Spatio-Temporal Fusion in Graph Neural Network for Streaming Knowledge Tracing

Yi-Fei Wen[1], Hang Liang[1], Carl Yang[2], Wen-Bo Xie[3], Yajun Du[1], Xianyong Li[1], and Yan-Li Lee[1✉]

[1] School of Computer and Software Engineering, Xihua University, Chengdu 610039, China
{wenyifei77, lianghang}@stu.xhu.edu.cn, {duyajun, lixy}@mail.xhu.edu.cn,
yanlicomplex@gmail.com✉
[2] Department of Computer Science, Emory University, Atlanta, Georgia, USA
j.carlyang@emory.edu
[3] School of Computer Science, Southwest Petroleum University, Chengdu 610500, China
wenboxie@swpu.edu.cn

Abstract. Knowledge tracing is an important research area in personalized education, aiming to predict students' future performance based on their historical interaction data. Existing knowledge tracing models face limitations in spatio-temporal information modeling. In the spatial dimension, the dynamic interactions between students and questions, along with peer effects among students, remain underexplored; In the temporal dimension, many existing works often ignore the modeling of student's learning rhythm and the spatial structure hidden in temporal information. To address the above issues, this work focuses on the realistic scenario of streaming data, provides an incremental **S**patio-**T**emporal Fusion **S**treaming **K**nowledge **T**racing model, referred to as **STSKT**. Specifically, we first incrementally construct a dynamic heterogeneous graph to capture complex spatial relationships, including dynamical student-question interaction and peer effects. We then enhance temporal modeling from two aspects: proposing a rhythm-aware representation for learning rhythms, and capturing the hidden spatial structure by transforming recent interaction sequences into structured graphs. Finally, a three-channel fusion strategy is designed to capture the local relevance, global disparity, and holistic information for prediction. Experiments on four real-world datasets show that STSKT outperforms ten baselines, achieving improvements of 8.52% in AUC and 12.80% in ACC. Further analysis demonstrates that our key designs in graph construction, temporal modeling, and fusion mechanism effectively enhance prediction performance.

Keywords: Knowledge tracing · Educational data mining · Graph Neural Networks · Dynamic graph

1 Introduction

Knowledge Tracing (KT), as a core support tool in personalized education, plays an important role in cognitive diagnosis and instructional feedback [1], and has attracted growing attention from both academia and industry. The core task of KT is to accurately model the evolving knowledge states of students. The evolution of students' knowledge states often exhibits significant spatio-temporal dependencies: in the spatial dimension, the evolution of knowledge states is not only affected by knowledge structure, but also driven by collaborative patterns shaped by peer effects among student

groups; In the temporal dimension, both the learning rhythm and the continuity of learning process affect the student's knowledge states. Meanwhile, students' knowledge states are also affected by the temporal proximity effect: intensive practice within a short period enhances retention, while prolonged absence of exposure to related content leads to forgetting. These factors jointly determine the dynamic changes in students' knowledge states and give rise to complex spatio-temporal interactions. While temporal and spatial signals offer valuable insights into students' learning processes, effectively fusing them remains a core challenge in accurately assessing knowledge states.

In recent years, numerous models have been proposed to capture the temporal and spatial signals in students' knowledge learning processes. Among them, the most representative approaches are sequential KT models and graph-based KT models. In sequential KT models, existing models mainly utilize Recurrent Neural Networks (RNNs) [2] and Transformer-based architectures [3, 4] to model temporal information in response sequences. Some models also employ self-attention mechanisms [5] to highlight critical historical interactions or introduce the Ebbinghaus forgetting curve [6–8] to mitigate performance degradation in long sequences. In graph-based KT models, various graph structures have been constructed to enrich representations of cognitive learning behavior and question structures, including skill dependency graphs [9], question-skill bipartite graphs [10], hierarchical exercise graphs [11], and dynamic question-answering graphs [12]. However, the above models have three main limitations: (i) In temporal representation, existing models often overlook students' learning rhythms and fail to capture the spatial structure embedded in temporal sequences; (ii) In modeling spatial structures, many models rely on static graphs, making it difficult to capture the dynamically evolving interactions between students and questions. Moreover, peer collaboration driven by the "peer effect" is often neglected; and (iii) Most existing KT models are not well-suited for streaming data scenarios, as they typically rely on a student's complete interaction history for modeling rather than dynamically updating with real-time data.

To address the above challenges, this paper proposes a Spatio-Temporal Streaming Evolving Streaming Knowledge Tracing model (STSKT for short), designed for streaming data scenarios. In terms of spatial topology modeling, STSKT considers the co-occurrence relationships of questions on skills to alleviate the challenges of sparse question representations at the knowledge level. By leveraging the temporal co-occurrence of students on the same questions, collaborative edges are constructed between students to address the challenge of capturing peer effects under limited data. Additionally, by modeling student-question interactions within a limited time window, response sequences are transformed into structured spatio-temporal graphs, which not only reduce the sparsity of dynamic graphs but also improve the efficiency of subsequent model training. In terms of temporal modeling, STSKT embeds the time intervals between student interactions in a periodic manner, enabling the model to capture the rhythm of student responses over time. Finally, a three-channel fusion strategy is proposed, incorporating local relevance, global disparity, and holistic information, to model the student's answering performance on a given question.

The main contributions of this paper are summarized as follows:

- We propose a spatio-temporal streaming KT framework for streaming data, which can dynamically capture and fuse the spatio-temporal information of student-question interactions.
- We propose a dynamic heterogeneous graph construction method to more accurately model the associations among questions, capture peer effects without rely-

ing on explicit labels, and characterize the structural information embedded in the user–question interaction sequences.
- We propose a rhythm-aware temporal representation method and a three-channel fusion strategy to model students' learning rhythm and enhance the representation of spatio-temporal information.
- Experimental results demonstrate that STSKT outperforms baselines in prediction performance.

