CogWave-KT: Multiscale Cognitive Volatility Modeling for Knowledge Tracing

Dawei Tan[1], **Zhenqiang Yu**[1], **Yuncheng Jiang**[1,2,3,*]
[1]School of Computer Science, South China Normal University, Guangzhou, 510631, China
[2]Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents of Ministry of Education (South China Normal University), Guangzhou, 510631, China
[3]School of Artificial Intelligence, South China Normal University, Foshan, 528225, China
[*]Corresponding author:ycjiang@scnu.edu.cn

Abstract

Knowledge Tracing (KT) aims to predict students' future performance based on their historical interaction data. With the advancement of attention mechanisms, attention-based KT models have achieved significant improvements in predictive performance. However, most existing attention-based KT models typically employ deterministic cognitive embeddings to represent students' knowledge states. Although such representations effectively capture the stability of cognitive processes, they fail to account for the inherent volatility in students' learning behaviors, thereby limiting the capacity for modeling authentic learning dynamics. To address this limitation, we propose a novel knowledge tracing model with enhanced capability for modeling cognitive volatilities—CWKT. Specifically, we model students' interaction representations as Gaussian distributions to capture the intrinsic long-term volatility in the learning process. We further apply a frequency decomposition to the mean of the Gaussian distribution, enabling us to extract local volatility information. To model the distributional transitions of students' knowledge states during learning, we introduce a Wasserstein Self-Attention mechanism. Moreover, an attention penalty module is incorporated to mitigate the model's potential overemphasis on long-term volatility, thereby improving the overall stability of predictions. Extensive experiments conducted on four public educational datasets demonstrate that the proposed model exhibits significant advantages in capturing cognitive volatilities.

Keywords: Knowledge Tracing, Gaussian, Wasserstein

Introduction

In the era of web-based learning, where large-scale online education platforms such as Coursera and other MOOCs have gained substantial traction, Knowledge Tracing (KT) has become a pivotal research area for modeling and understanding learner behavior. As a fundamental task in educational data mining and intelligent tutoring systems, KT seeks to infer and predict students' mastery levels of specific knowledge components in real time throughout the learning process. With the rapid evolution of online education platforms and intelligent learning technologies, vast amounts of learner interaction data are being continuously generated, offering a robust foundation for advancing personalized and adaptive education. However, how to efficiently and accurately extract cognitive patterns and learning behaviors from such complex data remains a critical challenge in the field of educational artificial intelligence. It is important to note that learning is inherently a long-term, dynamic process characterized by volatility. During each interaction, students may exhibit significant non-stationarity—commonly referred to as volatility—due to factors such as volatility in cognitive states, emotional changes, shifts in learning strategies, or knowledge transfer, i.e., the ability to apply previously acquired knowledge, skills, or strategies to novel contexts, tasks, or problems. This interaction-level volatility is not only prevalent but also plays a crucial role in accurately modeling students' knowledge states. Therefore, capturing such learning volatility from complex sequential interaction data has become a key challenge in enhancing both the performance and interpretability of KT models.

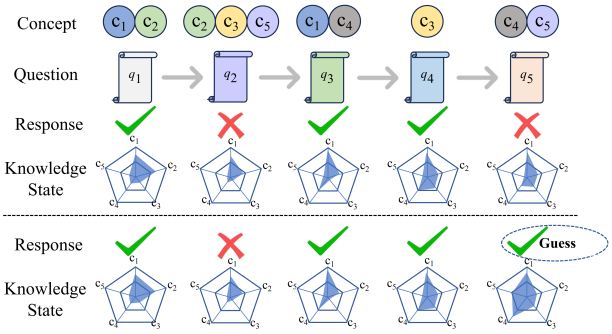

Figure 1: Impact of local volatility on knowledge state estimation.

In recent years, several prominent Knowledge Tracing (KT) models have been proposed, including LPKT (Shen et al., 2021), ATKT (Guo et al., 2021), and simpleKT (Liu, Liu, Chen, Huang, & Luo, 2023), which leverage memory mechanisms, attention mechanisms, and generalization strategies to advance model architecture and enhance predictive performance. However, these methods commonly assume a smooth temporal evolution of students' cognitive states. They typically adopt static or fixed cognitive embeddings to represent student knowledge, thereby limiting their capacity to capture and exploit the rich, dynamic information embedded in individual interactions. Consequently, such models struggle to accurately reflect subtle variations in concept understanding, diverse problem-solving strategies, and transient phenomena such as careless errors, lucky guesses, or conceptual confusion—ultimately hindering their ability to model complex learning dynamics and attain higher predictive accuracy.

In real-world educational scenarios, students' responses are influenced by a variety of complex factors, leading to

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

volatilities in the estimation of their knowledge states. These volatilities can be attributed to both long-term and short-term sources. On the one hand, long-term volatility may reflect differences in students' conceptual understanding, shifts in problem-solving strategies, or the transfer of knowledge across questions. On the other hand, short-term volatility is induced by recent and occasional behaviors, such as answering incorrectly due to carelessness (occasional errors) or getting a correct answer by chance without actual mastery (lucky guesses), as depicted in Figure 1. These multidimensional sources of volatility impose more stringent requirements on the accurate modeling of learners' knowledge states and pose substantial challenges to the robustness and adaptivity of KT models.

