DSCA-KT: DUAL-STATE CONTEXT-AWARE MODEL FOR DYNAMIC KNOWLEDGE TRACING

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge Tracing (KT) aims to model students' evolving knowledge states from historical learning interactions. Recently, dynamic KT methods have gained increasing attention for their ability to model continuously evolving interaction sequences. However, existing approaches still struggle to distinguish transient cognitive fluctuations from stable knowledge-acquisition patterns and often produce unreliable predictions in cold-start scenarios with limited interaction histories. To address these challenges, we propose DSCA-KT, a Dual-State Context-Aware framework for dynamic knowledge tracing. Specifically, we design context-aware recurrent aggregators that capture high-frequency cognitive variations, enabling the model to better characterize transient learning behaviors while preserving long-term knowledge trends. In addition, we introduce a context modulator that generates student- and question-specific contextual representations and incorporates prior structural knowledge to alleviate the cold-start issue. Extensive experiments on several widely used benchmark datasets demonstrate that DSCA-KT consistently outperforms state-of-the-art baselines under both transductive and inductive settings.

1 INTRODUCTION

With the rapid growth of online education platforms, a large volume of educational records has been collected. To leverage those records in helping teachers understand students' learning progress, various knowledge tracing (KT) methods Corbett & Anderson (1994); Harvey & Hammer (1999), which analyze students' answers to a series of questions to track the evolution of their knowledge state over time, have been extensively studied. Traditional knowledge tracing methods typically model student interactions as fixed historical sequences using sequential, attention-based, or graph-based architectures Abdelrahman et al. (2023). However, real-world learning behaviors evolve continuously, motivating growing interest in dynamic knowledge tracing, in which interaction sequences grow incrementally rather than remain static. By modeling continuously evolving student-question interactions, dynamic KT methods provide a more realistic formulation of the learning process and have demonstrated improved predictive performance in educational applications.

Despite recent progress, existing dynamic KT methods still have a limited ability to model the context of evolving interaction sequences. First, current approaches struggle to distinguish transient cognitive fluctuations from stable knowledge acquisition trends. As illustrated in Figure 1, students often exhibit short-term cognitive behaviors, such as sudden insights after mastering a concept (e.g., answering q_3 and q_n) or accidental mistakes caused by carelessness (e.g., answering q_m). These behaviors manifest as high-frequency signals in the corresponding wavelet scalogram. However, most existing dynamic KT models rely on standard recurrent architectures to aggregate historical states. Such recurrent aggregators inherently behave as low-pass filters Rahaman et al. (2019), which tend to smooth out high-frequency cognitive signals and mix them with long-term knowledge trends, resulting in inaccurate estimation of students' actual knowledge states.

Second, dynamic KT models remain vulnerable to cold-start scenarios with sparse interaction histories. Unlike traditional methods trained on fixed-length sequences, dynamic KT operates on continuously growing interaction streams Agarwal et al. (2020), in which newly observed questions initially have only limited historical records.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Consequently, the dynamic context is often insufficiently characterized during the early stages of interaction. Moreover, valuable structural knowledge remains under-exploited. As shown in Figure 1, prerequisite dependencies among concepts strongly influence students' mastery trajectories, yet most existing methods fail to incorporate such structural relationships into dynamic state modeling. Therefore, *capturing transient cognitive fluctuations within dynamically growing sequences and characterizing student interaction history and knowledge dependencies under cold-start scenarios* are becoming the major challenges in dynamic KT.

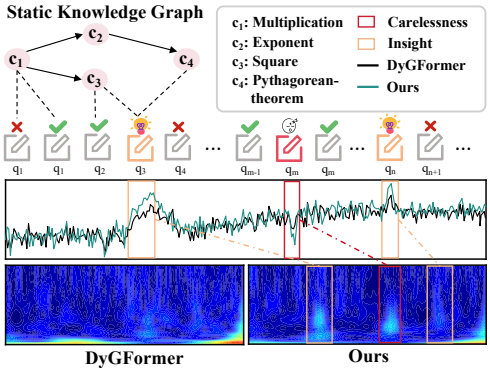


Figure 1: Illustration of the impact of knowledge dependencies under cold-start scenarios on model performance and comparison between DyGFormer and our method in capturing high-frequency shocks for knowledge tracing.

To address these challenges, we propose Dual-State Context-Aware model for Dynamic Knowledge Tracing (DSCA-KT). Our framework constructs a dynamic graph to jointly model the evolving states of students and questions, capturing both temporal and structural dependencies. To better preserve transient cognitive variations, we design a frequency-domain enhancer that isolates and amplifies informative high-frequency signals from interaction sequences. In addition, we introduce a context modulator equipped with a prior context generator, which injects structural priors from a static knowledge graph to improve robustness in question-side cold-start scenarios. By jointly modeling dynamic cognitive behaviors and structural contextual information, DSCA-KT achieves more reliable knowledge tracing performance. The main contributions of this work are summarized as follows:

- We propose context-aware recurrent aggregators that explicitly capture high-frequency transient cognitive fluctuations and sustain stable gradient flow, enhancing the model's sensitivity to instantaneous cognitive shifts.
- We design a context modulator that generates cognitive and temporal context features and injects prior context to mitigate question-side cold starts, modeling dynamic context.
- We conduct extensive experiments on public datasets over state-of-the-art baselines to show the effectiveness and efficiency of our proposed method.

2 Related Work

Traditional Knowledge Tracing. With the advancement of deep learning, knowledge tracing has evolved from probabilistic approaches to deep learning-based knowledge tracing (DLKT). Existing DLKT methods typically model student learning behaviors using fixed-length interaction sequences. Early studies mainly relied on recurrent neural networks (RNNs) to capture temporal dependencies, such as DKT Piech et al. (2015). To improve representation capacity and interpretability, subsequent works introduced memory-augmented architectures, including DKVMN Zhang et al. (2017) and SKVMN Abdelrahman & Wang (2019). Later approaches further incorporated auxiliary learning signals, such as temporal information, question relations, and cognitive diagnosis features Yeung & Yeung (2018); Lalwani & Agrawal (2019); Lee & Yeung (2019); Yeung (2019). To better model long-range dependencies, attention-based methods, including SAINT Choi et al. (2020a), AKT Ghosh et al. (2020), MIKT Sun et al. (2024), and FlucKT Hou et al. (2025), employed self-attention mechanisms to capture global interaction patterns. In parallel, graph-based methods leveraged structural relationships among questions and concepts. For example, GKT Nakagawa et al. (2019) constructs skill graphs to aggregate high-order relational information, while DGEKT Cui et al. (2024), SGKT Wu et al. (2022b), and DisenKT Li et al. (2025) further model heterogeneous or disentangled knowledge structures using graph neural networks and hyperbolic representations. Despite their effectiveness, these methods generally assume static interaction sequences and therefore cannot fully characterize the continuously evolving nature of real-world learning processes.

Dynamic Knowledge Tracing. Dynamic knowledge tracing aims to model the continuous evolution of student interactions over time. Early work by Agarwal et al. Agarwal et al. (2020) highlighted the

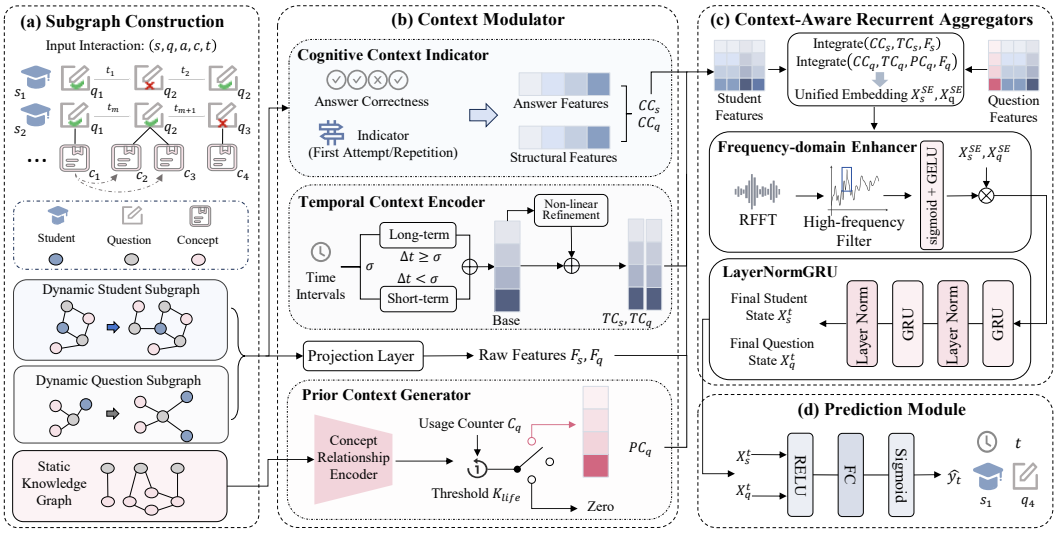


Figure 2: The overall framework and workflow of the proposed DSCA-KT.

