ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890058

Dynamically Causal-Enhanced Exercise Representations for Adaptive Knowledge Tracing

Yanhong Bai*
Lab of AI for Education
East China Normal University
Shanghai, China
Lucky_Baiyh@stu.ecnu.edu.cn

Jiabao Zhao†
Lab of AI for Education
East China Normal University
Shanghai, China
jbzhao@mail.ecnu.edu.cn

Tingjiang Wei*
Lab of AI for Education
East China Normal University
Shanghai, China
52275901031@stu.ecnu.edu.cn

Jinxin Shi
Lab of AI for Education
East China Normal University
Shanghai, China
jinxinshi@stu.ecnu.edu.cn

Liang He
School of Computer Science and Technology
East China Normal University
Shanghai, China
lhe@cs.ecnu.edu.cn

***Abstract*—Knowledge tracing assesses students' mastery and predicts future performance based on historical learning data. Traditional methods primarily rely on predefined static associations between concepts and exercises, which struggle to capture potential causal relationships and dynamic learning patterns, leading to reduced prediction accuracy and limited interpretability. To address these issues, this paper proposes a novel dynamic causal inference framework that integrates Gumbel-Softmax sampling with uncertainty estimation, transforming discrete causal relationships into differentiable continuous weights, and quantifying model uncertainty to enhance robustness against noisy data and improve interpretability. Additionally, inspired by item response theory, the model dynamically adjusts students' latent states by modeling the interaction between student ability and exercises difficulty. Experimental results on three widely-used benchmarks demonstrate that this method achieves state-of-the-art (SOTA) performance in prediction accuracy while also generating interpretable causal relationship weights that provide insights into knowledge acquisition patterns.**

***Index Terms*—knowledge tracing, causal inference, Gumbel-Softmax sampling, adaptive learning, personalized education**

I. Introduction

Knowledge tracing (KT) aims to predict students' future performance using historical learning data [1], [2]. KT methods are broadly categorized into traditional machine learning and deep learning approaches. Traditional methods, such as Bayesian knowledge tracing [3] and factor analysis models [3]–[5], focus on limited parameters, restricting their ability to model complex relationships. Deep learning approaches, including RNN-based models [6]–[8], and attention-based models [9], [10], offer significant advancements by capturing temporal dependencies and sequence-level interactions.

To enhance KT models' effectiveness and interpretability, understanding the mechanisms of knowledge acquisition is crucial. While deep learning excels at modeling complex temporal and sequential interactions, it often obscures causal relationships. Causal inference has become increasingly critical in KT, enabling the identification of relationships among learning interventions, exercises, and concepts to optimize learning processes. However, most existing approaches, from traditional static structures to advanced graph-based models [11]–[14], remain constrained by correlational modeling, hindering their ability to address the dynamic and nonlinear nature of learning processes. Static causal structures, such as Directed Acyclic Graphs (DAGs) employed in models like Causal GRU [15], have limited capacity to adapt to frequent shifts in students' knowledge states and variations in exercise difficulty. Similarly, interpretive models like IKT [16] offer causal insights but oversimplify the underlying relationships by employing Naive Bayes structures, which lack the capacity to model multivariate and nonlinear dependencies. While models like SKT [17] incorporate causal regularization to improve stability, they lack the flexibility to dynamically adjust causal weights in response to evolving learning trajectories. Advanced causal inference techniques, such as Granger causality [18], show potential in identifying causal effects but remain sensitive to noise and reliant on high-quality data. While TCKT [19] effectively addresses confounding factors and performs well under noise, its reliance on static global dictionaries and the lack of dynamic updates to causal relationships limit its adaptability to changing or personalized learning scenarios. Consequently, the lack of interpretable and dynamic causal mechanisms remains a key challenge in advancing knowledge tracing.

