Appendix of "A Fast-Adaptive Cognitive Diagnosis Framework for Computerized Adaptive Testing Systems"

Yuanhao Liu , **Yiya You** , **Shuo Liu** , **Hong Qian**[*] , **Ying Qian** and **Aimin Zhou**
Shanghai Institute of AI Education, and School of Computer Science and Technology, East China Normal University, Shanghai 200062, China
{51275901044, 10225102493, 51255901007}@stu.ecnu.edu.cn, {hqian, yqian, amzhou}@cs.ecnu.edu.cn

The appendix is organized as follows.

• Appendix A provides an overview of some of the notations used in this paper.

• Appendix B provides additional technical details in this paper, including the detailed task introduction and the time and space complexity analysis.

• Appendix C provides more detailed introduction of related work.

• Appendix D provides additional details about the experiments conducted in this paper, including the datasets, baselines and implementation details, additional experimental results on interpretability and different extraction modules, detailed hyperparameter analysis.

A Notation

The notations of this paper are shown in Table 1.

B Additional Technical Details

B.1 Details about Task Introduction

Here we provide Figure 1 for the further explanation on the CAT and CD task.

B.2 Details about Time and Space Complexity Analysis of FACD

Here we provide time and space complexity analysis of our FACD as follow:

Time Complexity Analysis. We can divide our time complexity analysis FACD framework into two parts, that is Dynamic Collaborative Diagnosis Module and Dynamic Personalized Diagnosis Module:

• **Dynamic Collaborative Diagnosis Module**: In this module, we construct a dynamic student-question-knowledge graph with three kinds of node types in node set $\mathcal{V}$ and two kinds of edge types edge sets $\mathcal{E}$. Since we do not use non-linear activation or feature transformations common in GNNs, this complexity is straightforward to compute as $O(2|\mathcal{E}|Ld)$, where $|\mathcal{E}|$ is the number of edges, L is number of graph layers and d is the dimension of the embeddings.

• **Dynamic Personalized Diagnosis Module**: In this module, we use a GRU layer and a self attention components. The GRU's complexity is$O(4Td^2)$ and self-attention's complexity is $O(2(Td^2 + T^2d))$, making the total complexity $O(6Td^2 + 2T^2d)$, where T denotes the current CAT time step and d is the dimension of the embeddings.

Given $T_{max} = 15$ and $d \leq 64$ for better performance in our experiment analysis, so $Td \ll |\mathcal{E}|$. So the overall complexity is approximately $O(2|\mathcal{E}|Ld)$. Comparatively, ORCDF's complexity is $O(4|\mathcal{E}|Ld)$ due to flipped graph representation convolution, and RCD's complexity is $O(2|\mathcal{E}|LZ^2)$ where Z is the number of knowledge ($d \ll Z$). Therefore, our algorithm demonstrates superior theoretical time efficiency. Notably, during training, our framework processes 27,988 response logs from 874 students in just 2.23 seconds, making it highly suitable for online CAT systems. Detailed comparisons are provided in Section 5.4 Inference Time Comparison in the main paper.

Space Complexity Analysis. We can also divide our space complexity analysis FACD framework into two parts:

• **Dynamic Collaborative Diagnosis Module**: In this module, space complexity mainly comes from the feature matrix of each node in the dynamic graph, the degree matrix, and the edge index. The feature matrix of nodes has a complexity of $O(|\mathcal{V}|d)$, where $|\mathcal{V}|$ is the number of nodes and d is the feature dimension. The degree matrix, being sparse, has a reduced complexity of $O(|\mathcal{V}|)$, and the edge index has a complexity of $O(2|\mathcal{E}|)$, where $|\mathcal{E}|$ is the number of edges.