To address the aforementioned challenges, we propose CWKT, a novel knowledge tracing model that captures multi-scale volatility in students' interaction behaviors. In CWKT, each interaction is represented as a Gaussian embedding whose mean encodes the student's baseline cognitive level and whose covariance quantifies both long-term volatility. Figure 2 provides an illustrative comparison between conventional deterministic embeddings and the proposed Gaussian representations, demonstrating how Gaussian embeddings more effectively capture volatility in learning dynamics under varying degrees of interaction volatility.

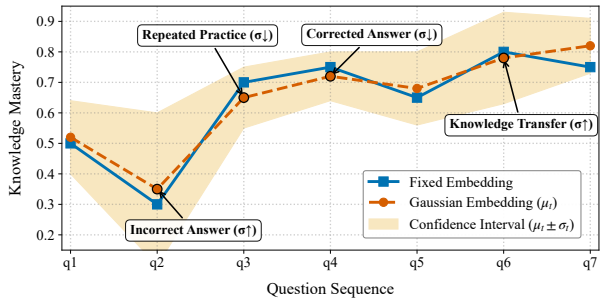

Figure 2: Comparison of fixed and gaussian embeddings

To address local volatility in the learning process—such as occasional mistakes or lucky guesses—CWKT employs a frequency decomposition technique (van den Oord et al., 2016). This method extracts dynamic cognitive features across multiple frequency bands from students' sequential interactions, thereby enabling a more com- prehensive characterization of local cognitive volatility. Our main contributions are summarized as follows:

- We propose modeling students' interaction behaviors as Gaussian distributions, where the mean and covariance, respectively, capture the baseline cognitive level and long-term cognitive volatility. Furthermore, we propose an frequency decomposition to extract fine-grained local cognitive volatility from sequential interactions.
- We design an attention penalty mechanism that guides the model to appropriately balance attention between long-term and local volatility during training, thereby enhancing its ability to model students' dynamic cognitive states with improved robustness.
- We conduct extensive experiments on four public educational datasets, demonstrating that our proposed model outperforms several strong baseline models in prediction performance. Additional analyses further validate the model's effectiveness in capturing volatility within student interactions.

Related Work

Prior to the advent of deep learning, Knowledge Tracing (KT) was predominantly modeled using probabilistic and Bayesian approaches (Corbett & Anderson, 1994; Pavlik et al., 2009; Yudelson et al., 2013), which represented students' mastery states as latent variables and updated them through principled probabilistic inference. With the rise of deep learning, KT models have evolved along several major directions. Sequential models, such as DKT (Piech et al., 2015) and its variants (Liu, Liu, Chen, Huang, Gao, et al., 2023; Shen et al., 2021; Yeung & Yeung, 2018), formulate KT as a sequence modeling problem and encode students' evolving knowledge states using recurrent or related architectures. Memory-augmented models, including DKVMN (Zhang et al., 2017) and its extensions (Abdelrahman & Wang, 2019, 2022), explicitly maintain external memory slots to track mastery over latent concepts. Graph-based models (Nakagawa et al., 2019; H. Tong et al., 2022; S. Tong et al., 2020) leverage graph neural networks to model structured dependencies among concepts and propagate knowledge states across the concept graph. More recently, attention-based and Transformer-style models, such as SAKT (Pandey & Karypis, 2019) and SAINT (Choi et al., 2020), have employed self-attention mechanisms to capture long-range dependencies in interaction sequences, with subsequent works incorporating temporal biases to better handle sequence sparsity and irregularity (Liu, Liu, Chen, Huang, & Luo, 2023; Yin et al., 2023).

Methodology

We present the overall framework of CWKT in Figure 3, comprising five key modules: (1) Input Encoding: models each student interaction as a Gaussian embedding, with mean and covariance capturing stable cognitive state and long-term volatility; (2) Local Volatility Extraction: extracts short-term volatility from the interaction sequence; (3) Wasserstein Self-Attention: leverages Wasserstein distance to compare Gaussian embeddings and track temporal knowledge evolution; (4) Attention Penalty: regularizes attention to balance local and long-term volatility; (5) Prediction: a two-layer fully connected network predicts future student performance from the integrated cognitive representation.

Input Encoding Module

Inspired by STOSA (Fan et al., 2022) and UKT (Cheng et al., 2025), we model each interaction as a Gaussian embedding with mean and covariance, capturing stable cognitive state