Specifically, we aim to address the following key questions: Q1) How can Gumbel-Softmax causal adjustment dynamically infer causal relationships between exercises and concepts? Q2) How can uncertainty estimation contribute to improving the accuracy and robustness of causal inference? Q3) How can an Item Response Theory (IRT) Adapter Layer refine the mod-

This work was supported by the National Natural Science Foundation of China (Grant number [62207013].

* These authors contributed equally to this work.

† Corresponding Author.

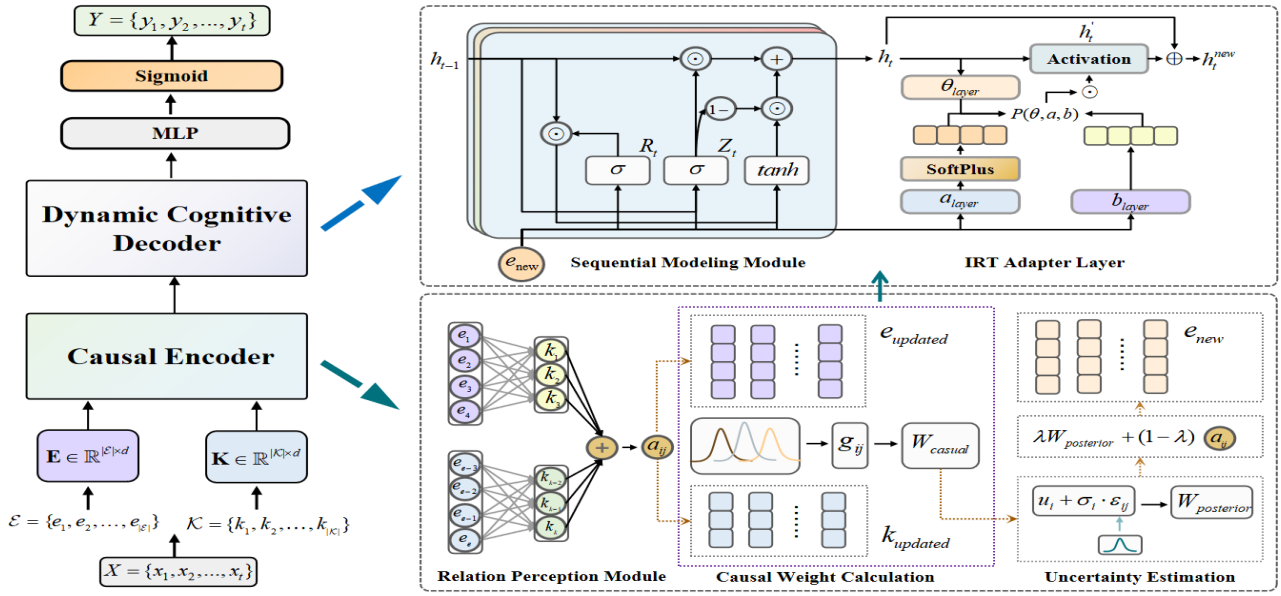

Fig. 1: The framework of this paper.

eling of students' knowledge states and improve personalized prediction accuracy?

To address these challenges, we propose the CausalKT, consisting of a causal encoder and a dynamic cognitive decoder. The causal encoder leverages Gumbel-Softmax sampling to transform discrete causal selection into a continuous optimization problem, allowing dynamic adjustment of causal weights and overcoming the static assumptions in existing methods. It also integrates Bayesian uncertainty estimation, enhancing robustness against noisy data by quantifying uncertainty in causal weights. The dynamic cognitive decoder combines temporal sequence information and an IRT adapter layer to adjust the student's learning state, adapting to changes in ability and exercise difficulty. Experimental results demonstrate that CausalKT outperforms existing methods across multiple datasets, showcasing its potential in personalized education.

