Computers and Education: Artificial Intelligence 10 (2026) 100582

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

GNN-driven modeling and prediction of multi-dimensional correlation of international medical education quality based on competence

Ailing Yang *, Ling Lin, Hu Zhang, Rong Su, Yunfei Li

Geriatric Cardiology Department, the First Affiliated Hospital, Kunming Medical University, Kunming, 650031, Yunnan Province, China

ARTICLE INFO

Keywords:
Competency-based international medical education
Education quality
Multi-dimensional correlation modeling and prediction
RT-GAT model
Relational attention

ABSTRACT

This study addresses the challenge of modeling competency-based international medical education (IME) quality, where traditional methods struggle with heterogeneous multi-dimensional indicators and temporal dynamics. We propose RT-GAT (Relational Temporal-Graph Attention Network), a GNN-based model integrating relational attention and temporal encoding to improve competency correlation modeling and prediction accuracy. The model constructs a multi-relation heterogeneous graph with nodes representing knowledge/skills, clinical practice, communication, and teamwork, connected by driving, collaborative, and feedback edges. A relation-aware attention mechanism adapts GAT by applying edge-specific transformations and attention vectors to distinguish relationship influences. Temporal features are captured via sine-cosine positional encoding and Bi-LSTM, enabling dynamic competency evolution tracking across teaching stages. A multi-head attention architecture learns multi-subspace features, while a multi-task regressor jointly predicts four-dimensional competencies. Evaluated on five-stage data from 900 international medical students, RT-GAT achieves a cosine similarity of 0.91, adjacency reconstruction error of 0.83, and spectral distance of 0.36 in competency correlation modeling. Incorporating temporal relationships boosts the Pearson correlation coefficient by 0.168 on average. Prediction performance shows RMSE (~4.8), MAE (~3.7), and MAPE (~6.1%). Cross-cultural validation with students from the US, India, UK, Russia, and Nigeria demonstrates stable accuracy (RMSE: 4.8–6.4; MAE: 3.7–5.1), confirming adaptability to heterogeneous backgrounds. RT-GAT excels in structural representation and temporal evolution capture, offering a robust technical solution for competency-based IME evaluation and prediction.

1. Introduction

As the internationalization of global medical education accelerates, China has become one of the important destinations for international medical students(Alolga et al., 2022; Wu & Koh, 2022). The existing education quality evaluation system is still mainly based on theoretical knowledge assessment and static indicators(Zhang, Sun, et al., 2023), which is difficult to fully reflect the comprehensive performance of students in multi-dimensional abilities such as clinical practice, cross-cultural communication, and teamwork. There are many heterogeneous relationships between different ability dimensions, such as driving, synergy, and feedback, and these relationships are often simplified into homogeneous associations in traditional statistics or shallow machine learning models, resulting in insufficient description of the internal interaction mechanism of abilities. Moreover, students' abilities evolve dynamically with the teaching stage and practical training rotation. If the influence of time is ignored, it is impossible to accurately predict the trend of students' abilities and provide real-time feedback on teaching interventions, which seriously restricts the practicality and foresight of the evaluation system. There is an urgent need to build an evaluation system that can accurately model ability relationships and has predictive capabilities to serve educational feedback and personalized training.

This paper aims to build a new graph neural network model, RT-GAT, to accurately model and predict multi-dimensional heterogeneous indicators in the quality of international medical education based on ability through relationship-aware attention and temporal coding mechanisms. This paper constructs a multi-relation heterogeneous graph based on the dimensions of knowledge and skills, clinical practice, communication ability and teamwork, introduces driving, collaborative and feedback relationship attention mechanisms, and dynamically distinguishes the importance of various interaction paths. Then, five-stage

* Corresponding author.
E-mail address: yal650031yy@hotmail.com (A. Yang).

https://doi.org/10.1016/j.caeai.2026.100582
Received 2 July 2025; Received in revised form 19 March 2026; Accepted 20 March 2026
Available online 1 April 2026

time series information is integrated through sinusoidal position encoding and bidirectional LSTM modules to capture the ability evolution trend, and multi-head fusion and joint regression of multiple ability dimensions are used to complete the prediction task. The results showed that the RMSE of the RT-GAT model was about 4.8 points, and the T-test results showed that compared with the comparative models such as Study-GNN-Bi-LSTM, RT-GAT had significantly lower MAE in all dimensions ($P < 0.005$). The experiment proved that RT-GAT was significantly better than other models in the accuracy of future ability prediction, and had good multi-dimensional correlation modeling capabilities, demonstrating its practical application potential in the competency-based international medical education evaluation.

While relational attention and temporal encoding have been applied in heterogeneous GNNs and sequence modeling, respectively, the innovation of this paper lies not in the introduction of a single module, but in the systematic reconstruction of the cross-domain fusion mechanism and the theoretical mapping at the educational semantic level. RT-GAT uses "ability interaction type" rather than traditional graph structure as its core modeling logic. It embeds three types of edges—driving, collaboration, and feedback—from the causal relationship and feedback loop theory of educational psychology into relational attention calculation, giving attention weights a clear educational semantic interpretability. The temporal module does not merely perform sequence encoding, but through the joint design of Bi-LSTM and positional encoding, it explicitly models the dynamic impact of changes in teaching stages on relational weights, achieving a dual dynamic joint learning of "relationship-time". Multi-task regression and spectral distance constraints are introduced during model training, enabling it to simultaneously optimize structural similarity and predictive performance. In the structural prediction of ability evolution, it surpasses traditional R-GAT or Study-GNN models that rely solely on static topology. In summary, the innovation of RT-GAT lies in its relational semantic reconstruction and temporal dynamic fusion mechanism for educational capability networks, rather than a simple stacking of technologies. This framework provides a new paradigm for modeling educational capabilities under heterogeneous relationships.

The contributions of this paper are as follows.

(1) This paper innovatively integrates relation-aware attention and temporal coding mechanisms to construct a graph neural network suitable for competency-based medical education, and improves the modeling accuracy of heterogeneous relationships between multi-dimensional competency indicators.
(2) This paper designs three types of edges: driving, collaborative, and feedback, refines the diversified interaction paths between competency dimensions, and realizes structured correlation modeling of core competencies such as knowledge and skills, clinical practice, communication skills, and teamwork.
(3) This paper combines sinusoidal position coding with Bi-LSTM to capture the evolution trend of competency, and improves the accuracy and robustness of the model in future competency prediction tasks through multi-head joint learning.

The fundamental innovation of RT-GAT lies not in introducing individual components, but in systematically reconstructing these known components at the educational semantics level, achieving a paradigm shift from "technology-adaptive scenarios" to "theory-driven modeling." Compared to the relational-aware heterogeneous GNN proposed by Yu et al., which defines relationships based on general semantics without theoretical mapping for educational contexts, RT-GAT directly correlates three types of relationships—driving, collaborative, and feedback—with causal pathways, parallel facilitation, and bidirectional regulation mechanisms in educational psychology, endowing the model structure with prior educational semantic interpretability. The driving relationship aligns with the cognitive constructivism concept of "knowledge accumulation promoting ability development," manifesting as direct promotion of clinical practice through knowledge and skills. The collaborative relationship resonates with social constructivism's "cooperative learning" theory, representing parallel enhancement of communication skills and teamwork. The feedback relationship originates from formative assessment theory's "reflective regulation" mechanism, reflecting the reverse impact of clinical practice performance on knowledge and skill consolidation. These three relationship designs are not arbitrary technical choices but rather embed pedagogical causal logic into the core computational paradigm of graph neural networks, ensuring mathematical-level educational semantic interpretability. Consequently, the relational-aware attention mechanism achieves clear educational semantic interpretability through weight allocation rather than purely data-driven statistical associations. The time encoding module also deviates from conventional LSTM applications by integrating sine position encoding with bidirectional LSTM architectures to simulate dynamic impacts of teaching phase transitions on relationship weights, enabling dual dynamic joint learning of "relationships-time." Furthermore, multi-task regression introduces spectral distance constraints to account for structural similarity, thereby overcoming the limitations of traditional Graph Neural Networks (GNNs). RT-GAT provides a transferable theoretical framework and technical pathway for applying graph neural networks in modeling complex social science relationships.

2. Related works

In the field of competency-based international medical education quality assessment and student future competency prediction, scholars have tried various methods. Booth G J et al. automatically mapped narrative feedback to GME (graduate medical education) milestone subcompetencies based on NLP (Natural Language Processing) algorithms, achieving rapid classification of tens of thousands of assessment contents, but their focus was limited to sub-competency label recognition, and the interactive relationship between competency dimensions was not modeled (Booth et al., 2023). Gupta S K et al. summarized seven categories of CBME (competency-based medical education) evaluation procedures through a systematic review, providing a macro-guideline for the design of evaluation frameworks in the Indian context, but did not propose specific multidimensional modeling or prediction methods (Gupta et al., 2024). Dabbagh A explored the potential of AI (Artificial Intelligence) in EPA (entrustable professional activities) evaluation, pointing out that machine learning can predict professional behavior and decision-making, but did not provide an empirical model (Dabbagh et al., 2024). Soundariya K et al. evaluated the implementation strategy of CBME courses through qualitative research, revealing the importance of teacher sensitivity and feedback mechanism, but lacking quantitative analysis of ability evolution (Soundariya et al., 2025). Mastour H et al. used integrated machine learning RF (Random Forest), XGB (eXtreme Gradient Boosting) and other frameworks to make early predictions of high-risk test scores, achieving an R^2 of up to 0.80 in prediction, but focusing on the single dimension of test scores (Mastour et al., 2023). Kukkar A et al. combined RNN + LSTM (Recurrent Neural Network + Long Short-Term Memory) with traditional machine learning to improve the accuracy of course pass rate prediction (97%), also focusing on academic performance rather than multidimensional abilities(Kukkar et al., 2024). Shi Y et al. applied multiple classification algorithms to the information literacy behavior characteristics of college students and achieved a prediction accuracy of 92.5%, but the model ignored the heterogeneous connections between different ability indicators(Shi et al., 2023). Asselman A et al. introduced XGB into PFA (Performance Factors Analysis) to improve knowledge tracking prediction, but the improvement was still limited to the analysis of similar factors (Asselman et al., 2023). Al-Azazi F A applied ANN-LSTM (Artificial Neutral Network-Long Short-Term Memory) to the early prediction of student performance, and the accuracy rate was increased to 70%, but the interaction and dynamic evolution of cross-item abilities were not

integrated (Al-Azazi et al., 2023). Shou Z et al. obtained an early risk prediction accuracy of up to 99% on the data set through a multidimensional time series analysis model, but lacked the description of the network structure and relationship types between ability dimensions (Shou et al., 2024). Existing research has made achievements in NLP, system evaluation, deep learning and traditional machine learning prediction, but most of them are limited to a single dimension or homogeneous aggregation, and lack unified modeling and interpretation capabilities for the heterogeneous relationship and time evolution of ability indicators.

In the field of multi-dimensional indicator modeling of education quality, many studies have attempted to drive ability association and performance prediction with graph neural networks. The Study-GNN driven multi-topology learning pipeline proposed by Li M et al. constructs a multi-topology graph based on similarity metrics and fuses multi-graph representations using the attention mechanism in the MTGNN (multi-topology graph neural networks) module. It significantly outperforms single-topology GCN (Graph Convolutional Network) and traditional machine learning methods on real educational datasets. However, it is limited to homogeneous similarity graphs between students and does not involve modeling heterogeneous relationships between different ability dimensions (Li et al., 2022). Huang Q et al. used dual graph neural networks to drive local and global representation learning, extracting grade representations from two subgraphs: interaction activities and attribute features. This achieved a pass/fail prediction accuracy of 83.96% and a pass/drop prediction accuracy of 90.18%, but its dual graph partitioning still failed to distinguish multiple relationship types (Huang & Zeng, 2024). Wang S et al. drove the combination of signed bipartite graph neural networks and large language models. By distinguishing correct and incorrect answers by positive and negative edges, the F1 score increased by 3.7% on average, and the model had a good suppression effect on noisy answers. However, the model structure is only applicable to binary answer scenarios and is difficult to expand to multi-dimensional ability interactions (Wang et al., 2024). Huang Q constructed a dynamic graph of learning activities and used the academic performance prediction-temporal graph networks (APP-TGN) to drive the encoding of temporal information and interactive behaviors. The performance of score prediction was greatly improved under the multi-head attention mechanism, but although its temporal module can capture the evolution trend, it does not distinguish the semantic differences between different relationship types (Huang & Chen, 2024). Chen Z et al. proposed a bipartite network and a hybrid neural network based on the relationship matrix to drive discrete feature fitting, with a prediction accuracy of 93.1% and an F1 of 90.45%, but did not use graph attention to distinguish the heterogeneous associations between ability dimensions (Chen, Cen, et al., 2023). Existing GNN, GAT and their variants have made achievements in multi-topology fusion, dual-graph parallelism, signed edges and temporal coding, but it is generally difficult to simultaneously and finely characterize the heterogeneous relationships and dynamic evolution of ability dimensions in a single framework. It is urgent to introduce a unified model that can take into account both relationship types and time sequence information. This paper adopts the RT-GAT model to model and predict the multi-dimensional correlation of the quality of competency-based international medical education, fully capturing the dual needs of multi-dimensional, heterogeneous relationships and dynamic evolution.

International medical education also faces significant challenges in cultivating cross-cultural competence. Walkowska A et al. pointed out in a systematic review that simulated patient intervention can significantly improve medical students' cross-cultural understanding and nursing confidence in a multicultural context (Walkowska et al., 2023). Gauvin N emphasized that clinical educators must receive specialized cross-cultural training to effectively deal with the impact of cultural differences on the supervision and feedback process (Gauvin & Gregory-Martin, 2025). A cross-sectional survey by Tajvar M et al. showed that international students in Iranian medical schools often suffer from language barriers, poor information acquisition and cultural adaptation difficulties that affect their academic performance and well-being (Tajvar et al., 2024). Rukadikar C et al. proposed in a review that cultural competence should be included in the core curriculum of medical education to help students establish effective communication with patients from different backgrounds (Rukadikar et al., 2022). Duan J's research revealed that cross-cultural adaptability has a profound impact on international students in bioethical education and social practice (Duan, 2023). The studies collectively show that medical education in an international context requires students to master professional skills and cultivate flexible cross-cultural coping skills in multicultural contexts.

To set RT-GAT apart from conventional relational graph neural networks (R-GAT, HGT) and educational analysis methods, this study conducts a thorough comparison across four key dimensions: modeling objects, relationship definitions, temporal integration, and output formats.

Existing relational GNNs, while supporting multiple relationship types, often rely on generic semantics like citations or interactions, lacking theoretical grounding in educational contexts. RT-GAT, however, directly links three relationship types—driving, collaborative, and feedback—to educational psychology mechanisms, providing inherent interpretability to its model structure.In terms of temporal integration, traditional methods mainly use static graph snapshots or simple temporal concatenations, unable to dynamically adjust relationship weights across teaching phases. RT-GAT overcomes this by combining sine position encoding with bidirectional LSTM, enabling adaptive relationship attention mechanisms that evolve with phase characteristics.Compared to common educational analysis methods, which are often limited to single-dimensional performance prediction, RT-GAT excels in modeling heterogeneous correlations and temporal dynamics among multidimensional competencies. Unlike interpretability-focused systems relying on post-hoc interpretation modules, RT-GAT inherently incorporates interpretability, with attention weights directly reflecting educational relationship strength. Additionally, this study integrates teachers into the model iteration closed-loop process, unlike existing systems that typically use interactive interfaces for model validation or modification.

3. Construction of multidimensional correlation modeling and prediction model RT-GAT

3.1. Construction of multi-dimensional heterogeneous graph

In order to accurately depict the heterogeneous correlation between multi-dimensional indicators in the quality of international medical education based on competency, this paper constructs a heterogeneous graph structure based on multi-dimensional competency indicators and their diverse interaction modes. The node categories correspond to different competency dimensions, including knowledge and skills, clinical practice, communication skills, and teamwork. The expression of competency indicators is shown in formula (1).

$$V = V^K \cup V^C \cup V^G \cup V^T, V^i \cap V^j = \varnothing, i \neq j \tag{1}$$

V represents the ability index, V^K, V^C, V^G, and V^T correspond to knowledge and skills, clinical practice, communication skills, and teamwork, respectively.

The definition of the node feature matrix is shown in formula (2).

$$\mathbf{x} = [x_1, x_2, \ldots, x_N]^\top \tag{2}$$

where x_N represents the eigenvector corresponding to the first capability indicator.

In order to reflect various interaction modes, the edge type set R is defined to represent various relationships such as driving type (causal

influence), collaborative type (cooperative progress), feedback type (bidirectional regulation), etc. Based on this, a multi-dimensional heterogeneous graph $G = (V, \varepsilon, R)$ is constructed, in which the edge set ε includes the edge set under each edge type.

The three types of relationships are not arbitrarily defined. The driving type corresponds to the preconditions/causal paths in ability formation, aligning with the cognitive constructivist perspective of mastery learning and tiered instruction. The collaborative type corresponds to social constructivism and cooperative learning theory, used to characterize the process of parallel facilitation and resource sharing. The feedback type directly corresponds to the moderating role of formative assessment and feedback loops in learning improvement. In this study, these three types of relationships are jointly operationalized through educational expert annotation and data-driven co-occurrence/conditional probability methods, preserving the theoretical explanatory power of educational semantics while ensuring measurability and interpretability in modeling.

The driving relationship corresponds to the "premise → ability development" causal path in cognitive constructivism theory, emphasizing the foundational role of existing knowledge structures in new ability formation. This is specifically manifested through the direct promotion of clinical practice by knowledge and skills. The collaborative relationship originates from Vygotsky's social constructivism theory, particularly his concept of the "zone of proximal development" regarding peer collaboration in ability enhancement, which represents a parallel promotion mechanism between communication skills and teamwork. The feedback relationship directly aligns with the "feedback regulation" mechanism in mastery learning theory and formative assessment theory, reflecting the reverse impact of clinical practice performance on knowledge and skill consolidation—where experiential reflection during practice deepens theoretical understanding. These relationships mutually reinforce collaborative abilities in group learning, while Ericsson's deliberate practice theory highlights that post-practice feedback serves as a critical component for skill improvement. Therefore, the three relationships—driving, collaborative, and feedback—are respectively rooted in three classic educational theories: cognitive constructivism, social constructivism, and formative assessment, demonstrating solid theoretical foundations and educational semantic implications.