importance of interaction recency in modeling dynamic learning behaviors. DIMKT Shen et al. (2022) further incorporated dynamically changing question difficulty conditioned on student ability. More recently, dynamic graph-based approaches have been proposed to capture the structural evolution of interaction streams. DyGKT Cheng et al. (2024) models student-question interactions using a continuous-time dynamic graph framework, while IPDGKT Xiao et al. (2026) introduces information-processing theory to distinguish short-term activation from long-term memory retention. Although these methods improve the modeling of temporal and structural evolution, they still have limited ability to characterize the dynamic learning context. In particular, existing approaches often fail to explicitly distinguish transient cognitive fluctuations from stable knowledge trends, and they rarely incorporate structural prior knowledge to address cold-start scenarios with sparse interaction histories.

To the best of our knowledge, our work is the first attempt to model dynamic context using a frequency-domain method that captures students' cognitive states for dynamic knowledge tracing. Based on our context modulator and context-aware recurrent aggregators, DSCA-KT can not only model cognitive context but also mitigate the cold-start problem.

3 Method

3.1 Problem Definition & Overview of Our Solution

We model the student learning process as a dynamic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V}$ encompasses students $\mathcal{S}$, questions $\mathcal{Q}$, and knowledge concepts $\mathcal{C}$. An interaction event at timestamp t is represented as an edge $e_t = (s, q, a, c, t) \in \mathcal{E}$. Here, $s \in \mathcal{S}$ denotes the student, $q \in \mathcal{Q}$ is the question attempted, and $c \in \mathcal{C}$ represents the underlying knowledge concept. The label $a \in \{0, 1\}$ indicates the correctness of the response. As time progresses, the graph evolves into a historical subgraph $\mathcal{G}^t = \{e_i | t_i < t\}$, reflecting the cumulative learning trajectory. The objective of knowledge tracing is to update the student's latent knowledge mastery, h_s^t, based on $\mathcal{G}^t$. Formally, given a new question q_{t+1} and its associated concept c_{t+1} at time $t+1$, the model aims to predict the probability $P(r_{t+1} = 1 | q_{t+1}, c_{t+1}, \mathcal{G}^T)$ of a correct answer r_{t+1}.

Figure 2 illustrates the framework of our DSCA-KT method. Given an interaction (s, q, a, c, t), we first sample the one-hop neighbors for student s and question q up to time t from the dynamic graph, respectively, forming two interaction sequences denoted as S_s^t and S_q^t, which represent the student s's historical response sequence and the question q's historical attempt sequence, respectively. Next, we prepare context-aware input features for these two sequences produced by the context modulator, and leverage prior knowledge from a static knowledge graph to initialize the aggregator for question q. The two processed feature sequences are fed into the context-aware recurrent aggregators. Then the aggregators, comprising a frequency-domain enhancer and a LayerNormGRU, process their

respective sequences. Finally, the aggregators output the final representations of students s and questions q at time t, which are used for the knowledge tracing task.

3.2 SUBGRAPH CONSTRUCTION

Extracting relevant context from past interactions is a key step toward modeling the states of both the student and the question. The student and question together form the context for knowledge tracing, and changes in their states will affect the final predictions. Therefore, for each interaction(s, q, a, c, t) to be predicted, we extract first-order historical neighbors before timestamp t from two complementary views: student historical sequence and question historical sequence, denoted by $S_s^t = \{s, q', a', c', t' | t' < t\}$ and $S_q^t = \{s', q, a', c', t' | t' < t\}$. These two neighbor sequences will later be used to encode the student's knowledge state and the question's feature. To maintain a consistent tensor shape within a batch B, we leave unused positions padded with zeros.

In addition, to alleviate the question-side cold-start problem in a dynamic graph, we construct a Static Knowledge Graph (SKG) for the subsequent generation of prior contexts, denoted as a heterogeneous graph $\mathcal{G}_{static} = (\mathcal{V}_Q \cup \mathcal{V}_C, \mathcal{E})$. Here, $\mathcal{V}_Q$ and $\mathcal{V}_C$ represent the sets of question and concept nodes, respectively. The edge set $\mathcal{E}$ comprises two types of pedagogical relations: i) *question-concept inclusion* $e_{q,c}$ that connects a question q to a concept c if solving q requires the mastery of c, and ii) *concept-concept prerequisite* e_{c_i,c_j} that connects concept c_i to c_j if c_i is a pedagogical prerequisite for understanding c_j. $\mathcal{G}_{static}$ is time-invariant and serves as a foundational knowledge base that encodes both the composition of questions and the hierarchical structure of concepts. For the detailed construction of concept-concept prerequisites, please refer to the Appendix A.6.

3.3 CONTEXT MODULATOR

To comprehensively characterize the dynamic learning context, we design a context modulator that transforms raw interactions into rich, context-aware features. This module explicitly captures cognitive and temporal context and leverages a prior context generator to ensure robust representations even in cold-start scenarios.

Cognitive Context Indicator. Based on the extracted neighbor sequences, we construct the cognitive context as high-dimensional feature matrices that comprise the interaction outcome and the historical cognitive state. We formulate the interaction outcome matrices to capture status trends from both student and question perspectives. For the student s, we generate the answer performance matrix $\mathbf{x}_{ans}^s \in \mathbb{R}^{N \times d}$ by projecting the entire historical answering outcomes. Similarly, for the question q, we construct $\mathbf{x}_{ans}^q \in \mathbb{R}^{N \times d}$ from its historical answering records. These matrices are obtained via:

$$\mathbf{x}_{ans}^s = \mathbf{a}_s \mathbf{W}_{ans} + \mathbf{b}_{ans},\ \mathbf{x}_{ans}^q = \mathbf{a}_q \mathbf{W}_{ans} + \mathbf{b}_{ans}, \tag{1}$$

where $\mathbf{a}_s, \mathbf{a}_q \in \mathbb{R}^{N \times 1}$ represent the answer label vectors from the student's and question's neighbor sequences, respectively. $\mathbf{W}_{\text{ans}} \in \mathbb{R}^{1 \times d}$ and $\mathbf{b}_{\text{ans}} \in \mathbb{R}^d$ are the learnable projection parameters. We also model the historical cognitive state. Specifically, we differentiate between different cognitive phases, whether an interaction represents initial learning or repeated practice, by computing the structural state indicators for each interaction step i within the sequence. Similarly, we model the question status to indicate whether the question is encountered by the same student or by different students. The x_{ctx}^s and x_{ctx}^q are computed as follows:

$$\mathbf{x}_{ctx}^s = m_{s_i} \mathbf{W}_{st} + \mathbf{b}_{st},\ \mathbf{x}_{ctx}^q = m_{q_i} \mathbf{W}_{st} + \mathbf{b}_{st}, \tag{2}$$

where $m_{q_i} \triangleq \mathbb{I}(q_i = q) \in \{0, 1\}$ is a repetition indicator denoting whether the question appears for the first time, and $m_{s_i} \triangleq \mathbb{I}(s_i = s) \in \{0, 1\}$ indicates whether the historical record belongs to the current student. $\mathbf{W}_{st} \in \mathbb{R}^{1 \times d}$ and $\mathbf{b}_{st} \in \mathbb{R}^d$ are the shared projection parameters.

$$\mathbf{CC}_s = \mathbf{x}_{ans}^s + \mathbf{x}_{ctx}^s,\ \mathbf{CC}_q = \mathbf{x}_{ans}^q + \mathbf{x}_{ctx}^q, \tag{3}$$

where $\mathbf{CC}_s, \mathbf{CC}_q \in \mathbb{R}^d$ contain integrated cognitive-contextual information on question familiarity and student identity consistency.