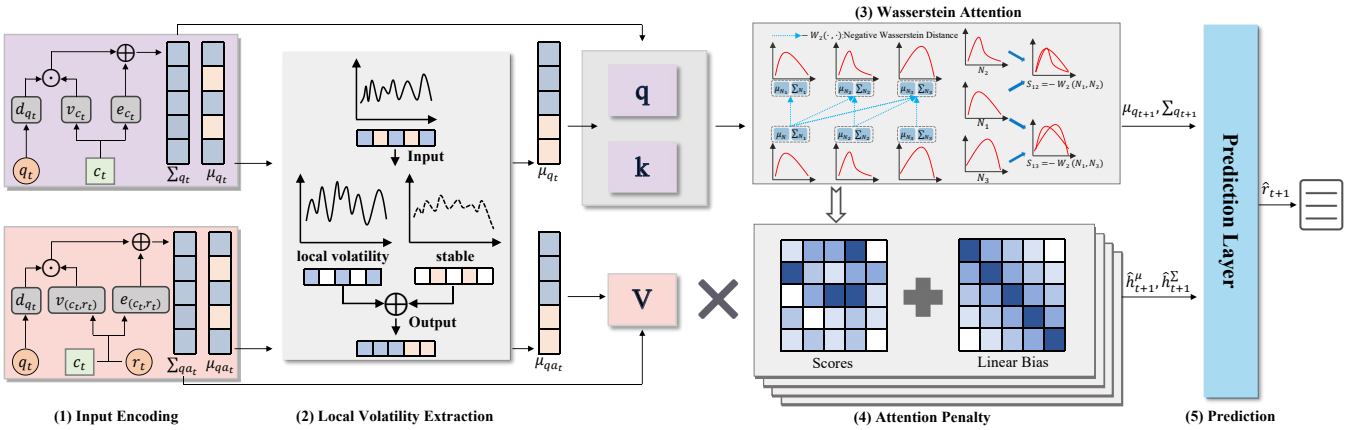

Figure 3: The framework of CWKT

and long-term volatility. Responses are mapped to associated Knowledge Concepts (KCs), with independent per-KC evaluations for fine-grained mastery, yielding a multi-dimensional elliptical representation of student cognition. The embeddings are defined as:

$$\mu_{c_t} = W_\mu^q \cdot e_{c_t},\ \Sigma_{c_t} = W_\Sigma^q \cdot e_{c_t} \tag{1}$$

$$\mu_{q_t} = \mu_{c_t} \oplus d_{q_t} \odot v_{c_t},\ \Sigma_{q_t} = \Sigma_{c_t} \oplus d_{q_t} \odot v_{c_t} \tag{2}$$

Here, e_{c_t} is the n-dimensional one-hot embedding of knowledge component c_t, and $W_\mu^q, W_\Sigma^q \in \mathbb{R}^{d \times n_c}$ are learnable linear transformations. d_{q_t} denotes a learnable question difficulty embedding. v_{c_t} captures the variation of KCs, and $v_{(c_t,r_t)}$ represents the KC-response variation for question q_t under response r_t. μ_{q_t} and Σ_{q_t} are the mean and covariance embeddings of the question.

$$\mu_{c_t,r_t} = W_\mu^r \cdot \mathrm{e}_{(\mathrm{c_t},\mathrm{r_t})},\ \Sigma_{c_t,r_t} = W_\Sigma^r \cdot \mathrm{e}_{(\mathrm{c_t},\mathrm{r_t})} \tag{3}$$

$$\mu_{qa_t} = \mu_{c_t,r_t} \oplus d_{q_t} \odot v_{(c_t,r_t)} \tag{4}$$

$$\Sigma_{qa_t} = \Sigma_{c_t,r_t} \oplus d_{q_t} \odot v_{(c_t,r_t)} \tag{5}$$

where $\mathrm{e}_{(c_t,r_t)}$ represents the joint embedding of c_t and the corresponding response r_t , and the matrices $W_\mu^r, W_\Sigma^r \in \mathbb{R}^{d \times (n_c+2)}$ are learnable linear transformations. The vector μ_{qa_t} and Σ_{qa_t} denote the mean and covariance embeddings of the student-question interaction. The operators $\odot$ and $\oplus$ denote element-wise multiplication and addition, respectively, while n_c indicates the total number of KCs.

Local Volatility Extraction Module

Frequency decomposition, effective for cognitive feature extraction (Zhou et al., 2023), enables Knowledge Tracing to jointly model stable learning patterns and local volatility (Hou et al., 2025). We apply convolutional filters of size k to the mean vectors of questions and interactions (μ_{q_t}, μ_{qa_t}) to extract low-frequency, stable components, and obtain high-frequency residuals reflecting local volatility; these components are then adaptively fused to produce multi-scale cognitive representations. Formally, $\mu_{q_t}, \mu_{qa_t} \in \mathbb{R}^{N \times L \times D}$.

$$\mu'_{q_t} = P\left(\mu_{q_t}, [0,2,1]\right),\ \mathcal{S}\left[\mu_{q_t}\right] = C\left(\mu'_{q_t}\right) \tag{6}$$

$$\mathcal{S}'\left[\mu_{q_t}\right] = P\left(\mathcal{S}\left[\mu_{q_t}\right], [0,2,1]\right) \tag{7}$$

Here, N, L, and D denote batch size, sequence length, and embedding dimension. P and C are the permutation and causal convolution operations, while $\mathcal{S}[\cdot]$ and $\mathcal{F}[\cdot]$ extract stable and fluctuating cognitive components, respectively.

$$\mathcal{F}\left[\mu_{q_t}\right] = \mu_{q_t} - \mathcal{S}'\left[\mu_{q_t}\right] \tag{8}$$

$$\mu_{q_t} = \mathcal{S}'\left[\mu_{q_t}\right] + \tau\mathcal{F}\left[\mu_{q_t}\right] = (1-\tau)\mathcal{S}'\left[\mu_{q_t}\right] + \tau\mu_{q_t} \tag{9}$$

where τ is a learnable fusion coefficient. We illustrate the fusion on μ_{q_t}, noting that μ_{qa_t} undergoes an analogous process. This decomposition and fusion effectively extract local cognitive volatility from students' interaction states.