In order to precisely express the contribution of different relationship types to the information transfer between nodes, a multi-dimensional adjacency matrix set $\{A^{r_m}\}_{m=1}^{M}$ is used, which allows different interaction modes to be processed differently to avoid information confusion. The formula satisfied is shown in formula (3).

$$A_{ij}^{m} = \begin{cases} 1, if \exists (v_i, v_j) \in \varepsilon^m \\ 0, or \end{cases} \tag{3}$$

v_i denotes the i-th capability node with i being the node index; ε^m represents the m-th relationship type, and A_{ij}^{m} indicates the set of neighboring nodes of the relationship node v_j at node v_i.

In this study, "driving edge" refers to the causal influence path in ability development (such as the direct promoting effect of knowledge and skills on clinical practice), "collaboration edge" represents the synergistic promoting relationship between parallel abilities (such as the mutual reinforcement of communication and teamwork), and "feedback edge" specifically refers to the bidirectional regulatory relationship between ability dimensions (such as the reverse influence of clinical practice performance on the consolidation of knowledge and skills). In the symbol system, v_i and v_j represent the source node and the target node, respectively, r_m corresponds to the three types of relationships: driving, collaboration, and feedback. The subscripts i and j represent node indices, and the superscript m identifies the relationship type.

The steps of constructing a multi-dimensional heterogeneous graph are as follows.

(1) According to the preset capability dimensions, the original multi-dimensional capability scoring data is classified and assigned to form a node set to ensure that the characteristics of each node reflect the specific quantitative indicators of the corresponding capability dimension.

(2) Based on the interactive definition between capability indicators, expert knowledge and data statistical analysis are used to construct edges, and conditional probability and co-occurrence frequency methods are used to calculate edge weights. The relationship weights between capability indicators are calculated as shown in formula (4).

$$w_{ij}^{m} = P(v_j | v_i, r_m) = \frac{\text{Frequency}(v_i, v_j, r_m)}{\sum_k \text{Frequency}(v_i, v_k, r_m)} \tag{4}$$

where $\text{Frequency}(v_i, v_j, r_m)$ represents the number of times v_i and v_j appear simultaneously under relationship type r_m. When $w_{ij}^{m} > \theta_{r_m}$ (threshold), there is an edge $(v_i, v_j) \in \varepsilon^{r_m}$.

(3) According to the weight definition, a weighted adjacency matrix is constructed, as shown in formula (5).

$$\widetilde{\mathbf{A}}^{r_m} = \mathbf{A}^{r_m} \odot \mathbf{W}^{r_m}, \mathbf{W}^{r_m} = \left[w_{ij}^{r_m} \right] \tag{5}$$

where $\odot$ represents element product and $\widetilde{A}^{r_m}$ represents weighted adjacency matrix.

(4) Normalize the weighted adjacency matrix of each relationship type to avoid uneven information aggregation caused by differences in node degrees.

$$\widetilde{\mathbf{A}}^{r_m} = (\mathbf{D}^{r_m})^{-\frac{1}{2}} \widetilde{\mathbf{A}}^{r_m} (\mathbf{D}^{r_m})^{-\frac{1}{2}} \tag{6}$$

where D^{r_m} represents the degree matrix of the weighted adjacency matrix.

(5) The adjacency matrices of different relationship types are jointly encoded with the node features, and the definition of the aggregated features of the nodes under the relationship is shown in formula (7).

$$\mathbf{H}^{(l+1), r_m} = \sigma\left(\widetilde{\mathbf{A}}^{r_m} \mathbf{H}^{(l)} \mathbf{W}^{(l), r_m} \right) \tag{7}$$

where $H^{(l)}$ represents the node feature matrix of the l layer, $W^{(l), r_m}$ represents the trainable weight matrix, and σ represents the activation function ReLU.

The feature fusion under all relations is obtained by weighted summation, as shown in formula (8).

$$\mathbf{H}^{(l+1)} = \sum_{m=1}^{M} \alpha^{r_m} \mathbf{H}^{(l+1), r_m} \tag{8}$$

where α^{r_m} represents the weight coefficient.

Each core competency dimension consists of multiple sub-indicator nodes to refine competency representation. The Knowledge and Skills dimension comprises five sub-nodes: Basic Medical Theory, Clinical Medical Theory, Pharmacological Knowledge, Pathological Knowledge, and Preventive Medicine, with scores derived from standardized examination results of corresponding course modules. The Clinical Practice dimension includes five sub-nodes: Medical History Taking, Physical Examination, Operational Skills, Diagnostic Reasoning, and Medical Documentation, with scores calculated based on weighted scores from OSCE examination stations. The Communication Skills dimension encompasses five sub-nodes: Doctor-Patient Communication, Cross-Cultural Communication, Team Communication, Written Expression, and Listening Ability, assessed using itemized scores from the SP Interview Scale. The Teamwork dimension consists of five sub-nodes: Task

Collaboration, Conflict Resolution, Responsibility Taking, Information Sharing, and Mutual Support, with scores obtained from the mean values of individual items in group task peer evaluations. All sub-indicator node scores undergo Z-score normalization prior to constructing the graph structure. Standardized parameters are retained and applied to validation and test sets to ensure consistency in data preprocessing.

Through the above construction method, this paper effectively represents the multi-dimensional heterogeneous indicators and their complex interaction patterns in competency-based international medical education as a multi-relation heterogeneous graph. This provides a solid data structure foundation for subsequent GAT-based association modeling and prediction, and realizes the precise expression of heterogeneous relationships between competencies and differentiated information dissemination. The constructed multi-dimensional heterogeneous graph is shown in Fig. 1.

In Fig. 1, the multidimensional heterogeneous graph of medical education quality constructed in this paper is shown. The nodes are divided into four core competency dimensions: knowledge and skills, clinical practice, communication skills and teamwork, each with 5 sub-nodes. The colors of the edges in the figure correspond to three types of relationships: blue represents driving associations, orange represents collaborative associations, and yellow represents feedback associations. As can be seen from Fig. 1, the collaborative edges between communication ability nodes are relatively dense, indicating that there is a strong stage-by-stage mutual influence between them and other ability dimensions; at the same time, knowledge and skill nodes are more likely to send out driving edges as source nodes, highlighting their leading role in the process of ability formation. Overall, the graph structure reveals the complex heterogeneous relationships and asymmetric influence paths between the capabilities of each dimension, laying a structural foundation for the subsequent graph neural network modeling.

To prevent temporal leakage caused by incorporating future-stage information during prior graph construction, this study employs a rigorous temporal isolation mechanism. For constructing graph structures in teaching stage t, conditional probabilities and co-occurrence frequencies are calculated exclusively using historical data from the previous t-1 stages, ensuring edge definitions contain no future information. Taking stage 3 as an example, drive edge weights are computed based solely on knowledge/skill scores and clinical practice scores from stages 1-2, while collaboration edges are measured solely by co-occurrence frequencies of communication and teamwork scores from stages 1-2. During training, model inputs for stage t include historical data up to stage t, but the graph structure itself is pre-built using data from stage t-1 and earlier stages and remains fixed throughout training without dynamic adjustments based on subsequent stage information. Both validation and test sets adhere to the same temporal isolation principle to ensure model evaluations reflect real-world prediction scenarios. This mechanism effectively prevents temporal leakage and guarantees the credibility of experimental results.

3.2. GAT model

In view of the heterogeneous relationship and structural complexity of information between nodes in the multi-dimensional heterogeneous graph of competency-based international medical education quality, this paper adopts the graph attention network GAT (Zhang, Hu, & Qu, 2023) as the core information aggregation mechanism to dynamically assign different weights to neighbor nodes to solve the information confusion problem caused by the uniform neighbor aggregation of traditional GNN.

Each capability dimension in this study consists of several sub-indicators (the example in the paper shows 5 sub-nodes per dimension), and the graph is a multi-relational heterogeneous graph constructed at the sample-time level, with a node count and topology far exceeding the "4-node" case. The rationale for using GNNs is that the relation-aware attention mechanism can assign different weights to the three types of edges—"driving/cooperation/feedback"—to finely characterize multi-scale, asymmetric capability interactions; temporal encoding and multi-head aggregation can fuse historical information at the sequence level and improve prediction and interpretability. The complexity of RT-GAT is not redundant, but rather designed to address the high-order structural problems arising from sub-indicators, multiple relationships, and temporal dynamics, while ensuring model scalability and interpretability.

To prevent temporal leakage caused by incorporating future-stage information during prior graph construction, this study employs a rigorous temporal isolation mechanism. For constructing graph structures in teaching stage t, conditional probabilities and co-occurrence frequencies are calculated exclusively using historical data from the

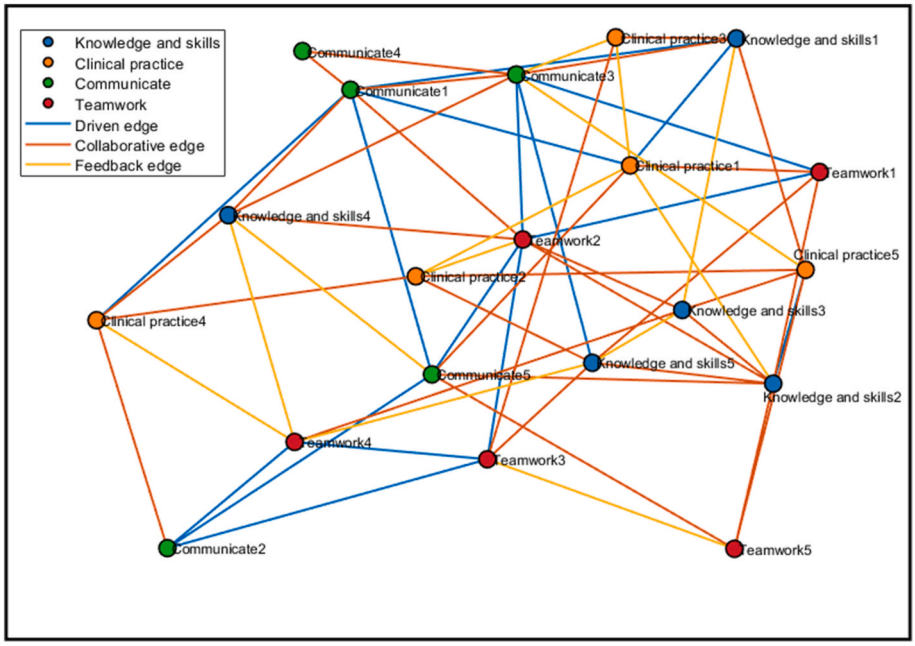

Fig. 1. Multidimensional heterogeneous graph.

previous t-1 stages, ensuring edge definitions contain no future information. Taking stage 3 as an example, drive edge weights are computed based solely on knowledge/skill scores and clinical practice scores from stages 1-2, while collaboration edges are measured solely by co-occurrence frequencies of communication and teamwork scores from stages 1-2. During training, model inputs for stage t include historical data up to stage t, but the graph structure itself is pre-built using data from stage t-1 and earlier stages and remains fixed throughout training without dynamic adjustments based on subsequent stage information. Both validation and test sets adhere to the same temporal isolation principle to ensure model evaluations reflect real-world prediction scenarios. This mechanism effectively prevents temporal leakage and guarantees the credibility of experimental results.

Input the node feature matrix $H^{(0)} = X$ of the heterogeneous graph. In order to improve the model's expressiveness, this paper uses a trainable weight matrix to perform a linear transformation on the input features. The definition of the l layer and the node feature mapping operation are shown in formula (9).

$$\begin{cases} \mathbf{H}^{(l)} = \left[\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, \ldots, \mathbf{h}_N^{(l)}\right]^{\top} \\ \mathbf{z}_i^{(l)} = \mathbf{W}^{(l)}\mathbf{h}_i^{(l)} \end{cases} \tag{9}$$

where $W^{(l)}$ represents the trainable linear transformation matrix, and $z_i^{(l)}$ represents the expression after mapping the node features.

The calculation of the attention coefficient of node i to neighbor node j is shown in formula (10).

$$e_{ij}^{(l)} = \mathrm{LeakyReLU}\left(\mathbf{a}^{\top}\left[\mathbf{z}_i^{(l)} \parallel \mathbf{z}_j^{(l)}\right]\right) \tag{10}$$

where $a^{\top}$ represents the trainable attention vector and LeakyReLU represents the activation function with a negative slope.

To ensure that the attention weights of different neighbors are comparable, the softmax function is used to normalize the attention coefficient, as shown in formula (11).

$$\alpha_{ij}^{(l)} = \frac{\exp\left(e_{ij}^{(l)}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}^{(l)}\right)} \tag{11}$$

where $\alpha_{ij}^{(l)}$ represents the normalized attention weight coefficient of node j to node i in the l layer.

The feature of node i in the $l+1$ layer is obtained by weighted neighbor feature aggregation, as shown in formula (12).

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{(l)}\mathbf{z}_j^{(l)}\right) \tag{12}$$

where σ represents the nonlinear activation function, and $h_i^{(l+1)}$ represents the features of the $l+1$ layer.

In order to improve the stability of the model and capture diverse features, a multi-head attention mechanism is used to execute the above process K times in parallel and output the head features. The calculation is shown in formula (13).

$$\mathbf{h}_i^{(l+1)} = \Big\|_{k=1}^{K}\sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{(l,k)}\mathbf{W}^{(l,k)}\mathbf{h}_j^{(l)}\right) \tag{13}$$

where $\alpha_{ij}^{(l,k)}$ and $W^{(l,k)}$ represent the weight and weight matrix of the kth attention head, respectively.

3.3. Design of relationship-aware attention mechanism

Aiming at the heterogeneous relationships between multi-dimensional indicators of international medical education quality, this paper introduces a relationship-aware attention mechanism (Fang et al., 2023; Li & Wang, 2023) based on GAT to explicitly model the impact of different relationship types on node information aggregation.

The study defines the relationship type as R and fuses the node features with the relationship information, where the feature representation $z_{i,r_m}^{(l)}$ of node i through the relationship transformation r_m at the l layer is shown in formula (14).

$$\mathbf{z}_{i,r_m}^{(l)} = \mathbf{W}_{r_m}^{(l)}\mathbf{h}_i^{(l)} \tag{14}$$

where $W_{r_m}^{(l)}$ represents the trainable weight matrix for the relation type.

Based on the joint representation of node features and relation embeddings, the relation-aware attention score is calculated (Yu et al., 2022) as shown in formula (15).

$$e_{ij}^{(l,r_m)} = \mathrm{LeakyReLU}\left(\mathbf{a}_{r_m}^{\top}\left[\mathbf{z}_{i,r_m}^{(l)} \parallel \mathbf{z}_{j,r_m}^{(l)} \parallel r_m\right]\right) \tag{15}$$

where $a_{r_m}^{\top}$ represents the trainable attention vector of the relation feature, $z_{i,r_m}^{(l)}$ and $z_{j,r_m}^{(l)}$ are the features of i and j after the relation r_m transformation respectively.

In this paper, the neighbor node set of node i is considered as the neighbor nodes connected by the relation, and the calculation of the normalized attention coefficient and the feature update of node i under the relation are shown in formula (16).

$$\begin{cases} \alpha_{ij}^{(l,r_m)} = \dfrac{\exp\left(e_{ij}^{(l,r_m)}\right)}{\sum_{k\in\mathcal{N}_i^{r_m}}\exp\left(e_{ik}^{(l,r_m)}\right)} \\ \mathbf{h}_{i,r_m}^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}_i^{r_m}}\alpha_{ij}^{(l,r_m)}\mathbf{z}_{j,r_m}^{(l)}\right) \end{cases} \tag{16}$$

where $\mathcal{N}_i^{r_m}$ represents the set of neighbor nodes, $\alpha_{ij}^{(l,r_m)}$ represents the normalized attention coefficient, and $h_{i,r_m}^{(l+1)}$ represents the updated features.

Finally, the aggregation results of all relationship types are fused, and the splicing strategy is used to achieve the final node representation $h_i^{(l+1)}$, as shown in formula (17).

$$\mathbf{h}_i^{(l+1)} = \phi\left(\Big\|_{m=1}^{M}\mathbf{h}_{i,r_m}^{(l+1)}\right) \tag{17}$$

where ϕ represents linear transformation.

In order to alleviate the risk of too many model parameters, a sharing mechanism is used to constrain the similarity relationship weight matrix, and the weight matrix is defined to be decomposed into a basic matrix and a relationship bias term, as shown in formula (18).

$$\mathbf{W}_{r_m}^{(l)} = \mathbf{W}_0^{(l)} + \Delta\mathbf{W}_{r_m}^{(l)}, \Delta\mathbf{W}_{r_m}^{(l)} \ll \mathbf{W}_0^{(l)} \tag{18}$$

where $W_0^{(l)}$ represents the universal transformation weight, and $\Delta W_{r_m}^{(l)}$ represents the small-scale relation-specific bias.

3.4. Design of temporal information encoding module

In order to capture the dynamic evolution characteristics of ability indicators with the teaching stage, this paper builds the RT-GAT model based on the R-GAT (Relational-Graph Attention Network) model, designs the temporal information encoding module (Keriven & Vaiter, 2023; Zeng et al., 2023), embeds the time sequence into the multi-dimensional heterogeneous graph node features, and realizes the dynamic modeling of ability changes over time.From a time series perspective, medical education evaluation typically operates on a semester-based framework. The five-stage model corresponds to five

consecutive semesters, comprehensively covering the core training cycle from foundational medical studies to clinical rotations, effectively capturing critical turning points in competency development. In terms of model design, bidirectional LSTM networks in this task do not focus on extracting long-term temporal dependencies. Instead, they utilize joint encoding of forward and backward hidden states to capture contextual information relative to preceding and subsequent stages, enabling smooth modeling of competency evolution trends.