Temporal Context Encoder. Time intervals between interactions act as a critical proxy for the temporal context of knowledge tracing. Short intervals indicate continuous engagement or hot knowledge states, whereas long intervals often signal potential forgetting. To capture these dynamics,

we propose a temporal context encoder. This mechanism first decouples temporal patterns by interval length, then employs a nonlinear refinement module to adaptively calibrate the temporal representation. Specifically, we bifurcate the processing of time intervals Δt based on a threshold τ. We project the time intervals into a high-dimensional space through two distinct learnable kernels, designed to capture immediate responsiveness and long-term memory decay, respectively:

$$\mathbf{h}_{base} = \phi(\mathbf{W}_S(\Delta t \odot \mathbf{M})) + \phi(\mathbf{W}_L(\Delta t \odot (\mathbf{1} - \mathbf{M}))), \tag{4}$$

where $\mathbf{M} \in \{0, 1\}$ is a mask indicator, with $\mathbf{M} = 1$ if $\Delta t \leq \tau$ and 0 otherwise. $\mathbf{W}_S, \mathbf{W}_L$ are the projection matrices for short-term and long-term patterns, and $\phi(\cdot)$ denotes the activation function. We feed the base embedding $\mathbf{h}_{base}$ into a non-linear transformation block $\mathcal{F}(\cdot)$, which consists of an MLP augmented with GELU activation and layer normalization.

$$\mathcal{F}(\mathbf{h}_{base}) = \text{LayerNorm}(\mathbf{W}_2(\text{Dropout}(\text{GELU}(\mathbf{W}_1\mathbf{h}_{base})))), \tag{5}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times 2d}$ projects the features into a higher-dimensional latent space to capture subtle temporal nuances, and $\mathbf{W}_2 \in \mathbb{R}^{2d\times d}$ projects them back. Finally, to balance the stability of the foundational bi-scale features with the flexibility of the refined context, we employ a learnable gating parameter β to control the injection of the refinement signal:

$$\mathbf{TC}_s = \mathbf{h}_{base} + \beta \cdot \mathcal{F}(\mathbf{h}_{base}). \tag{6}$$

Similarly, we apply the same method to capture temporal information $\mathbf{TC}_q$ for question states.

Prior Context Generator. To address challenges in the cold-start scenario, we employ a concept relationship encoder to learn structure-aware representations from the SKG. We implement the Concept Relationship Encoder (CRE) using a two-layer relational graph convolutional network Ye et al. (2019). The first layer projects node embeddings into a hidden space activated by a ReLU, while the second layer maps them to the final output dimension d. This architecture encodes heterogeneous dependencies and naturally aggregates information for multi-concept questions via relation-specific message passing. For each question q, the encoder aggregates the embeddings of its associated concepts to generate a static prior embedding $\mathbf{x}^q_{prior} \in \mathbb{R}^d$:

$$\mathbf{x}^q_{prior} = \text{CRE}(\mathcal{G}_{static}, q). \tag{7}$$

As the model aggregates dynamic interactions for the question, the learned dynamic representations render the static prior redundant. To prevent over-reliance on priors and facilitate a smooth transition to dynamic learning, we design a lifecycle-fading mechanism. We maintain a global usage counter c_q that tracks the cumulative occurrence count of question q throughout the training process:

$$PC_q = \mathbb{I}(c_q < K_{life}) \cdot \sigma\left(\mathcal{F}(\mathbf{x}^q_{prior})\right) \odot \mathbf{x}^q_{prior}, \tag{8}$$

where $\mathbb{I}(\cdot)$ is the life-cycle indicator function, and $\sigma(\mathcal{F}(\cdot))$ represents a learnable gating network that adaptively regulates the injection intensity.

3.4 Context-Aware Recurrent Aggregators

To capture high-frequency transient cognitive fluctuations and maintain stable gradient flow, we first employ a frequency-domain enhancer to isolate and amplify high-frequency cognitive shifts via spectral reweighting, and then stabilize temporal propagation with a specialized residual normalization architecture, LayerNormGRU.

Frequency-domain Enhancer. Inspired by FNet Lee-Thorp et al. (2022), FEDformer Zhou et al. (2022), and TimesNet Wu et al. (2022a), which demonstrated the efficacy of frequency-domain methods in deep sequence modeling, we propose a frequency-domain enhancer to capture transient dynamics in student learning behaviors. First, we integrate cognitive ($\mathbf{CC}_s$), temporal ($\mathbf{TC}_s$), and raw features, which encode the fundamental attributes of questions and concepts, contexts (F_s) into a unified embedding space. Specifically, all context features are projected to the same hidden dimension d and fused via element-wise summation:

$$\mathbf{X}^{Emb}_s = F_s + \mathbf{CC}_s + \mathbf{TC}_s. \tag{9}$$

To isolate high-frequency transient fluctuations, we transform the embedding into the frequency domain using the Real Fast Fourier Transform (RFFT) computed along the temporal dimension

L, yielding a frequency-domain representation of length $L' = \lfloor L/2 \rfloor + 1$. We specifically focus on the high-frequency band to measure the intensity of sudden state changes. To explicitly filter out low-frequency background trends, we retain only the top 30% of frequency bands (using a high-frequency cutoff ratio $\gamma = 0.7$). We compute a global frequency-domain intensity vector $\bar{\mathbf{h}}$:

$$\bar{\mathbf{h}} = \frac{1}{|\mathcal{F}_h|} \sum_{f \in \mathcal{F}_h} \left| \mathrm{RFFT}(\mathbf{X}_s^{Emb}) \right|, \tag{10}$$

where $\mathcal{F}_h$ denotes the set of high-frequency indices and $|\cdot|$ represents the complex magnitude. This frequency-domain signature is projected to generate a time-domain sensitivity gate $\mathbf{g}$, which adaptively amplifies the original embeddings:

$$\mathbf{g} = \sigma\big(\mathbf{W}_2 \, \mathrm{ReLU}(\mathbf{W}_1 \bar{\mathbf{h}})\big), \; \mathbf{X}_s^{SE} = \mathbf{X}_s^{Emb} \odot \mathbf{g}, \tag{11}$$

where $\odot$ denotes element-wise multiplication. By reweighting inputs by their frequency-domain properties, the model preserves sudden cognitive shifts before the recurrent aggregator. We also provide a detailed theoretical Spectral Learning Analysis in Appendix A.4.

LayerNormGRU. The LayerNormGRU employs a stacked architecture with residual connections.

$$\widehat{\mathbf{H}} = \mathrm{LayerNorm}\big(\mathrm{GRU}(\mathbf{X}_s^{SE})\big), \; \mathbf{X}_s^t = \mathrm{LayerNorm}\big(\mathrm{GRU}(\widehat{\mathbf{H}})\big) + \widehat{\mathbf{H}}, \tag{12}$$

where $\widehat{\mathbf{H}}$ represents the intermediate latent state. The residual connection allows gradients to flow unimpeded through the layers He et al. (2016). This design effectively suppresses gradient decay while preserving the high-frequency details extracted by the frequency-domain enhancer Zhang & Sennrich (2019), ensuring that the final state $\mathbf{X}_s^t$ robustly captures both long-term mastery trends and short-term behavioral nuances. Similarly, the question states that $\mathbf{X}_q^t$ is derived via the same process using F_q, CC_q, TC_q.

3.5 Prediction Module

After obtaining $\mathbf{X}_s^t$ and $\mathbf{X}_q^t$ at timestamp t from the aggregators, the final phase is to forecast the interaction outcome. We formulate this task as estimating the probability of a correct response based on the compatibility between the student's mastery level and the question's difficulty:

$$\hat{y}_t = \sigma\left(\mathbf{W}_{pred} \cdot \phi\big([\mathbf{X}_s^t \oplus \mathbf{X}_q^t]\big) + \mathbf{b}_{pred}\right), \tag{13}$$

where $\oplus$ denotes the concatenation operation, $\sigma(\cdot)$ is the Sigmoid activation function, and $\phi(\cdot)$ represents the hidden layers of the MLP with non-linear activation (e.g., ReLU). $\mathbf{W}_{pred}$ and $\mathbf{b}_{pred}$ are the learnable weights and bias of the output layer. The scalar $\hat{y}_t$ represents the predicted probability that student s answers question q correctly at time t. We optimize all learnable parameters in the network by minimizing the binary cross-entropy loss between the predicted probability $\hat{y}_t$ and the answer r_t as follows:

$$\mathcal{L} = -\sum_{t \in \mathcal{T}} \left(r_t \log \hat{y}_t + (1 - r_t) \log(1 - \hat{y}_t)\right), \tag{14}$$

where $\mathcal{T}$ represents the set of all training interactions. The model is trained end-to-end using the Adam optimizer.

