{d_s}{H}$ denotes the dimension of each self-attention head, $H$ stands for the number of self-attention heads, and $softmax(\cdot)$ and $concat(\cdot)$ denotes the row-wise softmax function and concatenation operation, respectively. Specifically, the multi-head mechanism enables the model to implicitly learn representation from different aspects. The output $\mathbf{X}^S_l$ is then passed into a FNN layer, as defined below

$$\begin{aligned}
&\widetilde{\mathbf{X}}^S_l = LayerNorm\left(\mathbf{X}^S_{l-1} + \mathbf{X}^S_l\right),\\
&\mathbf{X}^S_l = LayerNorm\left(\sigma\left(\widetilde{\mathbf{X}}^S_l\mathbf{W}_{l,1} + \mathbf{b}_{l,1}\right)\mathbf{W}_{l,2} + \mathbf{b}_{l,2}\right),
\end{aligned} \tag{4}$$

where $\mathbf{W}_{l,1} \in \mathbb{R}^{d_s \times d}$, $\mathbf{b}_{l,1} \in \mathbb{R}^d$, $\mathbf{W}_{l,2} \in \mathbb{R}^{d \times d_s}$, and $\mathbf{b}_{l,2} \in \mathbb{R}^{d_s}$ are trainable parameters, $d$ is the hidden dimension, $LayerNorm(\cdot)$ denotes the LayerNorm operation [47], and $\sigma(\cdot)$ denotes the activation function (e.g., GELU [48]).

Unlike NLP tasks [46], [49], which preserve the structural information of chain-structured language by inputting position encoding in the transformer, graph-structured data is difficult to generalize such operations. Therefore, we devise relation encoding that relates to the associative distance information between students (i.e., the edge feature matrix $\mathbf{Z}^S$ in $\mathbf{G}^S$) to inject the inductive bias in the self-attention module for learning student representations.

*Relation Encoding:* For each pair of student nodes $s_u$ and $s_v$ in the student-side relation sub-graph $\mathbf{G}^S$, we first obtain the edge feature $z^S_{u,v}$ between them, which is calculated from the association distance, and then encode it as an attention scalar $b^S_{u,v}$ in a relation attention matrix $\mathbf{B}^S = \{b^S_{u,v}\}_{u,v} \in \mathbb{R}^{N_s \times N_s}$ as below

$$b^S_{u,v} = sigmoid\left(z^S_{u,v}\mathbf{h}^S_1 + \mathbf{b}^S_1\right)^T\mathbf{h}^S_2 + b^S_2, \tag{5}$$

where $z^S_{u,v}$ is the edge feature value between node $s_u$ and $s_v$, $\mathbf{h}^S_1, \mathbf{b}^S_1, \mathbf{h}^S_2 \in \mathbb{R}^d$ and $b^S_2 \in \mathbb{R}$ are the trainable projection parameters, and $sigmoid(\cdot)$ denotes the sigmoid activation function. We then introduce structural information by adding the encoded relation attention matrix to the attention score component of the self-attention module, which is formulated as follows:

$$\mathbf{A}^S_{l,k} = softmax\left(\frac{\mathbf{Q}^S_{l,k}(\mathbf{K}^S_{l,k})^T}{\sqrt{d_h}} + \mathbf{B}^S\right). \tag{6}$$

Different from the previous structural encoding strategies that encode path features between node pairs in graph transformer architectures [40], [41], [50], our proposed method utilizes the respective edge attributes on the same interaction path between student nodes to calculate the association distance, tailored for relationship mining among students (especially within groups) in educational scenarios. Essentially, the introduction of $\mathbf{B}^S$ into a single transformer layer actually involves serving relation encoding as guiding information, enabling each student node within a group to adaptively attend to other nodes and thus more effectively aggregate intra-group influences for learning student representations. In addition to intra-group effects, we consider that the association of students between groups is beneficial for students' perceptions of ability, e.g., two students who are friends although not in the same group usually have similar study habits. Thus, we further propose inter-group information enhancement to promote the learning of student representations.

*Inter-Group Information Enhancement:* For any student node $s_u \in \mathbf{G}^S$, we calculate its behavioral similarity $d_{u,v'}$ to other out-group nodes $s'_v \in \hat{\mathbf{G}}^S$ using the association distance in (1), where $\hat{\mathbf{G}}^S = \mathcal{G}^S \setminus \mathbf{G}^S$ denotes the sub-graph of the student-side relation graph $\mathcal{G}^S$ excluding $\mathbf{G}^S$. We then obtain the similarity set of between node $s_u$ and all nodes outside the group, i.e., $\{d_{u,v'}\}_{v' \in \hat{\mathbf{G}}^S}$. As shown in Fig. 2(b), we select top-k nodes from $\hat{\mathbf{G}}^S$ with the highest behavioral similarity to $s_u$ to construct its first-order neighbors. Simultaneously, we add an abstract node in $\mathbf{G}^S$ for signifying the group and connecting it to all student nodes in the group. With this strategy, the representations of student nodes receive not only local influence from the intra-group but also are enhanced by the global information of the inter-group. Furthermore, the group node, as a high-level element, enables the perception of a global perspective and the aggregation of productive information to promote the improvement of students' representations.

*2) Exercise-Side Graph Transformer:* Here, we leverage a transformer to model the exercise-side relation graph by introducing relation encoding that implies information about the association of exercises. Similar to the student representation learning, which is optimized at the group level, we implement the exercise-side graph transformer by sampling the subgraph $\mathbf{G}^E$ from the whole exercise-side relation graph $\mathcal{G}^E$ to learn the exercise representations. For any two exercise nodes $e_i$ and $e_j$

in $\mathcal{G}^E$, we encode the association distance between them into attention scalar $b^E_{i,j}$ as the relation encoding

$$b^E_{i,j} = sigmoid\left(z^E_{i,j}\mathbf{h}^E_1 + \mathbf{b}^E_1\right)^T \mathbf{h}^E_2 + b^E_2, \tag{7}$$

where $z^E_{i,j}$ is the edge feature value between node $e_i$ and $e_j$, $\mathbf{h}^E_1, \mathbf{b}^E_1, \mathbf{h}^E_2 \in \mathbb{R}^d$ and $b^E_2 \in \mathbb{R}$ are the trainable parameters. We then obtain the relation encoding matrix $\mathbf{B}^E = \{b^E_{i,j}\}_{i,j} \in \mathbb{R}^{N_e \times N_e}$, and $N_e$ denotes the number of exercise nodes within the group. To incorporate it into the self-attention module, we rewrite the attention score matrix $\mathbf{A}^E_{l,k} \in \mathbb{R}^{N_e \times N_e}$ in the transformer layer applied to learn exercise representations as follows:

$$\mathbf{A}^E_{l,k} = softmax\left(\frac{\mathbf{Q}^E_{l,k}\left(\mathbf{K}^E_{l,k}\right)^T}{\sqrt{d_h}} + \mathbf{B}^E\right), \tag{8}$$

where $\mathbf{Q}^E_{l,k} \in \mathbb{R}^{N_e \times d_h}$ and $\mathbf{K}^E_{l,k} \in \mathbb{R}^{N_e \times d_h}$ are the query matrix and key matrix. Notably, the definition here is similar to (3). The introduction of this guided information enables transformer architecture to expediently aggregate enriched information from the exercise-side relation graph and potentiate the learning of the exercise representations.

### C. Cognitive Diagnosis Module

The goal of the cognitive modeling module is to train the RDGT jointly using group and student response records on the exercises and model the group-dimensional cognitive interactions by predicting the groups' exercising performance.

*1) Group Representation Aggregation:* As shown in the Fig. 2(c), for the group $o \in \mathcal{O}$, we denote the student-side node embedding matrix outputted by the proposed relation encoding-based graph transformer as $\hat{\mathbf{X}} \in \mathbb{R}^{(1+|o|)\times d_s}$, which includes an abstract group node and $|o|$ student nodes. Although the virtual group node learns extensive information, the actual group representation supposedly is formed by aggregating the features of all students in the group. Therefore, referring to [13], [51], we employ an attention mechanism to aggregate the representations of students in the group which reflects that each student contributes differently to the group's ability

$$\mathbf{x}^O_o = \sum_{\mathbf{x}^S_j \in \hat{\mathbf{X}}^S} \lambda_j \mathbf{x}^S_j, \tag{9}$$

where $\mathbf{x}^S_j \in \mathbb{R}^{d_s}$ is the $j$th node representation, $\mathbf{x}^S_0$ stands for the learned representation of the virtual group node, and $\lambda_j$ presents the contribution weight of the $j$th node

$$\tilde{\lambda}_j = ReLU\left(\left(\mathbf{x}^S_j\right)^T \mathbf{W}^k + \mathbf{x}^S_0 \mathbf{W}^q\right)\mathbf{h},$$
$$\lambda_j = \frac{exp\left(\tilde{\lambda}_j\right)}{\sum_{1 \le j' \le |o|} exp\left(\tilde{\lambda}_{j'}\right)}. \tag{10}$$

where $\mathbf{W}^k, \mathbf{W}^q \in \mathbb{R}^{d_s \times d}$ are the key matrix and query matrix of the attention layer, and $\mathbf{h} \in \mathbb{R}^d$ is the weight vector for projecting attention scores.

*2) Interaction Layer:* To model the complicated interactions of group-exercise and student-exercise for more reliable cognitive diagnosis, we adopt the widely utilized neural interaction function [9] in our model. It can seamlessly integrate with non-linear neural network layers, and its capability to model high-dimensional interactive elements (e.g., group, student, and exercise) enables the acquisition of extensive knowledge and the presentation of interpretable information. In this work, the interaction layer consists of the interaction function and the diagnosis layer, defined as follows:

$$\begin{cases} \hat{r}_{se} = \mathrm{MLP}_S\left(\mathcal{T}_S\left(\mathbf{x}^S_s, \mathbf{x}^E_e\right)\right) \\ \hat{y}_{oe} = \mathrm{MLP}_O\left(\mathcal{T}_O\left(\mathbf{x}^O_o, \mathbf{x}^E_e\right)\right) \end{cases}, \tag{11}$$

where $\hat{r}_{se}$ and $\hat{y}_{oe}$ denote the predicted response results of student $s$ and group $o$ on the exercise $e$, $\mathrm{MLP}_S$ and $\mathrm{MLP}_O$ are two different MLP networks, and $\mathcal{T}_*$ indicates the neural interaction function [9]

$$\begin{cases} \mathcal{T}_S\left(\mathbf{x}^S_s, \mathbf{x}^E_e\right) = \mathcal{Q}_e \circ \left(\mathbf{x}^S_s - \mathbf{h}_{diff}\right) \times \mathbf{h}_{disc} \\ \mathcal{T}_O\left(\mathbf{x}^O_o, \mathbf{x}^E_e\right) = \mathcal{Q}_e \circ \left(\mathbf{x}^O_o - \mathbf{h}_{diff}\right) \times \mathbf{h}_{disc} \end{cases}, \tag{12}$$

where the $\mathbf{h}_{diff} = \mathbf{x}^E_{e\,[:-1]} \in \mathbb{R}^{d_e - 1}$ and $\mathbf{h}_{disc} = \mathbf{x}^E_{e\,[-1]} \in \mathbb{R}$ are two exercise factors representing difficulty and discrimination, which are split from the exercise representation $\mathbf{x}^E_e$, $\circ$ and $\times$ denote the element-wise product and the multiplication operation, respectively, and $\mathcal{Q}_e$ denotes the knowledge concept attribute corresponding to $e$ originates from the $Q$-matrix $\mathcal{Q}$.

*3) Loss Function:* In the training phase, we jointly evaluate the performance of the predicted student-exercise interaction and group-exercise interaction. We believe that predicting the students' exercising performance with label information from $\mathcal{U}$ is beneficial for the training of the group-level diagnosis task, which makes the group and student information more coupled and alleviates the interaction sparsity problem. Specifically, for each student group $o$, we adopt the cross-entropy loss function for students' exercising performance prediction as follows:

$$\mathcal{L}^{stu}_o = -\sum_{(s,e,r_{se}) \in \mathcal{U}_o \cup \mathcal{F}_o} \hat{r}_{se} \log\ r_{se} + (1 - \hat{r}_{se}) \log(1 - r_{se}), \tag{13}$$

where $\mathcal{U}_o \subset \mathcal{U}$ and $\mathcal{F}_o \subset \mathcal{F}$ stand for the student's interaction records related to the group $o$. Then, we choose the mean square error loss (MSE) function for predicting the groups' correct rate for a given exercise

$$\mathcal{L}^{grp}_o = \sum_{(o,e,r_{oe}) \in \mathcal{F}_o} \left(\hat{y}_{oe} - y_{oe}\right)^2, \tag{14}$$

Finally, we obtain the complete optimization objective function by summing the loss functions of the above two objectives with weight coefficients to balance the scale

$$\mathcal{L}_o = \frac{1}{|\mathcal{F}_o|}\mathcal{L}^{grp}_o + \frac{\gamma}{|\mathcal{U}_o \cup \mathcal{F}_o|}\mathcal{L}^{stu}_o. \tag{15}$$

where $\gamma$ is the weight coefficient to control the influence of auxiliary student-exercise interaction data.