Wasserstein Self-Attention Module

Dot-product attention in traditional KT models captures only linear vector similarity and overlooks distributional features of Gaussian embeddings, such as probabilistic distance, support overlap, and variance. We adopt a Wasserstein distance–based attention (Clement & Desch, 2008; Fan et al., 2023) to directly measure similarity between Gaussian distributions. Moreover, separate positional embeddings $P^\mu, P^\Sigma \in \mathbb{R}^{n_c \times d}$ are introduced for mean and covariance, enabling enriched sequence embeddings that capture both temporal and distributional dynamics of student interactions and their associated knowledge concepts.

$$\widetilde{\mu}_{q_t} = \left[\mu_{q_{t_1}} + P^\mu_{t_1},\ \mu_{q_{t_2}} + P^\mu_{t_2},\ \cdots,\ \mu_{q_{t_n}} + P^\mu_{t_n}\right] \tag{10}$$

$$\widetilde{\Sigma}_{q_t} = \left[\Sigma_{q_{t_1}} + P^\Sigma_{t_1},\ \Sigma_{q_{t_2}} + P^\Sigma_{t_2},\ \cdots,\ \Sigma_{q_{t_n}} + P^\Sigma_{t_n}\right] \tag{11}$$

$$\widetilde{\mu}_{qa_t} = \left[\mu_{qa_{t_1}} + P^\mu_{t_1},\ \mu_{qa_{t_2}} + P^\mu_{t_2},\ \cdots,\ \mu_{qa_{t_n}} + P^\mu_{t_n}\right] \tag{12}$$

$$\widetilde{\Sigma}_{qa_t} = \left[\Sigma_{qa_{t_1}} + P^\Sigma_{t_1},\ \Sigma_{qa_{t_2}} + P^\Sigma_{t_2},\ \cdots,\ \Sigma_{qa_{t_n}} + P^\Sigma_{t_n}\right] \tag{13}$$

$$\mu_{q_t} = \widetilde{\mu}_q\ ,\ \mu_{qa_t} = \widetilde{\mu}_{qa} \tag{14}$$

Additionally, to ensure the positive definiteness of the covariance, we use the Exponential Linear Unit (ELU) activation function to map the covariance inputs to the range $[-1, +\infty]$

$$\Sigma_{q_t} = ELU\left(diag\left(\widetilde{\Sigma}_q\right)\right)+1, \Sigma_{qa_t} = ELU\left(diag\left(\widetilde{\Sigma}_{qa}\right)\right)+1 \tag{15}$$

We use the 2-Wasserstein distance $W_2(\cdot,\cdot)$ and apply it in knowledge retrieval. The computation method is as follows.

$$\begin{aligned} \mathrm{S_{t+1}} &= -W_2^2(Q_t, Q_{a_t}) \\ &= -\Big(\left\|\mu_q - \mu_{qa}\right\|_2^2 + \mathrm{tr}\big(\Sigma_q + \Sigma_{qa} - 2\big(\Sigma_q^{\frac{1}{2}}\Sigma_{qa}\Sigma_q^{\frac{1}{2}}\big)^{\frac{1}{2}}\big)\Big) \end{aligned} \tag{16}$$

Attention Penalty Module

Standard attention often overweights long-term cognitive volatility, underrepresenting local volatility that reflects short-term student responses. This can fragment the model's view of cognition, reducing coherence and generalization. Inspired by FoLiBiKT (Im et al., 2023), we add a linear bias to attention scores as a regularizer, forming the Attention Penalty Module g:

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1j} \\ b_{21} & b_{22} & \cdots & b_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ b_{i1} & b_{i2} & \cdots & b_{ij} \end{bmatrix} \tag{17}$$

$$b_{ij} = -2^{-8\frac{n}{H}} \cdot |i-j| \tag{18}$$

where $2^{-8\frac{n}{H}}$ represents the coefficient used to adjust the attention scores of the n-th attention head among H total attention heads.

$$g\,(S, B, M) = Softmax\{(S \oplus B) \odot M\} \tag{19}$$

where S, B, and M denote attention scores, linear bias, and the causal mask, respectively, with Softmax$\{\cdot\}$ representing the softmax operation. M ensures the model cannot access future interactions. Implementation details are as follows:

$$m_{ij} = \begin{cases} 1, & \text{if } i \geq j \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

where m_{ij} denotes the element in the i-th row and j-th column of the matrix M.

Prediction

In this section, we first retrieve the knowledge state at time step t+1, and then employ a two-layer fully connected neural network to finely model the knowledge state. The implementation is as follows:

$$\begin{aligned} Q_{t+1}^{\mu} &= \mu_{q_{t+1}}, \quad & Q_{t+1}^{\Sigma} &= \Sigma_{q_{t+1}}, \\ K_j^{\mu} &= \mu_{q_j}, & K_j^{\Sigma} &= \Sigma_{q_j}, \\ V_j^{\mu} &= \mu_{qa_j}, & V_j^{\Sigma} &= \Sigma_{qa_j} \end{aligned} \tag{21}$$

where $j = 1, \cdots, t$. While the 2-Wasserstein distance follows the same formulation as in Equation (16), we adopt its square root form here to explicitly quantify the cognitive volatility between the predicted state and the observed interaction.