3.6 Computation Cost

Let E be the number of interactions, L be the neighbor sampling size, and d be the hidden dimension.

Time Complexity. The computation of DSCA-KT consists of three parts. Context modulation: linear projections for features operate in $O(E \cdot L \cdot d)$. The frequency-domain enhancer efficiently captures high-frequency dynamics using the RFFT with complexity $O(L \log L)$. The total cost is $O(E \cdot L \log L \cdot d)$. Recurrent Aggregation: The GRU-based aggregation processes sequences in $O(E \cdot L \cdot d^2)$. The overall time complexity is $O(E \cdot L \cdot d^2)$, as the $L \log L$ term is negligible compared to matrix multiplications, which is linear with respect to the dataset scale E.

Space Complexity. Storage overhead includes dynamic interaction features $O(E \cdot d)$ and the static knowledge graph structure ($O(|\mathcal{V}_{SKG}| \cdot d + |\mathcal{E}_{SKG}|)$). Although SKG introduces additional parameters, its size is constant and negligible compared to the continuously growing interaction history ($|\mathcal{E}_{SKG}| \ll E$). Thus, overall memory consumption remains linear, ensuring DSCA-KT's scalability with large datasets. We have compiled a comparison of the methods used by all models in our paper, along with their time and space complexity details in 3.6.

4 EXPERIMENT

To evaluate the performance of our proposed DSCA-KT framework, we conduct extensive experiments to address the following Research Questions (RQs):

RQ1 (Superiority) Can DSCA-KT achieve better prediction performance than SOTA KT models?

RQ2 (Effectiveness) How does each component of DSCA-KT contribute to its overall effectiveness?

RQ3 (Generalization) Does our module enhance cold-start performance across other backbones?

4.1 EXPERIMENTAL SETUP

Datasets and Baselines. We examined five public datasets, including ASSIST12 Feng et al. (2009), ASSIST17, EdNet-KT1 Choi et al. (2020b), MoocRadar Yu et al. (2023a) and XES3G5M Liu et al. (2023b). To evaluate our model's performance, we compared it against 12 state-of-the-art methods, including: i) sequential modeling methods (i.e., DKT Piech et al. (2015), IEKT Long et al. (2021), MIKT Sun et al. (2024)) ii) attention based methods (i.e., AKT Ghosh et al. (2020), simpleKT Liu et al. (2023a), DTransformer Yin et al. (2023)), FlucKT Hou et al. (2025) iii) graph based methods (i.e., GIKT Yang et al. (2021), DASKT Sun et al. (2025)), and iv) dynamic methods (i.e., DIMKT Shen et al. (2022), DyGFormer Yu et al. (2023b), DyGKT Cheng et al. (2024)).

Table 1: Dataset statistics.

Dataset	Student	Question	Concept	Interaction
ASSIST12	29,018	53,091	265	2,711,813
ASSIST17	1,709	2,803	101	864,713
EdNet-KT1	30,000	12,272	188	26,056,109
MoocRadar	14,226	2,513	5,600	12,100,066
XES3G5M	18,066	7,652	865	5,549,184

Evaluation Metrics. To comprehensively evaluate our model's performance in predicting students' binary answers (correct or incorrect), we use two widely used metrics in knowledge tracing Shen et al. (2024): AUC (Area Under the ROC Curve) and AP (Average Precision). This metric measures the model's ability to distinguish between positive (correct) and negative (incorrect) samples. A higher AUC indicates the model's ability to prioritize correct over incorrect predictions across various thresholds. AP measures the model's ability to rank positive samples higher than negative ones, summarizing precision across different recall levels.

Training and Evaluation Settings. We adopt the two evaluation settings established in Cheng et al. (2024): transductive and inductive. In the former setting, all student and question nodes are visible during training. Conversely, in the latter setting, a subset of question nodes and their interactions is withheld from the training set to evaluate the model's generalization to unseen items. In both settings, the dataset is partitioned into training, validation, and test sets at 80%, 10%, and 10%, respectively. To prevent data leakage, the SKG training excludes all unseen question nodes and contains only training questions and concepts. During inference, we derive the prior representation of an unseen question by aggregating the learned embeddings of its associated concepts. We fix the embedding dimension to 256 and the dropout rate to 0.1. The initial learning rate is set to 5×10^{-4} and the weight-decay (L_2) coefficient to 1×10^{-4}. All experiments are conducted on a workstation equipped with an Intel Core i7-14900K CPU and an NVIDIA GeForce RTX 4090 GPU.

4.2 PERFORMANCE COMPARISON (RQ1)

To evaluate the predictive performance of our method, we calculate the mean and standard deviation of AUC and AP across 5 folds and compare them with 12 baseline models. Table 2 shows the overall performance of all models in the five datasets. The results reveal several key observations. i) Our proposed model, DSCA-KT, outperforms all state-of-the-art models on the main evaluation metric, AUC, with an average improvement of 1.12%. On the secondary evaluation metrics, AP, our model outperforms almost all baselines. Sequential, attention-based, and static graph models are trained and evaluated on fixed-length answer records, which leads to performance degradation as sequence length increases. Our model performs well on dynamic graphs, effectively modeling and adapting to changes in students' states and question difficulty, and is more aligned with real long-term learning scenarios. ii) Inductive experiments demonstrate that when students attempt previously unseen questions, model performance does not degrade significantly. This indicates that our context modulator effectively captures foundational conceptual features relevant to the problem domain. Consequently, even when a

Table 2: Performance gains of DSCA-KT over the strongest baseline are statistically significant across all AUC evaluations and for AP in configurations where DSCA-KT achieves superior performance under a two-sided paired t-test ($p < 0.05$).