In summary, in the modeling of group-exercise interactions, we also utilize the interaction data between students and the

TABLE II
THE STATISTICS OF ALL DATASETS

| Statistics | ASSISTment12 | NIPS-Edu | SLP-math | SLP-biology | SLP-physics |
|---|---|---|---|---|---|
| #Students | 1,078 | 2,415 | 4,173 | 3,052 | 4,314 |
| #Groups | 111 | 177 | 172 | 111 | 170 |
| #Exercises | 11,564 | 921 | 226 | 120 | 115 |
| #Knowledge concepts | 182 | 86 | 40 | 22 | 37 |
| Avg. group size | 14.31 | 14.54 | 24.26 | 27.49 | 25.37 |
| Avg. responses per student | 78.64 | 149.33 | 125.70 | 111.13 | 107.11 |
| Avg. responses per group | 9.57 | 35.81 | 73.77 | 55.32 | 81.44 |

exercises within and outside the group for joint training. This co-optimization strategy conforms better to the graph structure we have constructed and enables deeper mining of group features.

## V. EXPERIMENTS

In this section, we conduct extensive experiments on five real-world datasets with the aim of validating the effectiveness and superiority of our model. Specifically, we will answer the following research questions to unfold the experiments.

*RQ1:* What about the effectiveness and superiority of the proposed model in the group-level performance prediction task?

*RQ2:* Can the introduction of student-exercise interactions $\mathcal{U}$ as auxiliary data effectively model the exercising behavior of groups?

*RQ3:* What are the benefits of each component including the dual-side relation graph, inter-group information enhancement, and attention-based group representation aggregation in our model?

*RQ4:* How do the hyper-parameters influence the effectiveness of the proposed RDGT?

*RQ5:* Can our approach identify the representative individuals and uncover potential collaborations among students?

### A. Experimental Setting

*1) Dataset Description:* In this paper, we conducted experiments on three public education benchmarks: ASSISTment12 [52], NIPS-Edu [53], and SLP [54]. ASSISTment12 dataset is collected from ASSISTments online tutoring service system and contains student exercising data for the school year, which has been widely used in cognitive diagnosis tasks. NIPS-Edu dataset comes from a diagnosis question competition (i.e., the NeurIPS 2020 Education Challenge), where students' answer records to mathematics questions are provided. As for SLP, is a dataset collected from an online learning platform called Smart Learning Partner (SLP), which intentionally records learners' data from multiple dimensions and subjects to provide rich content. Specifically, we selected three sub-datasets in the SLP corresponding to three different subjects as experimental data: SLP-math, SLP-biology, and SLP-physics. All datasets above contain group labels (i.e., the class to which the students belong), and students from the same class share the same label category. We followed MGCD [13] to construct two types of response records without overlap for each dataset, i.e., student-exercise logs and group-exercise logs, where the former indicates students' responses on exercises, while the latter denotes all students' answers to the same exercise within the same group and takes the correct rate as the response result. To ensure reasonableness, we screened out the groups with few than three students and fewer than two response logs, and notably, students whose response counts are substantially greater than the average are also eliminated. The statistics of five datasets are shown in Table II.

*2) Baseline Approaches and Evaluation Metrics:* To verify the effectiveness of our proposed RDGT framework, we compared it with several baselines. Specifically, the selected comparison approaches fall into two categories. One category is the representative CD methods (such as IRT, MIRT, MF, and NeuralCD) and a state-of-the-art GCD method (i.e., MGCD). The details are displayed as follows:

- *IRT* [6]: IRT is one of the most popular cognitive diagnosis methods, which performs unidimensional modeling of student profiles and exercise attributes by a linear function.
- *MIRT* [7]: MIRT is a multidimensional extension of the IRT model that models the characteristics of students and exercises from multiple dimensions.
- *MF* [55], [56]: MF is a latent factor model aiming at predicting students' exercising performance by factoring score matrix and can obtain the latent trait vectors of students and exercises.
- *NeuralCD* [9]: NeuralCD is one of the most representative deep learning-based CD models. It leverages neural networks to explore and model the high-order and complex interaction between students and exercises.
- *MGCD* [13]: MGCD is a state-of-the-art GCD framework that models group-exercise responses from a multi-task learning perspective to alleviate the interaction sparsity problem.

In addition, we selected several competitive graph representation learning methods as another category of baselines including RGCN, SignedGCN, HGT, GraphTrans, and GATv2. Specifically, they were introduced for representation learning and followed by a neural diagnosis layer (11) for adapting the group performance prediction.

- *RGCN* [57]: RGCN is a classical heterogeneous graph learning method using multiple edge relations to model. We constructed group-student-exercise interaction graphs and used this strategy for information aggregation and representation learning.
- *SignedGCN* [58]: SignedGCN leverages balance theory to jointly model node associations in signed networks from positively connected sets and negatively connected sets.