2 Related Works

Knowledge tracing (KT), which aims to model the evolution of student knowledge states, has progressed significantly from traditional probabilistic models to deep learning models. Early foundational work, such as Bayesian Knowledge Tracing (BKT) [13], employs Hidden Markov Models to track mastery. However, its reliance on predefined skill mappings limits its flexibility. The advent of deep learning introduces models like DKT [2], which uses Recurrent Neural Networks (RNNs) to capture long-term dependencies in learning sequences. Further advancements, such as memory-augmented networks like DKVMN [14], enhance the modeling of knowledge storage and update mechanisms. To improve interpretability, some models incorporate cognitive theories; for instance, XKT [15] integrates multidimensional item response theory. Other works focus on fine-grained dynamic factors, such as temporal cross-skill influences (HawkesKT [16]) and peer group effects (STHKT [17]). While these models enhance performance and interpretability, they often struggle to capture complex relational structures inherent in learning.

In recent years, Graph Neural Networks (GNNs) have shown significant promise in KT by explicitly modeling the relationships among skills or questions. GKT [9] is a pioneering work that models knowledge dependencies as a graph structure. Subsequent models refine this approach, such as GIKT [10], which introduces a question-skill bipartite graph, and HGKT [11], which constructs hierarchical exercise graphs to capture deeper dependencies. More recent research shifts towards capturing the dynamic nature of these relationships. DyGformer [18] proposes a dynamic graph Transformer architecture. DyGKT [12] further advances this by modeling the continuous-time dynamic growth of student behavior with a question-answering graph. However, existing graph-based models often use static graphs or do not fully capture the complex spatio-temporal dynamics inherent in the learning process.

3 Preliminaries

To predict the response outcome r_i of the interaction (s_i, q_i) at time t_i—where $r_i = 1$ if user s_i answers question q_i correctly, and $r_i = 0$ otherwise—we propose a spatio-temporal fusion streaming KT model that fuses both spatial and temporal information. The key concepts and problem formalization definitions are introduced below.

Definition 1 (Student interaction stream). *Let $\mathcal{X}_t = \{x_1, x_2, \cdots, x_t\}$ denote the student interaction stream up to time t, where each interaction $x_i = \{s_i, q_i, t_i, r_i\}$ consists of a student ID, a question ID, a time step, and a response label. Let $\Delta\mathcal{X}_T = \mathcal{X}_T \setminus \mathcal{X}_t$ denote the newly arrived data between time steps t and T. This stream arrives continuously over time, reflecting the real-time learning behaviors of students.*

Definition 2 (Dynamic heterogeneous graph). *Given a student interaction stream $\mathcal{X}_t$, we define a dynamic heterogeneous graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$, where the node set $\mathcal{V}_t$ includes students and questions, and the edge set $\mathcal{E}_t$ comprises three types of edges: (i) question co-occurrence edge e^o, indicating questions that share at least one common concept; (ii) collaboration association edge e^c, capturing the collaborative relationship between students who attempt the same question within a recent time window; (iii) temporal proximity edge e^p, representing consecutive interactions made by the same student. As new data $\Delta\mathcal{X}_T$ arrives at time T, the graph is updated from $\mathcal{G}_t$ to a new graph $\mathcal{G}_T$ accordingly. The construction details are provided in Section 4.1.*

Definition 3 (Problem formalization). *Given a student interaction sequence $\mathcal{X}_T$, our core task is to predict the score r_{T+1} of student s_{T+1} on question q_{T+1} at time $T+1$. We evaluate the effectiveness of the model by measuring the accuracy of its predictions on students' question responses.*

Notably, the response sequence $\mathcal{X}_T$ contains interaction records from multiple students within the same time window. Unlike traditional KT models that require collecting the complete set of a student's interaction data, this work dynamically updates a heterogeneous graph based on streaming data to refine the representations of both students and questions. When distinguishing two students in a time-independent context, we denote them as s_i and s_j, and similarly, q_u and q_v represent two different questions.

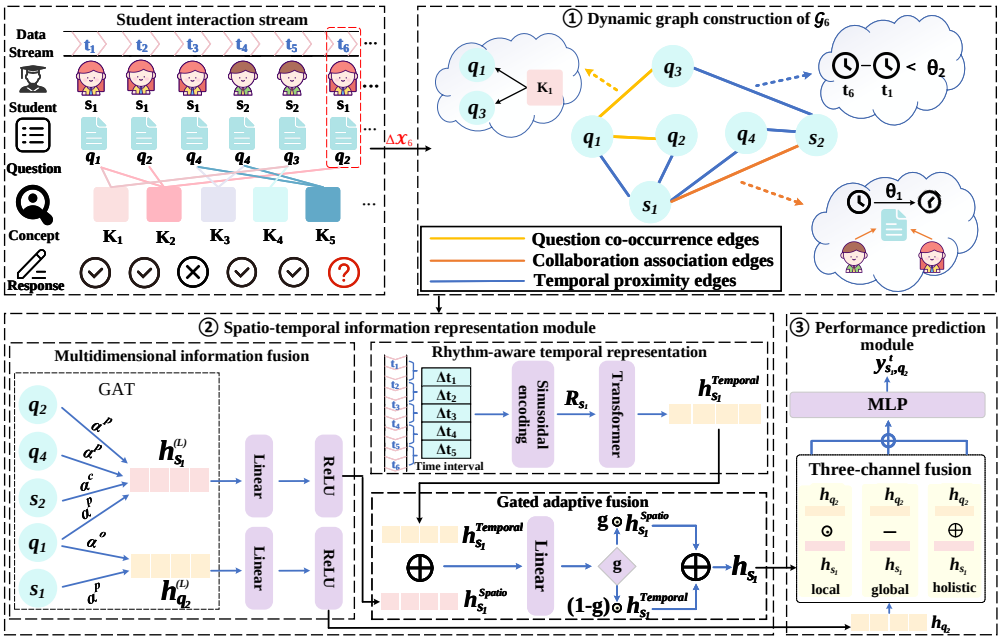

Fig. 1: Overview of the STSKT framework.