$$W_{t+1,j}^2 = W_2^2\big(\mathcal{N}(Q_{t+1}^{\mu}, Q_{t+1}^{\Sigma}), \mathcal{N}(K_j^{\mu}, K_j^{\Sigma})\big) \tag{22}$$

$$\alpha_{t+1,j} = \frac{\exp\big(-W_{t+1,j}^2\big)}{\sum_{i=1}^{t}\exp\big(-W_{t+1,i}^2\big)}, \tag{23}$$

where $\mathcal{N}(\cdot,\cdot)$ denotes a Gaussian distribution and $\alpha_{t+1,j}$ represents the normalized attention weight computed based on the Wasserstein distance. The index i enumerates all historical time steps from 1 to t and is used in the softmax denominator to ensure proper normalization of all past keys.

$$\hat{h}_{t+1}^{\mu} = \sum_{j=1}^{t} g(\alpha_{t+1,j}, B, M) \cdot V_j^{\mu} \tag{24}$$

$$\hat{h}_{t+1}^{\Sigma} = \sum_{j=1}^{t} g(\alpha_{t+1,j}, B, M) \cdot V_j^{\Sigma} \tag{25}$$

$$h_{t+1}^{\mu} = W_3 \cdot \mathrm{ReLU}\left(W_1 \cdot \left[\hat{h}_{t+1}^{\mu}\,;\, \mu_{q_{t+1}}\right] + b_1\right) + b_2 \tag{26}$$

$$\begin{aligned} h_{t+1}^{\Sigma} = \mathrm{ELU}\big(W_5 \cdot \mathrm{ReLU}\big(\\ W_4 \cdot [\hat{h}_{t+1}^{\Sigma}\,;\, \Sigma_{q_{t+1}}] + b_3\big) + b_4\big) + 1 \end{aligned} \tag{27}$$

where $W_1, W_4 \in \mathbb{E}^{d\times 2d}$, $W_3, W_5 \in \mathbb{E}^{d\times d}$ and $b_1, b_2, b_3, b_4 \in \mathbb{E}^{d\times 1}$ are all trainable parameters. Finally, a fully connected neural network layer is employed to make the prediction. The prediction function is optimized by minimizing the binary cross-entropy loss between the predicted response $\widehat{r}_{t+1}$ and the ground truth response r_{t+1}. The formulation is as follows:

$$\widehat{r}_{t+1} = \psi\left(\mathbf{w}^{\top} \cdot \mathrm{ReLU}\left(h_{t+1}^{\mu}\,;\, h_{t+1}^{\Sigma}\right) + b\right) \tag{28}$$

$$\mathcal{L} = -\sum_{t}\left(r_{t+1} \cdot \log\left(\widehat{r}_{t+1}\right) + (1 - r_{t+1}) \cdot \log\left(1 - \widehat{r}_{t+1}\right)\right) \tag{29}$$

where ψ denotes the Sigmoid function. $\mathbf{w} \in \mathbb{E}^{d\times 1}$ is a weight matrix, and b is a scalar bias term. $\mathbf{w}$ and b are all trainable parameters.

Experiments

Following (Liu et al., 2022), 80% of the data were used for five-fold cross-validation and 20% for final testing. Sequences shorter than three interactions were discarded, and remaining sequences were padded or truncated to length 200. Models were trained with Adam for up to 200 epochs, using early stopping after ten non-improving validation epochs. Learning rate was 1e-4, batch size 80, embedding dimension 256, and kernel size $k = 5$. Attention heads and transformer blocks were selected from [4,8] and [2,4], respectively. Gradient clipping with global norm 0.2 was applied to prevent explosion. Experiments ran on an NVIDIA A4000 GPU, and performance was evaluated with AUC and ACC.

Models	AUC				ACC			
	AS2009	AL2005	BD2006	NIPS34	AS2009	AL2005	BD2006	NIPS34
DKT	0.7541±0.0014	0.8149±0.0011	0.8015±0.0008	0.7689±0.0002	0.7297±0.0011	0.8097±0.0005	0.8553±0.0002	0.7032±0.0004
DKVMN	0.7629±0.0009	0.8054±0.0011	0.7983±0.0009	0.7673±0.0004	0.7266±0.0007	0.8027±0.0007	0.8545±0.0002	0.7016±0.0005
GKT	0.7666±0.0012	0.8110±0.0009	0.8046±0.0008	0.7689±0.0024	0.7290±0.0015	0.8088±0.0008	0.8511±0.0004	0.7014±0.0028
AKT	0.7853±0.0017	0.8306±0.0019	0.8208±0.0007	0.8033±0.0003	**0.7429±0.0015**	0.8124±0.0011	0.8587±0.0005	0.7323±0.0005
ATKT	0.7470±0.0008	0.7995±0.0023	0.7889±0.0008	0.7665±0.0003	0.7332±0.0009	0.7998±0.0019	0.8555±0.0002	0.7013±0.0002
Dtransformer	0.7725±0.0013	0.8188±0.0025	0.8093±0.0009	0.7994±0.0003	0.7304±0.0012	0.8043±0.0021	0.8555±0.0007	0.7295±0.0007
simpleKT	0.7744±0.0018	0.8254±0.0003	0.8160±0.0006	0.8035±0.0000	0.7315±0.0011	0.8067±0.0011	0.8567±0.0010	0.7328±0.0001
FoLiBiKT	0.7620±0.0011	0.8310±0.0010	0.8199±0.0008	0.8029±0.0012	0.7362±0.0015	0.8045±0.0021	0.8531±0.0003	0.7322±0.0008
QIKT	0.7812±0.0011	0.8118±0.0011	0.7937±0.0013	0.8006±0.0007	0.7362±0.0015	0.8045±0.0021	0.8531±0.0003	0.7305±0.0005
FlucKT	0.7818±0.0004	0.8316±0.0010	0.8247±0.0028	0.8016±0.0020	0.7373±0.0012	0.8143±0.0041	0.8562±0.0017	0.7313±0.0019
extraKT	0.7773±0.0009	0.8313±0.0021	0.8235±0.0006	0.8041±0.0003	0.7360±0.0015	0.8110±0.0009	0.8605±0.0012	0.7340±0.0004
UKT	0.7857±0.0014	0.8338±0.0012	0.8178±0.0009	0.8030±0.0004	0.7380±0.0021	0.8139±0.0011	0.8531±0.0006	0.7316±0.0004
CWKT(ours)	**0.7905±0.0013**	**0.8438±0.0007**	**0.8284±0.0001**	**0.8074±0.0005**	0.7414±0.0006	**0.8231±0.0003**	**0.8612±0.0003**	**0.7352±0.0006**