TABLE III
EXPERIMENTAL RESULTS ON GROUP PERFORMANCE PREDICTION WITHOUT $\mathcal{U}$

| Datasets | ASSISTment12 | | NIPS-Edu | | SLP-math | | SLP-biology | | SLP-physics | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE |
| IRT | 0.2514 | 0.1996 | 0.2346 | 0.1875 | 0.1989 | 0.1504 | 0.2278 | 0.1714 | 0.2047 | 0.1616 |
| MIRT | 0.2297 | 0.1970 | 0.2442 | 0.1973 | 0.2495 | 0.1992 | 0.2183 | 0.1662 | 0.2256 | 0.1738 |
| MF | 0.2360 | 0.1768 | 0.2111 | 0.1706 | 0.1822 | 0.1456 | 0.2247 | 0.1695 | 0.1982 | 0.1526 |
| NCD | 0.2221 | 0.1694 | 0.2107 | 0.1654 | 0.1774 | 0.1154 | 0.1963 | 0.1461 | 0.1869 | 0.1481 |
| MGCD | 0.2124 | 0.1598 | 0.1955 | 0.1507 | 0.1789 | 0.1167 | 0.1879 | 0.1371 | 0.1815 | 0.1403 |
| RGCN | 0.1984 | 0.1603 | 0.2253 | 0.1742 | 0.2015 | 0.1538 | 0.2059 | 0.1584 | 0.1907 | 0.1494 |
| SignedGCN | 0.1896 | 0.1346 | 0.2038 | 0.1569 | 0.1847 | 0.1423 | 0.1967 | 0.1473 | 0.1838 | 0.1432 |
| HGT | 0.1911 | 0.1556 | 0.1988 | 0.1603 | 0.1826 | 0.1501 | 0.1943 | 0.1366 | 0.1896 | 0.1453 |
| GraphTrans | 0.1845 | 0.1349 | 0.1931 | 0.1474 | 0.1719 | 0.1114 | 0.1855 | 0.1382 | 0.1769 | 0.1345 |
| GATv2 | 0.1836 | 0.1401 | 0.1939 | 0.1485 | 0.1689 | 0.1001 | 0.1837 | 0.1353 | 0.1799 | 0.1389 |
| RDGT-Int | 0.1969 | 0.1411 | 0.1924 | 0.1476 | 0.1670 | 0.1064 | 0.1866 | 0.1378 | 0.1758 | 0.1366 |
| RDGT | **0.1747** | **0.1266** | **0.1819** | **0.1371** | **0.1566** | **0.0954** | **0.1738** | **0.1297** | **0.1643** | **0.1128** |

The bold values indicate the best experimental results, and underlined markings indicate suboptimal results.

We introduce the student-exercise interaction results as signed information.

- *HGT* [45]: HGT designs node- and edge-type dependent parameters to characterize the heterogeneous self-attention mechanism. We introduce it to model the student-exercise graph.
- *GraphTrans* [59]: GraphTrans integrates GNN modules and self-attention models for mining local and global information. We implemented GraphTrans by introducing self-attention modules to learn node representations on the student-exercise graph.
- *GATv2* [60]: GATv2 works as an improvement of GAT [36], which achieves a universal approximator attention function by modifying the operation order of neighbor information aggregation. We introduce it into the modeling of the student-exercise graph for learning the representations.

It is worth noting that since the traditional CD approaches fail to provide a solution for group-level diagnosis with student-exercise response data, the above baselines based on the traditional CD are improved by learning student representations separately and then aggregating them into group representations on average for GCD task rather than learning groups as individual units.

Group-level cognitive diagnosis is essentially a regression task whose principal form is exhibited by predicting the correct rate of a group for a given exercise. Thus, to evaluate the performance of all methods, we used two popular metrics, i.e., root mean square error (RMSE) and mean absolute error (MAE). Referring to [13], [16], in our experiments, we randomly split the group-exercise interaction data of each dataset into two parts in the ratio of 8:2 as the training set and testing set, respectively. Meanwhile, we divide 90% as train data and 10% as validation data respectively from the training set.

*3) Parameter Settings:* We implemented all models with PyTorch by Python and conducted our experiments on a Linux server with two Nvidia GeForce GTX 1080Ti GPUs. All models were tuned to have the best performance to ensure fairness.

To set up the training process, we initialized all network parameters with Xavier initialization [61]. Each parameter is sampled from $U(-\sqrt{2/(n_{in}+n_{out})}, \sqrt{2/(n_{in}+n_{out})})$, where $n_{in}$ and $n_{out}$ denote the numbers of neurons feeding in and feeding out, respectively. We use the Adam algorithm [62] as the optimizer, where the learning rate was searched in [0.001, 0.005, 0.01, 0.015, 0.02]. The number of diagnosis layers $L$ (11) is set to 2 and the corresponding dimensions are 128 and 1, respectively. The coefficient $\gamma$ was searched in [1e-4, 1e-3, 1e-2, 1e-1, 1]. The value of k in the inter-group information enhancement module is set to 5. We adopt the cosine similarity as the similarity calculation function $sim(\cdot)$ (1) and (2).

### *B. Performance Comparison (RQ1 and RQ2)*

To answer RQ1, in this part, we validate the superiority of the proposed RDGT. Specifically, we first conducted the groups' performance prediction experiments in the above five datasets that only contain group-exercise responses $\mathcal{F}$. Table III shows the experimental results of the proposed models' performance compared with the baselines. We highlighted the best results of all models in boldface and underlined the suboptimal results. According to the results, there are several observations. First, our model has significant improvements over the baseline model on all datasets. Especially, compared to the state-of-the-art method MGCD, our model has an average 0.02 performance improvement on all datasets in terms of both metrics and reaches even a 0.04 improvement on the ASSISTment12 dataset. Second, MGCD shows clear advantages over baselines with improved group diagnostics using traditional CD methods, but it does not outperform the graph-based baseline on several datasets such as ASSISTment12 and SLP-math. This demonstrates the effectiveness of introducing graph structures for modeling group-level diagnosis. Finally, SignedGCN exhibits superiority compared to RGCN, which demonstrates the feasibility of incorporating student-exercise response results as edge features into the graph structure for modeling interactions and learning representations.