4 The STSKT Framework

This section presents the detailed architecture of STSKT for streaming data scenarios. As shown in Fig.1, when predict the response correctness of student s_1 to question q_2 at

time step t_6, we first use the dynamic graph construction module to update a dynamic heterogeneous graph $\mathcal{G}_6$ for the interaction (s_1, q_2) based on the streaming response sequence $\mathcal{X}_6$. Next, the spatio-temporal information representation module incorporates three mechanisms–multidimensional information fusion, rhythm-aware temporal representation, and gated adaptive fusion, to derive comprehensive spatio-temporal representations of student s_1. The performance prediction module adopts a three-channel fusion approach that integrates local relevance, global disparity, and overall interaction feature to support accurate prediction of the response correctness of student s_i to question q_2. A detailed description of the three modules is presented in the following sections.

4.1 Dynamic Heterogeneous Graph Construction

This section introduces how the three types of edges are updated in the dynamic heterogeneous graph, evolving from $\mathcal{G}_t$ to $\mathcal{G}_T$, given the newly arrived data $\Delta\mathcal{X}_T$. The specific methods are as follows.

Question Co-occurrence Edges: Existing research often represents question information based on the relationships between questions and skills [19] or between questions and students [20]. Nevertheless, the semantic representations of questions remain insufficient due to the sparsity of the aforementioned interactions. In general, questions that share the same skill tags or question types (e.g., multiple-choice, fill-in-the-blank) tend to exhibit a high degree of similarity in terms of both knowledge assessment and required cognitive abilities. Therefore, we construct co-occurrence edges between questions in the dynamic heterogeneous graph. An edge is created between questions q_i and q_j if they share the same skill or belong to the same question type, and the edge weight $w^{\mathrm{o}}_{q_i q_j} \geq 0.3$. The edge weight is defined as follows:

$$w^{o}_{q_i q_j} = \frac{|K_{q_i} \cap K_{q_j}|}{|K_{q_i} \cup K_{q_j}|} + \lambda \cdot \mathbb{I}(\tau_{q_i} = \tau_{\mathrm{qj}}), \tag{1}$$

where K_{q_i} represents the set of skills associated with question q_i, τ_{q_i} denotes the question type of question q_i, and $\mathbb{I}$ is an indicator function that takes the value of 1 when two questions belong to the same type and 0 otherwise. λ is a hyperparameter. As new data $\Delta\mathcal{X}_T$ continuously arrives, the number of question co-occurrence edges in the dynamic heterogeneous graph also increases accordingly.

Collaboration Association Edges: The collaborative effect among peers has been demonstrated as a significant factor in improving student performance [21]. However, such relationships cannot be explicitly obtained even when student social network data are available. It has been shown that students working on the same task within a similar time frame are more likely to be influenced by shared instructional resources or discussion environments, leading to cognitive co-evolution [22]. Motivated by this finding, we view this "time-task" co-occurrence as a potential source of collaboration signals. An edge is created between students s_i and s_j if they interact with the same question within the same time window of length θ_1, and the collaboration intensity weight $w^{c}_{s_i s_j} \geq 0.1$ (to prevent outdated social associations from interfering with the current state). The collaboration intensity weight is defined as follows:

$$w^{\mathrm{c}}_{s_i s_j} = \log\left(1 + |\Gamma(s_i) \cap \Gamma(s_j)|\right) \exp\left(-\frac{\Delta t_{s_i s_j}}{\theta_1}\right), \tag{2}$$

where $\Gamma(s_i)$ represents the set of questions answered by student s_i in the response sequence, and $\Delta t_{s_i s_j}$ represents the time interval between the last time two students answered the same question ($\Delta t_{s_i s_j} \leq \theta_1$). As new data $\Delta \mathcal{X}_T$ continuously arrives, the number of collaboration association edges between students in the dynamic heterogeneous graph also increases accordingly.

Temporal Proximity Edges: Students' performance on recently attempted questions often provides a more accurate and timely indication of their current knowledge state and cognitive level. To highlight the importance of recent interactions, previous studies [6] have introduced the Ebbinghaus forgetting curve, using temporal decay mechanisms to model how knowledge retention diminishes over time and to emphasize more recent learning behaviors. However, such approaches focus solely on the importance of recent interactions in linear time for assessing students' knowledge states, while overlooking the contextual information formed by the co-occurrence of questions within recent interactions. In practice, questions presented within the same time window are often related to the same knowledge unit or instructional stage, and thus tend to share similar semantics. Moreover, students' learning states and environmental conditions are typically consistent during these interactions. Therefore, modeling and leveraging the spatial contextual information among questions within the same time window enables a more context-aware representation of a student's knowledge state, leading to improved learner modeling. Specifically, an edge is created between student s_i and question q_j if student s_i interacts with question q_j within the time window θ_2. The edge weight is defined as follows:

$$w^p_{s_i q_j} = \exp\left(-\frac{t_n - t_j}{\theta_2}\right), \tag{3}$$

where t_n and t_j denote the time of student s_i's most recent answer and the most recent answer to question q_j, respectively. As new data $\Delta \mathcal{X}_T$ continuously arrives, the temporally proximity edges in the dynamic heterogeneous graph are also updated accordingly.

4.2 Spatio-Temporal Information Fusion for Student Representation

In this section, we first propose a multidimensional information fusion mechanism to embed the spatial structure of students and questions based on the dynamic heterogeneous graph. Next, we introduce a rhythm-aware temporal representation method to capture students' learning rhythms from their interaction sequences. Finally, a gated adaptive fusion mechanism is employed to obtain the fused spatio-temporal representation of students. The detailed process is described as follows:

Spatial Feature Extraction To extract spatial features from the constructed dynamic heterogeneous graph, we design a Graph Attention Network (GAT). Standard GAT calculates attention solely based on node features, ignoring any explicit edge importance. However, the edges in our graph carry meaningful, pre-computed weights (w^o, w^c, w^p) that represent crucial relational information. To leverage this, we modify the standard attention mechanism to be edge-weighted. This allows the model to prioritize neighbors that are not only similar in the feature space but also structurally important according to our graph construction logic. Accordingly, the edge-weighted attention coefficient is defined as:

$$\alpha^*_{v_i v_j} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T\left[\mathbf{W_1}\mathbf{e}_{v_i} \parallel \mathbf{W_1}\mathbf{e}_{v_j}\right]\right) w^*_{v_i v_j}\right)}{\sum_{v_k \in \Gamma(v_i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^T\left[\mathbf{W_1}\mathbf{e}_{v_i} \parallel \mathbf{W_1}\mathbf{e}_{v_k}\right]\right) w^*_{v_i v_k}\right)}, \tag{4}$$

where v_i denotes a student node or a question node, and $\mathbf{e}_{v_i}$ is its initial embedding obtained via a linear transformation. $\mathbf{a}$ and $\mathbf{W}_1$ are the trainable parameters. $\Gamma(v_i)$ denotes the set of neighbors of v_i, and $w^*_{v_i v_j}$ denotes the edge weight between v_i and v_j, where "*" denotes the edge type: "o" for question co-occurrence edge, "c" for collaboration association edge, and "p" for temporal proximity edge. Accordingly, the multidimensional spatial structure embedding of nodes in the dynamic heterogeneous graph is defined as

$$\mathbf{h}^{(l+1)}_{v_i} = ||^H_{h=1}\sigma\left(\sum_{v_j \in \Gamma(v_i)} \alpha^*_{v_i v_j} \mathbf{W}^h_2 \mathbf{h}^{(l)}_{v_j}\right), \tag{5}$$

where, "||" denotes the concatenation operation, σ is the sigmoid function, $\mathbf{W}^h_2$ is the trainable parameter matrix of the h-th head, and $\mathbf{h}^{(l)}_{v_j}$ is the feature representation of node v_j in layer l. Finally, the Linear+ReLU non-linear mapping function is used to project the final layer representations of the student $\mathbf{h}^{(L)}_{s_i}$ and the question $\mathbf{h}^{(L)}_{q_j}$ to the same space, as follows,

$$\mathbf{h}^{Spatio}_{s_i} = f_s\left(\mathbf{h}^{(L)}_{s_i}\right), \tag{6}$$

and

$$\mathbf{h}^{Spatio}_{q_j} = f_q\left(\mathbf{h}^{(L)}_{q_j}\right). \tag{7}$$

Temporal Answering Patterns Modeling A student's learning rhythm can reflect their learning state (e.g., fatigue or focus), making it a key factor in assessing knowledge mastery. The time interval between a student's responses serves as an effective indicator of this rhythm. To help the model capture both short-term fluctuations and long-range temporal dependencies, we apply sinusoidal encoding to the time intervals. This encoding projects each interval into a multi-frequency representation, enabling the model to learn temporal patterns across different scales. Specifically, given a target student s_i from the newly arrived data, we extract the most recent M interactions of s_i and encode the time intervals between them as follows:

$$\mathbf{R}_{s_i} = [\phi(\Delta t_{s_i,1}), \phi(\Delta t_{s_i,2}), \cdots, \phi(\Delta t_{s_i,M-1})], \tag{8}$$

where $\Delta t_{s_i,j} = t_{s_i,j+1} - t_{s_i,j}$ denotes the time interval between the $(j+1)-th$ and $j-th$ interactions, and the encoding function ϕ is defined as:

$$\begin{aligned}\phi(\Delta t_{s_i,j}) = [&\sin(\omega_1 t_j + \tfrac{\Delta t_j}{T_{\max}}), \cos(\omega_1 t_j + \tfrac{\Delta t_j}{T_{\max}}), \cdots, \\ &\sin(\omega_{d/2} t_j + \tfrac{\Delta t_j}{T_{\max}}), \cos(\omega_{d/2} t_j + \tfrac{\Delta t_j}{T_{\max}})]^T \\ &+ \mathbf{W}_3 \mathbf{h}_{q_j} + \mathbf{b}_1,\end{aligned} \tag{9}$$

where $\mathbf{W}_3$, $\mathbf{b}_1$ are trainable parameters, $T_{max} = \max_j \Delta t_j$, $\omega_k = 1/10000^{2(k-1)/d}$ with even dimension d. Subsequently, we apply a Transformer to the time interval encoding sequence $\mathbf{R}_{s_i}$ of the target student s_i to obtain a representation of the student's learning rhythm, which is defined as follows:

$$\mathbf{h}^{Temporal}_{s_i} = \text{Transformer}(\mathbf{R}_{s_i}). \tag{10}$$

Representation Fusion Students' spatial features and learning rhythms jointly influence and characterize their future knowledge states. Clearly, due to individual differences, these factors affect students in different ways. Accordingly, the contributions of spatial topology and learning rhythms to knowledge state assessment should also vary. To this end, we introduce a gated adaptive fusion mechanism that flexibly and accurately captures their personalized effects. We first employ a trainable weight matrix $\mathbf{W}_4$ to model high-order nonlinear interactions between the spatial representation and the temporal learning rhythm representation, thereby quantifying their respective contributions to the overall student representation, as defined below,

$$\mathbf{g}_{s_i} = \sigma(\mathbf{W}_4[\mathbf{h}_{s_i}^{Spatio} || \mathbf{h}_{s_i}^{Temporal}]), \tag{11}$$

Subsequently, the gating vector $\mathbf{g}_{s_i}$ is used to adaptively fuse the spatial and temporal representations of students. The final spatio-temporal representation of student s_i is given by

$$\mathbf{h}_{s_i} = \boldsymbol{g}_{s_i} \odot \mathbf{h}_{s_i}^{Spatio} + (1 - \boldsymbol{g}_{s_i}) \odot \mathbf{h}_{s_i}^{Temporal}, \tag{12}$$

where $\odot$ denotes the dot product operation.