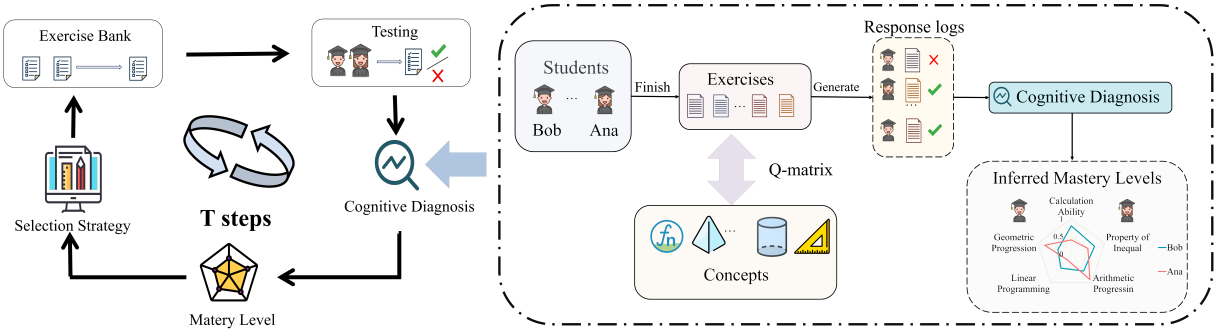
• **Dynamic Personalized Diagnosis Module**: In this module, the space complexity stems from the model parameters of the GRU and self-attention modules. The space complexity of the GRU is $O(2d^2 + Td)$, and self-attention mechanism is $O(Td^2 + T^2)$, where T is the current time steps.

Therefore, the dominant space complexity is from the feature matrix of the graph nodes, which is $O(|\mathcal{V}|d)$. In fact, this space complexity is not significantly different from that of traditional neural CD methods (such as NCD) that do not use graphs, and it does not impose a significant burden on the online CAT environment. To further help with understanding, we also provide a comparison of the parameter count of our model with other neural and graph CD used in our experiments. To further help with understanding, we also provide a comparison of the parameter size of our model on NeurIPS2020 dataset with other neural and graph CD used in our experiments in Table 2.

[*] Corresponding Author.

Table 1: Notations of this paper.

Symbols	Descriptions
S, E, C	The student set, the question set, the concept set
s, e, r, c	The student, the question, the response, the concept
C_i, J_i	Student i's candidate questions set, evaluated questions set
$R_{t,i}$	The response sequence of student i on step t
$\hat{Mas}_{t,i}$	Student i's estimated mastery level on step t
Mas_i^*, R_i^*	Student i's true mastery level, student i's actual responses
$\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$	The dynamic student- question-concept graph at time step t
$\mathcal{V}$	Nodes including students, questions and concepts
$\mathcal{E}_t, \mathcal{E}_t^{se}, \mathcal{E}^{ec}$	Total edges between the nodes on the step t, the interactions between students set S and questions set E, the relationships between questions set E and concepts set C
$\mathbf{A}_t, \mathbf{A}_t^{se}, \mathbf{A}^{ec}$	The adjacency matrix of $\mathcal{G}_t$, the adjacency matrix of student-question subgraph, the adjacency matrix of question-concept subgraph
$\mathbf{Z}_{s,e,c}^{(0)}$	Initial ID embeddings for students, questions and knowledge concepts
$\mathbf{Z}_{s,e,c}^{(1)}$	Dynamic collaborative representation for students, questions and knowledge concepts
$\mathbf{Z}_{s,e,c}^{(2)}$	Dynamic personalized representation for students, questions and knowledge concepts
η_t	Smoothing weight for the dynamic collaborative features at time step t
$\mathbf{L}_{t,s_i}$	The personalized question embedding sequences for student t at time step t
$\mathbf{h}_{t,s_i}$	Hidden states of GRU at the time step t for student s_i
$\mathbf{H}_{t,s_i}, \mathbf{H}_{t,e_j}, \mathbf{H}_{t,c_k}$	Ultimate fused representation for student s_i, question e_j and knowledge conceptc_k
$\alpha_{s,e,c}$	Normalized weights of students', questions' and knowledge concepts' embeddings for ultimate fused representation



(a) Computerized Adaptive Testing **(b) Cognitive Diagnosis**

Figure 1: An illustrative example of computerized adaptive testing and cognitive diagnosis.

Table 2: Parameter sizes for different CDMs.