TABLE IV
EXPERIMENTAL RESULTS ON GROUP PERFORMANCE PREDICTION WITH $\mathcal{U}$

| Datasets | ASSISTment12 | | NIPS-Edu | | SLP-math | | SLP-biology | | SLP-physics | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE |
| IRT | 0.2520 | 0.2036 | 0.2298 | 0.1836 | 0.1957 | 0.1427 | 0.2162 | 0.1659 | 0.1977 | 0.1535 |
| MIRT | 0.2429 | 0.2001 | 0.2395 | 0.1897 | 0.2376 | 0.1863 | 0.2207 | 0.1684 | 0.2061 | 0.1624 |
| MF | 0.2267 | 0.1697 | 0.2088 | 0.1678 | 0.1807 | 0.1396 | 0.2093 | 0.1618 | 0.1905 | 0.1496 |
| NCD | 0.2171 | 0.1718 | 0.2048 | 0.1587 | 0.1765 | 0.1123 | 0.1885 | 0.1457 | 0.1818 | 0.1437 |
| MGCD | 0.2012 | 0.1539 | 0.1990 | 0.1540 | 0.1793 | 0.1184 | 0.1824 | 0.1426 | 0.1762 | 0.1359 |
| RGCN | 0.1939 | 0.1577 | 0.2036 | 0.1554 | 0.1833 | 0.1385 | 0.1971 | 0.1482 | 0.1854 | 0.1431 |
| SignedGCN | 0.1807 | 0.1315 | 0.1984 | 0.1537 | 0.1764 | 0.1159 | 0.1898 | 0.1384 | 0.1775 | 0.1362 |
| HGT | 0.1861 | 0.1548 | 0.2117 | 0.1648 | 0.1812 | 0.1368 | 0.1893 | 0.1353 | 0.1813 | 0.1346 |
| GraphTrans | 0.1822 | 0.1623 | 0.1905 | 0.1452 | 0.1688 | 0.1094 | 0.1809 | 0.1344 | 0.1692 | 0.1268 |
| GATv2 | 0.1808 | 0.1386 | 0.1899 | 0.1425 | 0.1655 | 0.1069 | 0.1821 | 0.1467 | 0.1710 | 0.1325 |
| RDGT-Int | 0.1928 | 0.1391 | 0.1899 | 0.1438 | 0.1649 | 0.1037 | 0.1778 | 0.1346 | 0.1745 | 0.1323 |
| RDGT | **0.1715** | **0.1234** | **0.1792** | **0.1325** | **0.1531** | **0.0946** | **0.1693** | **0.1264** | **0.1605** | **0.1081** |

The bold values indicate the best experimental results, and underlined markings indicate suboptimal results.

In particular, we also implemented an intuitive comparison model, called RDGT-Int, which models the student associations within groups in an intuitive manner to learn representations. Concretely, we explicitly modeled the behavioral similarity between each student pair from the student-exercise exercising data within groups, which is computed from the response records on the same exercises for both. Then, we calculated the average similarity within the group and used it as a threshold to filter out student pairs above this value. Immediately after, we explicitly added edges between these student pairs and used a two-layer GCN [34] to learn the representation of student nodes for the final group performance prediction. As can be observed from the experimental results, RDGT-Int shows suboptimal performance on several datasets such as NIPS-Edu and SLP-math, which proves the validity and necessity of mining the potential associations between students

In addition, to verify the feasibility as well as the validity of the student-exercise response data $\mathcal{U}$ for our model, we further conducted experiments by adding $\mathcal{U}$ (details in Section V-A1) to the training process. The results as illustrated in Table IV. It can be observed that with the inclusion of $\mathcal{U}$ as auxiliary data, the performance of our proposed model has a certain improvement and remains significantly superior, especially with respect to MGCD modeled from a multi-task perspective. This demonstrates that our model is effective for deeply mining the correlation between the two types of instances and mitigating group-exercise interaction sparsity.

### C. Ablation Study (RQ3)

To answer RQ3, we conduct two ablation experiments to investigate the effectiveness of the proposed relation-guided graph transformer model and the attention-based group representation aggregation strategy, respectively. Due to the limited space, we use three datasets including ASSISTment12, NIPS-Edu, and SLP-math in the ablation experiments.

*1) Investigation of Relation-Guided Graph Transformer:* To investigate the feasibility and effectiveness of the relation-guided graph transformer model, an ablation experiment is conducted to observe the contributions of each component. We propose three variants of RDGT: *RDGT-S*, *RDGT-S+*, and *RDGT-SE*. *RDGT-S* means that only the student-side relation graph is used for modeling. *RDGT-S+* denotes modeling with student-side relation graph and inter-group information. *RDGT-SE* indicates the concurrent utilization of the dual-side relation graphs (student-side and exercise-side) while discarding inter-group information. As shown in Fig. 3, compared to RDGT, several variants suffer relative performance degradation on three datasets with respect to both metrics, especially RDGT-S shows the most pronounced decrease tendency. The experimental results demonstrate the effectiveness of the dual-side relation graph and the inter-group information enhancement module for group-level modeling and diagnosis.

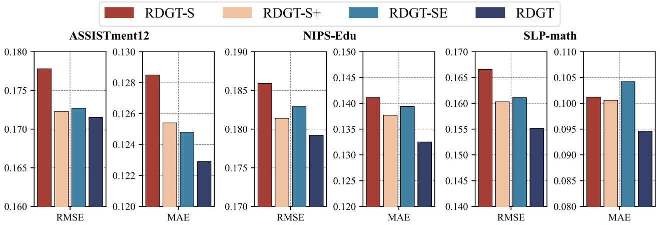


Fig. 3. Performance comparison of different modules.

*2) Investigation of Attention-Based Aggregation:* In order to observe the effectiveness of the attention-based group representation aggregation mechanism, we compared it with several aggregation strategies, including average aggregation, max pooling, and min pooling. As illustrated in Fig. 4, the implemented attention-based aggregation method achieves noticeable performance improvements over other strategies in terms of both rmse and mae metrics. It can be observed that the performance of both max pooling and min pooling strategies decreases quite significantly, indicating that the group structure embodies rich information while specific individuals cannot simply replace the group representation.

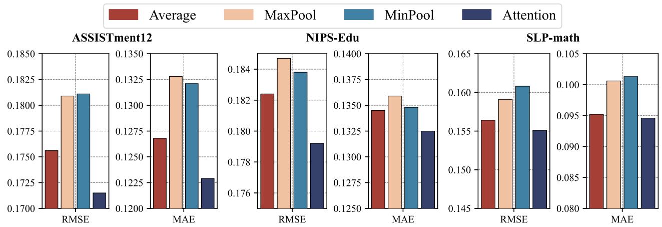
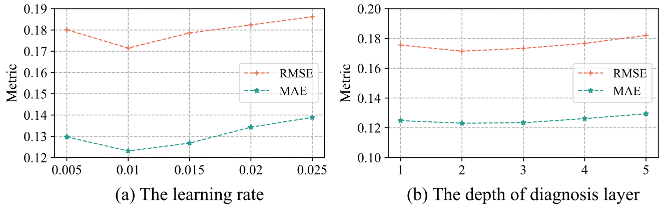


Fig. 4. Performance comparison of different aggregation strategies of group representation.



(a) The learning rate

(b) The depth of diagnosis layer

Fig. 5. Influence of the learning rate and the depth of diagnosis layer.