II. Problem Definition

Knowledge tracing predicts a student's future performance based on their learning history. Let $\mathcal{E} = \{e_1, e_2, \ldots, e_{|\mathcal{E}|}\}$ be the set of exercises and $\mathcal{K} = \{k_1, k_2, \ldots, k_{|\mathcal{K}|}\}$ the set of concepts, where each exercise e_i is linked to one or more concepts k_j. A student's learning trajectory is a sequence $\mathcal{T}_s = \{(e_1, r_1), (e_2, r_2), \ldots, (e_T, r_T)\}$, with $r_i \in \{0, 1\}$ indicating the correctness of their response. The task is to model the student's knowledge state and predict their probability of correctly answering the next exercise e_{T+1}, i.e., $P(r_{T+1} = 1 \mid \mathcal{T}_s, e_{T+1})$.

III. Method

As illustrated in Fig. 1, the proposed CausalKT model consists of two components: a Causal Encoder and a Dynamic Cognitive Decoder. The Causal Encoder infers dynamic causality between exercises and concepts using Gumbel-Softmax sampling, with uncertainty estimation to improve robustness under noise. The Dynamic Cognitive Decoder enhances knowledge state modeling through sequential learning and an IRT adapter layer. This framework allows CausalKT to capture complex causal structures and accurately predict student performance.

A. Causal Encoder

The Causal Encoder defines causal inference as identifying directed edges between exercises and concepts, indicating influence on mastery or outcomes. By framing this as a discrete variable selection problem, Gumbel-Softmax sampling converts it into continuous optimization. The encoder consists of Three modules: Embedding Representation, Relation Perception, Causal Weight Calculation, allowing dynamic and robust inference of causal relationships.

1) Embedding Representation: The embedding module maps the exercise set $\mathcal{E} = \{e_1, e_2, \ldots, e_{|\mathcal{E}|}\}$ and concept set $\mathcal{K} = \{k_1, k_2, \ldots, k_{|\mathcal{K}|}\}$ into high-dimensional vectors for relation modeling and causal inference. The embeddings for concepts k_j and exercises e_i are linearly transformed as follows:

$$k'_j = W_k k_j, \quad e'_i = W_e e_i \tag{1}$$

where W_k and W_e are learnable transformation matrices.

2) Relation Perception: The relationship between exercise e_i and concept k_j is represented by an adjacency matrix $A \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{K}|}$, where $A_{ij} = 1$ if exercise e_i involves concept k_j, and $A_{ij} = 0$ otherwise. We apply a modified Graph Attention Network (GAT) that aggregates the features of related concepts k_j for each exercise e_i. The embeddings of e_i and k_j, denoted as $e'_i \in \mathbb{R}^d$ and $k'_j \in \mathbb{R}^d$ respectively, are concatenated as $[e'_i || k'_j]$, where $||$ represents vector concatenation. The

attention score is computed using a learnable weight vector $a \in \mathbb{R}^{2d}$. The adjacency matrix A is used as a mask to ensure only valid connections are considered, followed by Softmax normalization to obtain the attention weights α_{ij}:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^\top [e'_i \,||\, k'_j])) \cdot A_{ij}}{\sum_{k \in \mathcal{N}(e_i)} \exp(\text{LeakyReLU}(a^\top [e'_i \,||\, k'_k])) \cdot A_{ik}} \quad (2)$$

The updated concept embedding $k_{\text{updated},i}$ and the updated exercise embedding $e_{\text{updated},i}$ are computed as follows:

$$k_{\text{updated},i} = \sum_j \alpha_{ij} k'_j, \quad e_{\text{updated},i} = k_{\text{updated},i} + e'_i \odot E \quad (3)$$

where $\odot$ denotes element-wise multiplication, and $E \in \mathbb{R}^d$ is a trainable parameter.