Table 1: Overall performance of CWKT and baselines

Dataset	interactions	students	KCs	Avg. interactions/student
AS2009	337,415	4217	123	80.1
AL2005	607,021	574	112	1057.5
BD2006	1,817,458	1,145	493	1587.3
NIPS34	1,382,678	4,918	948	281.8

Table 2: Statistics of the four real-world datasets

Models	AL2005	BD2006	NIPS34
CWKT	0.8438±0.0007	0.8284±0.0001	0.8074±0.0005
CWKT w/o W	0.8320±0.0012	0.8122±0.0014	0.8022±0.0009
CWKT w/o AP	0.8411±0.0010	0.8233±0.0005	0.8055±0.0007
CWKT w/o LVE	0.8405±0.0011	0.8225±0.0007	0.8034±0.0002
CWKT w/o AP & LVE	0.8223±0.0025	0.8101±0.0030	0.7981±0.0027

Table 3: AUC comparison of ablation experiments

Datasets

Experiments were conducted on four datasets, with key statistics presented in Table 2.

Baselines

To assess the proposed model's effectiveness, we benchmark it against ten state-of-the-art KT methods: DKT (Piech et al., 2015), DKVMN (Zhang et al., 2017), GKT (Nakagawa et al., 2019), AKT (Ghosh et al., 2020), ATKT (Guo et al., 2021), FoLiBiKT (Im et al., 2023), DTransformer (Yin et al., 2023), simpleKT (Liu, Liu, Chen, Huang, & Luo, 2023), QIKT (Chen et al., 2023), extraKT (Li et al., 2024), FlucKT (Hou et al., 2025) and UKT (Cheng et al., 2025).

Overrall Performances

As shown in Table 1, our model consistently outperforms existing KT methods. On AL2005, it achieves AUC 0.8438 (+1.00% over UKT) and 0.92% ACC gain; on BD2006, AUC rises from 0.8178 to 0.8284 (+1.06%) with 0.81% ACC improvement. Compared to extrakt, which ignores volatility, AUC improves by 1.25–1.3% with similar ACC gains. Unlike UKT, which models uncertainty via Gaussian embeddings but neglects short-term volatility, our approach captures both long- and short-term volatility , yielding superior, generalizable KT performance.

Ablation Study

To assess the contribution of individual components, we evaluate four ablated variants of CWKT (Table 3). "w/o AP" removes the Attention Penalty module, "w/o LVE" excludes the Local Volatility Extraction module, and "w/o AP&LVE" eliminates both. In "w/o W", the Wasserstein distance is replaced with dot-product similarity to test whether distribution-level alignment is necessary. All ablated variants exhibit clear performance degradation: removing AP or LVE weakens the modeling of local cognitive volatility, while substituting Wasserstein attention substantially reduces the model's ability to capture both mean shifts and covariance structures. These results confirm the complementary roles of AP, LVE, and Wasserstein attention in robust knowledge tracing.

Attention Analysis

To examine the effects of AP and LVE on attention behavior, we visualize the attention distributions of different CWKT variants in Figure 4. In the full CWKT model, first-layer attention is strongly concentrated along the diagonal, indicating an initial focus on short-term local cognitive volatility, such as consecutive errors or abrupt performance changes. This pattern is induced by the residual convolution in LVE, which performs a low–high frequency decomposition and amplifies high-frequency components corresponding to local volatility. As the network deepens, attention gradually shifts toward a more global distribution, while AP suppresses low-relevance long-term volatility, enabling a balanced integration of local and long-term cognitive signals. In contrast, CWKT w/o AP&LVE exhibits diffuse attention even in early layers, with