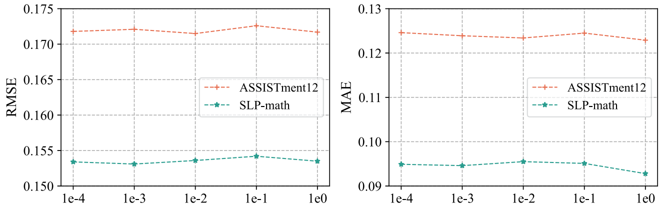


Fig. 6. Influence of the weight coefficient $\gamma$ on ASSISTment12 and SLP-math datasets.

### D. Parameter Sensitivity Analysis (RQ4)

This part focuses on pointing out which hyper-parameters affect our model. The analyzed hyper-parameters mainly include the learning rate, the depth of diagnosis layers, and the weight coefficient $\gamma$. Due to limited space, we mainly show the experimental results on the ASSISTment12 and SLP-math datasets. Specifically, we searched for the proper value in a small interval and set the learning rate as {0.005, 0.01, 0.015, 0.02, 0.025}, and reported the performance of RDGT with five values {1, 2, 3, 4, 5} of the depth of diagnosis layers. As shown in Fig. 5(a), we observe that 0.01 is sufficient for the learning rate. With the increase of the rate, the performance shows a trend of rising first and then falling, reaching the optimal value at 0.01. As shown in Fig. 5(b), the model reaches the best performance when the depth of diagnosis layers is 2. With either too few or too many diagnosis layers, the performance of RDGT declines, demonstrating that too few diagnosis layers cannot model the complex interactions and too many layers are prone to over-fitting. Meanwhile, to investigate the effect of different values of $\gamma$ on the performance of RDGT, we set the value list of $\gamma$ to be {1e-4, 1e-3, 1e-2, 1e-1, 1}, and the results are shown in Fig. 6. It can be observed that the model achieves optimal performance when the weight coefficient $\gamma$ is 1e-2 and 1e-3 in the ASSISTment12 and SLP-math datasets, respectively, and that either too large or too small weights affect the final performance.

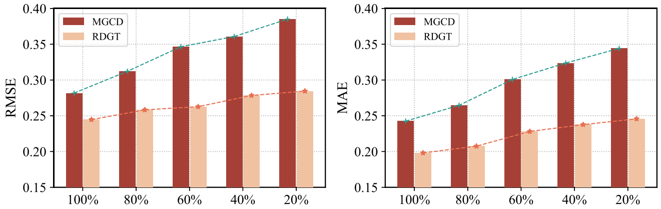


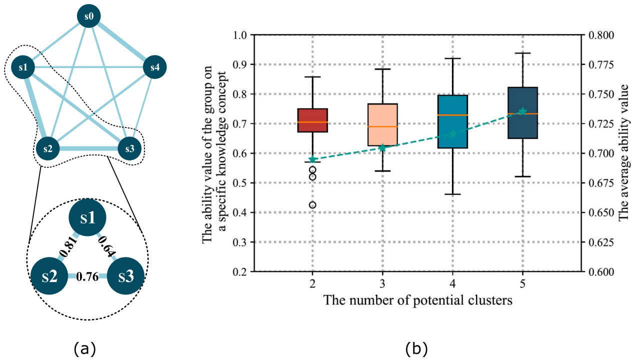
Fig. 7. Performance study of representative individuals of different proportions.

### E. Case Study (RQ5)

In this subsection, two case studies were performed on the ASSISTment dataset to answer RQ5.

*1) Case Study 1: Representative Individual Performance Study of Different Proportions:* In this case study, we tried to investigate whether the student representations learned by RDGT are excellent and whether the information they contain can help us identify representative individuals in the group. Thus, we selected students proportionally from each group in the testing set and then aggregated them into group representations for the group performance prediction. Specifically, for each group on the testing set, we obtained $k$ clusters denoting subgroups of different ability levels by clustering similarities in the representations of group members learned from the training process, and then we proportionally selected students who are close to the cluster center from each subgroup and aggregated their representations. In particular, groups with extremely few student members in the test set are excluded due to the requirements of clustering, and we set k to 3 considering the group size. The experimental results are shown in Fig. 7. We have the following observations: 1) As the percentage of students drops from 100% to 20%, there is a nearly 35% decline in MGCD in terms of RMSE, while our RDGT only declines by 16%, demonstrating the obvious advantage of our model in this task scenario as well. This clear contrast can also be observed in the trend of the broken line in the figure. 2) When the size of the student population decreased by 40%, the performance of RDGT decreased by only 7%, which proves that our model is effective in identifying representative individuals and learning representations. It is worth noting that similar to the computerized adaptive testing [3] in the individual assessment, this is essentially a group-level adaptive testing task, that is, how to accurately diagnose group ability with only a fraction of the students in the group taking the exercises. Simultaneously, it can support teachers to tailor their teaching to student groups. Our model demonstrates impressive potential and superiority under such goals.

*2) Case Study 2: Individual Influence Mining and Analysis:* In addition, a case study of group proficiency observations was performed in order to identify potential collaborations between students and representative groups. Specifically, we used the learned inter-student attention scores, which are extracted from the last self-attention layer in the RDGT, as edge features on the student relation graph. This characteristic reveals potential associations and collaborative information among students. Then, we performed graph clustering based on the spectral clustering method [63] on the student relationship graph using this edge



(a)

(b)

Fig. 8. Mining and analysis of individual influence: (a) Illustration of collaborative information between individuals within a group; (b) Demonstration of group proficiency on a specific knowledge concept in relation to the number of potential clusters.

feature. As shown in Fig. 8(a), the group contains five students, and the thickness of the edge between each student pair indicates the learned association weight. It can also be observed that the edge weights between nodes within the subgroup are larger than those outside the subgroup. This demonstrates the existence of potential subgroups within the group and that potential associations between students contribute to the mining of these subgroups. Further, we also tried to explore whether the ability of the group is somehow related to this number of potential subgroups. We then performed a statistical information on the groups after clustering, and the results are shown in Fig. 8(b). It can be observed that groups with a higher number of potential clusters have a tendency to be more capable. This finding is beneficial for the mining and identification of representative groups, as well as for facilitating other types of group level diagnosis tasks (e.g., the mining of talented teams and the discovery of outstanding teacher groups).