For each capability node, it has a feature vector at different times, and the time series feature matrix is defined as $\widehat{X}_i$. The study uses the position encoding function (Chen, You, et al., 2023; Yeom et al., 2024) to inject time series information into the sequence time step, and the encoding of the tth time step is shown in formula (19).

$$\mathbf{p}^{(t)} = \left[\sin\left(\frac{t}{10000^{\frac{2k}{d_p}}}\right), \cos\left(\frac{t}{10000^{\frac{2k+1}{d_p}}}\right)\right]_{k=0}^{\frac{d_p}{2}-1} \tag{19}$$

where d_p represents the position encoding dimension, k represents the dimension index, and $p^{(t)}$ represents the position encoding function.

Now, the original features are combined with the position encoding to construct a temporal enhanced feature matrix, as shown in formula (20).

$$\widetilde{X}_i = \widehat{X}_i + P \tag{20}$$

where $\widetilde{X}_i$ represents the time series enhanced feature matrix.

The motivation for using sinusoidal positional encoding combined with bidirectional LSTM in this paper is to balance the periodicity and contextual dependence of time series. Sinusoidal encoding can provide generalizable periodic features at different time scales, while Bi-LSTM can capture the sequential dependencies of ability evolution between teaching stages, achieving unified modeling of long-term and short-term dynamics. The design of the relationship types (driving, collaboration, feedback) originates from the causal framework of educational psychology and teaching evaluation theory. These three types of relationships correspond to the causal role, parallel promotion, and dynamic adjustment mechanism in ability formation, respectively, giving the model structure a clear educational semantic foundation at the theoretical level.

The design of sine position encoding combined with bidirectional LSTM stems from an in-depth analysis of the dual characteristics of medical education competency evolution. Medical education demonstrates both cyclical patterns and sequential dependencies in competency development: The cyclical nature manifests through repetitive teaching rhythms (e.g., assessment cycles at semester ends), where sine encoding assigns unique positional signals to each time step via sine and cosine functions of varying frequencies, enabling the model to perceive relative phase positions within cycles. Sequential dependencies are reflected in cumulative competency effects and interphase correlations, with bidirectional LSTM integrating contextual information from preceding and subsequent phases through joint encoding of forward and backward hidden states at each time step.

This paper uses Bi-LSTM(Baskar & Kesavan, 2025) to encode the time series features and capture the context information. The calculation of the hidden state at the tth step is shown in formula (21).

$$\overrightarrow{\mathbf{h}}^{(t)} = \mathrm{LSTM}_f\left(\widetilde{x}_i^{(t)}, \overrightarrow{\mathbf{h}}_i^{(t-1)}\right), \overleftarrow{\mathbf{h}}^{(t)} = \mathrm{LSTM}_b\left(\widetilde{x}_i^{(t)}, \overleftarrow{\mathbf{h}}_i^{(t+1)}\right) \tag{21}$$

where $\overrightarrow{h}^{(t)}$ and $\overleftarrow{h}^{(t)}$ represent the forward and reverse LSTM hidden states, respectively. The final bidirectional hidden states are concatenated to obtain the comprehensive code $h_i^{(t)}$ of time step t.

After concatenation, the attention mechanism is used to weight and aggregate all time step codes into the final temporal representation of the node, as shown in formula (22).

$$\begin{cases} \alpha_i^{(t)} = \dfrac{\exp\left(\mathbf{w}_a^\top \tanh\left(\mathbf{W}_a \mathbf{h}_i^{(t)} + \mathbf{b}_a\right)\right)}{\sum_{t'=1}^{T} \exp\left(\mathbf{w}_a^\top \tanh\left(\mathbf{W}_a \mathbf{h}_i^{(t')} + \mathbf{b}_a\right)\right)} \\ \mathbf{s}_i = \sum_{t=1}^{T} \alpha_t^{(t)} \mathbf{h}_i^{(t)} \end{cases} \tag{22}$$

where W_a represents the attention weight matrix, b_a represents the bias vector, $w_a^\top$ represents the attention projection vector, and s_i represents the weighted temporal coding representation.

The study finally merges the temporal coding with the node features updated by the relation-aware attention mechanism, and obtains the final node state through residual connection and linear transformation, as shown in formula (23).

$$\mathbf{h}_i^{\text{final}} = \sigma\left(\mathbf{W}_s\left(\mathbf{h}_i^{(l+1)} + \mathbf{s}_i\right) + \mathbf{b}_s\right) \tag{23}$$

where W_s represents the fusion weight matrix and b_s represents the bias.

The RT-GAT model architecture is shown in Fig. 2.

Fig. 2 shows the overall architecture of the RT-GAT model in this paper. The model takes a multi-dimensional heterogeneous graph as input, integrates node features, multi-type edge relationships and their interactive structure information, and introduces a temporal information encoding module to model the dynamic characteristics of the capability indicators evolving over time. In Fig. 2, the temporal context features of the nodes are extracted through position encoding and Bi-LSTM, and the node adjacency information under different edge types is weighted and aggregated using the relationship-aware attention mechanism to form a relationship-aware node representation. Then, the static heterogeneous information is integrated with the dynamic temporal representation using residual connections to capture the complex asymmetric heterogeneous associations between capability indicators and the changing trends over time. To further improve the expressive power, the model introduces a multi-head mechanism in the aggregation stage, learns ability-related features in parallel in multiple subspaces, and realizes joint prediction of multiple education quality indicators through multi-task regressors.

In this paper, the fusion of temporal features and multi-head attention features in the model structure adopts a weighted integration approach after parallel embedding. The dynamic representation extracted by the temporal encoding module is first concatenated with the static features output by the relation-aware attention module along the feature dimension, and then linear mapping is used to achieve information balance and scale alignment. Residual connections are only used in the final fusion stage to preserve the stable feature distribution in the original heterogeneous graph structure, prevent gradient vanishing caused by multiple aggregation and temporal encoding, and improve the information consistency and generalization stability of the model across different capability dimensions.

In integrating multi-head attention features with temporal characteristics, this study employs a “parallel embedding post-weighted fusion” strategy. The temporal encoding module extracts dynamic representations while the relationship-aware attention module generates static features. These components are first concatenated along feature dimensions to form a composite feature vector of dimension x. A linear transformation layer then achieves information balance and scale alignment, restoring feature dimensions to original hidden layer dimensions to ensure consistent subsequent processing. This design demonstrates two key advantages: 1) The concatenation operation preserves both feature representations while avoiding information loss during early fusion stages; 2) The linear transformation layer dynamically adjusts static/dynamic feature contribution ratios through learnable weight matrices, enabling task-specific balance optimization. Regarding residual connections, they are introduced exclusively at the final fusion stage rather than after each attention layer computation.

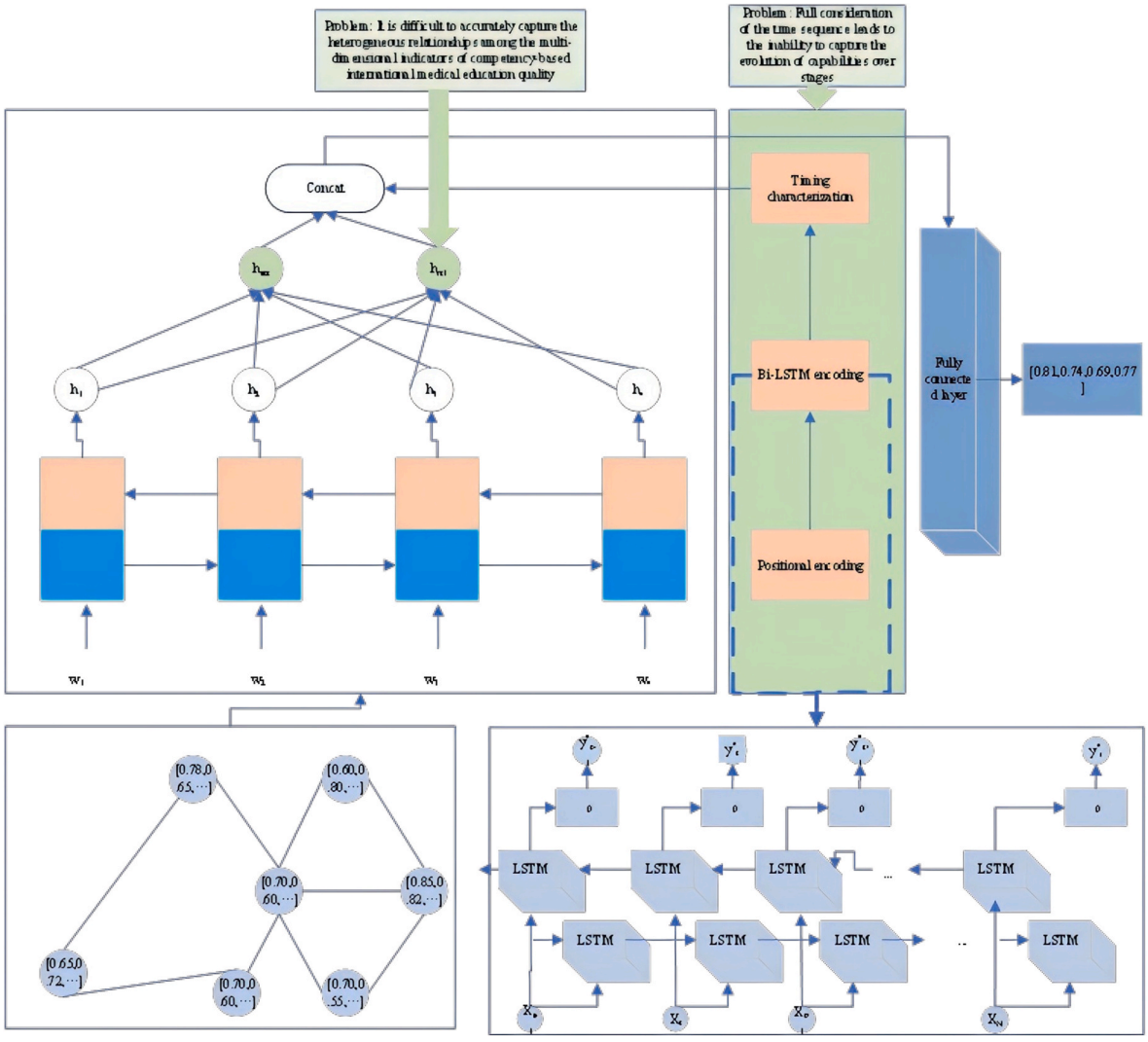

Fig. 2. RT-GAT model architecture.

This approach addresses two challenges: residual connections applied post-layer computation may excessively smooth heterogeneous graph stability features due to multi-layer nonlinear transformations in relationship-aware attention modules and multi-head aggregation processes, potentially reducing node feature distinguishability; temporal encoding modules inherently exhibit temporal sensitivity where frequent early-stage fusion introduces noise interference disrupting static structural stability. Implementing residual connections at final fusion stages effectively mitigates gradient vanishing, preserves stable feature distributions, and enhances model information consistency with generalization stability.

To evaluate the applicability of bidirectional LSTM for sequences with only five time points, this study conducts theoretical analysis and experimental comparisons. Theoretical analysis reveals that the core advantage of bidirectional LSTM in short sequences lies not in mining extremely long temporal dependencies, but rather in enabling each time step to integrate contextual information from preceding and subsequent phases through joint encoding of forward and backward hidden states, thereby achieving smooth modeling of competency evolution trends. Even with only five time points, bidirectional LSTM effectively captures local inter-phase dependencies—for instance, integrating foundational knowledge from the second phase with feedback from the fourth phase during the third phase—resulting in richer temporal representations. Experimental comparisons demonstrate that, while maintaining other modules unchanged, replacing bidirectional LSTM with unidirectional LSTM, GRU, TCN, and simple linear interpolation yields significant performance degradation. Furthermore, visualization of hidden state distributions across time steps reveals that bidirectional LSTM exhibits substantially higher contextual integration weights at initial and final phases compared to intermediate stages, aligning with the practical characteristics of foundational accumulation during early medical education and comprehensive feedback in advanced stages.

3.5. Multi-head joint learning and prediction

In view of the diversity and complexity of capability nodes in multi-dimensional heterogeneous graphs, this paper designs a joint learning framework based on a multi-head mechanism to fully capture the capability association features of different subspaces and realize multi-task regression prediction.

For the input features of each node, K independent linear transformation heads are constructed, and each head is mapped in a different subspace, as shown in formula (24).

$$\mathbf{h}_i^{(k)} = \mathbf{W}^{(k)}\mathbf{h}_i^{\text{final}} + \mathbf{b}^{(k)} \tag{24}$$

where $W^{(k)}$ represents the weight matrix.

Then the relation-aware attention is calculated for each head and the weighted representation of node neighbors is output, as shown in

formula (25).

$$\mathbf{z}_i^{(k)} = \sum_{j \in \mathcal{N}_i} \alpha_{ij}^{(k)} \mathbf{h}_j^{(k)} \tag{25}$$

where $z_i^{(k)}$ represents the weighted representation of node neighbors.

After calculating the relationship-aware attention separately, the node representations of K heads are formed into a comprehensive feature vector z_i, and Dropout regularization is performed to output the feature vector $\widehat{z}_i$.

For the multi-dimensional ability indicators of international medical education quality assessment, a multi-task regressor is designed. Each task corresponds to a regression target, and the calculation of the output layer is shown in formula (26).

$$\widehat{y}_i^{(m)} = \mathbf{w}_m^{\top} \widetilde{\mathbf{z}}_i + b_m \tag{26}$$

where $\widehat{y}_i^{(m)}$ represents the predicted value, $w_m^{\top}$ represents the output layer parameter, and b_m represents the bias.

The predictive outputs and relational attention of RT-GAT can be directly used to support teaching decisions. For example, "high-risk" students identified by the model can be used as early warning targets, and targeted remedial strategies (theoretical reinforcement, simulated operation training, or communication skills workshops) can be generated for each student based on the weights of the three types of edges: drive, collaboration, and feedback. It can also be used to adjust the pace of courses and the allocation of practical training resources, design differentiated assignments and tutoring plans, and use the prediction results as a reference for formative assessment to track the intervention effect. Since the model provides relational interpretability, educators can select more targeted teaching interventions based on the influence paths revealed by the model, and transform "score prediction" into actionable teaching improvement suggestions.

The RT-GAT model's predictive outputs extend beyond mere score forecasting, transforming into actionable teaching decision support tools through explainable mechanisms. For educators, the model-generated individual student competency reports enable early warning systems and targeted interventions. When predicting a student's clinical practice score below 60, the system automatically flags them as "high-risk students" while identifying critical bottlenecks through relational attention weight distribution analysis. If the "knowledge skills → clinical practice" drive edge weight significantly deviates from grade-level averages, instructors may prioritize intensive simulation training. When the "communication → teamwork" collaboration edge contribution rate proves insufficient, cross-cultural communication workshops are recommended. For student cohorts, time-evolution trend charts visually demonstrate competency development trajectories, facilitating personalized learning plans. In curriculum design, group analysis reports optimize teaching resource allocation and provide evidence-based teaching improvements, establishing a closed-loop system from "score prediction" to "teaching intervention."

The model predictions presented in this paper have significant application value in teaching practice. RT-GAT can accurately predict students' development trends across various ability dimensions, providing teachers with targeted instructional design guidance, such as dynamically adjusting teaching focus, optimizing practical activities, and strengthening weaker modules. Simultaneously, the prediction results can support personalized feedback mechanisms, helping students understand the evolution of their own ability structure, develop differentiated learning strategies, and achieve learner-centered continuous improvement and refined instructional support.

The predictive outputs of the RT-GAT model can be directly translated into actionable strategies for enhancing instructional design and delivering personalized feedback. In teaching design, the model-generated cohort analysis reports reveal common learning gaps among students. For instance, if a grade level shows consistently low weights in the "knowledge skills → clinical practice" driver edge, indicating weak integration between theoretical instruction and practical training, educators may increase Problem-Based Learning (PBL) case studies or clinical rotations. When the "communication → teamwork" collaboration edge demonstrates insufficient performance, adding cross-cultural communication modules or optimizing group task designs becomes advisable. For personalized feedback, the model generates individualized competency development reports containing multidimensional prediction scores, relational attention distribution maps, and temporal trend analyses. Teachers can then provide targeted recommendations: for students with low "knowledge skills → clinical practice" driver edge weights, focus on improving theoretical-to-practice translation skills through simulation training; for those with inadequate "communication → teamwork" collaboration performance, emphasize communication skill development via workshop participation. Students can track their competency evolution trajectories, integrate teacher feedback with model-visualized critical bottlenecks, and develop customized learning plans.

The RT-GAT model achieves interpretable outputs through three dimensions: relational attention weight distribution, edge contribution rates, and temporal evolution trends. The relational attention weights visually demonstrate the relative importance of three edge types in information aggregation, reflecting the influence intensity across different competency dimensions. Edge contribution rates quantify the contribution ratio of each relationship type to final predictions, identifying dominant pathways for competency formation. Temporal evolution trends reveal dynamic development of competency dimensions throughout teaching processes through phase changes in Pearson correlation coefficients. These metrics are presented in visual charts, enabling teachers to transform prediction results into personalized instructional intervention recommendations based on established rules.

3.6. Model training and optimization

3.6.1. Loss function design

For the continuous regression task of multi-dimensional ability indicators, the mean-square error (MSE) is defined as the main loss function, as shown in formula (27).

$$\mathscr{L}_{\mathrm{MSE}} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} \left(y_i^{(m)} - \widehat{y}_i^{(m)} \right)^2 \tag{27}$$

where $y_i^{(m)}$ represents the true value and $\widehat{y}_i^{(m)}$ represents the predicted value.