3) Causal Weight Calculation: To address the discrete selection challenge in causal structure learning, we innovatively introduce Gumbel-Softmax sampling. Unlike traditional soft attention mechanisms, Gumbel-Softmax combines stochastic noise for weight selection, preserving differentiability while enhancing the model's discriminative ability in causal inference. First, we generate a logits matrix $\mathbf{L}$, where L_{ij} represents the initial association score between updated concepts and exercises. By adding Gumbel noise $\mathbf{g}$ and applying the Gumbel-Softmax operation, we sample the causal weights:

$$L_{ij} = k_{\text{updated},i} \cdot e^T_{\text{updated},j} \quad (4)$$

$$W_{\text{causal},ij} = \text{Softmax}\left(\frac{L_{ij} + g_{ij}}{\tau}\right) \quad (5)$$

Here, g_{ij} is drawn from a Gumbel(0, 1) distribution, and τ is the temperature parameter that controls the smoothness of the sampling process. This mechanism offers flexibility in handling discrete selections, making it well-suited for complex relationships in causal inference tasks.

4) Bayesian Uncertainty Estimation: To enhance robustness against noise and uncertainty, we compute the mean and variance of the causal weights. Uncertainty weights are then sampled from the posterior distribution:

$$W^{\text{posterior}}_{\text{causal},ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}) \quad (6)$$

where μ_{ij} and σ_{ij} are the mean and standard deviation of the causal weights.

5) Final Exercise Representation: The combined attention weight $W_{\text{combined},ij}$ is a weighted sum of the posterior causal weight $W^{\text{posterior}}_{\text{causal},ij}$ and the original attention weight a_{ij}, with λ controlling their balance. The concept embedding $k_{\text{new},i}$ and exercise embedding $e_{\text{new},i}$ are then updated using the combined weights.

$$W_{\text{combined},ij} = \lambda W^{\text{posterior}}_{\text{causal},ij} + (1 - \lambda) a_{ij} \quad (7)$$

$$k_{\text{new},i} = \text{ReLU}\left(W_{\text{combined},ij} \cdot k'_i + b_{\text{f}}\right) \quad (8)$$

$$e_{\text{new},i} = \text{ELU}\left(W_{\text{reduce}} \cdot [k_{\text{new},i} \| \left(k_{\text{new},i} \odot e'_i\right)]\right) \quad (9)$$

This approach combines the benefits of causal inference and traditional attention mechanisms for both concept and exercise embeddings, using a dimensionality reduction matrix $\mathbf{W}_{\text{reduce}}$and ELU activation.

B. Dynamic Cognitive Decoder

The Dynamic Cognitive Decoder uses causal relationships from the Causal Encoder and student's historical responses to predict future performance. It refines hidden states by combining updated exercise representations with past learning states. Key components include the Sequential Modeling Module, IRT Adapter Layer, and Prediction Generation Module.

1) Sequential Modeling Module: For each time step t, the hidden state h_t is computed by a GRU:

$$h_t = \text{GRU}(X_t, h_{t-1}) \quad (10)$$

where X_t is the input at time step t.

2) IRT Adapter Layer: This module leverages IRT to model student ability and exercise difficulty, integrating student ability with exercise-specific causal information to dynamically adjust the learning state. The student's ability θ is extracted from h_t and the exercise's difficulty discrimination a and difficulty b are derived from $e_{\text{new},i}$:

$$\theta = W_\theta h_t + b_\theta \quad (11)$$

$$a = \text{Softplus}(W_a e_{\text{new},i} + b_a) + \epsilon, \quad b = W_b e_{\text{new},i} + b_b \quad (12)$$

The Softplus function ensures the discrimination parameter is positive, as required by the IRT model. The probability P of a correct answer is computed using the IRT model, and the final updated hidden state h^{new}_t is:

$$P = \frac{1}{1 + \exp(-a(\theta - b))} \quad (13)$$

$$h'_t = \text{ReLU}\left(W_{\text{fusion}} \cdot [h_t \| e_{\text{new},i}] + b_{\text{fusion}}\right) \cdot P \quad (14)$$

$$h^{\text{new}}_t = h_t + h'_t \quad (15)$$

3) Prediction Generation Module: The final prediction is generated based on the personalized hidden state h^{new}_t through a fully connected layer:

$$\hat{y}_t = \sigma(W_2 \cdot \text{ReLU}(W_1 [h^{\text{new}}_t \,||\, e_{\text{new},i}])) \quad (16)$$

where σ is the sigmoid function. The loss function is binary cross-entropy:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right) \quad (17)$$