4.3 Prediction Module

To accurately predict the probability that a student answers a question correctly, it is crucial to comprehensively model the interaction between the student's knowledge state and the question's characteristics. Relying on a single similarity measure may overlook important aspects of their relationship. Therefore, we design a three-channel fusion method to capture multiple perspectives of this interaction. Specifically, at the local level, we employ element-wise multiplication to capture fine-grained alignment between the student and question representations, highlighting detailed knowledge matching. At the global level, Euclidean distance is used to measure their overall discrepancy in the representation space, capturing broader differences in knowledge scope. Moreover, to preserve holistic information and enable the modeling of more complex nonlinear relationships, we incorporate vector concatenation. From the representations of these three channels, we derive the comprehensive interaction representation between the student and the question as follows:

$$\mathbf{h}_{s_i q_i} = MLP\left([\mathbf{h}_{\mathrm{s_i}} \odot \mathbf{h}_{q_i}; \|\mathbf{h}_{\mathrm{s_i}} - \mathbf{h}_{q_i}\|; \mathbf{h}_{\mathrm{s_i}} || \ \mathbf{h}_{q_i}]\right), \tag{13}$$

where

$$\mathbf{h}_{q_i} = \mathrm{MLP}_{q_i}\left(\left[\mathbf{h}_{q_i}^{Spatio} \| \mathbf{e}_{q_i}\right]\right), \tag{14}$$

and $\mathbf{e}_{q_i}$ is the initial embedding representation of question q_i. Accordingly, at the time step i, the probability of the student s_i answering the question q_i correctly is defined as follows:

$$\hat{y}_t(s_i, q_i) = \sigma\left(\mathbf{W}_5 \mathrm{ReLU}\left(\mathbf{W}_6 \mathbf{h}_{s_i q_i} + \mathbf{b}_2\right) + \mathbf{b}_3\right), \tag{15}$$

where $\mathbf{W}_5$,$\mathbf{W}_6$,$\mathbf{b}_2$,$\mathbf{b}_3$ are trainable parameters. If the predicted probability $\hat{y}_t(s_i, q_i)$ that the student answers the next question correctly is greater than or equal to 0.5, the response is considered correct; otherwise, it is considered incorrect. The corresponding loss function for the model is represented as follows:

$$\mathcal{L} = -\sum_{(s_i, q_i)} [r_{s_i q_i} \log \hat{y}_t(s_i, q_i) + (1 - r_{s_i q_i}) \log(1 - \hat{y}_t(s_i, q_i))] + \beta \|\Theta\|_2. \tag{16}$$

5 Experiment

This section evaluates the algorithmic performance of STSKT on four real-world datasets. To guide the experiments, we formulate the following research questions:

- **RQ1**: How does the overall performance of STSKT compare with the baselines?
- **RQ2**: Is the proposed dynamic heterogeneous graph construction method effective?
- **RQ3**: Do the components in the node representation and information fusion methods work effectively?

5.1 Experiment Preparation

Datasets Our experiments are conducted on four public datasets (see Table 1 for statistics): (i) **Assist12**[4], this dataset is collected from the ASSISTments platform, contains students' practice logs on concept-based exercises from the 2012–2013 academic year. We retain only the records annotated with knowledge concepts to ensure data quality; (ii) **Assist17**[5], this dataset is released for the 2017 ASSISTments Data Mining Competition, includes complete learning histories from students in four schools. Its multi-school structure provides diverse learning contexts, which are useful for analyzing interaction patterns in real educational settings; (iii) **Junyi**[6], this dataset is collected from Junyi Academy, provides fine-grained timestamps and response durations for each question, enabling analysis of students' learning rhythms and short-term behavioral patterns; (iv) **EdNet**[7], this dataset is gathered from the Santa AI tutoring system in South Korea, covers a large-scale population of users across multiple platforms and serves to evaluate the model's generalization and robustness in real-world online learning environments.

Table 1: Basic statistics of all datasets.

Dataset	**# Students**	**# Questions**	**# Concepts**	**# Interactions**
Assist12	25.3k	50.9k	245	2621.3k
Assist17	1.7k	3.2k	102	942.8k
Junyi	175.4k	0.7k	40	25670.2k
EdNet	685.4k	12.3k	141	35083.7k

Baselines We compare STSKT against ten baselines, categorized as: (i) classic deep learning KT models (DKT [2], AKT [7]); (ii) attention and feature-based models (SAKT [5], IEKT [23], DIMKT [24]); (iii) static graph-based models (GKT [9], GIKT [10]); and (iv) dynamic or temporal-aware models (HawkesKT [16], DyGFormer [18], DyGKT [12]).

Implementation Details We split each dataset into training, validation, and test sets using an 8:1:1 ratio. All models are trained using the ADAM optimizer[25] with a learning rate of 0.0001 and a batch size of 1024. The hidden dimension is set to 64.

[4] https://sites.google.com/site/assistmentsdata/home/2012-13-school-data-with-affect
[5] https://sites.google.com/view/assistmentsdatamining/dataset
[6] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198
[7] https://github.com/riiid/ednet

STSKT involves four hyper-parameters, namely λ, θ_1, θ_2 and β. We obtain the optimal settings of hyperparameters through a traversal method. In this paper, we set λ=0.4, θ_1=180 minutes, and θ_2=30 minutes, and β is tuned via grid search over {1e-5, 1e-4, 1e-3, 1e-2}, with the best value selected on the validation set. We use Area Under the Curve (AUC) and Accuracy (ACC) as our evaluation metrics.

5.2 Performance Comparison (**RQ1**)

Table 2 presents the performance comparison between STSKT and all baselines across the four datasets in terms of AUC and ACC, respectively. Overall, STSKT consistently achieves the best results on both evaluation metrics. On average, GIKT, IEKT, DIMKT, and DyGKT achieve the best results after STSKT in terms of AUC. Regarding ACC, DyGKT closely follows STSKT but still falls short. More specifically, compared with DyGKT, the closest competitor, STSKT yields improvements of 4.26% in AUC and 3.06% in ACC. To further assess the statistical significance, we conduct the Mann-Whitney U test, which shows that STSKT significantly outperforms all baselines on both metrics ($p < 10^{-3}$).

Table 2: AUC and ACC of STSKT and the ten baselines on four datasets. The best performance and the second performance results are denoted in bold and underlined, respectively.