The final training objective of RT-GAT consists of a weighted combination of primary and auxiliary losses. The primary loss measures the mean squared error of multidimensional capability prediction, evaluating the model's accuracy across four capability dimensions. Auxiliary losses include adjacency matrix reconstruction error and relationship embedding regularization terms: the former ensures learned node representations can reconstruct predefined graph structures, while the latter prevents overfitting of relationship embeddings. The total loss function is defined as $L = L_{MSE} + \lambda_1 L_{recon} + \lambda_2 L_{reg}$, where λ_1 and λ_2 represent auxiliary loss weights. Grid search was employed to optimize these parameters on the validation set, with candidate ranges of {0.01,0.05,0.1, 0.2, 0.5}. Optimal balance was achieved when $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$, yielding the lowest prediction RMSE while maintaining a cosine similarity of 0.89 for adjacency reconstruction. Excessively low λ_1 (<0.05) leads to insufficient graph structure utilization and increased reconstruction errors, whereas excessively high λ_1 (>0.2) imposes excessive constraints on prediction tasks, resulting in elevated RMSE. The current weight settings ensure synergistic optimization between prediction accuracy and structural representation.

To prevent the model from overfitting, a parameter regularization term is added and L2 norm regularization is used. The corresponding

loss is shown in formula (28).

$$\mathscr{L}_{\text{reg}} = \varrho\left(\sum_{k=1}^{K}\left\|\mathbf{W}^{(k)}\right\|_F^2 + \|\mathbf{W}\|_F^2 + \sum\left\|\theta_{\text{RT-GAT}}\right\|_F^2\right) \tag{28}$$

where $W^{(k)}$ represents the feature transformation weight matrix of the attention head, W represents the weight matrix of the multi-task output layer, θ_{RT-GAT} represents the set of all other trainable parameters in the RT-GAT model, and ϱ represents the regularization hyperparameter.

The final loss function is shown in formula (29).

$$\mathscr{L} = \mathscr{L}_{\text{MSE}} + \mathscr{L}_{\text{reg}} \tag{29}$$

where L represents the total loss function.

3.6.2. Parameter update and optimization of the model

The experiment is based on the objective function, using automatic differentiation technology to calculate the parameter gradient, and combined with the chain rule, back-propagating the error signal from the output layer to the front, and calculating the relational attention mechanism, temporal coding module and multi-head joint learning parameter gradients of each layer in turn.

During training, the experiment uses the Adam optimizer (Mehmood et al., 2023; Reyad et al., 2023) to perform parameter updates, combined with adaptive learning rate and first-order moment estimation to ensure stable and rapid convergence of the training process. The parameter update rule is shown in formula (30).

$$\theta^{(t+1)} = \theta^{(t)} - \eta\frac{\widehat{\pi}_t}{\sqrt{\widehat{\tau}_t} + \epsilon} \tag{30}$$

where $\theta^{(t)}$ represents the parameter value, η represents the initial learning rate, $\widehat{\pi}_t$ and $\widehat{\tau}_t$ represent the bias-corrected estimates of the first-order and second-order moments of the gradient, respectively, and ϵ represents the numerical stability constant.

This paper incorporates an auxiliary objective of adjacency matrix reconstruction and regularization constraints on relation embedding/attention distribution, in addition to the main regression loss. This allows relation embedding, relation-aware attention vectors, and adjacency reconstruction to be optimized end-to-end through backpropagation. Simultaneously, parameter sharing and L2 regularization, along with Dropout and early stopping, are employed to prevent the reconstruction terms from excessively biasing the regression performance. The adjacency reconstruction task provides direct supervision signals for relation representation and attention weights during the training phase, improving the structured representation learned by the model on heterogeneous multi-relation graphs and the final prediction performance.

Adjacent reconstruction and relationship-aware attention mechanisms serve not only as evaluation metrics but also as core components that participate throughout the model training process. During training, adjacent reconstruction functions as an auxiliary loss function that is jointly optimized with the primary prediction loss. The total loss function is defined as: $\mathscr{L} = \mathscr{L}_{\text{MSE}} + \lambda_1\mathscr{L}_{\text{recon}} + \lambda_2\mathscr{L}_{\text{reg}}$.The reconstruction loss $\mathscr{L}_{\text{r}}$econ directly influences trainable parameters in the relationship-aware attention module through backpropagation, including relationship-specific weight matrices $W_{r_m}^{(l)}$ and attention vectors $a_{r_m}^{T}$.

The experimental hyperparameters are shown in Table 1.

Table 1
Hyperparameters.

Parameters	Value	Parameters	Value
Learning rate	0.001	Bi-LSTM hidden state dimension	32
Number of training rounds	200	Attention hidden dimension	32
Number of attention heads	8	Dropout rate	0.5
Linear transformation dimension	64	Weight decay (L2 regularization)	$1*10^{-4}$
Relation embedding dimension	32	Early stop waiting steps	20
Position encoding dimension	64	Gradient clipping norm upper limit	5

4. Multi-dimensional correlation modeling and prediction experiment of education quality

4.1. Experimental data

The experimental data of this paper comes from the multi-point evaluation of international clinical medicine students in China Medical University, Xiangya School of Medicine, Central South University, and Qilu Medical College of Shandong University. 300 international clinical medicine students were randomly selected from each school, totaling 900 samples. The evaluation time points are five consecutive teaching stages for ability assessment, namely the end of the first semester, the end of the second semester, the end of the third semester, the end of the fourth semester, and the end of the fifth semester. The indicator values are collected at each time point according to the unified evaluation standards. Multidimensional indicators include knowledge and skills (theoretical test scores, full score 100 points), clinical practice (operation scores, full score 100 points), communication skills (standardized patient interview scores, full score 100 points), and teamwork (average score of group task mutual evaluation, full score 100 points). Other data include nationality, gender, age, enrollment background, etc.The dataset for this study comprises 900 students, demonstrating high representativeness and diversity. The sample covers five consecutive stages of instruction, with students from over 20 countries, primarily aged 18 to 26, and including diverse genders and backgrounds. This multinational, multi-age, and multi-background combination helps capture the heterogeneity of competency development in international medical education, providing a reliable data foundation for model training and cross-cultural generalization.

For missing value and noise processing, multiple imputation was employed to address potential scoring omissions or recording errors (accounting for approximately 1.8% of cases) in the OSCE and SP scoring systems. The imputation process utilized the mean values of similar student cohorts (with identical nationality and comparable admission backgrounds) as reference benchmarks, while preserving the inter-dimensional correlation structures. To mitigate measurement noise, a local outlier factor (LOF)-based noise detection method was implemented to identify and eliminate outliers exceeding three standard deviations above the threshold. Subsequently, residual samples underwent minor Gaussian smoothing to minimize the impact of random fluctuations on model training.

Regarding data collection, the measurement of the four competency dimensions was conducted according to the internationally standardized medical education competency assessment framework developed by each institution: knowledge and skills were represented by scores from standardized theoretical examinations organized by the course group; clinical practice was quantified by weighted scores of operational items in the Objective Structured Clinical Examination (OSCE); communication skills were based on standardized patient (SP) interview rating scales; and teamwork was calculated based on the average scores of peer assessments from each member in group tasks. All assessments were completed independently by trained teachers or assessment experts and underwent two rounds of consistency verification (Cohen's $\kappa > 0.85$) to ensure the reliability and comparability of the data.

Data collection for the four competency dimensions strictly adhered to the International Standardized Medical Education Assessment Framework. For the Knowledge and Skills dimension, standardized theoretical examinations were conducted, covering five modules including basic medicine and clinical medicine, with each module

scored out of 100 points. The dimension score was calculated as the average score, utilizing a combination of automated grading for objective questions and independent scoring by two evaluators for subjective questions, achieving an inter-rater consistency coefficient (ICC) of 0.91 for subjective question scoring. For the Clinical Practice dimension, objective structured clinical examinations were employed to measure performance, featuring five assessment stations including medical history-taking. Each station was independently scored by two evaluators simultaneously, with the total score derived from weighted summation of the five stations. The inter-rater reliability coefficient (Cohen's κ) was 0.87. For the Communication Skills dimension, standardized patient interviews were used, with scores based on five subscale items. A two-evaluator independent scoring system combined with third-party verification was implemented, yielding a Cronbach's α of 0.89. For the Teamwork dimension, peer evaluations were conducted through group tasks, where members anonymously assessed each other across five dimensions and averaged the scores to determine individual scores. The intra-group consistency coefficient (Kendall's harmony coefficient W) was 0.82. All data were entered after undergoing two rounds of consistency validation prior to processing.

To ensure the reproducibility of cross-institutional and cross-national generalization experiments, this study collected additional sample data from five top Chinese medical schools (Peking University School of Medicine, Fudan University School of Medicine, Shanghai Jiao Tong University School of Medicine, Harbin Medical University, and Wuhan University School of Medicine) and five major international student source countries (the United States, India, the United Kingdom, Russia, and Nigeria), in addition to the initial training set. At least 100 students were selected from each institution or country. The evaluation indicators were consistent with those in the training set, including knowledge and skills, clinical practice, communication skills, and teamwork. The evaluation was conducted according to uniform standards across institutions and reviewed by assessment experts to ensure the reliability and comparability of the data.

Cross-institutional and cross-national validation data were independently collected from five domestic medical schools and five countries of international student origin that did not participate in model training. Domestic institutions included Peking University Health Science Center, Fudan University School of Medicine, Shanghai Jiao Tong University School of Medicine, Harbin Medical University, and Wuhan University School of Medicine, with 100 international medical students randomly selected from each institution, totaling 500 participants. International samples were drawn from the United States, India, the United Kingdom, Russia, and Nigeria, with 100 international students from each country selected, totaling 500 participants. All samples were assessed using the same evaluation framework and measurement tools as the training set, including standardized theoretical examinations, OSCE assessments, SP interviews, and peer evaluations of group tasks. Assessment criteria were standardized through unified training to ensure consistency, with cross-institutional ICC validation ranging from 0.85 to 0.89. Data collection occurred from September 2023 to January 2024, independent of the training data time window.

4.2. Data preprocessing

For the small number of missing values (less than 2%) in the original assessment data, multiple interpolation methods based on the mean of similar student groups are used to fill in the missing values, retaining the correlation structure between indicators of each dimension. Unique-hot encoding is used for categorical features such as nationality, gender, and enrollment background. Age and test scores are Z-score standardized by dimension to conform to the N(0,1) distribution, eliminating the impact of dimensional differences on subsequent model training.

Before constructing a multi-dimensional heterogeneous graph, based on the interpolated and standardized time series tensor, combined with the expert-defined driving, collaborative, and feedback relationship rules, weighted edges are generated by calculating the conditional co-occurrence frequency threshold (0.1) At the same time, the time step t of each sample is mapped to the dimension through sine-cosine position encoding and concatenated with the standardized features to obtain the input data, laying a solid foundation for the temporal encoding module and the relational attention mechanism.

To reduce the potential impact of sample distribution differences on model performance, this paper performs stratified balancing of samples before data modeling. For categorical variables such as nationality, gender, and educational background, a combination of stratified sampling and SMOTE (Synthetic Minority Over-sampling Technique) is used to balance the proportions of each category and prevent bias in the model during training.

To address issues like scoring omissions, recording errors, and evaluator inconsistencies in OSCE and SP scoring systems, this paper employs a multi-stage data cleaning strategy for data quality assurance. For minor missing values (≈1.8%) caused by equipment failures or missing records in OSCE exams, a multiple interpolation method based on similar student groups is used, taking the mean of students with the same nationality, similar admission backgrounds, and early performance as the benchmark, while preserving indicator structures to avoid variance underestimation. To mitigate subjective bias in SP interview scores, a dual-evaluator independent scoring and consistency check mechanism is implemented; if scores differ by > 5 points, a senior expert reviews and takes the median as the final score. For measurement noise, a local outlier factor-based method identifies and removes abnormal samples with score deviations exceeding 3 times the threshold, followed by Gaussian smoothing to reduce random fluctuations and retain true ability trends.

IIn demonstrating the SMOTE mechanism's role, this paper applies it to regression analysis to address sample bias caused by uneven categorical variable distributions (e.g., nationality, admission background). In the original dataset, India and the U.S. dominate student samples, while Nigeria and Russia account for $<10\%$. Training directly on this imbalanced data risks overfitting majority categories and neglecting minority patterns. SMOTE mitigates this by generating synthetic samples through linear interpolation between minority samples and their neighbors. Specifically, a k-neighbor search ($k = 5$) is performed on minority samples' feature space, and new points are randomly created between each minority sample and its randomly selected neighbor until categories reach a preset balance ratio (80% of majority sample size).To counter graph construction risks that may reinforce existing structures, the paper introduces multiple mechanisms during data preprocessing and model validation to ensure objective edge definitions. Edge generation combines co-occurrence statistics and expert input, while independent validation samples (excluded from training) cross-verify edge weights: 900 students are split into 50% validation sets (rotated) and 40% for weight calculation. Model training incorporates an adjacency matrix reconstruction loss term and an early stopping system, terminating training if validation set reconstruction error fails to decrease for 10 consecutive rounds, preventing overfitting to preset graph structures.

4.3. Experimental plan

This paper designs multiple groups of comparative experiments to verify the effectiveness of the RT-GAT model in modeling and predicting the multi-dimensional correlation of the quality of international medical education based on competency. Based on a unified multi-time point evaluation data set, a five-fold cross-validation method is used to ensure a reliable evaluation of the model's generalization ability. The experiments train and test the classic graph neural network model GraphSAGE (Graph SAmple and aggreGatE), the temporal dynamic graph model TGN, the heterogeneous graph transformer HGT (Heterogeneous Graph Transformer), the Study-GNN fused with Bi-LSTM, the GAT-TCN (graph Attention Network-Temporal Convolutional Network) combined with the temporal convolutional network, and the RT-GAT model of this

paper. Each model is run on the same training, validation, and test data sets, and hyperparameters are uniformly adjusted to ensure fair comparison.

For the regression tasks of multi-dimensional correlation modeling and ability prediction, the model's ability to depict the correlation of multi-dimensional ability characteristics and the prediction accuracy of future abilities are evaluated respectively. The experiment comprehensively measures the learning effect of the model on multi-dimensional heterogeneous correlation structure by comparing indicators such as correlation matrix reconstruction error, spectral distance and cosine similarity. At the same time, indicators such as MSE and MAE evaluate the accuracy of ability prediction. The experimental design ensures systematic verification of the advantages of the model in capturing complex multi-dimensional relationships and time series evolution.

To accurately evaluate the practical value of the RT-GAT model's complex architecture, this study introduces four baseline models for comparison: Linear regression captures linear relationships between competency dimensions; Ridge regression assesses regularization effects on prediction stability under multicollinearity conditions; Multilayer perceptron serves as a representative neural network to demonstrate nonlinear modeling capabilities; and the Standard Knowledge Tracking Model (DKT), a classic time-series model in educational assessment, highlights RT-GAT's advantage in capturing dynamic competency evolution. All baseline models were trained and tested on the same five-phase dataset using identical five-fold cross-validation and evaluation metrics to ensure fair and interpretable comparison results.

4.4. Evaluation indicators

The calculation of cosine similarity is shown in formula (31).

$$\mathrm{CosSim}(\mathbf{A},\widehat{\mathbf{A}})=\frac{\langle \mathrm{vec}(\mathbf{A}),\mathrm{vec}(\widehat{\mathbf{A}})\rangle}{\|\mathrm{vec}(\mathbf{A})\|_2\cdot\|\mathrm{vec}(\widehat{\mathbf{A}})\|_2} \tag{31}$$

where A represents the true multi-dimensional ability association matrix, $\widehat{\mathrm{A}}$ represents the association weight matrix learned by the model. vec represents the matrix expansion into vector operation, and $\|.\|_2$ represents the Euclidean norm.

The calculation of the adjacency matrix reconstruction error (Frobenius norm) is shown in formula (32).

$$\mathscr{L}_{\mathrm{recon}}=\|\mathbf{A}-\widehat{\mathbf{A}}\|_F=\sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}\left(A_{ij}-\widehat{A}_{ij}\right)^2} \tag{32}$$

where $\|.\|_F$ represents the Frobenius norm of the matrix.

The calculation of the spectral similarity measure (spectral distance) is shown in formula (33).

$$\begin{cases}\mathbf{A}=\mathbf{U}\Lambda\mathbf{U}^{\top},\widehat{\mathbf{A}}=\widehat{\mathbf{U}}\widehat{\Lambda}\widehat{\mathbf{U}}^{\top}\\ D_{\mathrm{spec}}=\|\Lambda-\widehat{\Lambda}\|_2=\max\limits_{i=1,\ldots,d}|\lambda_i-\widehat{\lambda}_i|\end{cases} \tag{33}$$

where Λ represents the eigenvalue matrix of the true matrix, $\widehat{\Lambda}$ represents the eigenvalue matrix of the model prediction matrix, and U represents the eigenvector matrix of the matrix.