Setting	Model	AUC (%)					AP (%)				
		ASSIST2012	ASSIST2017	EdNet-KT1	XES3G5M	MoocRadar	ASSIST2012	ASSIST2017	EdNet-KT1	XES3G5M	MoocRadar
Transductive	DKT	68.00±0.81	70.98±0.29	72.08±0.13	54.02±3.08	72.68±0.36	80.71±0.73	60.86±0.62	84.49±0.09	83.30±1.30	84.00±0.28
	AKT	65.57±3.65	68.29±0.26	73.84±0.06	58.09±0.50	74.06±0.73	79.03±2.57	59.66±0.10	87.42±0.04	85.38±0.31	84.84±0.54
	SimpleKT	72.78±1.22	66.67±0.66	77.49±0.10	67.58±0.92	75.18±0.92	83.44±0.81	58.65±1.06	88.53±0.09	89.21±0.39	85.44±0.75
	DTransformer	77.38±0.12	74.31±0.08	74.59±0.01	79.44±0.06	78.21±0.02	84.01±0.05	65.41±0.05	87.87±0.05	91.32±0.03	86.68±0.02
	GIKT	77.54±0.06	72.32±0.18	74.23±0.03	62.75±1.24	70.88±0.87	85.34±0.26	62.38±0.16	87.12±0.11	88.96±0.96	83.65±0.34
	DASKT	78.59±0.06	79.91±0.13	77.67±0.09	76.33±0.03	72.25±0.36	87.77±0.03	71.85±0.18	86.98±0.10	91.35±0.04	84.36±0.25
	IEKT	73.42±0.01	73.10±0.26	76.62±0.12	63.43±0.59	74.12±1.11	83.83±0.00	63.93±0.31	86.41±0.13	87.27±0.16	84.62±0.55
	MIKT	78.34±0.06	77.91±0.03	76.48±0.05	80.72±0.08	79.01±0.07	87.36±0.13	68.15±0.02	84.38±0.21	92.55±0.06	84.94±0.06
	FlucKT	78.46±0.23	75.19±0.28	75.69±0.46	80.45±0.09	79.24±0.38	87.56±0.18	69.35±0.12	84.69±0.14	92.41±0.11	85.01±0.07
	DIMKT	69.14±0.20	74.45±0.32	73.02±0.17	68.91±0.08	74.25±0.06	80.88±0.16	53.95±0.14	87.98±0.03	85.46±0.31	84.68±0.12
	DyGFormer	76.67±0.29	77.48±0.12	78.07±0.10	78.22±0.03	79.68±0.14	86.03±0.19	68.17±0.14	87.26±0.08	93.38±0.02	**87.86±0.14**
	DyGKT	78.04±0.19	79.92±0.11	77.51±0.01	80.82±0.92	75.15±1.37	87.60±0.08	71.81±0.30	87.95±0.03	94.47±0.32	84.38±0.71
	DSCA-KT	**81.45±0.98**	**81.66±0.05**	**79.50±0.05**	**82.56±0.15**	**80.89±0.28**	**88.89±0.56**	**73.34±0.02**	**88.98±0.03**	**95.71±0.08**	87.71±0.17
Inductive	DKT	66.89±0.68	70.85±0.30	68.92±0.20	53.91±2.94	70.09±0.31	80.10±0.62	60.56±0.62	83.04±0.12	84.24±1.27	81.59±0.31
	AKT	64.35±3.87	68.12±0.27	71.27±0.21	58.17±0.39	71.45±0.91	78.37±2.67	59.39±0.11	86.40±0.15	86.32±0.21	82.51±0.68
	SimpleKT	72.45±1.06	66.52±0.64	77.20±0.09	67.67±0.97	72.96±1.15	83.25±0.72	58.43±1.03	88.01±0.08	89.94±0.41	83.37±0.96
	DTransformer	77.96±0.19	75.02±0.07	74.81±0.03	79.01±0.05	78.01±0.03	84.12±0.02	65.89±0.02	87.33±0.04	90.97±0.05	86.18±0.21
	GIKT	77.01±0.01	71.30±1.36	74.85±0.04	60.75±1.01	70.32±0.36	84.89±0.10	62.01±0.32	87.31±0.15	86.81±0.63	82.73±0.56
	DASKT	79.93±0.09	80.13±0.10	77.02±0.07	75.93±0.12	72.03±0.41	87.80±0.04	71.96±0.11	86.32±0.19	91.01±0.09	84.18±0.29
	IEKT	72.93±0.01	73.04±0.28	76.11±0.10	63.63±0.76	71.76±1.56	83.52±0.01	63.76±0.33	85.79±0.13	88.19±0.21	82.53±1.00
	MIKT	77.21±0.21	76.23±0.14	76.01±0.18	79.92±0.02	79.03±0.12	86.46±0.08	68.01±0.44	83.59±0.35	91.58±0.03	84.86±0.16
	FlucKT	78.03±0.54	74.82±0.29	75.61±0.41	80.27±0.14	79.16±0.33	87.13±0.15	69.31±0.22	84.22±0.16	92.40±0.20	84.87±0.07
	DIMKT	69.02±0.27	74.01±0.33	72.98±0.14	68.92±0.13	74.28±0.07	80.39±0.11	53.66±0.11	87.53±0.09	85.48±0.21	84.29±0.16
	DyGFormer	76.98±0.25	77.59±0.12	77.57±0.13	78.36±0.04	79.40±0.21	86.23±0.17	68.22±0.13	87.16±0.15	93.73±0.02	87.16±0.21
	DyGKT	78.12±0.33	79.78±0.20	77.46±0.01	80.93±0.86	74.75±1.46	87.73±0.08	71.84±0.31	88.12±0.05	94.77±0.07	83.56±0.80
	DSCA-KT	**81.07±0.86**	**81.38±0.11**	**79.01±0.11**	**82.36±0.13**	**80.72±0.26**	**89.02±0.73**	**73.65±0.01**	**88.88±0.06**	**95.86±0.04**	**87.98±0.15**

new question involves knowledge concepts absent from a student's prior interaction history, the model can still generate accurate predictions by leveraging these generalized contextual representations.

4.3 Ablation Study (RQ2)

To assess the effectiveness of each design component in the DSCA-KT model, we constructed four variants for ablation experiments. Specifically, for variant DSCA-KT w/o FE and w/o LNG, we removed the frequency-domain enhancer and LayerNormGRU, respectively. Instead, we aggregate using a single GRU. For variant DSCA-KT w/o TC, we removed the nonlinear transformation block and retained only the bi-scale interval projection. For variant DSCA-KT w/o PC, we removed the concept relationship encoder, allowing the model to aggregate from scratch without prior knowledge. We evaluated the performance of different variants on the five datasets mentioned above and presented the results in Figure 3. Compared with the complete DSCA-KT model, most ablation variants perform worse across the five datasets. Specifically, the variant without LNG also decreases performance, and the standard deviation increases. This indicates that LayerNormGRU is crucial for stabilizing training dynamics and ensuring consistent model performance. The variant without PC shows a slight decrease, but the model remains competitive. This indicates that our superior performance in the inductive setting is not solely due to access to external static graph data. Even without these priors, the core architecture effectively captures the characteristics of concepts, demonstrating the robustness of our framework relative to baselines that do not use static graphs.

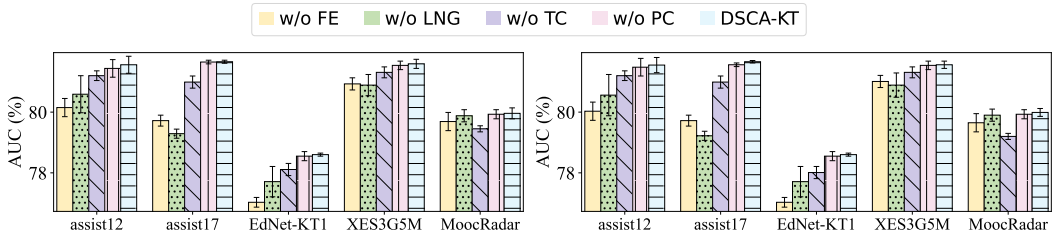


Figure 3: AUC for inductive and transductive ablation study on 5 datasets. .

To further demonstrate the effectiveness of the frequency-domain enhancer, we conducted a systematic analysis showing it can isolate meaningful cognitive events from noise. Specifically, we analyze the

average spectral distribution and gating activations across 1,000 randomly sampled test sequences from the ASSIST12 dataset. As shown in Figure 4, our learned gating mechanism does not uniformly amplify the high-frequency band. Instead, it exhibits highly selective, non-linear filtering: For the low- to mid-frequency components, the average gating values consistently remain high, ensuring the baseline cognitive state is uncorrupted. Within the high-frequency band, the average gating weight for general background fluctuations is strongly suppressed to 0.12, effectively filtering out meaningless noise. In contrast, specific high-frequency indices that strongly correlate with sharp cognitive transitions trigger localized spikes in gating activations with an average of 0.84. To investigate how the FE responds to different temporal interaction patterns, we conduct a controlled spectral perturbation analysis using the synthetic interaction sequences in Appendix A.5. We also present a case study in Appendix A.3 to demonstrate the effectiveness of our method. Overall, each module in the DSCA-KT model plays a critical role.

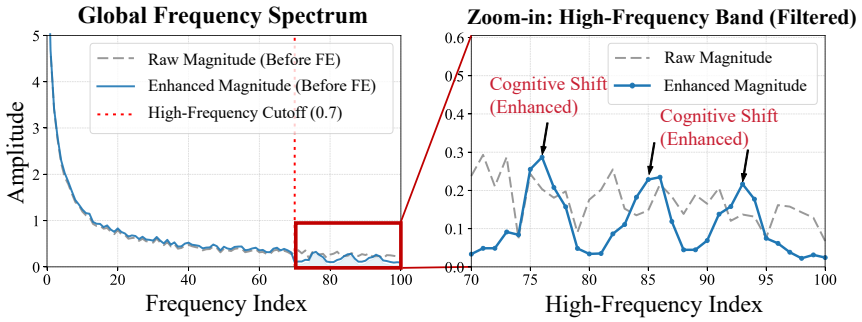


Figure 4: Average spectral distribution and gating activations across 1,000 sampled test sequences.

4.4 Generalization Study (RQ3)

To verify the generalization of our method, we extended the prior context generator to three representative baselines. Specifically, we integrated the prior context (PC) into these backbones by fusing the structure-aware concept embeddings generated by our static graph encoder with their original item initializations. Figure 5 illustrates the performance comparison over the first 12 interactions with new questions. In the initial phase, variants equipped with the PC module consistently outperform their original counterparts. This indicates that the static knowledge graph effectively compensates for the lack of historical interaction data, providing a high-quality warm start for the models. These results confirm that our prior context generator can be integrated into various KT architectures to significantly boost their performance in cold-start scenarios without altering their core mechanisms.