## VI. Conclusion

In this paper, we proposed a novel group-level cognitive diagnosis model, namely **R**elation-guided **D**ual-side **G**raph **T**ransformer (RDGT), which performs group representation learning by adaptively mining the intrinsic relation of the students within the group. Specifically, we first constructed dual-side relation graphs, i.e., student-side and exercise-side, from the group-student-exercise heterogeneous interaction data for explicitly modeling associations between students and exercises, respectively. Moreover, we implemented two improved graph transformers by introducing relation encoding to better capture the holistic information about the dual-side graphs including node and edge features for representation learning. Then, we designed a cognitive diagnosis module for learning the groups' proficiency in specific knowledge concepts, which includes an attention-based aggregation strategy and a hybrid loss. Finally, extensive experiments on real-world datasets clearly demonstrated the effectiveness of our model and two case studies revealed that our model can be utilized to identify representative groups and potential collaborations among students. We hope this work could lead to further studies on GCD.

## Acknowledgment

This work was accomplished when Xiaoshan Yu was an intern in the Career Science Lab, BOSS Zhipin supervised by Chuan Qin and Hengshu Zhu.

## References

[1] G. Jeckeln et al., "Face identification proficiency test designed using item response theory," 2021, *arXiv:2106.15323*.

[2] S. Yang et al., "Cognitive diagnosis-based personalized exercise group assembly via a multi-objective evolutionary algorithm," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 3, pp. 829–844, Jun. 2023.

[3] H. Bi et al., "Quality meets diversity: A model-agnostic framework for computerized adaptive testing," in *Proc. IEEE Int. Conf. Data Mining*, 2020, pp. 42–51.

[4] W. Xu and Y. Zhou, "Course video recommendation with multimodal information in online learning platforms: A deep learning framework," *Brit. J. Educ. Technol.*, vol. 51, no. 5, pp. 1734–1747, 2020.

[5] S. Yang, X. Yu, Y. Tian, X. Yan, H. Ma, and X. Zhang, "Evolutionary neural architecture search for transformer in knowledge tracing," 2023, *arXiv:2310.01180*.

[6] S. E. Embretson and S. P. Reise, *Item Response Theory*. London, U.K.: Psychology Press, 2013.

[7] M. D. Reckase, "Multidimensional item response theory models," in *Multidimensional Item Response Theory*, Berlin, Germany: Springer, 2009, pp. 79–112.

[8] J. De La Torre, "DINA model and parameter estimation: A didactic," *J. Educ. Behav. Statist.*, vol. 34, no. 1, pp. 115–130, 2009.

[9] F. Wang et al., "Neural cognitive diagnosis for intelligent education systems," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6153–6161.

[10] W. Gao et al., "RCD: Relation map driven cognitive diagnosis for intelligent education systems," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 501–510.

[11] F. Wang et al., "NeuralCD: A general framework for cognitive diagnosis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8312–8327, Aug. 2023.

[12] E. Hammar Chiriac, "Group work as an incentive for learning–students' experiences of group work," *Front. Psychol.*, vol. 5, pp. 558–567, 2014.

[13] J. Huang et al., "Group-level cognitive diagnosis: A multi-task learning perspective," in *Proc. IEEE Int. Conf. Data Mining*, 2021, pp. 210–219.

[14] Q. Liu, "Towards a new generation of cognitive diagnosis," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 4961–4964.

[15] L. V. DiBello, L. A. Roussos, and W. Stout, "31A review of cognitively diagnostic assessment and a summary of psychometric models," *Handbook Statist.*, vol. 26, pp. 979–1030, 2006.

[16] J. Li et al., "HierCDF: A Bayesian network-based hierarchical cognitive diagnosis framework," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 904–913.

[17] S. Khoshraftar and A. An, "A survey on graph representation learning methods," 2022, *arXiv:2204.01855*.

[18] Q. Guo et al., "A survey on knowledge graph-based recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3549–3568, Aug. 2022.

[19] S. Bourigault, C. Lagnier, S. Lamprier, L. Denoyer, and P. Gallinari, "Learning social network embeddings for predicting information diffusion," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, 2014, pp. 393–402.

[20] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017.

[21] M. Chen et al., "A trend-aware investment target recommendation system with heterogeneous graph," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[22] G. Lv, Z. Hu, Y. Bi, and S. Zhang, "Learning unknown from correlations: Graph neural network for inter-novel-protein interaction prediction," 2021, *arXiv:2105.06709*.

[23] W. Wang et al., "Group-aware long-and short-term graph representation learning for sequential group recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1449–1458.

[24] Y. Rong, W. Huang, T. Xu, and J. Huang, "DropEdge: Towards deep graph convolutional networks on node classification," 2019, *arXiv:1907.10903*.

[25] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3/5, pp. 75–174, 2010.

[26] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2010, pp. 243–252.

[27] L. Dai et al., "Enterprise cooperation and competition analysis with a sign-oriented preference network," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 774–782.

[28] R. Zha et al., "Career mobility analysis with uncertainty-aware graph autoencoders: A job title transition perspective," *IEEE Trans. Computat. Social Syst.*, early access, Feb. 17, 2023, doi: 10.1109/TCSS.2023.3239038.

[29] W. Wang et al., "Incorporating link prediction into multi-relational item graph modeling for session-based recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2683–2696, Mar. 2023.

[30] R. Burioni and D. Cassi, "Random walks on graphs: Ideas, techniques and results," *J. Phys. A: Math. Gen.*, vol. 38, no. 8, pp. 45–82, 2005.

[31] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[32] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," 2017, *arXiv:1709.05584*.

[33] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[34] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[35] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–17.

[36] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–12.

[37] X. Wang et al., "Heterogeneous graph attention network," in *Proc. World Wide Web Conf.*, 2019, pp. 2022–2032.

[38] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 793–803.

[39] J. Zhao, X. Wang, C. Shi, B. Hu, G. Song, and Y. Ye, "Heterogeneous graph structure learning for graph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 4697–4705.

[40] E. Min et al., "Transformer for graphs: An overview from architecture perspective," 2022, *arXiv:2202.08455*.

[41] C. Ying et al., "Do transformers really perform badly for graph representation?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 28877–28888.

[42] L. Chanussot et al., "Open catalyst 2020 (OC20) dataset and community challenges," *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021.

[43] E. Min et al., "Masked transformer for neighhourhood-aware click-through rate prediction," 2022, *arXiv:2201.13311*.

[44] Y. Fan et al., "Heterogeneous temporal graph transformer: An intelligent system for evolving android malware detection," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 2831–2839.

[45] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proc. Web Conf.*, 2020, pp. 2704–2710.

[46] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[47] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[48] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[49] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," 2018, *arXiv:1803.02155*.