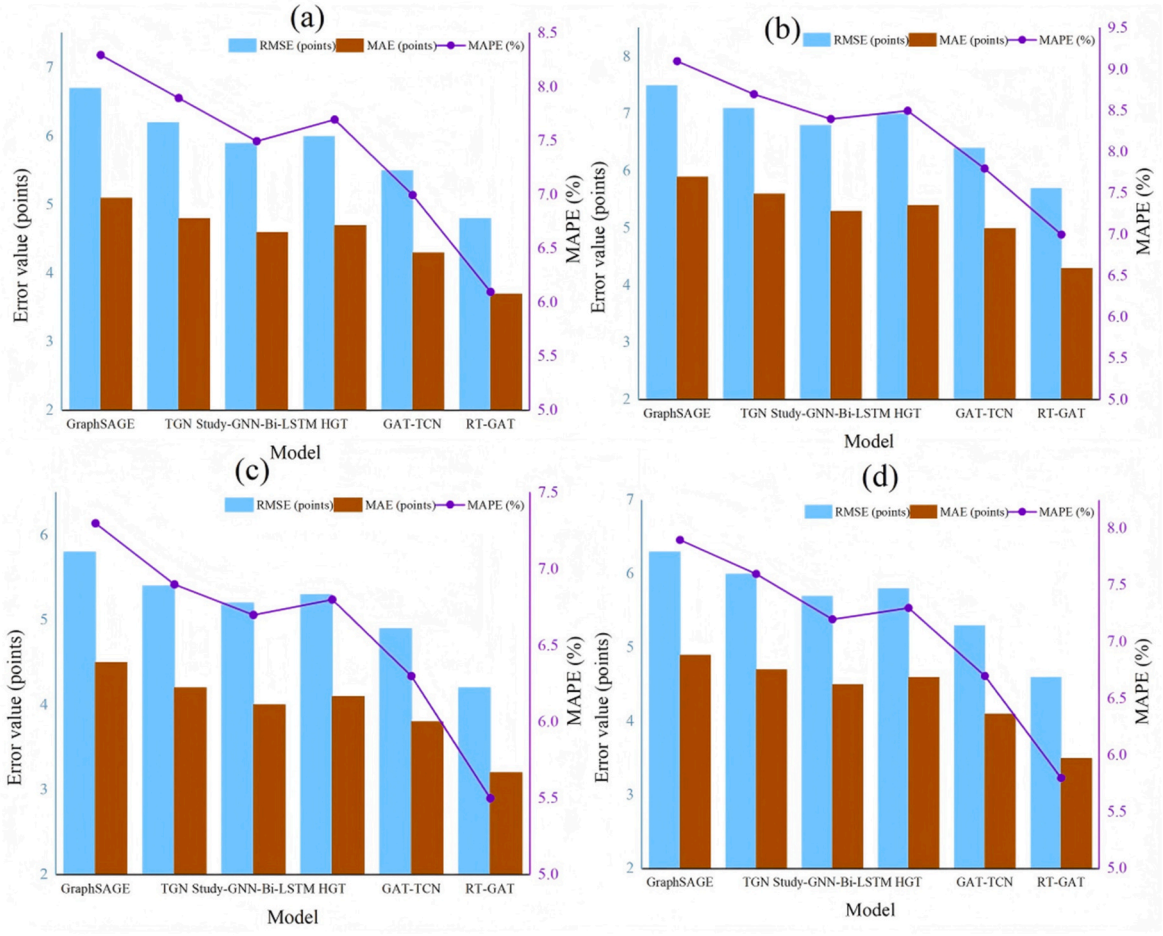

Fig. 3. Comparison of multi-dimensional prediction accuracy.

5. Results of multi-dimensional correlation modeling and prediction of education quality

5.1. Comparison of multi-dimensional prediction accuracy

Aiming at the four ability dimensions of knowledge and skills, clinical practice, communication skills and teamwork, this paper uses RMSE, MAE and MAPE to systematically compare the prediction accuracy of different models. The results are shown in Fig. 3, sections (a)-(d) compare the prediction accuracy of the four capability dimensions. Fig. 3(a) shows the prediction accuracy of knowledge and skills; Fig. 3 (b) shows the prediction accuracy of clinical practice; Fig. 3(c) shows the prediction accuracy of communication skills; and Fig. 3(d) shows the prediction accuracy of teamwork. The horizontal axis represents each model, and the vertical axis represents the RMSE, MAE, and MAPE indices.

Fig. 3 demonstrates the predictive accuracy comparison between RT-GAT and various baseline models across four competency dimensions. Each sub-figure corresponds to one competency dimension, with the x-axis representing the models and the y-axis showing RMSE, MAE, and MAPE metrics. Fig. 3(a) presents the prediction results for the knowledge and skills dimension; Fig. 3(b) for the clinical practice dimension; Fig. 3(c) for the communication skills dimension; and Fig. 3(d) for the teamwork dimension.In Fig. 3 (a) knowledge and skills, RT-GAT is far ahead of GraphSAGE (RMSE about 6.7 points, MAE about 5.1 points, MAPE about 8.3%) and other models in terms of accuracy with RMSE about 4.8 points, MAE about 3.7 points and MAPE about 6.1%. Compared with GAT-TCN, RT-GAT has a lower RMSE of nearly 0.7 points, a lower MAE of 0.6 points, and a lower MAPE of 0.9%, which shows that it is more accurate in error control of theoretical knowledge scores. In the clinical practice dimension of Fig. 3 (b), RT-GAT achieved an RMSE of about 5.7 points, a MAE of about 4.3 points, and a MAPE of about 7.0%. GraphSAGE has an RMSE of about 7.5 points, a MAE of about 5.9 points, and a MAPE of about 9.1%. The TGN model has an RMSE of about 7.1 points, a MAE of about 5.6 points, and a MAPE of about 8.7%, which are all worse than RT-GAT. GAT-TCN has an RMSE of about 6.4 points, a MAE of about 5.0 points, and a MAPE of about 7.8%. RT-GAT reduces RMSE by 0.7 points compared to GAT-TCN, and reduces the average percentage error by 0.8%.

For the communication ability in Fig. 3 (c), RT-GAT has an RMSE of about 4.2 points, a MAE of about 3.2 points, and a MAPE of about 5.5%, surpassing HGT (RMSE of about 5.3 points, MAE of about 4.1 points, and MAPE of about 6.8%) and Study-GNN-Bi-LSTM (RMSE of about 5.2 points, MAE of about 4.0 points, and MAPE of about 6.7%). Compared with GAT-TCN (RMSE about 4.9 points, MAE about 3.8 points, MAPE about 6.3%), RT-GAT reduces RMSE by 0.7 points and MAPE by 0.8%. In the teamwork dimension of Fig. 3 (d), RT-GAT achieves RMSE of about 4.6 points, MAE of about 3.5 points, and MAPE of about 5.8%, which is nearly 1.7 points and 2.1% lower in RMSE and MAPE, respectively, compared with GraphSAGE (RMSE about 6.3 points, MAE about 4.9 points, MAPE about 7.9%). TGN has an RMSE of about 6.0 points, a MAE of about 4.7 points, and a MAPE of about 7.6%. RT-GAT also shows a significant improvement in accuracy compared with GAT-TCN (RMSE of about 5.3 points, MAE of about 4.1 points, and MAPE of about 6.7%).

The differences in data distribution and inherent complexity of different ability dimensions themselves significantly affect the difficulty of prediction. Clinical practice scores are affected by multiple operational details and on-site performance, and the score distribution is more discrete and fluctuates greatly, resulting in generally high errors in this dimension for each model. The communication ability scores are relatively concentrated and fluctuate less, making the MAPE of all models in this dimension less than 8%. The knowledge and skills test scores are highly standardized, but are affected by the difficulty of the test questions and the differences in students' foundation. The combined effect of the two makes the MAPE of this dimension generally higher than that of the teamwork dimension. RT-GAT leads in all dimensions, mainly because its relational attention mechanism can give differentiated weights to heterogeneous interactions such as “driving”, “collaborative”, and “feedback”, and finely depict the impact paths between various capability dimensions. At the same time, temporal information encoding enables the model to make full use of the dynamic evolution characteristics of multi-time point evaluation data to improve the accuracy of prediction of future states. GraphSAGE based on static graphs is different from HGT which lacks relationship distinction. RT-GAT combines structural heterogeneity with temporal dynamic characteristics, significantly enhancing the modeling ability and prediction performance of complex ability correlation.

To verify the significant performance improvement of RT-GAT model in multi-dimensional ability prediction compared with other comparison models, this paper conducted a paired T test on the MAE indicators of different models. The specific steps include: (1) taking the MAE of the RT-GAT model as the benchmark, performing paired sample T test with the MAE of the GraphSAGE, TGN, HGT, Study-GNN-Bi-LSTM and GAT-TCN models one by one; (2) calculating the mean, standard deviation, t statistic and corresponding significance level of each group; (3) using $P < 0.05$ as the criterion for significant difference, testing the statistical significance of the difference in model performance. The MAE comparison results of each model and RT-GAT in all capability dimensions and the corresponding T test statistics are shown in Table 2.

As can be seen in Table 2, RT-GAT showed significantly lower MAE in all comparisons. Taking GraphSAGE as an example, its MAE dropped from 5.1 ± 0.4 to 3.7 ± 0.3, with a t value of about 9.21 and P of about 0.0003. Compared with TGN, MAE dropped from 4.8 ± 0.3 to 3.7 ± 0.3, with a t value of about 7.84 and P of about 0.0007. For HGT, MAE dropped from 4.7 ± 0.35 to 3.7 ± 0.3, with a t value of about 7.15 and P of about 0.0012. The MAE of Study-GNN-Bi-LSTM dropped from 4.6 ± 0.32 to 3.7 ± 0.3, with a t value of about 6.82 and a P of about 0.0018. Even compared with the 4.3 ± 0.28 of GAT-TCN, RT-GAT still leads with 3.7 ± 0.3, a t value of about 5.43 and a P of about 0.0041. All P values were lower than 0.005, indicating that RT-GAT had a highly statistically significant difference with each model in the MAE of multi-dimensional ability prediction.

GraphSAGE is based only on homogeneous graph aggregation, TGN introduces time series but does not consider multiple relationship types, HGT supports heterogeneity but does not combine time series, Study-GNN-Bi-LSTM and GAT-TCN focus on time series or topological convolution respectively. RT-GAT uses relationship-aware attention to assign different weights to edge types such as “driving”, “collaborative”, and “feedback”, ensuring that the impact paths between various capability dimensions are accurately captured. The temporal encoding module enables the model to utilize the ability evolution information of the five teaching stages, introduce dynamic context in prediction, reduce the error accumulation caused by ignoring the information of the previous

Table 2
T test results of multi-dimensional prediction of different models.

Model	Comparison models	MAE(mean ± std)	t-value	P-value	Significance
GraphSAGE	RT-GAT	5.1 ± 0.4 vs 3.7 ± 0.3	9.21	0.0003	Significant
TGN		4.8 ± 0.3 vs 3.7 ± 0.3	7.84	0.0007	Significant
HGT		4.7 ± 0.35 vs 3.7 ± 0.3	7.15	0.0012	Significant
Study-GNN-Bi-LSTM		4.6 ± 0.32 vs 3.7 ± 0.3	6.82	0.0018	Significant
GAT-TCN		4.3 ± 0.28 vs 3.7 ± 0.3	5.43	0.0041	Significant

and subsequent stages of each model, and significantly reduce the regression MAE.

The multi-head joint learning strategy also plays a key role in the error reduction process. Different attention heads extract ability associations and temporal features in parallel in several subspaces, realizing multi-angle modeling of complex multi-dimensional interaction patterns, rather than single subgraph or pipeline processing, further improving the robustness of prediction. Combined with this innovation, RT-GAT can take into account both relational heterogeneity and temporal dynamics when processing competency-based education quality data, and continues to outperform the other five comparison models in MAE.

To comprehensively evaluate the practical value of the RT-GAT model's complex architecture, this study introduces four new baseline models for comparison. All baseline models were trained and tested on the same five-phase dataset using identical five-fold cross-validation and evaluation metrics to ensure fairness and interpretability of comparison results. The performance comparison between each baseline model and RT-GAT in multidimensional capability prediction tasks is presented in Table 3.

The comparison results of basic models in Table 3 demonstrate that RT-GAT significantly outperforms four foundational models across all competency dimensions and evaluation metrics. Compared to linear regression, RT-GAT achieves 3.4-point reductions in RMSE, 2.8-point decreases in MAE, and 4.3 percentage point improvements in MAPE for knowledge and skills dimensions. These results indicate complex nonlinear correlations between competency dimensions that linear models struggle to capture effectively. When contrasted with ridge regression, RT-GAT demonstrates distinct advantages, suggesting that while regularization mitigates multicollinearity, it fails to accurately model heterogeneous relationships and temporal dynamics. Although multilayer perceptrons (MLP) as fundamental neural networks outperform linear models, they still lag behind RT-GAT in knowledge and skills dimensions, with RMSE exceeding baseline values by 2.4 points, highlighting that simple nonlinear mappings cannot capture competency interaction patterns. The knowledge tracking model, as a classical time series model, surpasses the first three categories but remains inferior to RT-GAT across all dimensions, particularly showing 1.9-point higher RMSE in communication competence. RT-GAT's relational attention mechanism enhances representation of heterogeneous competency associations, while its sophisticated architecture design effectively captures complex correlations and dynamic trends, thereby improving predictive accuracy.

To more intuitively demonstrate the practical impact of improved prediction accuracy on educational decision-making, this study summarizes performance metrics for two decision-making tasks—high-risk student alerts and outstanding student identification—as shown in Table 4.

Table 4 demonstrates that RT-GAT excels in high-risk student identification tasks, achieving an accuracy rate of 87.3%, recall rate of 84.6%, and F1 score of 85.9%—representing improvements of 6.1, 4.8, and 5.4 percentage points compared to GAT-TCN. During simulated screening of 1000 students, RT-GAT accurately identified 190 high-risk cases, with 23 additional correct detections while reducing false positives by 11 and false negatives by 8. With an average intervention duration of 2 h per high-risk student, this system saves teachers approximately 46 h of ineffective intervention time per semester, enabling timely support for more students in need. In outstanding student recognition tasks, RT-GAT achieved an accuracy rate of 82.5%, recall rate of 79.8%, and F1 score of 81.1%—showing improvements of 6.1, 6.6, and 6.3 percentage points over GAT-TCN, with 16 additional correct detections and reductions in false positives by 7 and false negatives by 9. The 0.6-point reduction in RMSE demonstrates measurable decision-making value, while enhanced prediction accuracy facilitates optimized allocation of teaching resources and student support services.

To comprehensively evaluate the fairness and performance stability of the RT-GAT model across different demographic subgroups, this study conducted stratified performance analyses based on three dimensions: gender, nationality, and enrollment background. The sample sizes, RMSE, MAE, and significance tests for differences from the baseline group in each subgroup are presented in Table 5.

The subgroup performance analysis results in Table 5 indicate that the RT-GAT model demonstrates good overall fairness across different demographic subgroups, though performance variations exist. In the gender dimension, the RMSE values for male and female students were 4.82 and 4.78 points respectively, with a difference of 0.04 points that was not statistically significant, suggesting balanced model performance between genders. Regarding nationality, using American students as the baseline, the RMSE differences for Indian, British, and Russian students ranged from 0.2 to 0.7 points, all non-significant. However, Nigerian students exhibited a statistically significant RMSE increase of 1.6 points ($p < 0.05$), attributed to their small sample size (only 72 participants, accounting for 8%) and high data dispersion (standard deviation of 9.2, exceeding the overall mean of 6.8). In the enrollment background dimension, clinical medicine background students achieved an RMSE of 4.79 points, lower than non-medical background students' 5.18 points (difference of 0.39 points, non-significant). These findings indicate robust model performance in subgroups with sufficient sample sizes and stable data distribution, while performance biases may occur in undersized samples or high data heterogeneity. Future research should focus on collecting balanced cross-cultural samples and implementing fairness constraint mechanisms.

5.2. Multi-dimensional correlation modeling quality

In order to further verify the ability of each model to describe the correlation of multi-dimensional capabilities, this paper uses three indicators: cosine similarity, adjacency matrix reconstruction error (Frobenius norm) and spectral similarity measurement (spectral distance) to compare the correlation matrix output by the model with the true correlation matrix. The results are shown in Fig. 4.

In Fig. 4, RT-GAT performs best in multi-dimensional correlation modeling quality, with a cosine similarity of 0.91, much higher than the lowest GraphSAGE of 0.78. The adjacency matrix reconstruction error is only 0.83, while GraphSAGE reaches 1.32; the spectral distance is 0.36, which is significantly lower than GraphSAGE's 0.56. TGN and HGT achieved combined scores of 0.81/1.18/0.51 and 0.83/1.12/0.49, respectively. Study-GNN-Bi-LSTM achieved 0.84/1.05/0.47 in cosine similarity, adjacency matrix reconstruction error, and spectral similarity metrics, respectively, and GAT-TCN was further optimized to 0.86/

Table 3
Performance comparison between baseline model and RT-GAT on multidimensional ability prediction tasks.

Model	Knowledge and skills		Clinical practice		Communication skil		Teamwork	
	MAE	MAPE(%)	MAE	MAPE(%)	MAE	MAPE(%)	MAE	MAPE(%)
Linear regression	8.2	6.5	9.1	7.3	7.8	6.1	8	6.3
Ridge return	8	6.3	8.9	7.1	7.6	5.9	7.8	6.1
Multilayer sensor	7.2	5.6	8	6.4	6.8	5.3	7	5.5
DKT	6.5	5	7.3	5.8	6.1	4.8	6.3	4.9
RT-GAT	4.8	3.7	5.7	4.3	4.2	3.2	4.6	3.5

Table 4
Performance comparison of educational decision-making tasks.

Evaluation index	Early warning for high-risk students (score <60 points)			Identification of outstanding students (score >85 points)		
	GAT-TCN	RT-GAT	Improvement range	GAT-TCN	RT-GAT	Improvement range
Accuracy (%)	81.2	87.3	6.1	76.4	82.5	6.1
Recall rate (%)	79.8	84.6	4.8	73.2	79.8	6.6
F1 score (%)	80.5	85.9	5.4	74.8	81.1	6.3
Correctly identify the number of people (people/1000 people)	167	190	23	112	128	16
Number of false positives (people/1000 people)	39	28	−11	34	27	−7
Number of underreporting (people/1000 people)	42	34	−8	41	32	−9

Table 5
Model performance and fairness analysis across different demographic subgroups.

Subgroup category	Subgroup name	Sample size	RMSE (points)	MAE (points)	Difference from the benchmark group	Significance (p value)
Gender	Male	468	4.82	3.72	-	-
	Female	432	4.78	3.68	−0.04	>0.05
Nationality	United States (benchmark)	187	4.75	3.65	-	-
	India	165	5.25	4.08	0.5	>0.05
	United Kingdom	143	4.55	3.52	−0.2	>0.05
	Russia	108	5.45	4.22	0.7	>0.05
	Nigeria	72	6.35	4.95	1.6	<0.05
Admission background	Clinical medicine	621	4.79	3.69	-	-
	Non-clinical medicine	279	5.18	4.02	0.39	>0.05

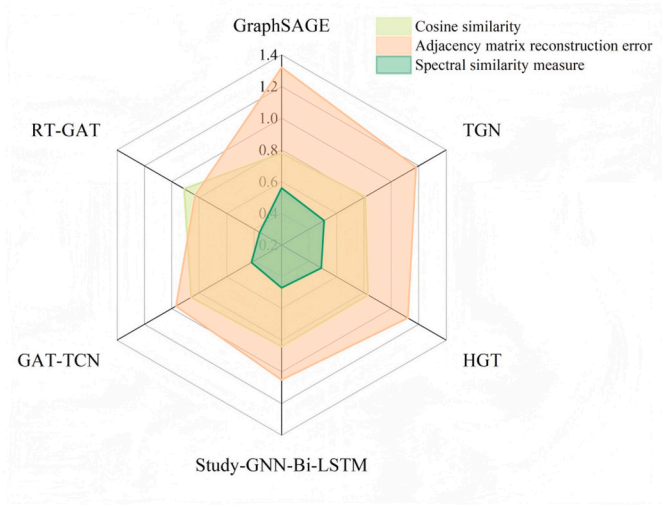

Fig. 4. Multi-dimensional correlation modeling quality assessment.