IV. Experiments

A. Datasets and Baselines

Datasets. We evaluate our model using three widely adopted real-world datasets for knowledge tracing, summarized in Table I. These include **ASSISTments2009 (ASSIST09)** and **ASSISTments2012 (ASSIST12)**, collected from the ASSISTments platform during the 2009-2010 and 2012-2013 school years, and **slepemapy.cz**, an adaptive platform for geographic knowledge.

Baselines. To validate the effectiveness of our model, we compare it against several baselines: **DKT [6]** (an

TABLE I: Dataset Statistics.

Dataset	Interactions	Exercises	Concepts	Learners	Avg. Attempts
ASSIST09	110.2k	16.9k	111	3.7k	7
ASSIST12	879.5k	50.9k	245	25.3k	17
slepemapy.cz	2877.5k	2.9k	1473	81.7k	992

LSTM-based sequential model), **DKVMN [7]** (a memory-augmented model), **SAKT [9]**(leveraging self-attention), **DKT+Forgetting [20]** (accounting for forgetting behavior), **DKT-IRT [21]** (incorporating IRT for better interpretability), **KTM [22]** (based on factorization machines), **GKT [23]**(utilizing graph structure), **AKT-R [24]** (employing monotonic attention), and **HawkesKT [25]** (focusing on temporal interactions between sessions).

B. Experimental Setup and Evaluation Metrics

We used 5-fold cross-validation, with 10% of each fold for hyperparameter tuning. Learning rates (0.001, 0.005, 0.01) and batch sizes (32, 64, 128) were tuned, and early stopping halted training after 5 iterations without improvement. Evaluation metrics included AUC and ACC, as the task involves binary classification.

C. Overall Performance

Each dataset represents a distinct knowledge tracing environment, from smaller platforms (ASSIST09) to large-scale student interactions (slepemapy.cz), ensuring the model's generalization. CausalKT consistently outperforms baseline models, achieving a notable AUC of 0.7911 and accuracy of 0.7453 on the ASSIST09, setting new state-of-the-art results, as shown in Table II.

TABLE II: Performance comparison of all KT methods. The best results are in bold.

Model	ASSIST09		ASSIST12		slepemapy.cz	
	AUC	ACC	AUC	ACC	AUC	ACC
DKT	0.7525	0.7247	0.7322	0.7371	0.7512	0.7820
DKVMN	0.7326	0.7189	0.7057	0.7294	0.7371	0.7859
DKVMN-E	0.6742	0.6946	0.6943	0.7255	0.7237	0.7847
SAKT	0.6894	0.6864	0.6912	0.7216	0.6739	0.7711
DKT-IRT	0.7565	0.7268	0.7365	0.7396	0.7546	0.7823
DKT+Forgetting	0.7573	0.7272	0.7462	0.7373	0.7574	0.7819
GKT	0.7489	0.7178	0.7345	0.7300	0.7501	0.7799
KTM	0.7353	0.7178	0.7514	0.7412	0.7421	0.7772
AKT-R	0.7519	0.7199	0.7649	0.7523	0.7547	0.7827
HawkesKT	0.7617	0.7305	0.7669	0.7475	0.7572	0.7823
Ours	**0.7911**	**0.7453**	**0.7886**	**0.7730**	**0.7813**	**0.8205**

D. Ablation Experiment

We performed ablation studies to answer three key questions. Fig. 2 summarize the results, validating our model's contributions: Q1 (Gumbel-Softmax): Removing this caused a significant performance drop, highlighting its role in capturing causal relationships. Q2 (Uncertainty estimation): Its removal caused a slight AUC drop, highlighting its contribution to enhancing robustness in noisy data environments. Q3 (IRT adapter): Removing the IRT adapter led to a clear performance decrease, underscoring its importance in balancing ability and difficulty for accurate predictions.