Datasets	Assist12		Assist17		Junyi		EdNet	
Metrics	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
DKT [2]	0.7289	0.7361	0.7213	0.6909	0.7234	0.8395	0.6836	0.6889
SAKT [5]	0.6911	0.7232	0.6839	0.6660	0.7464	0.8427	0.6719	0.6435
GKT [9]	0.7350	0.7143	0.8158	*0.7541*	0.7383	0.8420	0.6792	0.6132
AKT [7]	0.7707	0.7535	0.7501	0.7083	0.7836	0.8477	0.7003	0.6592
GIKT [10]	0.7672	0.7506	0.7791	0.7089	0.7840	0.8537	0.7351	0.7123
HawkesKT [16]	0.7559	0.7441	0.7033	0.6845	0.7609	0.8411	0.6869	0.7012
IEKT [23]	0.7618	0.7506	0.7812	0.7155	0.7911	0.8472	0.7405	0.6945
DIMKT [24]	0.7552	0.7453	0.7528	0.7033	*0.8166*	0.8628	*0.7431*	0.7096
DyGFormer [18]	*0.7721*	0.7542	0.7506	0.7078	0.8149	0.8577	0.6932	0.6791
DyGKT [12]	0.7034	*0.8317*	*0.8235*	0.7181	0.8076	*0.9146*	0.7334	*0.7885*
STSKT	**0.7842**	**0.8366**	**0.8662**	**0.8229**	**0.8276**	**0.9188**	**0.7603**	**0.7968**

5.3 Effectiveness Analysis of Dynamic Heterogeneous Graph Construction (**RQ2**)

To evaluate the effectiveness of the proposed dynamic heterogeneous graph, we adopt the static graph construction method STATIC from [9] to replace the dynamic graph used in STSKT, while keeping all other representation settings unchanged. As shown

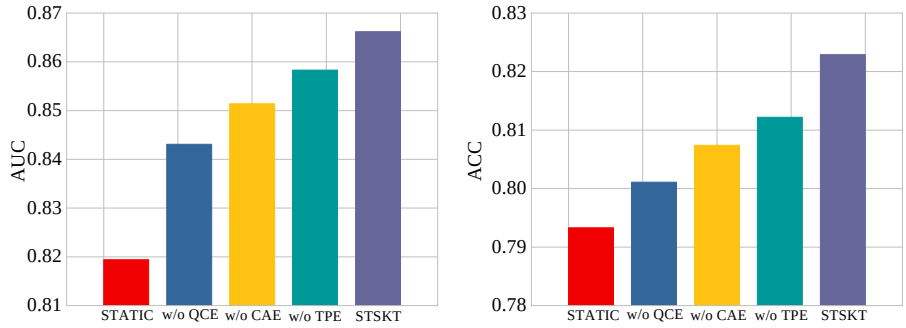

Fig. 2: Effectiveness analysis of dynamic heterogeneous graph construction on Assist17.

in Fig. 2, when the dynamic graph in STSKT is replaced by the static graph, the performance drops by 5.40% in AUC and 3.60% in ACC, demonstrating that the proposed dynamic graph structure significantly improves model performance. To further investigate which types of edges in the dynamic heterogeneous graph contribute most, we design three ablation experiments, each removing a specific type of edge while keeping the rest of STSKT unchanged: (i) removing question co-occurrence edges (w/o QCE); (ii) removing collaborative association edges (w/o CAE); (iii) removing temporal proximity edges (w/o TPE). As shown in Fig. 2, question co-occurrence edges provide the largest contribution: removing them leads to a performance drop of 2.67% in AUC and 2.65% in ACC. Collaborative association edges are the second most influential, with decreases of 1.71% in AUC and 1.88% in ACC when removed. Temporal proximity edges contribute relatively less, and their removal results in decreases of 0.91% in AUC and 1.30% in ACC.

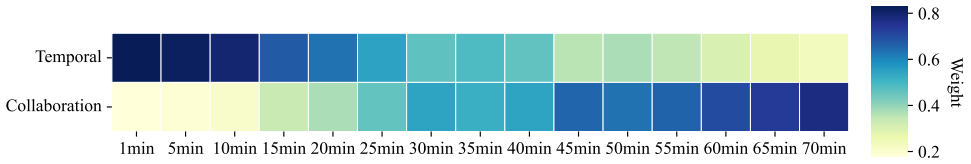

Fig. 3: Contributions of collaborative and temporal proximity edges to knowledge-state assessment across varying average answering intervals on Assist17. The x-axis denotes students' average answering interval, the y-axis denotes the two edge types, and color intensity corresponds to the attention weights from Equation (5).

The construction of the proposed dynamic heterogeneous graph is closely related to students' answering frequency: students with higher frequencies are more likely to form high-weight collaborative edges and temporal proximity edges. A key question is which contributes more to knowledge state assessment for students with varying answering frequencies. To address this question, we extract the corresponding attention weights from Equation (5) and analyze the differences between the two types of edges under varying answering intervals. As shown in Fig. 3, for students with low answering frequency (i.e., longer average answering intervals), collaborative relations between students play a more important role in knowledge-state assessment. In contrast, for students with high answering frequency, their recent answering history is more critical. Moreover, we find that compared with static graphs, dynamic heterogeneous graphs ex-

hibit clear advantages in assessing the knowledge states of students with non-continuous learning behaviors. As shown in Fig. 4, we classify students with answering intervals longer than 30 minutes as the non-continuous learning group. The results show that, for this group, both static and dynamic graphs experience performance degradation, but the decline is more pronounced in static graphs. Overall, the experiments confirm that the proposed dynamic heterogeneous graph not only improves prediction accuracy but also exhibits superior adaptability to non-continuous learning behaviors, aligning more closely with real-world learning patterns.

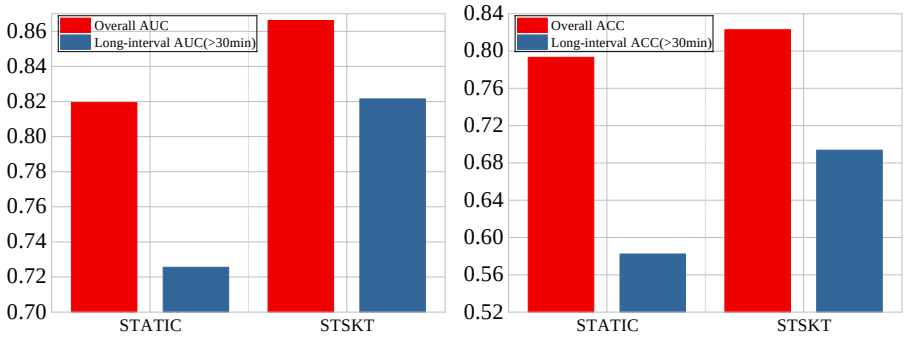

Fig. 4: Adaptability of the dynamic heterogeneous graph to non-continuous learning behaviors on Assist17.