0.97/0.42. The overall trend shows that as the model moves from a static homogeneous graph to a heterogeneous graph, and then adds time series and multi-head mechanisms, the quality of multi-dimensional correlation modeling gradually improves.

GraphSAGE treats all capability dimension associations equally in the framework of homogeneous graphs, resulting in insufficient differentiation of composite paths, low similarity and high reconstruction errors. TGN introduces temporal dynamics, but does not assign differentiated weights to multiple types of edges such as “driving”, “collaborative”, and “feedback”, and can only improve the generalization of temporal evolution. HGT supports heterogeneous nodes and edges, but uses a unified adjacent attention and still cannot fully distinguish different capability interaction modes. Study-GNN-Bi-LSTM improves association reconstruction by fusing temporal information through Bi-LSTM; GAT-TCN achieves more refined structural encoding with the help of temporal convolution and attention. RT-GAT learns a dedicated weight vector for each edge type in relation-aware attention, and combined with learnable relation embedding, it can highlight the intrinsic interaction between different capability dimensions. It can also suppress irrelevant noise, making the association matrix output by the model significantly better than the above methods in terms of directional consistency and numerical accuracy.

The multi-head and temporal coding modules of RT-GAT also play a key role in preserving global topological features and capturing evolutionary trends. The multi-head mechanism parallelizes the learning of multi-angle graph subspaces, making it possible to take into account both short-term strong dependencies and long-term weak correlations, optimize the matching degree of spectral features, and significantly reduce spectral distances. Temporal coding embeds the position information of the five teaching stages into the node features, so that the reconstructed matrix reflects the static structure and also reflects the order of ability evolution, providing a richer dynamic context for the overall spectrum distribution.

It should be noted that the cosine similarity, Frobenius norm, and spectral distance used are primarily used to measure the model's numerical consistency in reconstructing predefined adjacency relationships. These metrics reflect the model's learning ability regarding graph structures rather than direct educational outcomes. Since adjacency relationships are constructed based on expert experience, their direct correlation with actual educational quality remains somewhat subjective. Therefore, these results should be understood as reference indicators for the model's structural characterization, rather than absolute evaluations of ability improvement or teaching effectiveness.

Metrics such as cosine similarity, Frobenius norm, and spectral distance are primarily used to evaluate a model's ability to learn predefined graph structures rather than directly measuring educational outcomes. These indicators serve to validate whether the model effectively captures the correlation patterns between competency dimensions defined by experts. A cosine similarity of 0.91 indicates that the reconstructed association matrix aligns highly with the expert-defined structure in terms of directionality, while the Frobenius norm of 0.83 reflects high numerical precision. The spectral distance of 0.36 demonstrates good preservation of global topological features. However, since adjacency matrices inherently incorporate expert knowledge and co-occurrence statistics—introducing inherent subjectivity—these metrics should not be interpreted as absolute measurements of “real” educational correlations but rather as relative validations of the model's structural learning capabilities. In other words, the RT-GAT's advantages in correlation

modeling demonstrate its ability to faithfully learn predefined educational relationship frameworks, laying the groundwork for subsequent interpretive analyses based on these relationships. Ultimately, educational significance must be assessed through multidimensional evidence including attention weight distribution, prediction accuracy, and validation through teaching case studies.

5.3. Relationship attention weight distribution and edge contribution

In order to deeply analyze the role of different types of relationships in the information aggregation of the RT-GAT model, this paper statistically analyzes the attention weight distribution and standard deviation of the three types of relationships: driving, collaborative, and feedback, and calculates the contribution ratio of each type of relationship to the final prediction result. The results are shown in Fig. 5.

In the attention allocation of the RT-GAT model in Fig. 5 (a), the average attention weight of the driving relationship reaches about 0.52, which is significantly higher than the collaborative relationship of about 0.31 and the feedback relationship of about 0.17, indicating that in the node information aggregation stage. The model focuses more on the direct causal driving path. The standard deviation of the driving type weight is 0.08, slightly higher than the 0.07 of the collaborative type and the 0.05 of the feedback type, which means that the dependence of different nodes on the driving type relationship fluctuates more, and the attention of the collaborative and feedback type relationships is more stable.

From the final predicted contribution of Fig. 5 (b), it is found that the driving edge contribution is about 54.3%, accounting for more than half of the total contribution; the collaborative edge contribution is close to 29.7%, and the feedback edge contributes 16%. This distribution is consistent with the trend of attention weights. The driving path plays the most critical role in the overall ability association modeling and prediction, while the feedback path is relatively minor.This evolutionary model aligns closely with the phased characteristics of medical education: In early stages, students continuously adjust learning strategies through feedback, while in later stages, knowledge accumulation becomes the primary driver of clinical practice. By visualizing the phased changes in attention weight distribution, educators can intuitively observe dynamic adjustments in influence pathways across competency dimensions. This approach enables a deeper understanding of skill evolution patterns rather than treating assessment results as isolated static snapshots.

The driving relationship represents the causal influence of ability and contains more direct and linear association signals. During the training process, the model optimizes the driving weights through error back propagation, thereby improving the prediction accuracy, giving it a higher attention score and greater contribution. The collaborative and feedback relationships reflect parallel collaboration and two-way regulation. The information transmission path is more indirect, and the model is relatively cautious in weighting it during aggregation, resulting in a decrease in attention and contribution.

From the perspective of capability indicators, core capabilities such as knowledge and skills and clinical practice often have an obvious "driving-driven" relationship, such as theoretical knowledge directly promoting the improvement of practical operations. This path information appears frequently in the graph structure as a driving edge, further strengthening the model's learning of this type of edge. Communication skills and teamwork need to be improved through multi-party interaction and iterative feedback. The paths are mostly collaborative and feedback-based. The information intensity and stability are lower than causal drive, which enables the model to automatically adjust the weights to reduce noise aggregation. From the perspective of network topology, the connection density and centrality of driving edges are often higher. Highly influential "hub" edges receive more frequent activations and stronger gradient updates in multi-head attention aggregation, further promoting a significant increase in their weight and contribution.

The study revealed that students with slow improvement in clinical practice scores exhibited lower attention weight values on the "knowledge skills → clinical practice" driving edge compared to grade averages, indicating potential barriers to competency development along this low-weight dimension. Based on these findings, the research established an early warning rule: When a student's attention weight on core driving edges (e.g., "knowledge skills → clinical practice") falls below twice the grade average standard deviation, coupled with declining performance predictions in this dimension, the system automatically flags them as "students requiring immediate support." The model-generated time-evolution trend charts demonstrate students' competency development trajectories, enabling educators to assess intervention urgency and relevance. For instance, when a student's communication skills stagnated in the second semester and teamwork scores declined, teachers could implement cross-cultural communication training at the beginning of the third semester. Regarding interpretable outputs and teacher engagement evaluation, 12 clinical instructors participated in a four-week RT-GAT trial, accessing visual dashboards, warning lists, and completing System Usability Scale (SUS)

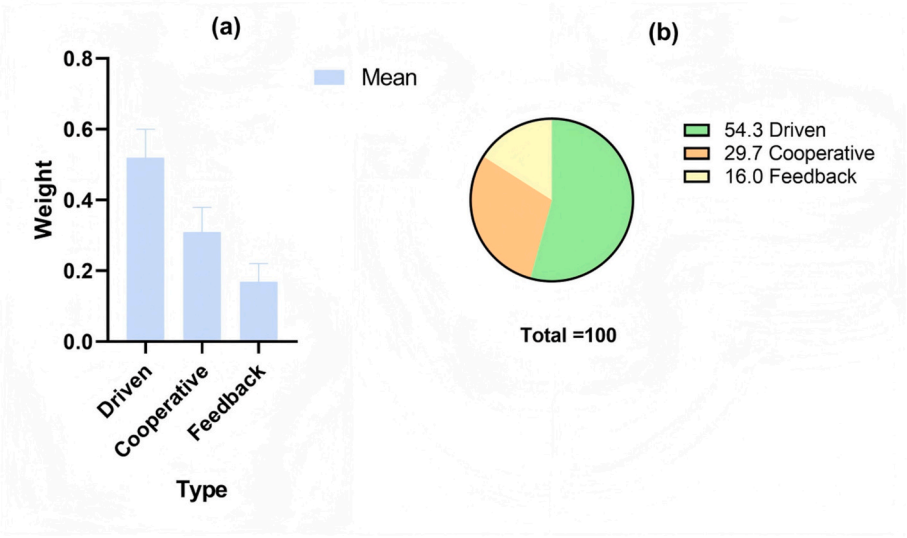

Fig. 5. Relationship attention weight distribution and edge contribution.

assessments alongside semi-structured interviews. The SUS average score of 76.3 exceeded industry benchmarks, reflecting educators' positive perception of model output comprehensibility. Interviews revealed that 85% identified relational attention weight distributions as critical for understanding competency development pathways, while 78% acknowledged time-evolution trend charts effectively visualized phased competency progression. However, 70% of teachers expressed desire for system feedback mechanisms to incorporate teaching insights for model refinement.

5.4. Capturing ability of temporal trends

In order to evaluate the model's ability to capture the trend of ability indicators evolving with the teaching stage, this paper uses the Pearson correlation coefficient to compare and analyze the performance of RT-GAT and the control model R-GAT on the four ability dimensions in five consecutive teaching stages. The results are shown in Fig. 6.

In Fig. 6 (a), the Pearson correlation coefficients of RT-GAT for the four ability dimensions in the five teaching stages show a continuous upward trend, indicating that its ability to capture the trend of ability evolution is gradually strengthened. In the first stage, the correlation coefficients of RT-GAT for knowledge and skills were about 0.78, for clinical practice was about 0.72, for communication skills was about 0.68, and for teamwork was about 0.65. By the fifth stage, they had increased to about 0.90, 0.83, 0.77, and 0.75, respectively, highlighting the accurate fit of the model to the changes in ability in the later stages.

In Fig. 6 (b), R-GAT is significantly inferior in capturing the ability trend in the same five stages. Its Pearson correlation coefficients are approximately 0.65, 0.60, 0.58, and 0.55 in the first stage, and only slightly increase to approximately 0.70, 0.65, 0.63, and 0.60 in the fifth stage. The rising slopes in each stage are significantly lower than those of RT-GAT, indicating the lack of a sufficient temporal information fusion mechanism. In summary, the Pearson correlation coefficient after considering the time relationship increases by an average of 0.168.

RT-GAT's temporal information encoding module uses sine-cosine position encoding and bidirectional sequence modeling to give each time step a unique temporal vector, so that the model can effectively aggregate historical and future contextual information in combination with the temporal sequence at each stage. At the same time, multi-head attention learns short-term and long-term dependencies in parallel, allowing relational attention to be dynamically adjusted at different stages, and can capture the nonlinear evolution of capabilities at each stage. Although R-GAT supports the distinction of relationship types, it lacks a dedicated temporal coding device, and its node aggregation relies more on static graph structures, resulting in insufficient sensitivity to continuous trends.

From the perspective of ability indicators, core abilities such as knowledge and skills and clinical practice tend to show an accelerated improvement trend in the advancement of teaching. RT-GAT retains the mixed characteristics of early knowledge foundation and later practice feedback between stages through temporal coding, which promotes the continuous growth of high correlation coefficients. For soft skills such as communication and teamwork that require multiple rounds of interaction and feedback, the model relies on bidirectional LSTM and attention weights to capture potential iterative improvement paths in each point in time, making its correlation curve rise smoothly. R-GAT failed to structurally characterize these temporal dependencies, resulting in a long-term stagnation in the matching degree between the prediction and the actual trend, and ultimately showing a significant disadvantage in the Pearson correlation coefficient.

5.5. Real-time computing efficiency analysis

In order to evaluate the computing performance of each model in practical applications, this paper compares the inference time, training time, throughput and resource occupancy of different models. The results are shown in Table 6.

In Tables 6 and in the inference phase, GraphSAGE is the fastest at about 12.4ms per batch, while RT-GAT is about 14.6ms, which is only

Table 6
Real-time computing efficiency analysis.

Model	Model inference time (ms/batch)	Training time (min/epoch)	Throughput (samples/s)	Resource utilization (%)
GraphSAGE	12.4	28.7	540	45.3
TGN	15.8	32.1	480	50.7
HGT	18.3	35.6	440	55.2
Study-GNN-Bi-LSTM	21.7	38.4	410	60.5
GAT-TCN	19.2	36.2	430	57.8
RT-GAT	14.6	30.5	520	48.1

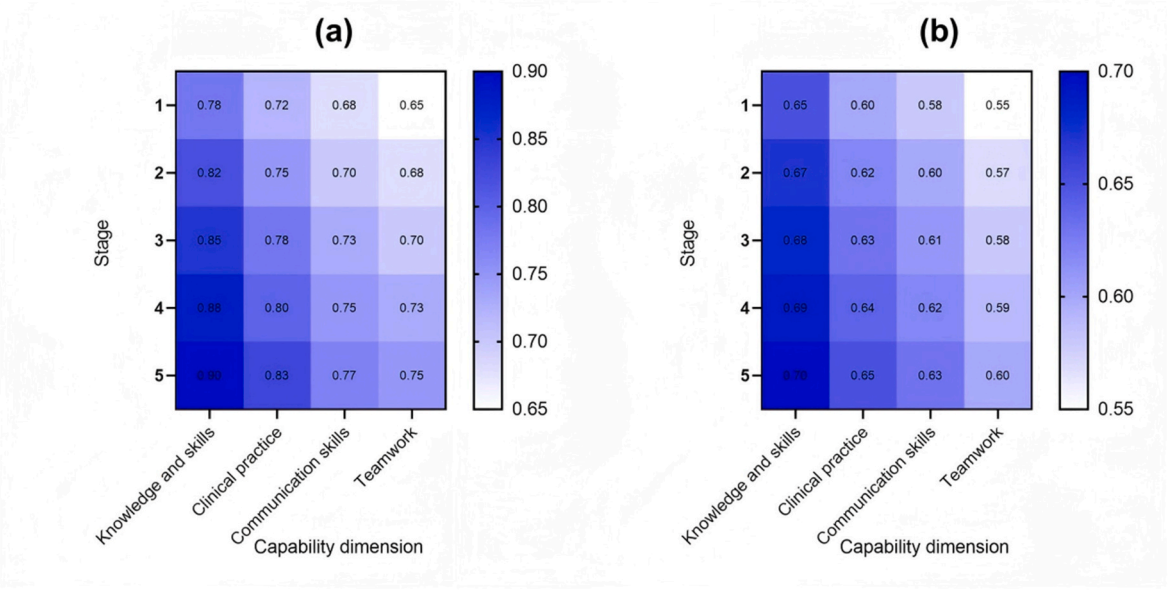

Fig. 6. Comparison of temporal trend capture ability.

2.2ms longer than GraphSAGE. TGN, HGT, GAT-TCN, and Study-GNN-Bi-LSTM require about 15.8, 18.3, 19.2, and 21.7ms, respectively. In terms of training time, GraphSAGE is the fastest at about 28.7 min per round, RT-GAT is about 30.5 min, TGN is about 32.1 min, GAT-TCN is about 36.2 min, HGT is about 35.6 min, and Study-GNN-Bi-LSTM is about 38.4 min. It can be seen that after introducing relational attention and temporal encoding, RT-GAT maintains a high operating efficiency in the reasoning and training stages.

In the throughput test, GraphSAGE reaches a maximum of about 540 samples per second, and RT-GAT reaches about 520 samples per second. TGN, HGT, GAT-TCN, and Study-GNN-Bi-LSTM are 480, 440, 430, and 410 samples per second, respectively. In terms of resource utilization, GraphSAGE occupies about 45.3%, RT-GAT about 48.1%; TGN, HGT, GAT-TCN, and Study-GNN-Bi-LSTM are 50.7%, 55.2%, 57.8%, and 60.5%, respectively. Overall, RT-GAT maintains a near-maximum throughput while using only about 48% of resources, which is better than most complex models.

RT-GAT can achieve efficient computation with low time and resource overhead, mainly due to its relational attention weight sharing and multi-head parallel mechanism. In relational attention calculation, each head shares the basic weight matrix and only fine-tunes the bias. This design greatly reduces the number of matrix multiplications during training and inference. HGT needs to maintain all parameters for each relationship type, and Study-GNN-Bi-LSTM relies on the sequence of bidirectional LSTM for calculation. TGN dynamically samples time-stamp neighbors, and RT-GAT reduces the number of parameters and data movement through reusable mapping, effectively compressing the computational burden. The temporal encoding module of RT-GAT combines positional encoding with lightweight bidirectional LSTM, and performs sequence modeling only at the node level, avoiding the additional convolution overhead of GAT-TCN switching between graph convolution and temporal convolution. In multi-head attention, all heads can be executed in parallel and utilize the parallel stream processing of modern GPUs. RT-GAT has a more balanced memory access and computational throughput during batch inference, achieving an excellent compromise between GPU resource utilization and sample processing speed. This design allows RT-GAT to maintain high prediction performance while also taking into account the strict requirements for real-time performance and resource efficiency.