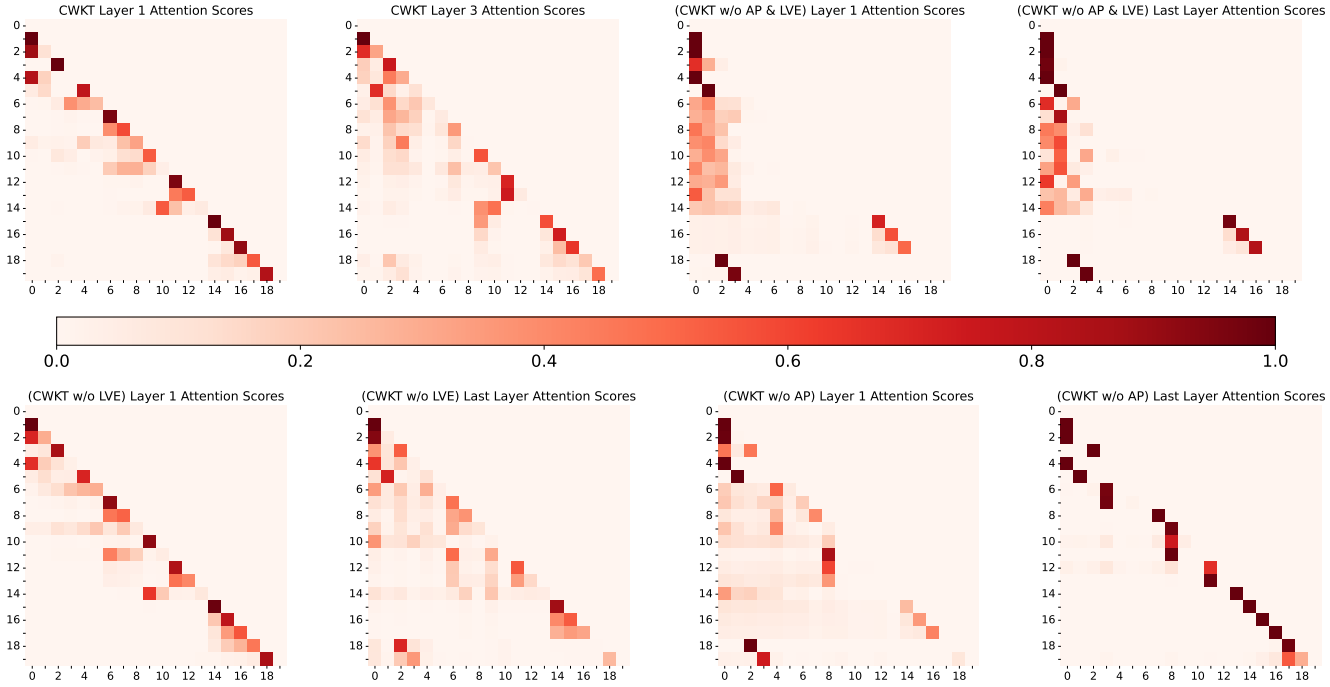

Figure 4: Attention scores in CWKT variants (w/o)

minimal structural evolution across depth, suggesting indiscriminate reliance on global historical trends without isolating salient local events. When only AP is removed (w/o AP), distinct localized attention clusters emerge at positions associated with consecutive errors or sudden gains, highlighting LVE's ability to capture and propagate local cognitive volatility through stable–volatile dual-channel modeling. By comparison, CWKT w/o LVE shows diagonal concentration in shallow layers but increasingly shifts attention toward long-term volatility at deeper layers, indicating compensation for the absence of explicit local volatility modeling.

Interpretability of Learned Covariance

Figure 5 compares two representative students from Algebra 2005 with similar average mastery but different covariance magnitudes. Despite comparable mean performance, the high-covariance student exhibits persistent covariance fluctuations across the sequence, accompanied by frequent alternations between correct and incorrect responses, indicating sustained behavioral instability. In contrast, the low-covariance student shows moderate fluctuations in early interactions, followed by a smoother trajectory as learning progresses, with occasional short-lived spikes that quickly decay. As the mean values are nearly identical, these differences cannot be explained by overall mastery. Instead, the learned covariance captures a distinct aspect of learning behavior beyond average performance, associated with the persistence of learning volatility over time, consistent with CWKT's design in which

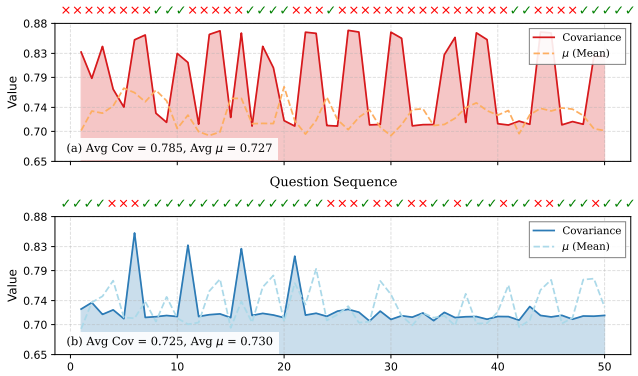

Figure 5: Visualization of Students in the Algebra 2005

covariance is accumulated through Gaussian embeddings.

Conclusion

We propose CWKT, a knowledge tracing model capturing student learning volatility using Gaussian embeddings parameterized by mean and covariance for long-term cognitive dynamics, combined with frequency decomposition and an attention-penalty mechanism to extract and balance local volatility. Experiments on multiple benchmarks show CWKT consistently surpasses state-of-the-art methods in detecting and adapting to behavioral volatility. Future work will integrate affective analysis to uncover psychological drivers, enhancing modeling of complex learning behaviors.

Acknowledgments

The works described in this paper are supported by The National Natural Science Foundation of China under Grant No. 62477015; Key Research and Development Program of Guangdong of China under Grant Nos. 2025B0101120004 and 2023B0303010004; The Innovation Team Project for Universities in Guangdong Province in China under Grant No. 2023KCXTD011.