5.4 Effectiveness Analysis of Components in Node Representation and Information Fusion Methods (RQ3)

This section evaluates the effectiveness of STSKT's key components, which include its modules for spatial feature extraction, temporal rhythm modeling, spatio-temporal fusion, and final prediction. The effectiveness of the four operations is evaluated in three experimental setups.

Table 3: Ablation study of STSKT's key components on Assist17.

Model Variant	**AUC**	**ACC**
STSKT (Full Model)	**0.8662**	**0.8229**
Analysis of Representation		
STSKT W/o Heterogeneous Info (STSKT-GCN)	0.8446	0.8002
STSKT W/o Rhythm-aware (STSKT-LSTM)	0.8307	0.7823
Analysis of Spatio-Temporal Fusion		
STSKT W/o spatial	0.7833	0.6853
STSKT W/o temporal	0.6569	0.5795
Analysis of Three-Channel Interaction		
STSKT W/o Local	0.8521	0.8093
STSKT W/o Global	0.8604	0.8167
STSKT W/o Holistic	0.8573	0.8124

First, to assess the impact of multi-dimensional information embedding and rhythm-aware operations on model performance, we conduct two replacements in STSKT: (i) the GAT method used for embedding heterogeneous information in the student spatial structure is replaced with GCN (STSKT-GCN), which ignores heterogeneous information; and (ii) the Transformer used to embed learning rhythm in the interaction sequence is replaced with LSTM (STSKT-LSTM), which ignores time intervals. As shown in Table 3, replacing the heterogeneous information embedding with GCN (STSKT-GCN) reduces the model's AUC and ACC by 2.2% and 2.3%, respectively; replacing the rhythm-aware embedding with LSTM (STSKT-LSTM) reduces them by 3.6% and 4.1%, respectively. This indicates that embedding heterogeneous information in the dynamic graph, together with incorporating learning rhythms from temporal information, can effectively enhance the predictive performance of the algorithm. Among them, the utilization of learning rhythms from temporal information contributes more significantly to the performance improvement.

Next, we analyze the contributions of temporal and spatial information in the gated adaptive fusion mechanism. As shown in Table 3, removing the spatial component in Equation (11) (W/o spatio) leads to decreases of 8.3% and 13.8% in AUC and ACC, respectively, while removing the temporal component (W/o temporal) results in drops of 20.9% and 24.3%, respectively. Overall, both temporal and spatial information are crucial for prediction performance, with the temporal information operations proposed in this work contributing more substantially to the improvement. We further examine the fusion coefficients of temporal and spatial information in Equation (11) (i.e., the mean of vector $\mathbf{g}$) for different student types. As shown in Fig. 5, students with longer average answering intervals receive higher attention on spatial topology information, whereas students with shorter intervals are assigned higher attention on temporal information.

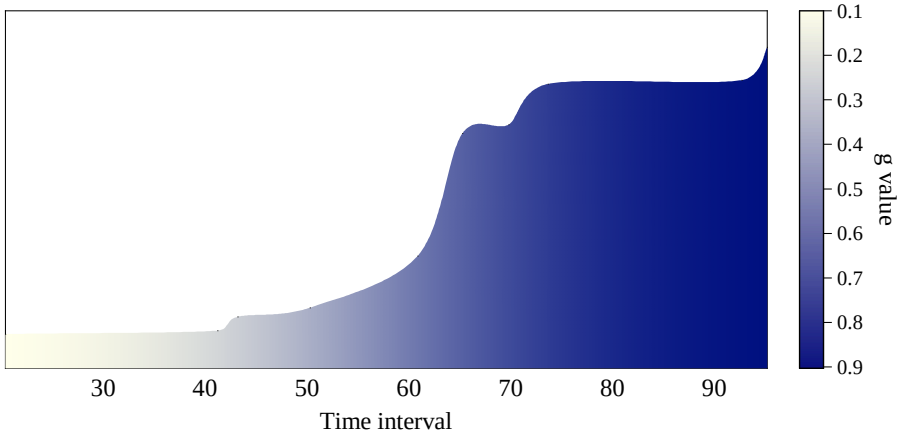

Fig. 5: Analysis of the fusion weights of spatio-temporal information for students with different characteristics. The color depth is proportional to the average value of the fusion weight vector.

Finally, we evaluate the effectiveness of the three-channel fusion method through three ablation experiments. Specifically, we remove the local matching channel (W/o Local), the global difference channel (W/o Global), and the overall information channel (W/o Holistic) from the interactive representation of students and questions. As shown in Table 3, excluding any of the three channels reduces the algorithm's predictive per-

formance, with the largest degradation caused by removing the local matching channel and the smallest by removing the global difference channel. The relatively minor role of the global difference channel may stem from its informational overlap with the local matching channel. When the discrepancy between student and question information is large, this difference can already be captured by the local matching channel, leading to redundancy. Only when the two are close does the global difference channel add value, as it captures the subtle sensitivity that the local matching channel alone cannot distinguish.

6 Conclusion

We propose STSKT, a spatio-temporal fusion streaming knowledge tracing model designed for realistic interactive data streams. Instead of retraining on all historical interactions, STSKT incrementally updates a dynamic heterogeneous graph from the current batch and models temporal dependencies within each student's most recent N interactions. In this dynamic heterogeneous graph, question co-occurrence edges, collaborative association edges, and temporal adjacency edges are constructed to enrich question representations, capture peer effects, and strengthen the spatial encoding of temporal dynamics. Temporal modeling further incorporates time intervals to reflect students' learning rhythms. To integrate these rich features, a gated mechanism provides adaptive fusion, dynamically weighing the contributions of a student's spatial topology and their temporal learning rhythm based on individual answering patterns. This is followed by a three-channel prediction module that models the student-question interaction from global, local, and holistic perspectives. Experimental results show that STSKT outperforms all baselines. Moreover, the fusion of spatio-temporal information can effectively improve the predictive performance of the model, with temporal information playing a more significant role. Furthermore, compared with static graphs, our dynamic heterogeneous graph yields greater accuracy and is particularly effective for students with discontinuous learning behaviors. In modeling students' spatial structures, question co-occurrence edges contribute the most, followed by collaborative association edges. Finally, all three channels of the prediction module are shown to be effective, with the local matching channel being the most critical.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 62507039 and 62576287), Sichuan Science and Technology Program (Grant No. 2025ZNSFSC0506), and the Yibin Science and Technology Program (No. 2023SF004). Carl Yang was not supported by any funds from China.