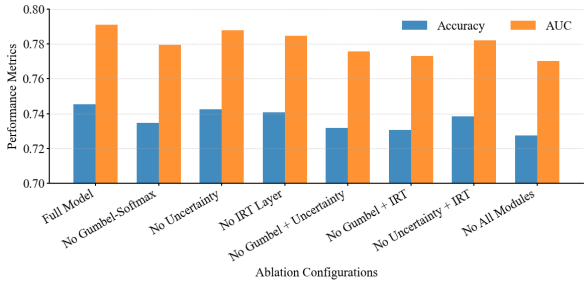

Fig. 2: Ablation Experiment.

E. Case Study

This study compares attention mechanisms and causal discovery algorithms in skill node selection, as shown in Fig. 3. Attention mechanisms focus on nodes linked to questions, while causal discovery can further identify key nodes (in green) with causal influence, including those missed by attention mechanisms. For example, nodes 33 and 42 were identified as crucial despite weak links, and node 7988 (connected by purple lines) showed indirect effects that attention mechanisms overlooked. Combining both methods provides a more comprehensive identification of relevant skill nodes, improving skill mapping and personalized learning.

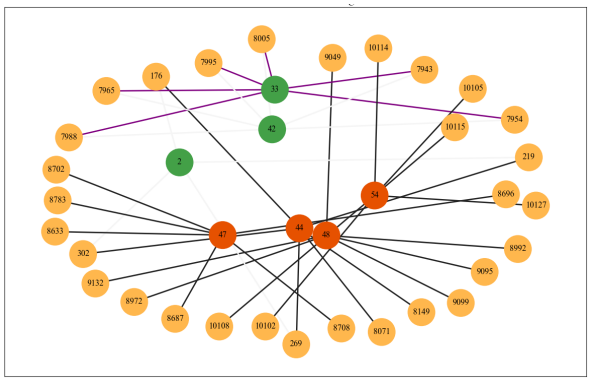

Fig. 3: Case Study.

V. Conclusion

This paper introduces CausalKT, a model combining Gumbel-Softmax sampling for dynamic causal inference with uncertainty estimation to enhance accuracy and robustness in knowledge tracing. The model outperforms baselines across multiple datasets by dynamically adjusting student learning states through an IRT adapter, improving personalized predictions. CausalKT offers a robust, interpretable solution for personalized education.

REFERENCES

[1] G. Abdelrahman, Q. Wang, and B. Nunes, "Knowledge tracing: A survey," ACM Comput. Surv., vol. 55, no. 11, pp. 1–37, 2023.

[2] Y. Bai, J. Zhao, T. Wei, Q. Cai, and L. He, "A survey of explainable knowledge tracing," Appl. Intell., pp. 1–32, 2024.

[3] F. B. Baker, *The Basics of Item Response Theory*. ERIC, 2001.

[4] H. Cen, K. Koedinger, and B. Junker, "Learning factors analysis–a general method for cognitive model evaluation and improvement," in *Proc. Int. Conf. Intell. Tutoring Syst.*, Springer, 2006, pp. 164–175.

[5] P. I. Pavlik Jr, H. Cen, and K. R. Koedinger, "Performance factors analysis–A new alternative to knowledge tracing," Online Submission, ERIC, 2009.

[6] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, *et al.*, "Deep knowledge tracing," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[7] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 765–774.

[8] M. Chen, Q. Guan, Y. He, Z. He, L. Fang, and W. Luo, "Knowledge tracing model with learning and forgetting behavior," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 3863–3867.

[9] S. Pandey and G. Karypis, "A self-attentive model for knowledge tracing," arXiv preprint, arXiv:1907.06837, 2019.