Models	NCD	FACD-NCD	ORCDF-NCD	RCD-NCD
Size	0.408MB	0.527MB	0.508MB	0.464MB

C Detailed Related Work

C.1 Computerized Adaptive Testing

CAT consists of two key components: a CDM and a selection strategy. Traditionally, IRT [Embretson and Reise, 2013] and NCD [Wang *et al.*, 2020a] are popular choices for CDM in CAT. But most of the research focuses on the selection strategy in CAT. The most commonly-used approach is Maximum Fisher Information (MFI) [Lord, 2012]. Another popular method, Kullback-Leibler Information (KLI) [Chang and Ying, 1996], computes an integral over an ability interval to choose questions. These heuristic algorithms are typically tailored to specific CDMs, such as IRT. To address this limitation, a model-agnostic algorithm, MAAT [Bi *et al.*, 2020], was introduced. It utilizes active learning for question selection and incorporates an additional module to enhance concept diversity. More recently, BECAT [Zhuang *et al.*, 2023] develop an expected gradient difference approximation to design a simple greedy selection strategy, which selects a question subset that closely matches the gradient of the full responses. The above selection strategies assume accurate CDM diagnosis, which are difficult to achieve in the early stage of CAT. Inaccurate CDM results can mislead the selection strategies, leading to unsuitable question choices and prolonging the overall CAT process. Current data-driven selection strategies, such as BOBCAT [Ghosh and Lan, 2021] and NCAT [Zhuang *et al.*, 2022a], pretrain strategies on large datasets and using the trained strategies for online question selection. Although this kind of selection strategies may help mitigate the effect of lack of data in the early stage of CAT, well-trained strategies still fail to effectively address the inaccuracy and instability of CDM's diagnosis in the early stage. Therefore, designing a CDM tailored to the CAT and providing timely and accurate diagnosis in the early stage is crucial.

C.2 Cognitive Diagnosis

Cognitive diagnosis is a vital field in educational psychology, which is used to infer students' mastery levels for each concept by applying either latent factor models, such as IRT [Embretson and Reise, 2013] and Multidimensional IRT (MIRT) [Sympson, 1978], or leveraging models that capture concept mastery patterns, like the Deterministic Input, Noisy And Gate model (DINA) [De La Torre, 2009] where 0 represents an unmastered state and 1 indicates mastery. With the advancements in deep learning, researchers have achieved considerable success in handling large-scale interactions. For example, NCD [Wang *et al.*, 2020a] employs multilayer perceptrons (MLPs) as the interaction function and represents mastery patterns as continuous variables between 0 and 1. Other approaches to extracting rich information from response logs include MLP-based models [Ma *et al.*, 2022; Wang *et al.*, 2023a], symbolic regression [Shen *et al.*, 2024b], graph attention networks [Gao *et al.*, 2021; Qian *et al.*, 2024; Shen *et al.*, 2024a; Wang *et al.*, 2023c] and Bayesian networks [Li *et al.*, 2022]. Recently, several CDMs have been developed to address cold-start issues within the field. Cross-domain Cognitive Diagnosis is a classic cold-start scenario, aimed at diagnosing students' cognitive levels in new domains by leveraging their interaction records across different subjects. TechCD [Gao *et al.*, 2023] and ZeroCD [Gao *et al.*, 2024] are aimed to solve this issue. The former uses a Knowledge Concept Graph to link knowledge across domains, achieving knowledge transfer, while the latter primarily utilizes few-shot learning to learn new domain knowledge from a limited number of samples. Another cold-start scenario arises in traditional static CD, where students have a limited amount of interaction records. Methods addressing this issue include BetaCD [Bi *et al.*, 2023] and AGCDM [Pei *et al.*, 2022], both of which are based on meta-learning. The former uses Bayesian structure learning, while the latter employs a Self-Attention Gate. It can be observed that CD encompasses different cold-start scenarios, and these scenarios differ from the early-stage adaptation problem in CAT. First, the CAT scenario does not involve cross-domain issues, as we assume that all problems stem from the same domain. Second, the CAT process requires CD to adapt to the dynamic process of CAT and model the relationships within questions sequences. In contrast, the cold-start problem in static CD typically does not consider dynamic relationships and lacks the ability to model the sequence of questions chosen by selection strategy in CAT. Additionally, meta-learning approaches often require substantial computational resources, which are not in line with the requirements for real-time assessment in CAT. Thus, static CD structures are not suitable for use in the dynamic process of CAT. At the same time, existing CDM studies rarely emphasize the adaptation to downstream tasks. As far as we know, there is no work that has specifically designed a CDM for the unique setting of CAT. Given CAT's sensitivity to CDM's modeling ability, such a targeted adaptation is highly necessary.

D Experimental Details

D.1 Details About Datasets

Datasets Source. Here, we provide the details about datasets source as follow:

- **NeurIPS2020** [Wang *et al.*, 2020b] originates from the NeurIPS 2020 Education Challenge and includes data from two school years (2018-2020). It includes detailed student responses to mathematics questions sourced from Eedi, a globally recognized, web-based online learning platform actively utilized by millions of students daily. The platform provides a rich set of student's online interactions, allowing for in-depth analysis of learning behaviors in a web-based educational environment.