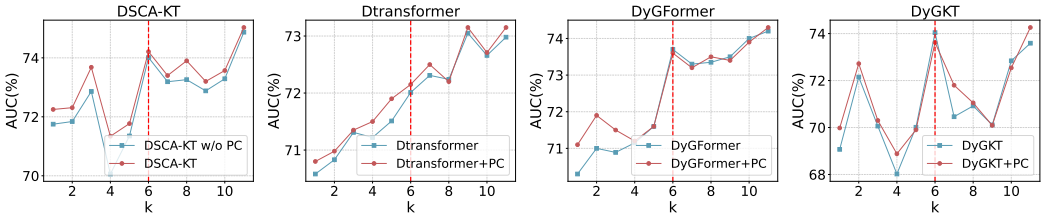


Figure 5: Comparison between models with and without prior context in terms of AUC for the first k interactions with a new question.

5 Conclusion

This paper proposed DSCA-KT, a dual-state context-aware model for dynamic knowledge tracing. To address the limitations of existing models in capturing transient cognitive fluctuations within dynamically growing sequences and characterizing student interaction history and knowledge dependencies in question-side cold-start scenarios, we designed context-aware recurrent aggregators and a multi-faceted context modulator. Experimental results confirmed the effectiveness of our method. In future work, we plan to deploy the DSCA-KT framework in real-world intelligent tutoring systems to further validate and enhance its robustness across diverse pedagogical environments.

REFERENCES

Ghodai Abdelrahman and Qing Wang. Knowledge tracing with sequential key-value memory networks. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 175–184, 2019.

Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

Deepak Agarwal, Ryan S Baker, and Anupama Muraleedharan. Dynamic knowledge tracing through data driven recency weights. In *Proceedings of the International Conference on Educational Data Mining (EDM)*, 2020.

Ke Cheng, Linzhi Peng, Pengyang Wang, Junchen Ye, Leilei Sun, and Bowen Du. Dygkt: Dynamic graph learning for knowledge tracing. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 409–420, 2024.

Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, and Jaewe Heo. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the ACM conference on Learning @ Scale (L@S)*, pp. 341–344, 2020a.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. Ednet: A large-scale hierarchical dataset in education. In *Proceedings of International Conference on Artificial Intelligence in Education (AIED)*, pp. 69–73. Springer, 2020b.

Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994.

Chaoran Cui, Yumo Yao, Chunyun Zhang, Hebo Ma, Yuling Ma, Zhaochun Ren, Chen Zhang, and James Ko. Dgekt: A dual graph ensemble learning method for knowledge tracing. *ACM Transactions on Information Systems*, 42(3):1–24, 2024.

M. Feng, N. T. Heffernan, and K. R. Koedinger. Addressing the assessment challenge in an intelligent tutoring system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19: 243–266, 2009.

Aritra Ghosh, Neil Heffernan, and Andrew S Lan. Context-aware attentive knowledge tracing. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2330–2339, 2020.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Mingliang Hou, Xueyi Li, Teng Guo, Zitao Liu, Mi Tian, Renqiang Luo, and Weiqi Luo. Cognitive fluctuations enhanced attention network for knowledge tracing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 14265–14273, 2025.

Amar Lalwani and Sweety Agrawal. What does time tell? tracing the forgetting curve using deep knowledge tracing. In *Proceedings of International Conference on Artificial Intelligence in Education (AIED)*, pp. 158–162. Springer, 2019.

Jinseok Lee and Dit-Yan Yeung. Knowledge query network for knowledge tracing: How knowledge interacts with skills. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pp. 491–500, 2019.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 4296–4313, 2022.

Jiawei Li, Shun Mao, Yixiu Qin, Feng Wang, and Yuncheng Jiang. Hyperbolic hypergraph transformer with knowledge state disentanglement for knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, 37(8):4677–4690, 2025.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simplekt: a simple but tough-to-beat baseline for knowledge tracing. *arXiv preprint arXiv:2302.06881*, 2023a.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. In *Proceedings of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 32958–32970, 2023b.

Ting Long, Yunfei Liu, Jian Shen, Weinan Zhang, and Yong Yu. Tracing knowledge state with individual cognition and acquisition estimation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 173–182, 2021.

Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT)*, pp. 156–163, 2019.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 505–513, 2015.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 5301–5310, 2019.

Shuanghong Shen, Zhenya Huang, Qi Liu, Yu Su, Shijin Wang, and Enhong Chen. Assessing student's dynamic knowledge state by exploring the question difficulty effect. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 427–437, 2022.

Shuanghong Shen, Qi Liu, Zhenya Huang, Yonghe Zheng, Minghao Yin, Minjuan Wang, and Enhong Chen. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, 17:1858–1879, 2024.

Jianwen Sun, Fenghua Yu, Qian Wan, Qing Li, Sannyuya Liu, and Xiaoxuan Shen. Interpretable knowledge tracing with multiscale state representation. In *Proceedings of the ACM Web Conference (WWW)*, pp. 3265–3276, 2024.

Xinjie Sun, Kai Zhang, Qi Liu, Shuanghong Shen, Fei Wang, Yuxiang Guo, and Enhong Chen. Daskt: A dynamic affect simulation method for knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1714–1727, 2025.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022a.

Zhengyang Wu, Li Huang, Qionghao Huang, Changqin Huang, and Yong Tang. Sgkt: Session graph-based knowledge tracing for student performance prediction. *Expert Systems with Applications*, 206:117681, 2022b.

Xiong Xiao, Shengyingjie Liu, Yue Li, Xiuling He, Jing Fang, and Yangyang Li. Information processing dynamic graph for knowledge tracing. *Information Processing & Management*, 63(1): 104331, 2026.

Yang Yang, Jian Shen, Yanru Qu, Yunfei Liu, Kerong Wang, Yaoming Zhu, Weinan Zhang, and 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 (ECML-PKDD)*, pp. 299–315. Springer, 2021.

Rui Ye, Xin Li, Yujie Fang, Hongyu Zang, and Mingzhong Wang. A vectorized relational graph convolutional network for multi-relational network alignment. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4135–4141, 2019.

Chun-Kit Yeung. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*, 2019.

Chun-Kit Yeung and Dit-Yan Yeung. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Annual ACM Conference on Learning at Scale (L@S)*, pp. 1–10, 2018.

Yu Yin, Le Dai, Zhenya Huang, Shuanghong Shen, Fei Wang, Qi Liu, Enhong Chen, and Xin Li. Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer. In *Proceedings of the ACM Web Conference (WWW)*, pp. 855–864, 2023. ISBN 9781450394161. doi: 10.1145/3543507.3583255. URL https://doi.org/10.1145/3543507.3583255.

Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, Juanzi Li, and Jie Tang. Moocradar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in moocs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 2924–2934, New York, NY, USA, 2023a. doi: 10.1145/3539618.3591898. URL https://doi.org/10.1145/3539618.3591898.

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. In *Proceedings of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 67686–67700, 2023b.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, (NeurIPS)*, pp. 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 765–774, 2017.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 27268–27286, 2022.

A Appendix

A.1 Parameter Sensitivity

The time-interval threshold τ and the high-frequency cutoff are key parameters because they capture essential signals for cognitive evolution and affect how we model dynamic learning contexts. In Figure 6, we examine how sensitive model performance is to variations in the time-interval threshold τ and the high-frequency cutoff on the ASSIST17 dataset in the transductive setting. For the time-interval threshold τ, both AUC and AP increase as τ grows, reaching their maximum at 24 hours, after which performance progressively declines. This is because the 24 hours balances the immediate reinforcement of hot knowledge with the slower decay of memory over time, consistent with the notion of cognitive evolution. For the high-frequency cutoff, predictive performance improves steadily as the truncation ratio increases, but falls off sharply once it exceeds 0.7. This performance drop occurs because a 0.7 cutoff effectively captures non-steady cognitive transitions, such as abrupt insights or mistakes, while a higher cutoff begins to remove crucial signal characteristics. Therefore, we set the time-interval threshold τ to 24 hours and the high-frequency cutoff to 0.7 in our experiments.

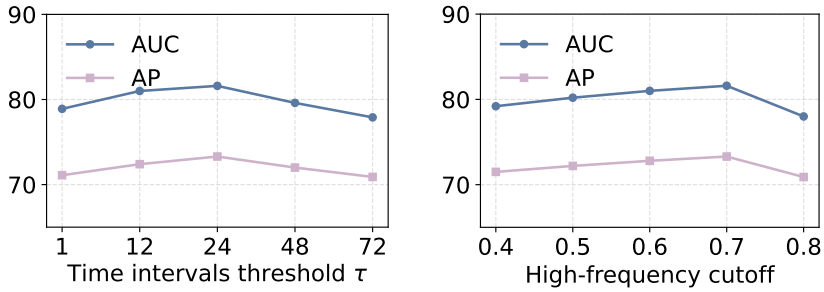


Figure 6: Performance comparison of different parameter settings.