[10] D. Shin, Y. Shim, H. Yu, S. Lee, B. Kim, and Y. Choi, "Saint+: Integrating temporal features for ednet correctness prediction," in *Proc. LAK21: 11th Int. Learn. Analytics Knowl. Conf.*, 2021, pp. 490–496.

[11] T. Wu and Q. Ling, "Fusing hybrid attentive network with self-supervised dual-channel heterogeneous graph for knowledge tracing," Expert Syst. Appl., vol. 225, p. 120212, 2023.

[12] Y. Yang, J. Shen, Y. Qu, Y. Liu, K. Wang, Y. Zhu, W. Zhang, and Y. Yu, "Gikt: a graph-based interaction model for knowledge tracing," in Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part I, Springer, 2021, pp. 299–315.

[13] G. Abdelrahman and Q. Wang, "Deep graph memory networks for forgetting-robust knowledge tracing," IEEE Trans. Knowl. Data Eng., vol. 35, no. 8, pp. 7844–7855, 2022.

[14] Q. Guan, F. Xiao, X. Cheng, L. Fang, Z. Chen, G. Chen, and W. Luo, "Kg4ex: An explainable knowledge graph-based approach for exercise recommendation," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, 2023, pp. 597–607.

[15] N. A. Kumar, W. Feng, J. Lee, H. McNichols, A. Ghosh, and A. S. Lan, 'A conceptual model for end-to-end causal discovery in knowledge tracing," in Proc. 16th Int. Conf. Educ. Data Mining, EDM 2023, Bengaluru, India, July 11–14, 2023. Int. Educ. Data Mining Soc., 2023.

[16] S. Minn, J. Vie, K. Takeuchi, H. Kashima, and F. Zhu, "Interpretable knowledge tracing: Simple and efficient student modeling with causal relations," in Proc. AAAI Conf. Artif. Intell., vol. 36, no. 11, pp. 12810–12818, 2022.

[17] J. Zhu, X. Ma, and C. Huang, "Stable knowledge tracing using causal inference," IEEE Trans. Learn. Technol., vol. 17, pp. 124–134, 2023.

[18] Z. Zhang and L. Wu, "Graph neural network-based bearing fault diagnosis using granger causality test," Expert Syst. Appl., vol. 242, p. 122827, 2024.

[19] C. Huang, H. Wei, Q. Huang, F. Jiang, Z. Han, and X. Huang, "Learning consistent representations with temporal and causal enhancement for knowledge tracing," Expert Syst. Appl., vol. 245, p. 123128, 2024.

[20] K. Nagatani, Q. Zhang, M. Sato, Y. Chen, F. Chen, and T. Ohkuma, "Augmenting knowledge tracing by considering forgetting behavior," in The World Wide Web Conf., WWW, ACM, 2019, pp. 3101–3107.

[21] G. Converse, S. Pu, and S. Oliveira, "Incorporating item response theory into knowledge tracing," in Int. Conf. Artif. Intell. Educ., Springer, 2021, pp. 114–118.

[22] J. Vie and H. Kashima, "Knowledge tracing machines: Factorization machines for knowledge tracing," in The Thirty-Third AAAI Conf. Artif. Intell., AAAI Press, 2019, pp. 750–757.

[23] H. Nakagawa, Y. Iwasawa, and Y. Matsuo, "Graph-based knowledge tracing: Modeling student proficiency using graph neural net," in Proc. Int. Conf. Learn. Represent., ICLR, 2019, pp. 1–8.

[24] A. Ghosh, N. T. Heffernan, and A. S. Lan, "Context-aware attentive knowledge tracing," in The 26th ACM SIGKDD Conf. Knowl. Discov. Data Min., ACM, 2020, pp. 2330–2339.

[25] C. Wang et al., "Temporal cross-effects in knowledge tracing," in WSDM 2021, The Fourteenth ACM Int. Conf. Web Search Data Min., ACM, 2021, pp. 517–525.