- **EDMCup2023** [Ethan Prihar, 2023] is derived from the competition of EDM Cup 2023 and this dataset captures millions of student actions on the ASSISTments platform, a widely used web-based system designed to support K-12 math learning. As with NeurIPS2020, the focus of this datasets is on analyzing online learning patterns in a real-time, web-driven educational context, highlighting personalized learning pathways and adaptive feedback mechanisms.

- **FrcSub** [DeCarlo, 2011; Tatsuoka, 1984] consists of scores from middle school students on fraction subtraction objective problems. It offers valuable insight into student mathematical understanding and skill gaps, complementing the web datasets by providing a more controlled, traditional assessment setting.

Together, these datasets offer a comprehensive foundation for analyzing student interactions, learning behaviors, and performance within web-based and traditional educational frameworks.

Datasets Filtering. For data filtering, in order to ensure that each selected student has enough question data to support his or her cognitive diagnosis during CAT process, we only select students who answered more than 40 questions.

Datasets Partition. For data partition, following the standard paradigm of CAT system [Bi *et al.*, 2020; Wang *et al.*, 2023b], we first divide the dataset into a pretrain dataset and a train dataset. The pretrain dataset is used to pretrain the CDMs, allowing them to initially learn some parameters that remain fixed during the CAT process. The train dataset is then used for the actual training of the CAT system. Since the students in the two datasets cannot overlap, we split the students between the pretrain and train datasets in an $80\%/20\%$ ratio. For the train dataset, the questions answered by each student are further divided into a candidate question set and evaluated question set. During the CAT process, questions are selected from the candidate question set for each student, and after each selection, the CDM's performance is evaluated on the test question set. We set an $80\%/20\%$ ratio of each student's questions for the candidate question set and test question set.

D.2 Baseline Details

The baselines of CD in the experiments is as follow:

- **IRT** [Embretson and Reise, 2013] uses just one scalar to describe the total mastery level of a student and employs logistic function as interaction function.

- **NCD** [Wang *et al.*, 2020a] applies the deep learning methodology to cognitive diagnosis. It denotes the mastery level of one student with embedding vectors in the dimension of concepts and employs Positive MLPs as the interaction function.

- **RCD** [Gao *et al.*, 2021] explores the intricate relationships among students, questions, and concepts by employing GAT to effectively model these connections and leveraging a multi-level attention network to integrate node-level relation aggregation inside each local graphs.

- **ORCDF** [Qian *et al.*, 2024] extract the representation by introducing a response graph and a response-aware graph convolution network. It integrates different response signals as different types of edges to resistant oversmoothing.

Table 3: The performance on DOA using BECAT as strategy and NCD as CDM. The best value is marked in bold.

Dataset	FrcSub			EDMCup2023			NeurIPS2020		
Step	5	10	15	5	10	15	5	10	15
OL	61.45	75.29	81.65	66.69	73.15	74.86	61.31	67.86	70.81
FA	**77.77**	**83.04**	**83.35**	**68.56**	**74.29**	**76.39**	**65.84**	**69.57**	**71.54**

The baselines of selection strategy in CAT in the experiments is as follow:

• **Random** indicates that at each step of the CAT process, the questions selected for students are chosen randomly, without other designs.

• **MAAT** [Bi *et al.*, 2020] is an active learning-based approach that selects questions based on the Expected Model Change (EMC) of the CDM. Additionally, it includes a supplementary module aimed at enhancing concept diversity during the test process.

• **BOBCAT** [Ghosh and Lan, 2021] is a computationally efficient, bilevel optimization-based framework for CAT that learns a data-driven question selection algorithm directly from training data, independent of the underlying student response model.

• **NCAT** [Zhuang *et al.*, 2022b] is based on reinforcement learning and employs an attention-driven DQN. It selects questions by sampling from the Boltzmann distribution of Q-values, aiming to control question exposure effectively.

• **BECAT** [Zhuang *et al.*, 2023] is a data-summary method, which selects a question subset that closely matches the gradient of the full responses with an expected gradient difference approximation.

It is worth noting that FACD focuses on designing CDM for diverse CAT methods. The selection stratrgies above have been selected for their classic nature, wide usage, and strong performance. Recent CAT methods, such as CCAT [Liu *et al.*, 2024], redefine CAT tasks—for instance, shifting from absolute diagnostic accuracy to relative student ranking—often targeting completely different goals. While valuable, it is beyond the scope of this work to incorporate such redefined CAT paradigms.