A.2 Computation Cost Comparison

To fully verify the complexity of our method, we compare DSCA-KT to the top 10 baselines in Table 3. Our approach achieves a good balance between efficiency and performance. Our method introduces additional modules that improve modeling capacity while maintaining near-equivalent efficiency. Compared to more expressive models, our approach achieves both better predictive performance and significantly lower computational overhead.

Table 3: Details and computation cost for ASSIST12.

Model	Params (MB)	GPU Mem (MB)	Training Time (s / epoch)	Inference Time (s / epoch)	AUC (Trans)
DKT	6.483	247.952	48.873	4.852	68.00
AKT	10.080	565.781	73.864	5.986	65.57
SimpleKT	3.634	437.251	79.267	6.554	72.78
DTransformer	9.138	303.547	78.367	6.452	77.38
GIKT	11.696	345.215	81.347	7.254	77.54
DASKT	11.023	280.265	78.385	6.254	78.59
IEKT	12.452	230.336	197.347	15.254	73.42
MIKT	7.632	345.215	62.339	5.054	78.34
FlucKT	10.651	525.012	79.431	6.156	78.46
DyGFormer	3.811	493.895	584.738	44.826	76.67
DyGKT	0.292	240.965	67.852	5.125	78.04
DSCA-KT (Ours)	1.193	260.352	70.637	5.651	81.45

A.3 Case Study

Figure 7 demonstrates the ability of the context modulator to disentangle structural dependencies from temporal noise. While the student's performance fluctuates, the underlying knowledge state remains logically consistent with the concept prerequisites. It confirms that DSCA-KT does not blindly fit the data; instead, it understands the cognitive structure governed by the context modulator. In particular, on question q_6, the student answered incorrectly. But DSCA-KT maintains a non-reduced mastery score. This indicates that our method successfully identified this interaction as a high-frequency noise rather than a true decline in knowledge state, thereby preserving the stability of the long-term trend.

A.4 Theory assumptions and proofs

We provide a spectral analysis to theoretically justify the necessity of the proposed Frequency-domain Enhancer. Specifically, we prove that standard recurrent aggregators inherently act as low-pass filters, inadvertently smoothing out transient cognitive fluctuations. Subsequently, we demonstrate how the frequency-domain gating mechanism in DSCA-KT provides an adaptive high-frequency gain to overcome this spectral bias. This analysis should be interpreted as a local spectral approximation rather than an exact characterization of nonlinear GRU dynamics.

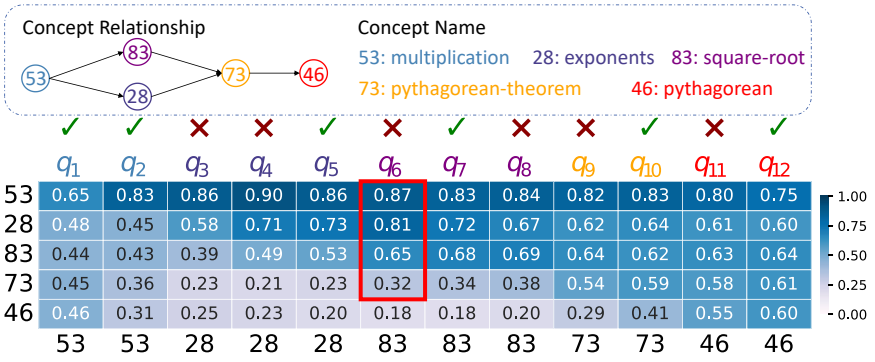


Figure 7: An example of a student's knowledge mastery degree of five concepts over 12 questions in ASSIST17. There are prerequisite relationships among the concepts. Questions that share the same concept are marked with the same color.

A.4.1 The Low-Pass Filtering Tendency of Recurrent Aggregators

Theorem 1. A standard recurrent neural network (RNN/GRU) inherently behaves as a low-pass filter, attenuating high-frequency signals associated with transient cognitive shifts.

Proof. Let the hidden state update of a generalized recurrent aggregator be defined as $h_t = \sigma(Wx_t + Uh_{t-1} + b)$, where $\sigma(\cdot)$ is a non-linear activation function. To conduct a spectral analysis, we linearize the system around a local equilibrium point using a first-order Taylor expansion. The system can be approximated as a discrete-time Linear Time-Invariant system:

$$h_t \approx Ax_t + Bh_{t-1}, \tag{15}$$

where $A = \frac{\partial h_t}{\partial x_t}$ and $B = \frac{\partial h_t}{\partial h_{t-1}}$ are the corresponding Jacobian matrices. Applying the Discrete-Time Fourier Transform (DTFT) to both sides yields:

$$H(\omega) = AX(\omega) + Be^{-j\omega}H(\omega). \tag{16}$$

Rearranging this equation, we obtain the frequency-domain transfer function $T(\omega)$ of the recurrent aggregator:

$$T(\omega) = \frac{H(\omega)}{X(\omega)} = (I - Be^{-j\omega})^{-1}A. \tag{17}$$

To evaluate the frequency response, we analyze the spectral norm $\|T(\omega)\|_2$. To ensure stable gradient propagation over long sequences and capture long-term memory, the dominant eigenvalues λ_i of the state transition matrix B are typically regularized or optimized to be close to 1 (i.e., $\lambda_i \to 1^-$). - Low-Frequency Response ($\omega \to 0$):** As the frequency approaches zero, $e^{-j\omega} \to 1$. The denominator term becomes $(I - B)$. Since $\lambda_i \approx 1$, the eigenvalues of $(I - B)$ approach 0, maximizing the norm of its inverse:

$$\|T(0)\|_2 \propto \frac{1}{|1 - \lambda_i|} \gg 1. \tag{18}$$

This indicates that low-frequency signals (representing stable, long-term knowledge accumulation) are passed with high gain.

High-Frequency Response ($\omega \to \pi$): As the frequency approaches the Nyquist limit π, $e^{-j\omega} \to -1$. The denominator term transitions to $(I + B)$:

$$\|T(\pi)\|_2 \propto \frac{1}{|1 + \lambda_i|} \approx \frac{1}{2}. \tag{19}$$

Compared to the low-frequency response, the gain for high-frequency components is drastically attenuated. Since $\|T(0)\|_2 \gg \|T(\pi)\|_2$, the transfer function $T(\omega)$ mathematically functions as a low-pass filter. Consequently, sudden, high-frequency cognitive fluctuations (e.g., moments of insight or careless mistakes) are irreversibly smoothed out during temporal aggregation.

A.4.2 Adaptive Spectral Compensation via DSCA-KT Gating

Proposition. The Frequency-domain Enhancer in DSCA-KT provides a monotonic, adaptive gain $G(\omega)$ for informative high-frequency bands, effectively compensating for the spectral bias of recurrent aggregators.

In the DSCA-KT framework, the input embedding is modulated by the Frequency-domain Enhancer before entering the recurrent aggregator:

$$X_s^{SE} = X_s^{Emb} \odot g,\ \overline{h} = \frac{1}{|\mathcal{F}_h|} \sum_{\omega_f \in \mathcal{F}_h} |X(\omega_f)|, \tag{20}$$

where the sensitivity gate is defined as $g = \sigma(W_2 \text{ReLU}(W_1 \overline{h}))$, and $\overline{h}$ represents the mean magnitude of the high-frequency band $\mathcal{F}_h$. Let an abrupt cognitive shift (e.g., mastering a concept suddenly) be modeled as a time-domain impulse or step function ΔX_{HF}. According to Fourier theory, such transient signals inherently inject significant energy into the high-frequency spectrum, causing $|X(\omega_{HF})| \uparrow$ for $\omega_{HF} \in \mathcal{F}_h$.