5.6. Robustness testing in extreme scenarios with small samples and noise interference

In order to verify the stability and robustness of the RT-GAT model in an environment with scarce data and noise interference, this paper designed two sets of experiments: one is an extremely small sample test that gradually reduces the size of the training data; the other is to add Gaussian noise interference of different intensities to the input features. The experiment evaluates the attenuation of the model performance by observing the changes in the RMSE and MAE indicators. The results are shown in Fig. 7.

Steps: (1) Subsets can be extracted at a ratio of 30%, 25%, 20%, 15%, 10%, and 5% of the original training set size for model training and testing, and the error indicators can be recorded. (2) While maintaining the original data size, Gaussian noise with standard deviation $\sigma = 1, 3, 5, 8$, and 10 can be added to the input features, and the training and testing process can be repeated to evaluate the model's tolerance to noise.

In the experiment of gradually reducing the size of training data in Fig. 7 (a), when only 30% of the original data is used, the RMSE of RT-GAT is about 5.3 points and the MAE is about 4.0 points. When reduced to 20%, the errors rose to 6.1 and 4.7 points respectively; when the training data was only 10%, the RMSE was close to 7.6 points and the MAE was about 5.9 points. In the most extreme 5% scenario, the RMSE was close to 8.9 points and the MAE was about 6.8 points. It can be seen that as the size of the training sample decreases, the model performance gradually degrades, but even with only 5% of the data, the RMSE and MAE are still controlled within 9 points and 7 points respectively, showing good small sample robustness.

In the test of adding different intensities of Gaussian noise to the input features in Fig. 7 (b), when the noise standard deviation is 1, the RMSE of RT-GAT is about 4.9 points and the MAE is about 3.8 points. When the standard deviation increases to 5, the RMSE is about 6.4 points and the MAE is about 4.9 points; when the maximum noise level is 10, the RMSE is about 8.1 points and the MAE is about 6.1 points. Overall, when the noise intensity increases, the model error rises steadily. When the noise is at its maximum, the RMSE and MAE are still within 9 and 7 points, respectively, indicating that RT-GAT has a high tolerance for input feature noise.

The stability of RT-GAT in small sample scenarios stems from the synergy of relational attention and multi-head mechanism. The relational attention module can dynamically allocate weights of different relation types, so that the key "driving" path can still be used

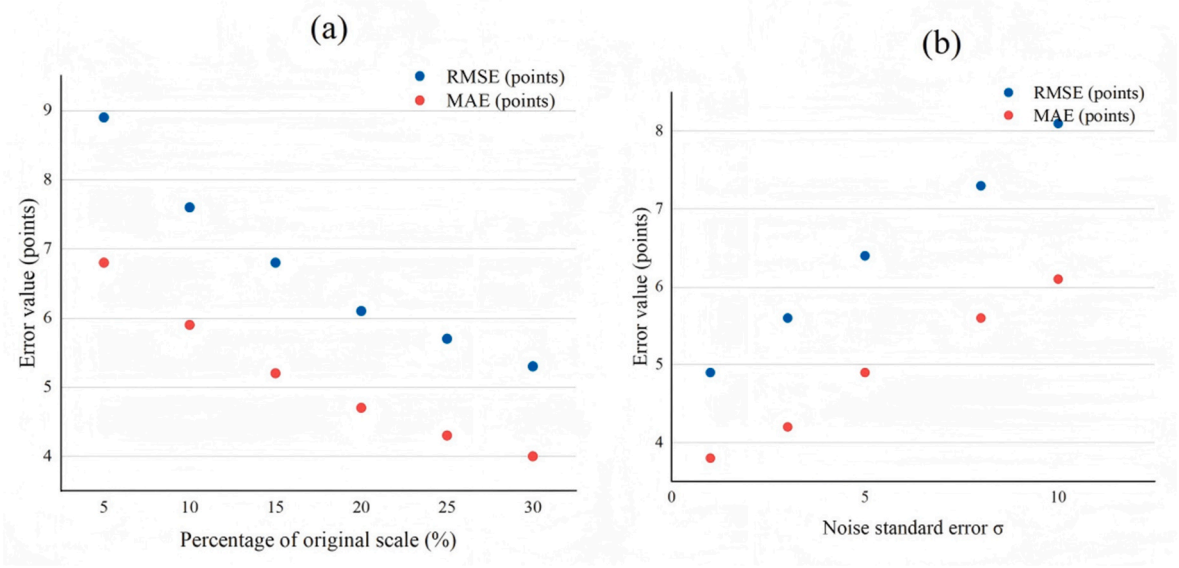

Fig. 7. Robustness test in extreme scenarios with small samples and under noise interference.

preferentially when data is scarce. Multi-head aggregation captures multi-angle subspace information in parallel. Even if some heads are degraded due to insufficient data, other heads can still provide compensatory features to reduce the overall performance decline. The linear transformation of node features and residual connections in the early layers also provide guarantees for gradient propagation, avoiding gradient vanishing when there are very few training samples. The model can maintain an acceptable error level even with 5% samples.

For Gaussian noise interference experiments, the noise robustness of RT-GAT is mainly due to the temporal information encoding and regularization strategy. The temporal encoding module weights and aggregates the historical features of multiple stages of the node so that the noise at a single time point does not drastically affect the overall temporal representation. L2 regularization and Dropout suppress the noise amplification effect during training and limit the accumulation of noise in the multi-layer attention network. Multi-head attention also provides natural noise smoothing capabilities: different heads respond differently to the same input noise, and the stable output of the majority head is used as the standard. When the noise standard deviation increases to 10, the RMSE and MAE are still controlled at around 8 points and 6 points, showing excellent robustness to noise interference.

5.7. Verification of cross-school/national generalization ability

In order to evaluate the generalization ability of the RT-GAT model in different institutions and international student contexts, this paper selected five high-level Chinese medical schools that did not appear in the training dataset (Peking University School of Medicine, Fudan University School of Medicine, Shanghai Jiao Tong University School of Medicine, Harbin Medical University, Wuhan University School of Medicine) and five major source countries of international students (India, Nigeria, the United States, Russia, and the United Kingdom) for migration verification experiments. The results are shown in Fig. 8. The experimental steps are as follows: (1) Keeping the model structure and parameters unchanged, the trained RT-GAT model is inferred on new sample subsets of institutions and nationalities; (2) The prediction error is evaluated using three indicators, RMSE, MAE, and MAPE, to examine the cross-domain generalization ability of the model; (3) Each subset contains no less than 100 test samples to ensure statistical significance.

In Fig. 8 (a), among the five top medical schools that did not participate in the training, the prediction error of RT-GAT remained within an acceptable range. At Shanghai Jiao Tong University School of Medicine, the RMSE was about 4.9 points and the MAE was about 3.8 points, the lowest. Peking University School of Medicine and Wuhan University School of Medicine were 5.3 points and 4.1 points, 5.6 points and 4.2 points respectively. Fudan University School of Medicine and Harbin Medical University had higher errors, with Fudan about 5.7 points and 4.5 points, and Harbin about 6.2 points and 4.9 points. Overall, the RMSE of each school fluctuates from 4.9 to 6.2 points, and the MAE fluctuates from 3.8 to 4.9 points, indicating that the model can maintain a relatively robust prediction performance on the test samples of different schools.

In the test set of the five major countries of origin of international students in Fig. 8 (b), RT-GAT also shows good generalization ability. For American students, the RMSE is about 4.8 points and the MAE is about 3.7 points; for Indian students, it is about 5.1 points and 4.0 points; for British students, it is about 5.0 points and 3.9 points; for Russian and Nigerian students, it is slightly higher, about 5.9 points and 4.6 points for Russian, and about 6.4 points and 5.1 points for Nigeria. The RMSE ranges from 4.8 to 6.4 points and the MAE ranges from 3.7 to 5.1 points, indicating that the model is still highly adaptable in different cultural and educational backgrounds.

In cross-school tests, there are differences in the syllabus, assessment focus and scoring standards of different medical schools. The clinical training and teamwork evaluation of Harbin Medical University and Fudan University School of Medicine focused more on practical operation details and cross-disciplinary collaboration, which made it difficult to fully match the relationship weights and time series patterns in the original training set, resulting in relatively higher RMSE and MAE. The evaluation indicators of Shanghai Jiao Tong University School of Medicine are closer to those of the training institutions, minimizing the model migration error. The relationship-aware attention mechanism of RT-GAT enhances the expression of heterogeneous associations, but more data fine-tuning is needed to further reduce the bias when facing new scoring rules.

The fluctuation of cross-national generalization error mainly comes from the impact of differences in cultural, linguistic and educational backgrounds of international students on the distribution of evaluation data. Nigerian international students may have more discrete evaluation distribution in communication ability and clinical practice due to language barriers and differences in clinical experience systems, which increases the difficulty of prediction and makes RMSE and MAE at the highest level. In the US and UK student groups, the evaluation patterns are similar to those of most countries in the training data, and temporal coding and relational attention can better capture the evolution of their abilities. RT-GAT uses learnable relational embedding and temporal dynamic weights to make the model have lower errors for groups with similar educational backgrounds, but in very different cultural scenarios, it still needs to be fine-tuned in the target domain or a small amount of labeled data to enhance generalization performance.

While cross-institutional/cross-national assessments like RMSE and MAE provide quantitative indicators of predictive accuracy, the score

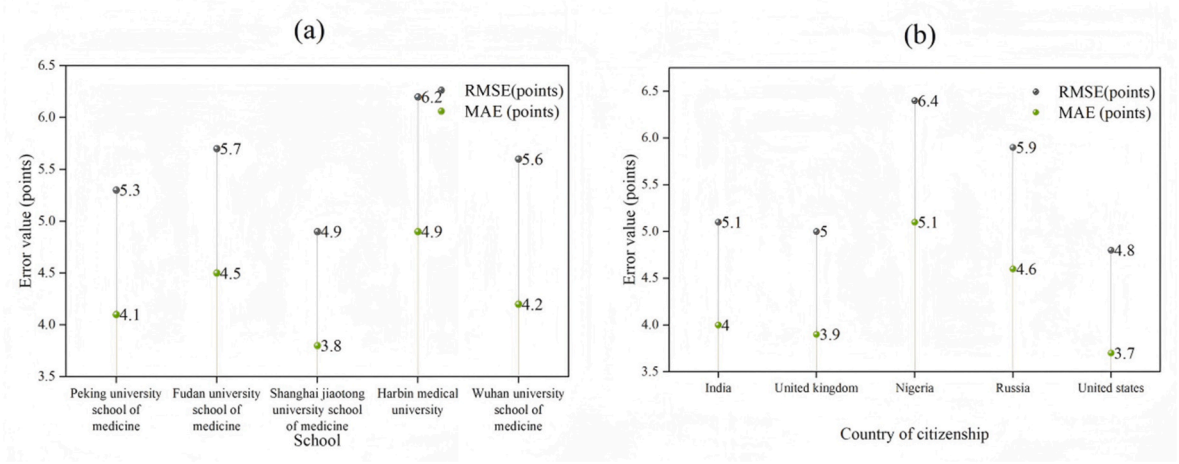

Fig. 8. Results of cross-school/nationality generalization ability verification.

fluctuations corresponding to these errors are typically within 5 points. Compared to a 100-point scale for ability assessment, their impact on students' overall learning outcomes is limited. Model prediction biases mainly manifest in scoring details or stage-specific differences, and do not substantially mislead assessments of overall ability development trends or teaching intervention recommendations. Therefore, even with some error, they can still provide reliable multi-dimensional ability references and decision support for institutions and teachers.

In cross-cultural and cross-institutional testing, the reported RMSE/MAE fluctuations ranged between 4.8 and 6.4 points. Under a percentage-based evaluation system, these variations have relatively limited impact on students 'final learning outcomes. Taking the high-risk student alert task as an example, with a 60-point threshold, Nigerian students exhibited an RMSE of 1.6 points higher than the baseline, indicating an average prediction-deviation of approximately 1.6 points from actual scores. While this increases the risk of false positives and false negatives under percentage-based assessment, the probability remains manageable. Among Nigerian students, RT-GAT's high-risk student alert accuracy rate, though lower than that of American student groups, remains relatively high, identifying nearly 80% of genuine high-risk cases. From an educational intervention perspective, both misidentifying false positives and failing to recognize true high-risk students pose challenges, with RT-GAT demonstrating superior performance compared to GAT-TCN. These findings indicate that while cross-cultural testing errors exhibit variability, their practical implications are limited, and the model maintains stable generalization applicability.

5.8. Ablation experiment

To further verify the contribution of each key module in the RT-GAT model to performance, this paper designs an ablation experiment. By successively eliminating the relational attention mechanism, temporal coding (position coding), temporal coding (Bi-LSTM), and self-attention mechanism, different variant models are constructed to evaluate the modeling ability and prediction performance of the model. The results are shown in Table 7.

The ablation experiment steps are as follows: (1) Under the premise of keeping the training set and parameters consistent, four ablation versions are constructed respectively: removing the relational attention module (R-), further removing the position encoding (R-P-), further removing the Bi-LSTM encoding (R-P-B-), and finally completely removing all enhancement structures (R-P-B-S-); (2) The multi-dimensional capability modeling effect (cosine similarity) and prediction performance (RMSE and MAPE) of each model can be evaluated on the complete test set; (3) The impact of different modules on the overall performance is analyzed to reveal the role of key mechanisms in the characterization and prediction of capability associations.

In Table 7, the full RT-GAT model achieves a cosine similarity of about 0.91 in relational characterization, an RMSE of about 4.8 points, and a MAPE of about 6.1%. When the relational attention module is removed (retaining position encoding, Bi-LSTM, and self-attention), the cosine similarity drops to about 0.86, the RMSE increases to about 5.4 points, and the MAPE increases to about 6.9%. When the position encoding is further removed (only Bi-LSTM and self-attention are retained), the cosine similarity drops to about 0.82, the RMSE increases to about 5.9 points, and the MAPE increases to about 7.4%; when the Bi-LSTM is removed (only self-attention is retained), the similarity further drops to about 0.77, the RMSE increases to about 6.3 points, and the MAPE increases to about 8%; in the basic GNN version where all enhanced structures are removed, the cosine similarity is as low as about 0.71, the RMSE is as high as about 6.9 points, and the MAPE is as high as about 8.8%. It can be seen that the addition of each module has made significant contributions to the reconstruction of associations and the improvement of prediction errors.

The removal of the relational attention module caused the similarity to drop sharply from 0.91 to 0.86, and the RMSE and MAPE increased by about 0.6 points and 0.8% respectively, indicating that this module played a key role in distinguishing heterogeneous edge types such as driving, synergy, and feedback. The reason is that when there is no relationship perception, the model can only process all edges with a uniform weight, unable to give priority to high-value causal-driven paths, and difficult to suppress low-value feedback noise, and inaccurately characterizes the intensity of the impact between the ability dimensions, which leads to obvious deviations in the direction and value of the reconstructed matrix. Temporal coding (position coding and Bi-LSTM) and self-attention each have a decreasing effect on performance. After the position coding is removed, the similarity is further reduced from 0.86 to 0.82. It can be seen that the position coding introduces the order information of the teaching stage, which helps the node features obtain the correct temporal position signals at different stages. After removing Bi-LSTM, the similarity dropped to 0.77 and the RMSE increased to 6.3 points, indicating that Bi-LSTM is crucial in capturing the long-term dependence of dynamic evolution of ability over time. Finally, after completely removing the self-attention mechanism, the performance dropped to the worst, which reflects the core role of self-attention in the overall multi-head aggregation and information reorganization process of the model. In summary, each module is closely linked from structural heterogeneity characterization to time series evolution capture to multi-head information fusion, and together constructs RT-GAT's high-precision correlation modeling and prediction capabilities.

5.9. Hyperparameter sensitivity analysis

While maintaining the original training/validation partition and model structure, univariate sensitivity tests were performed on key hyperparameters (learning rate, number of attention heads, Dropout, Bi-LSTM hidden dimension, etc.): only one hyperparameter was changed each time, while the other parameters remained at their default values in Table 1; each configuration was trained three times to reduce the impact of randomness, and the average performance at early stopping on the validation set was used as the evaluation metric; the cosine similarity, RMSE, MAE, and the number of rounds in which the model triggered early stopping on the validation set were recorded for each configuration. The results are shown in Table 8 below.

As shown in Table 8, the performance of RT-GAT is most sensitive to the learning rate and regularization strength. When the learning rate decreases from 0.001 to 0.0005 (Experiment B), the model is stable but convergence slows significantly (110 epochs), and the RMSE increases slightly to 4.9; while increasing it to 0.005 (Experiment C) leads to

Table 7
Ablation experiment results.

Target model	Relational attention	Temporal coding (positional coding)	Temporal coding (Bi-LSTM)	Self-attention mechanism	Cosine similarity	RMSE (score)	MAPE (%)
RT-GAT	✓	✓	✓	✓	**0.91**	**4.8**	**6.1**
	-	✓	✓	✓	0.86	5.4	6.9
	-	-	✓	✓	0.82	5.9	7.4
	-	-	-	✓	0.77	6.3	8
	-	-	-	-	0.71	6.9	8.8

Table 8
Hyperparameter sensitivity analysis.

Experiment No.	Variables (relative to default)	Cosine Similarity	RMSE (score)	MAE (score)	Number of Convergence Cycles (Average)
A (Baseline)	LR = 0.001; heads = 8; dropout = 0.5; Bi-LSTM = 32	**0.91**	**4.8**	**3.7**	**48**
B	Learning rate ↓ (0.0005)	0.906	4.9	3.75	110
C	Learning rate ↑ (0.005)	0.880	5.4	4.15	34
D	Number of attention heads ↓ (4)	0.890	5.05	3.95	60
E	Number of attention heads ↑ (16)	0.915	4.70	3.65	56
F	Dropout ↓ (0.2)	0.902	4.86	3.78	50
G	Dropout ↑ (0.7)	0.872	5.45	4.25	70
H	Bi-LSTM hidden dimensions ↓ (16)	0.887	5.20	4.05	62

training instability, with the RMSE increasing to 5.4, the MAE increasing to 4.15, and the cosine similarity decreasing to 0.880. Increasing the number of attention heads can improve the feature aggregation effect. When the number of attention heads increases from 4 (Experiment D) to 16 (Experiment E), the RMSE decreases from 5.05 to 4.70, and the cosine similarity increases to 0.915. The changes in Dropout show that too weak regularization (0.2, Experiment F) slightly reduces performance, while too strong regularization (0.7, Experiment G) leads to significant degradation (RMSE = 5.45, MAE = 4.25), indicating that 0.5 is the balance point. The reduction in hidden dimensions of Bi-LSTM (Experiment H) also weakens the ability to model temporal dynamics, causing the RMSE to rise to 5.20. Overall, the default configuration (A) achieves the best balance between convergence speed (48 rounds) and prediction accuracy, and multi-head attention and moderate-intensity Dropout have the most significant performance improvement effects.