[50] H. Li, D. Zhao, and J. Zeng, "KPGT: Knowledge-guided pre-training of graph transformer for molecular property prediction," 2022, *arXiv:2206.03364*.

[51] D. Cao, X. He, L. Miao, Y. An, C. Yang, and R. Hong, "Attentive group recommendation," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 645–654.

[52] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," *User Model. User-Adapted Interaction*, vol. 19, no. 3, pp. 243–266, 2009.

[53] Z. Wang et al., "Instructions and guide for diagnostic questions: The NeurIPS 2020 education challenge," 2020, *arXiv:2007.12061*.

[54] Y. Lu, Y. Pian, Z. Shen, P. Chen, and X. Li, "SLP: A multi-dimensional and consecutive dataset from k-12 education," in *Proc. 29th Int. Conf. Comput. Educ.*, 2021, pp. 261–266.

[55] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1257–1264.

[56] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," in *Proc. KDD Cup*, 2010, pp. 1–11.

[57] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. Eur. Semantic Web Conf.*, Springer, 2018, pp. 593–607.

[58] T. Derr, Y. Ma, and J. Tang, "Signed graph convolutional networks," in *Proc. IEEE Int. Conf. Data Mining*, 2018, pp. 929–934.

[59] Z. Wu, P. Jain, M. Wright, A. Mirhoseini, J. E. Gonzalez, and I. Stoica, "Representing long-range context for graph neural networks with global attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 13266–13279.

[60] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," 2021, *arXiv:2105.14491*.

[61] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[63] U. Von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, pp. 395–416, 2007.

**Xiaoshan Yu** received the BSc degree from Hefei Normal University, Hefei, China, in 2020. He is currently working toward the PhD degree with the School of Artificial Intelligence, Anhui University, Hefei, China. His current research interests include intelligent education, multi-objective optimization, neural architecture search, and graph learning.

**Chuan Qin** (Member, IEEE) received the PhD degree in computer science and technology from the University of Science and Technology of China (USTC), Hefei, China, in 2021. He is currently a senior researcher with BOSS Zhipin Career Science Lab (CSL) and a postdoctoral researcher with the PBC School of Finance, Tsinghua University. He has authored more than 30 journals and conference papers in the fields of natural language processing and recommender systems, including *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Computational Social Systems*, *ACM Transactions on Information Systems*, SIGKDD, SIGIR, WWW, ICDE, NeurIPS, AAAI, IJCAI, and ICDM. He has been honored with the Excellent Award from the President of the Chinese Academy of Sciences (2021), a nomination for the Baidu Scholarship (top 20 globally) (2021), and the Best Student Paper Award of SIGKDD-2018.

**Dazhong Shen** received the PhD degree in computer science and technology from the University of Science and Technology of China (USTC), Hefei, China, in 2021. He is presently a researcher with Shanghai Artificial Intelligence Laboratory. He has authored more than ten journal and conference papers in the fields of Generative Model and Natural Language Processing, including the *ACM Transactions on Information Systems (TOIS)*, *ACM Transactions on Management Information Systems (TMIS)*, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, Conference on Neural Information Processing Systems (NeurIPS), International World Wide Web Conferences (WWW), International Joint Conference on Artificial Intelligence (IJCAI), etc.

**Haiping Ma** received the BE degree from Anhui University, Hefei, China, in 2008, and the PhD degree from the University of Science and Technology of China, Hefei, China, in 2013. She is currently an associate professor with the Institutes of Physical Science and Information Technology, Anhui University, Hefei, China. Her current research interests include data mining and multi-objective optimization methods and their applications.

**Le Zhang** received the BE degree in software engineering from the Dalian University of Technology, Dalian, China, in 2016, and the PhD degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2022. He is currently a researcher with Baidu Research, Baidu Inc., Beijing, China. His general research interests include data mining and machine learning, with a focus on spatiotemporal modeling and its applications in business intelligence.

**Xingyi Zhang** (Senior Member, IEEE) received the BSc degree from Fuyang Normal College, Fuyang, China, in 2003, and the MSc and PhD degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2006 and 2009, respectively. He is currently a professor with the School of Artificial Intelligence, Anhui University, Hefei, China. His current research interests include unconventional models and algorithms of computation, evolutionary multi-objective optimization, and logistic scheduling. He is the recipient of the 2018 and 2021 IEEE Transactions on Evolutionary Computation Outstanding Paper Award and the 2020 IEEE Computational Intelligence Magazine Outstanding Paper Award.

**Hengshu Zhu** (Senior Member, IEEE) received the BE and PhD degrees in computer science from the University of Science and Technology of China (USTC), China, in 2009 and 2014, respectively. He is currently the head of BOSS Zhipin Career Science Lab (CSL). His general area of research is data mining and machine learning, with a focus on developing advanced data analysis techniques for innovative business applications. He has published prolifically in refereed journals and conference proceedings, such as *Nature Communications*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Mobile Computing*, *ACM Transactions on Information Systems*, SIGKDD, SIGIR, and NeurIPS. He served as the program co-chair of KDD CUP-2019 Regular ML Track, the industry chair of PRICAI-2022, the area chair of AAAI and IJCAI, and regularly as the (senior) program committee members in numerous top conferences. He was the recipient of the Distinguished Dissertation Award of CAS (2016), the Distinguished Dissertation Award of CAAI (2016), the Special Prize of President Scholarship for Postgraduate Students of CAS (2014), the Best Student Paper Award of KSEM-2011, WAIM-2013, CCDM-2014, and the Best Paper Nomination of ICDM-2014 and WSDM-2022. He is the senior member of ACM, CAAI, and CCF.

**Hui Xiong** (Fellow, IEEE) received the PhD degree in computer science from the University of Minnesota. He is a chair professor, associate vice president (Knowledge Transfer), and head of the AI Thrust with the Hong Kong University of Science and Technology (Guangzhou). His research interests span Artificial Intelligence, data mining, and mobile computing. He has served on numerous organization and program committees for conferences, including as program co-chair for the Industrial and Government Track for the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), program co-chair for the IEEE 2013 International Conference on Data Mining, general co-chair for the 2015 IEEE International Conference on Data Mining, and program co-chair of the Research Track for the 2018 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. He received several awards, such as the 2021 AAAI Best Paper Award and the 2011 IEEE ICDM Best Research Paper award. For his significant contributions to data mining and mobile computing, he was elected as a fellow of AAAS, in 2020.