Monotonicity of the Gate: An increase in $|X(\omega_{HF})|$ strictly increases the mean magnitude $\overline{h}$. Because the network is optimized end-to-end via Binary Cross-Entropy (BCE) to predict performance outcomes precisely, the projection weights W_1 and W_2 learn to preserve these critical cognitive shifts. Therefore, the partial derivative of the gate with respect to the high-frequency magnitude is positive for informative signals. This establishes g as an adaptive, signal-dependent gain $G(\omega)$:

$$\frac{\partial g}{\partial \overline{h}} > 0. \tag{21}$$

Spectral Compensation: The end-to-end transfer function of the aggregated system can be approximated as:

$$T_{DSCA}(\omega) \approx T_{RNN}(\omega) \cdot \text{diag}(g(\overline{h})). \tag{22}$$

When an informative transient behavior occurs ($\omega \in \mathcal{F}_h$), the gate g spikes. This localized amplification $G(\omega_{HF})$ directly counteracts the high-frequency attenuation factor $\frac{1}{|1+\lambda_i|}$ of $T_{RNN}(\omega)$. Therefore, rather than indiscriminately amplifying all noise, the Frequency-domain Enhancer acts as an adaptive spectral compensator. It precisely identifies and amplifies meaningful high-frequency cognitive shifts before low-pass aggregation, ensuring that transient learning behaviors are preserved alongside long-term knowledge trends.

A.5 Mechanistic Interpretability via Controlled Spectral Perturbation

We construct four illustrative synthetic interaction sequences (length = 12) corresponding to a single knowledge concept, where each sequence represents a distinct temporal learning pattern:

- Stable Mastery: consecutive correct responses, representing a stable mastered state;
- Careless Mistake (Slip): a single incorrect response embedded within correct responses, representing a transient anomaly;
- Sudden Mastery: a sharp transition from consecutive incorrect to consecutive correct responses, representing an abrupt state change;
- Random Noise: alternating correct and incorrect responses, representing highly unstable or unstructured behavior.

For each sequence, we visualize the interaction sequence, the magnitude of the localized spectral perturbation, the activation of the frequency gate g, and the resulting prediction dynamics $\hat{y}_{t+1}$. We further compare DSCA-KT against a baseline recurrent model (MIKT) without the FE module. This experiment aims to provide mechanistic evidence supporting the spectral interpretation.

As shown in Figure 8, for Stable Mastery, negligible spectral perturbation leads to consistently low gate activation, and both models maintain stable high-confidence predictions. For Careless Mistake, the localized perturbation spike causes the baseline RNN to exhibit a noticeable drop in predictions followed by a gradual recovery, while DSCA-KT shows substantially smaller fluctuations, indicating improved robustness to transient anomalies. For Sudden Mastery, the abrupt transition

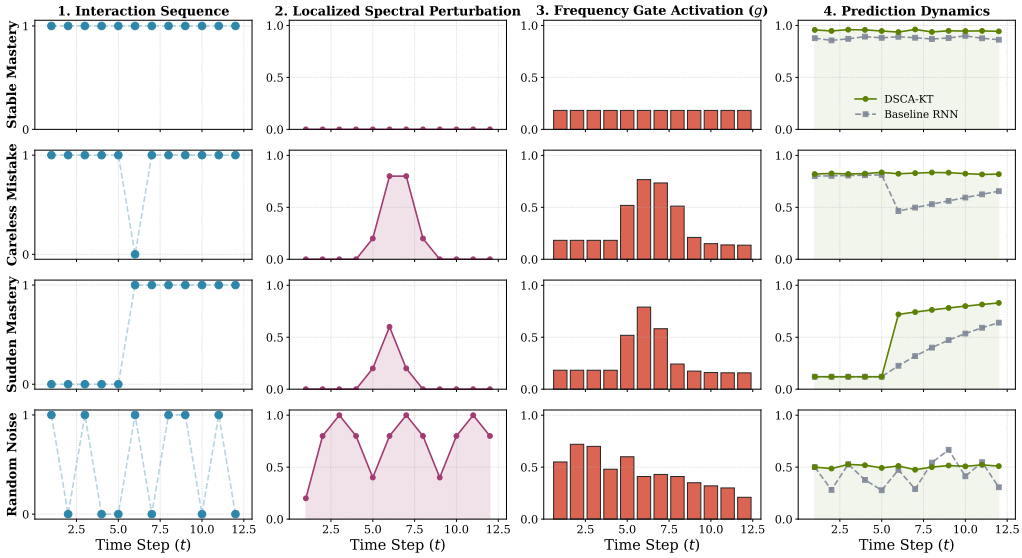


Figure 8: Controlled Spectral Perturbation Analysis.

strongly activates the frequency gate, enabling DSCA-KT to adapt more rapidly and produce a sharper prediction increase than the baseline RNN, suggesting reduced temporal oversmoothing. For Random Noise, DSCA-KT maintains relatively stable medium-confidence predictions, whereas the baseline RNN exhibits larger oscillations. This suggests that the gating mechanism dynamically regulates persistent unstable perturbations rather than indiscriminately amplifying all high-frequency signals.

Overall, these observations provide empirical evidence that the proposed FE module can distinguish between transient structured state changes and persistent unstructured fluctuations, thereby improving the responsiveness and robustness of recurrent KT dynamics.

A.6 Static Knowledge Graph Construction

The five benchmark datasets provide question-concept associations, but they do not directly provide the concept-concept prerequisite edges used in our experiments. We therefore constructed these edges using a two-stage, LLM-assisted annotation procedure.

LLM-Assisted Candidate Generation. In the first stage, we provided the LLM with the official concept names and available descriptions from each dataset and asked it to propose directed prerequisite candidates ($c_i \rightarrow c_j$). This stage used only concept metadata; no student responses, correctness labels, validation/test interactions, or model predictions were provided to the LLM. For this stage, we utilized GPT-4o (version: gpt-4o-2024-05-13) with the temperature set to 0 to ensure deterministic and reproducible outputs.

Human Verification In the second stage, three annotators reviewed the candidates according to a predefined criterion: an edge was retained only when mastery of c_i was judged to be pedagogically necessary or strongly expected before learning c_j. The annotators could remove unsupported candidates and add missing relations. Disagreements were resolved by voting. We acknowledge that prerequisite annotation may still involve pedagogical judgment and can vary across curricula; therefore, we do not treat these graphs as universally definitive.

Table 4: Dataset Statistics for Concept Prerequisite Graphs.

Dataset	Field	Exact Metadata Source	Concept Num	Final Edge Count
Assist2012	K-12 math	Official `ml_assist12_node.npy : skill`	265	225
Assist2017	K-12 math	Official `node.csv : node`	101	63
EdNet-KT1	English	Official `questions.csv : tags`	188	138
MoocRadar	Science courses	Official `kc_routes_map.json : cid`	5600	386
XES3G5M	Primary math	Official `concept_id_map.csv : mapped_id` + text	865	314

The table 4 summarizes the dataset-specific information used to construct the graphs. MoocRadar has a substantially larger and sparser concept vocabulary than the other datasets. The reported 5,600 concepts refer to the complete vocabulary, while prerequisite candidate generation was performed on the 1,457 concepts associated with at least three interactions.

Prompt Used for Candidate Generation System Role: You are an expert in educational domain knowledge, pedagogy, and knowledge graph engineering. Your task is to identify directed pedagogical prerequisite relationships between educational concepts based solely on their metadata.

Task Definition: Given a list of concepts with their names and descriptions, analyze all pairwise combinations and identify candidate directed prerequisite edges ($c_i \rightarrow c_j$). A directed prerequisite edge ($c_i \rightarrow c_j$) signifies that mastering concept c_i is logically or pedagogically necessary before learning concept c_j.

Evaluation Criteria: When determining if ($c_i \rightarrow c_j$) exists, strictly apply the following standards:

1. Pedagogical Precedence: Does learning c_j require foundational knowledge, definitions, or skill sets defined in c_i?
2. Asymmetry: Prerequisite relationships are strictly asymmetric. If $c_i \rightarrow c_j$, then $c_j \rightarrow c_i$ cannot exist.
3. Direct Dependency: Focus on primary candidate edges where c_i directly contributes to understanding c_j.
4. Metadata Grounding: Base your decision strictly on the logical meaning conveyed by the concept names and descriptions. Do not assume external unmentioned contexts.

A.7 BROADER IMPACTS

This work contributes to the development of dynamic knowledge tracing by improving robust cognitive state modeling on continuously growing interaction sequences. The proposed DSCA-KT framework can be applied to low-resource educational scenarios, such as newly deployed intelligent tutoring systems, personalized learning path recommendations, and adaptive assessment platforms. It helps reduce the reliance on extensive historical interaction data by leveraging structural priors, thereby improving the usability and generalization ability of knowledge tracing models in practical, cold-start applications. Since this work primarily focuses on methodological and technical improvements, and the public benchmark datasets used in our experiments do not involve sensitive personal information or pose privacy risks, we do not expect the proposed method to have any obvious negative societal impacts.