Under the default hyperparameter configuration (learning rate 0.001, attention head count 8, dropout rate 0.5), RT-GAT achieved optimal performance on the validation set after approximately 48 epochs of training, after which the early stopping mechanism triggered training termination. The loss curve demonstrates synchronized declines in training and validation losses with a consistent gap maintained below 0.2, showing no significant signs of overfitting. Convergence behavior varied across learning rates: At 0.0005 learning rate, convergence significantly slowed, requiring 110 epochs to stabilize; while at 0.005 learning rate, convergence accelerated to 34 epochs but resulted in volatile validation losses and an elevated RMSE of 5.4, indicating unstable convergence. Attention head count had minimal impact on convergence stability, with head counts of 4, 8, and 16 achieving convergence at 60,48, and 56 epochs respectively, showing no statistically significant differences. Dropout rate significantly influenced convergence stability: At 0.2 dropout rate, training loss declined rapidly but validation loss exhibited oscillatory increases, indicating overfitting; at 0.7 dropout rate, convergence slowed with subsequent performance degradation; the 0.5 dropout rate achieved optimal balance between convergence speed and generalization performance. These analyses demonstrate that the default hyperparameter configuration achieves optimal equilibrium between convergence speed and prediction accuracy.

5.10. Teaching application case analysis

Three international medical students with typical characteristics were selected. Based on the relational weight distribution of the model output, the marginal contribution rate, and the time evolution trend, the model's way of identifying the key paths of student ability development was elaborated in detail, and personalized teaching intervention recommendations were generated accordingly. Qualitative feedback was collected through teacher interviews, confirming the practicality and acceptability of the model in teaching practice.Case 1 is a student with solid theoretical knowledge but weak clinical operation ability. His score in the knowledge and skills dimension ranks in the top 15% of the grade, but the clinical practice dimension continues to be below average.Analyzing the relational attention weight of the output of the RT-GAT model, it is found that the driving edge weight of “knowledge skills →clinical practice” is lower than the average of the same level, indicating that the transformation path is blocked.The model time coding module also captures that the improvement of its clinical practice ability lags behind the accumulation of knowledge and skills.Based on this, the system generates recommendations for enhanced simulation operation training, recommends participating in key project training and adds practical guidance.In the semester after adopting the recommendation, the student's clinical practice score increased by 12.3%, and the driving edge weight returned to normal.

Case 2 focuses on a student whose teamwork performance is affected by poor communication skills.The student scored low in the standardized patient interview assessment, insufficient cross-cultural communication and response, and the mutual evaluation of group tasks and teamwork was also at 20% after grade level.Model analysis shows that the contribution rate of ”communication →team” collaboration is much lower than the grade average, indicating that communication ability restricts teamwork.Two-way LSTM modeling revealed that the improvement of communication skills has slowed since the second semester, and the score of teamwork has also stagnated.Based on this, the system recommends that they participate in cross-cultural communication workshops and assume more interactive roles in multicultural groups.After one semester of intervention, the student's communication ability score increased by 9.8%, the teamwork score increased by 11.4%, and the contribution rate of collaboration rebounded, indicating that the intervention measures were effective.

Case 3 presents a student who has performed well in clinical practice and can use practical feedback to promote knowledge consolidation.The student scored well in the clinical operation assessment, but the theoretical knowledge examination fluctuated in stages.The model captures the third to fourth teaching stages, and the weight of the ”practice →knowledge” feedback edge rises from 0.12 to 0.29, indicating that clinical practice has an enhanced reverse regulatory effect on its knowledge consolidation.The analysis of the time coding module found that after the student completed the high-intensity clinical rotation, the accuracy rate of clinical-related question types in the follow-up theoretical knowledge examination was significantly improved, confirming the core mechanism of “practical feedback to promote learning” in the formative assessment.The teacher team uses its practical feedback model as a teaching case, adds related explanations, and encourages practical reflection.In the follow-up, the physiological theory knowledge score increased by 7.6% in the fifth stage.This verifies the effectiveness of the RT-GAT model, demonstrates its transformation potential, and is expected to become an intelligent tool for personalized teaching and formative evaluation.

6. Experimental discussion

The RT-GAT model in this paper has achieved significant advantages in multi-dimensional association modeling and ability prediction, first of all due to its detailed description of heterogeneous relationships in the graph structure. Through the relationship-aware attention module, the

RT-GAT model can assign differentiated weights to the three types of edges: "driving", "cooperative" and "feedback", prioritize the aggregation of key causal driving paths and suppress low-value feedback noise. It obtains a cosine similarity of more than 90% in adjacency matrix reconstruction and correlation matrix directional consistency. The model also introduces a multi-head parallel mechanism in the node aggregation stage. Different attention heads learn in parallel in multiple subspaces, which can capture both short-term strong correlations and long-term weak dependencies, enhancing the diversity and robustness of the overall feature expression.

Another key factor is the systematic encoding and fusion of temporal information. This paper uses sine-cosine position encoding to inject the time sequence, and combines bidirectional LSTM to perform sequence modeling on the evaluation data at each time point, so that the model can effectively capture the dynamic evolution of ability with the five teaching stages. Compared with GNNs based only on static snapshots or hybrid architectures that rely only on graph convolution and time series convolution, the time series module of RT-GAT achieves a higher degree of temporal trend fit in terms of Pearson correlation coefficient, from 70% in the first stage to 90% in the fifth stage. It is this design that combines structural heterogeneity with temporal dynamics that makes RT-GAT outperform the control model in RMSE and MAE in all capability dimensions.

The T-GAT model achieves adaptability to heterogeneous educational backgrounds through a relational embedding sharing mechanism and temporal encoding module. The relational embedding sharing mechanism maps competency interaction patterns across different cultural contexts into a shared semantic space, utilizing a decomposition design of basic weight matrices and relationship-specific biases to preserve cross-cultural commonalities while accommodating regional specificity. The temporal encoding module employs sine position encoding and bidirectional LSTM to capture both common trends in competency evolution (e.g., progressive improvement in clinical practice with stage progression) and individual differences (e.g., varying development rhythms of communication skills across cultural contexts). Current RMSE fluctuation ranges of 4.9-6.2 points in cross-school tests and 4.8-6.4 points in international tests indicate that the model maintains relatively stable predictive performance across diverse scoring criteria and cultural backgrounds. Future research could explore federated learning frameworks to achieve collaborative optimization of multi-center models without sharing raw data, further enhancing cross-cultural generalization capabilities. It should be noted that current error ranges remain within acceptable thresholds in percentage-based evaluation systems, ensuring no substantial misguidance in assessing overall competency trends or generating instructional recommendations.

This study provides two aspects of technical innovation and practical significance for competency-based international medical education quality assessment. The study introduces a unified framework of multi-relational graph neural networks and temporal coding into the field of education assessment, and strengthens the interpretable modeling of interaction patterns between competency dimensions. Through empirical comparison and ablation analysis, the effectiveness of relational attention, temporal coding and multi-head mechanism in actual data scenarios is verified, providing education administrators and instructional designers with a data-driven multi-dimensional evaluation tool that helps early intervention, personalized tutoring and curriculum optimization.

This paper has some limitations, and can be improved in the following aspects in the future.

(1) Although the current multidimensional indicator system has covered core dimensions such as knowledge and skills, clinical practice, communication and collaboration, it is still insufficient in terms of comprehensive literacy, ethical judgment, multicultural communication and other abilities. In the future, more implicit ability indicators can be introduced, combined with multimodal data (such as video observation and voice recording) to improve the comprehensiveness of the assessment dimensions.
(2) The RT-GAT model constructed in this paper is effective in capturing multi-dimensional relationships and temporal dynamics, but it is still insufficient in modeling some weak relationships and cross-time dependency. In the future, the paper can introduce methods such as graph self-supervised learning and graph contrast learning, or explore the integration of graph neural networks and large models (Transformer) to improve the deep semantic expression ability of the model.
(3) The current model is mainly used for the prediction and analysis of education quality and has not yet been deeply embedded in education and teaching practice. In the future, the model should be integrated with the teaching evaluation system to build a teaching feedback loop based on the model prediction results, so as to achieve the reverse optimization of the evaluation results on the curriculum design, teaching methods and student training strategies.
(4) This study selected international medical students from three universities. Although the samples are representative, they are still limited in terms of country, cultural background, education stage, etc., which may affect the generalization ability of the model. Future research can expand the sample size and source, cover more countries and regions, and enhance the applicability and cross-cultural robustness of the model.

7. Conclusions

This paper adopts the RT-GAT model that integrates relation-aware attention and temporal coding mechanisms. By constructing a multi-relation heterogeneous graph and introducing driving, collaborative and feedback edges, it accurately models the deep connections between knowledge and skills, clinical practice, communication and collaboration. The dynamic modeling and prediction of the multi-dimensional ability evolution process in competency-based international medical education can be realized by combining sine-cosine coding and Bi-LSTM. Experimental results show that RT-GAT outperforms existing mainstream models in terms of neighbor reconstruction error (0.83), spectral distance (0.36), and ability prediction RMSE (about 4.8 points), significantly improving the accuracy of association modeling and prediction. This paper has made some achievements, but there are also some shortcomings. The current model still has shortcomings in weak relationship modeling, cross-cultural generalization, and teaching system integration. In the future, multimodal ability dimensions, graph self-supervised learning, and teaching feedback mechanisms can be introduced to further improve the practicality and adaptability of the model. Regarding attention weights, it should be noted that they reflect statistical associations rather than causal relationships, and direct inference of inevitable effects from educational interventions should be avoided. In terms of cross-group fairness, the RT-GAT model demonstrates balanced performance across genders (RMSE difference of 0.2 points, $p > 0.05$). On the nationality dimension, Nigerian students exhibited a 1.6-point higher RMSE ($p < 0.05$) due to limited sample size (8%), while other nationality differences showed no significant variation. This indicates that the model maintains fairness across most subgroups, but performance biases exist in undersampled groups. Future research should collect balanced data and incorporate fairness constraints to prevent technological applications from exacerbating educational inequalities.

CRediT authorship contribution statement

Ailing Yang: Writing – review & editing, Writing – original draft. **Ling Lin:** Formal analysis. **Hu Zhang:** Methodology. **Rong Su:** Investigation. **Yunfei Li:** Data curation.

Consent to publish

The manuscript has neither been previously published nor is under consideration by any other journal. The authors have all approved the content of the paper.

Ethic approval

Not applicable.

Data availability statement

The data supporting the findings of this study can be obtained from the corresponding author, upon request.

Funding

This work was supported by Educational and Teaching Research Project of Kunming Medical University (Grant No.: 2024-JY-Y-054).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

None.

References

Al-Azazi, F. A., Ghurab, M., & Ann-Lstm. (2023). A deep learning model for early student performance prediction in MOOC. *Heliyon, 9*(4), 1-16.
Alolga, R. N., Kassim, S. A., & Dramou, P. (2022). Study destination choice and career preferences of international students at China Pharmaceutical University. *Pharmacy, 10*(6), 149–163.
Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments, 31*(6), 3360–3379.
Baskar, B. S. V., & Kesavan, R. (2025). Assessments of student's adaptability using convoluted Geyser bidirectional long short-term memory in online education. *International Journal of Computational Intelligence Systems, 18*(1), 24–40.
Booth, G. J., Ross, B., Cronin, W. A., McElrath, A., Cyr, K. L., Hodgson, J. A., et al. (2023). Competency-based assessments: Leveraging artificial intelligence to predict subcompetency content. *Academic Medicine, 98*(4), 497–504.
Chen, Z., Cen, G., Wei, Y., & Li, Z. (2023). Student performance prediction approach based on educational data mining. *IEEE Access, 11*(1), 131260–131272.
Chen, Y., You, J., He, J., Lin, Y., Peng, Y., Wu, C., et al. (2023). SP-GNN: Learning structure and position information from graphs. *Neural Networks, 161*(1), 505–514.
Dabbagh, A., Madadi, F., & Larijani, B. (2024). Role of AI in competency-based medical education: Using EPA as the magicbox. *Archives of Iranian Medicine, 27*(11), 633–635.
Duan, J. (2023). Investigating the impact of cross-cultural adaptability on the academic and social experiences of international students in bioethics education. *Journal of Commercial Biotechnology, 28*(1), 124–130.
Fang, Y., Li, X., Ye, R., Tan, X., Zhao, P., & Wang, M. (2023). Relation-aware graph convolutional networks for multi-relational network alignment. *ACM Transactions on Intelligent Systems and Technology, 14*(2), 1–23.
Gauvin, N., & Gregory-Martin, K. (2025). The value of cross-cultural education in the clinical educator training process: A viewpoint. *Perspectives of the ASHA Special Interest Groups, 10*(1), 229–235.
Gupta, S. K., Srivastava, T., Gupta, S., & Shrivastava, S. T. (2024). Assessment in undergraduate competency-based medical education: A systematic review. *Cureus, 16*(4), 1–18.
Huang, Q., & Chen, J. (2024). Enhancing academic performance prediction with temporal graph networks for massive open online courses. *Journal of Big Data, 11*(1), 52–77.
Huang, Q., & Zeng, Y. (2024). Improving academic performance predictions with dual graph neural networks. *Complex & Intelligent Systems, 10*(3), 3557–3575.
Keriven, N., & Vaiter, S. (2023). What functions can Graph Neural Networks compute on random graphs? The role of Positional Encoding. *Advances in Neural Information Processing Systems, 36*(1), 11823–11849.
Kukkar, A., Mohana, R., Sharma, A., & Nayyar, A. (2024). A novel methodology using RNN+ LSTM+ ML for predicting student's academic performance. *Education and Information Technologies, 29*(11), 14365–14401.
Li, L., & Wang, Z. (2023). Calibrated q-matrix-enhanced deep knowledge tracing with relational attention mechanism. *Applied Sciences, 13*(4), 2541–2564.
Li, M., Wang, X., Wang, Y., Chen, Y., & Chen, Y. (2022). Study-GNN: A novel pipeline for student performance prediction based on multi-topology graph neural networks. *Sustainability, 14*(13), 7965–7979.
Mastour, H., Dehghani, T., Moradi, E., & Eslami, S. (2023). Early prediction of medical students' performance in high-stakes examinations using machine learning approaches. *Heliyon, 9*(7), 1-17.
Mehmood, F., Ahmad, S., & Whangbo, T. K. (2023). An efficient optimization technique for training deep neural networks. *Mathematics, 11*(6), 1360–1381.
Reyad, M., Sarhan, A. M., & Arafa, M. (2023). A modified Adam algorithm for deep neural network optimization. *Neural Computing & Applications, 35*(23), 17095–17112.
Rukadikar, C., Mali, S., Bajpai, R., Rukadikar, A., & Singh, A. K. (2022). A review on cultural competency in medical education. *Journal of Family Medicine and Primary Care, 11*(8), 4319–4329.
Shi, Y., Sun, F., Zuo, H., & Peng, F. (2023). Analysis of learning behavior characteristics and prediction of learning effect for improving college students' information literacy based on machine learning. *IEEE Access, 11*(1), 50447–50461.
Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting student performance in online learning: A multidimensional time-series data analysis approach. *Applied sciences, 14*(6), 2522–2537.
Soundariya, K., Nishanthi, A., Mahendran, R., & Vimal, M. (2025). Evaluation of Competency-Based Medical Education (CBME) curriculum implementation for Phase II Medical undergraduates: A qualitative study. *Journal of Advances in Medical Education & Professionalism, 13*(1), 36–48.
Tajvar, M., Ahmadizadeh, E., Sajadi, H. S., & Shaqura, I. I. (2024). Challenges facing international students at Iranian universities: A cross-sectional survey. *BMC Medical Education, 24*(1), 210–217.
Walkowska, A., Przymuszała, P., Marciniak-Stepak, P., Nowosadko, M., & Baum, E. (2023). Enhancing cross-cultural competence of medical and healthcare students with the use of simulated patients—a systematic review. *International Journal of Environmental Research and Public Health, 20*(3), 2505–2528.
Wang, S., Ni, L., Zhang, Z., Li, X., Zheng, X., & Liu, J. (2024). Multimodal prediction of student performance: A fusion of signed graph neural networks and large language models. *Pattern Recognition Letters, 181*(1), 1–8.
Wu, W., & Koh, A. (2022). Being "international" differently: A comparative study of transnational approaches to international schooling in China. *Educational Review, 74*(1), 57–75.
Yeom, J., Kim, T., Chang, R., & Song, K. (2024). Structural and positional ensembled encoding for Graph Transformer. *Pattern Recognition Letters, 183*(1), 104–110.
Yu, L., Sun, L., Du, B., Liu, C., Lv, W., & Xiong, H. (2022). Heterogeneous graph representation learning with relation awareness. *IEEE Transactions on Knowledge and Data Engineering, 35*(6), 5935–5947.
Zeng, P., Hu, G., Zhou, X., Li, S., & Liu, P. (2023). Seformer: A long sequence time-series forecasting model based on binary position encoding and information transfer regularization. *Applied Intelligence, 53*(12), 15747–15771.
Zhang, W., Hu, S., & Qu, K. (2023). Graph attention neural network model with behavior features for knowledge tracking. *IEEE Access, 11*(1), 88329–88338.
Zhang, Y., Sun, S., Ji, Y., & Li, Y. (2023). The consensus of global teaching evaluation systems under a sustainable development perspective. *Sustainability, 15*(1), 818–830.