References

Abdelrahman, G., & Wang, Q. (2019). Knowledge tracing with sequential key-value memory networks. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 175–184.

Abdelrahman, G., & Wang, Q. (2022). Deep graph memory networks for forgetting-robust knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 7844–7855.

Chen, J., Liu, Z., Huang, S., Liu, Q., & Luo, W. (2023). Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*, 14196–14204.

Cheng, W., Du, H., Li, C., Ni, E., Tan, L., Xu, T., & Ni, Y. (2025). Uncertainty-aware knowledge tracing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*, 27905–27913.

Choi, Y., Lee, Y., Cho, J., Baek, J., Kim, B., Cha, Y., Shin, D., Bae, C., & Heo, J. (2020). Towards an appropriate query, key, and value computation for knowledge tracing. *Proceedings of the Seventh ACM Conference on Learning at Scale*, 341–344.

Clement, P., & Desch, W. (2008). An elementary proof of the triangle inequality for the wasserstein metric. *Proceedings of the American Mathematical Society*, *136*(1), 333–339.

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

Fan, Z., Liu, Z., Peng, H., & Yu, P. S. (2023). Mutual wasserstein discrepancy minimization for sequential recommendation. *Proceedings of the ACM Web Conference*, 1375–1385.

Fan, Z., Liu, Z., Wang, Y., Wang, A., Nazari, Z., Zheng, L., Peng, H., & Yu, P. S. (2022). Sequential recommendation via stochastic self-attention. *Proceedings of the ACM Web Conference*, 2036–2047.

Ghosh, A., Heffernan, N., & Lan, A. S. (2020). Context-aware attentive knowledge tracing. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2330–2339.

Guo, X., Huang, Z., Gao, J., Shang, M., Shu, M., & Sun, J. (2021). Enhancing knowledge tracing via adversarial training. *Proceedings of the 29th ACM International Conference on Multimedia*, 367–375.

Hou, M., Li, X., Guo, T., Liu, Z., Tian, M., Luo, R., & Luo, W. (2025). Cognitive fluctuations enhanced attention network for knowledge tracing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*, 14265–14273.

Im, Y., Choi, E., Kook, H., & Lee, J. (2023). Forgetting-aware linear bias for attentive knowledge tracing. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3958–3962.

Li, X., Bai, Y., Guo, T., Zheng, Y., Hou, M., Zhang, B., Huang, Y., Liu, Z., Gao, B., & Luo, W. (2024). Extending context window of attention based knowledge tracing models via length extrapolation. *ECAI 2024*, 1479–1486.

Liu, Z., Liu, Q., Chen, J., Huang, S., Gao, B., Luo, W., & Wang, J. (2023). Enhancing deep knowledge tracing with auxiliary tasks. *Proceedings of the ACM Web Conference*, 4178–4187.

Liu, Z., Liu, Q., Chen, J., Huang, S., & Luo, W. (2023). Simplekt: A simple but tough-to-beat baseline for knowledge tracing. *arXiv preprint arXiv:2302.06881*.

Liu, Z., Liu, Q., Chen, J., Huang, S., Tang, J., & Luo, W. (2022). Pykt: A python library to benchmark deep learning based knowledge tracing models. *Advances in Neural Information Processing Systems*, *35*, 18542–18555.

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019). Graph-based knowledge tracing: Modeling student proficiency using graph neural network. *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 156–163.

Pandey, S., & Karypis, G. (2019). A self-attentive model for knowledge tracing.

Pavlik, J., Philip I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis—a new alternative to knowledge tracing. *International Conference on Artificial Intelligence in Education*, 531–538.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, *28*.

Shen, S., Liu, Q., Chen, E., Huang, Z., Huang, W., Yin, Y., Su, Y., & Wang, S. (2021). Learning process-consistent knowledge tracing. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1452–1460.

Tong, H., Wang, Z., Zhou, Y., Tong, S., Han, W., & Liu, Q. (2022). Introducing problem schema with hierarchical exercise graph for knowledge tracing. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 405–415.

Tong, S., Liu, Q., Huang, W., Huang, Z., Chen, E., Liu, C., Ma, H., & Wang, S. (2020). Structure-based knowledge tracing: An influence propagation view. *2020 IEEE International Conference on Data Mining (ICDM)*, 541–550.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio.

Yeung, C.-K., & Yeung, D.-Y. (2018). Addressing two problems in deep knowledge tracing via prediction-consistent regularization. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–10.

Yin, Y., Dai, L., Huang, Z., Shen, S., Wang, F., Liu, Q., Chen, E., & Li, X. (2023). Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer. *Proceedings of the ACM Web Conference*, 855–864.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized bayesian knowledge tracing models. *International Conference on Artificial Intelligence in Education*, 171–180.

Zhang, J., Shi, X., King, I., & Yeung, D.-Y. (2017). Dynamic key-value memory networks for knowledge tracing. *Proceedings of the 26th International Conference on World Wide Web*, 765–774.

Zhou, W., Sun, F., Jiang, Q., Cong, R., & Hwang, J.-N. (2023). Wavenet: Wavelet network with knowledge distillation for rgb-t salient object detection. *IEEE Transactions on Image Processing*, *32*, 3027–3039.