References

1. Amirreza Mehrabi, Jason W. Morphew, Brenda S. Quezada. Enhancing performance factor analysis through skill profile and item similarity integration via an attention mechanism of artificial intelligence. Frontiers in Education, 2024, 9: 1454319.
2. Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, Jascha SohlDickstein. Deep knowledge tracing. In: Proceedings of the 29th International Conference on Neural In formation Processing Systems, Montreal, Canada, 2015, pp. 505–513.

3. Shi Pu, Michael Yudelson, Lu Ou, Yuchi Huang. Deep knowledge tracing with transformers. In: Proceedings of the 21th International Conference on Artificial Intelligence in Education, Ifrane, Morocco, 2020, pp. 252–256.
4. Jiajun Cui, Zeyuan Chen, Aimin Zhou, Jianyong Wang, Wei Zhang. Fine-grained interaction modeling with multi-relational rtansformer for knowledge tracing. ACM Transactions on Information Systems, 2023, 41: 1–26.
5. Shalini Pandey, George Karypis. A self-attentive model for knowledge tracing. In: Proceedings of the 12th International Conference on Educational Data Mining, Montreal, Canada, 2019, pp. 384–389.
6. Jaap MJ Murre, Joeri Dros. Replication and analysis of Ebbinghaus' forgetting curve. PloS One, 2015, 10: 1-23.
7. Aritra Ghosh, Neil Heffernan, Andrew S Lan. Context-Aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 2020, pp. 2330–2339.
8. Yi-Fei Wen, Hang Liang, Carl Yang, Tao Zhou, Jia Liu, Yajun Du, Yan-Li Lee. Sequential contrastive learning for progressive knowledge tracing. Knowledge-Based Systems, 2025, 329: 114413.
9. Hiromi Nakagawa, Yusuke Iwasawa, Yutaka Matsuo. Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In: Proceedings of the 18th IEEE/WIC/ACM International Conference on Web Intelligence, Thessaloniki, Greece, 2019, pp. 156–163.
10. Yang Yang, Jian Shen, Yanru Qu, Yunfei Liu, Kerong Wang, Yaoming Zhu, Weinan Zhang, Yong Yu. GIKT: A graph-based interaction model for knowledge tracing. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Ghent, Belgium, 2020, pp. 299–315.
11. Hanshuang Tong, Zhen Wang, Yun Zhou, Shiwei Tong, Wenyuan Han, Qi Liu. Introducing problem schema with hierarchical exercise graph for knowledge tracing. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 2022, pp. 405–415.
12. Ke Cheng, Linzhi Peng, Pengyang Wang, Junchen Ye, Leilei Sun, Bowen Du. DyGKT: Dynamic graph learning for knowledge tracing. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 2024, pp. 409–420.
13. Albert T. Corbett, John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 1994, 4: 253–278.
14. Jiani Zhang, Xingjian Shi, Irwin King, Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 2017, pp. 765–774.
15. Chang-Qin Huang, Qiong-Hao Huang, Xiaodi Huang, Hua Wang, Ming Li, Kwei-Jay Lin. XKT: Towards explainable knowledge tracing model with cognitive learning theories for questions of multiple knowledge concepts. IEEE Transactions on Knowledge and Data Engineering, 2024, 36: 7308-7325.
16. Chenyang Wang, Weizhi Ma, Min Zhang, Chuancheng Lv, Fengyuan Wan, Huijie Lin, Taoran Tang, Yiqun Liu, Shaoping Ma. Temporal cross-effects in knowledge tracing. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, 2021, pp. 517–525.
17. Shuting Li, Shuanghong Shen, Yu Su, Xinjie Sun, Junyu Lu, Qi Mo, Zhenyi Wu, Qi Liu. STHKT: Spatiotemporal knowledge tracing with topological hawkes process. Expert Systems with Applications, 2025, 259: 125248.
18. Le Yu, Leilei Sun, Bowen Du, Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. Advances in Neural Information Processing Systems, 2023, 36: 67686–67700.
19. Jiangwei Yang, Tingnian He, Fucai Gao, Yang Yang. Deep knowledge tracing method based on enhancing knowledge graph embedding. In: Proceedings of the 7th International

Conference on Computer Information Science and Artificial Intelligence, New York, United States, 2024, pp. 576–580.
20. Hang Liang, Yi-Fei Wen, Yajun Du, Xiaoliang Chen, Tao Zhou, Yan-Li Lee. Interpretable knowledge tracing via fine-grained multi-feature attribution. Physica A: Statistical Mechanics and its Applications, 2026, 681: 131068.
21. Mehmet Diyaddin Yasar, Mustafa Erdogan, Veli Batdi, Ülkü Cinkara. Evaluation of cooperative learning in science education: A mixed-meta method study. European Journal of Science and Mathematics Education, 2024, 12: 411-427.
22. Meehyun Yoon, Jungeun Lee, Il-Hyun Jo. Video learning analytics: Investigating behavioral patterns and learner clusters in video-based online learning. The Internet and Higher Education, 2021, 50: 100806.
23. Ting Long, Yunfei Liu, Jian Shen, Weinan Zhang, Yong Yu. Tracing knowledge state with individual cognition and acquisition estimation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, 2021, pp. 173–182.
24. Shuanghong Shen, Zhenya Huang, Qi Liu, Yu Su, Shijin Wang, Enhong Chen. Assessing student's dynamic knowledge state by exploring the question difficulty effect. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 2022, pp. 427–437.
25. Diederik P. Kingma, Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412. 6980, 2014.