D.3 Implementation Details

All experiments are run on a Linux server with two 3.00GHz Intel Xeon Gold 6354 CPUs and one RTX3090 GPU. All the models are implemented by PyTorch [Paszke *et al.*, 2019]. We employ grid search to find the best hyperparameters using the validation set. And detailed analysis regarding the hyperparameters can be found in Appendix D.6.

As for the details on implementation of our baselines, the implementation of IRT and NCD comes from the public repository https://github.com/bigdata-ustc/EduCDM. For RCD, ORCDF, we adopt the implementation from the authors in https://github.com/bigdata-ustc/RCD, https://github.com/ECNU-ILOG/ORCDF. And for the implementation of MAAT, BOBCAT, NCAT and BECAT come from the public repository https://github.com/bigdata-ustc/EduCAT.

Table 4: The performance with different classic GNN frameworks using BECAT as strategy and NCD as CDM on EDMCup2023 dataset. The best value is marked in bold.

Dataset	EDMCup2023		
Step	5	10	15
GCN	73.94	76.74	79.24
GATv2	75.02	79.35	81.86
GT	75.75	79.95	81.71
FACD	**76.95**	**80.53**	**82.01**

D.4 Additional Experiment on Interpretability

We show the quantifiable interpretability metrics, such as DOA, for evaluating our FACD. Because DOA can not be used in IRT, so here we use the NCD for the CD backbone and BECAT for the selection strategy. As shown in table 3, FACD demonstrates fast adaptation to strong interpretability, especially evident in the early stages of CAT, which strengthen our work's contribution in terms of the fast adaption on interpretability.

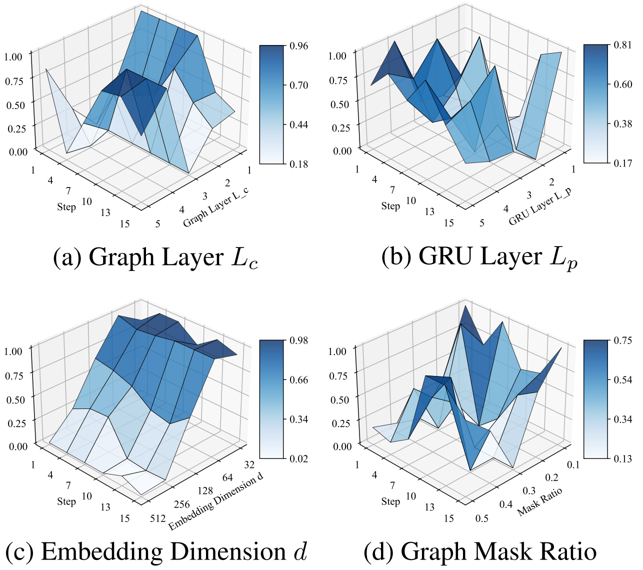
D.5 Additional Experiment on Different Extraction Modules.

Here we include classic GNN frameworks such as GCN [Kipf and Welling, 2016], GATv2 [Brody *et al.*, 2021], and GT [Wu *et al.*, 2021] to replace the lightweight GNN framework based on [He *et al.*, 2020] in our FACD. As shown in the Table 4, we conduct a performance comparison on the EDMCup2023 dataset using BECAT as the selection strategy and NCD as CDM, with AUC as the evaluation metric. The results demonstrate that, despite the simplicity of our GNN framework, it still outperforms traditional GNN frameworks like GCN, GATv2, and GT in terms of performance. This validates that our lightweight framework is both simple and efficient. Lightweight GNN frameworks have gained acceptance and adoption in various fields, such as recommendation systems, because they eliminate redundant components from traditional GNN frameworks (e.g., non-linear activation or feature transformations), resulting in significant improvements in speed and memory efficiency while maintaining theoretical guarantees on its performance. Hence, this makes it particularly well-suited for our online CAT systems.

D.6 Detailed Hyperparameter Analysis

Here, we provide the detailed analysis of hyperparameter experiment of FACD-NCD with BECAT in Figure 2. Notably, to better observe the impact of different hyperparameters on the model's performance, the data from each step has been standardized to ensure consistent scaling across different time step.

The Effect of Graph Layer L_c. The effect of graph layer L_c determines the number of graph layer in dynamic collaborative diagnosis module. As shown in Figure 2(a), it can be observed that when the number of graph layers is set at 1 and 2, the model achieves optimal performance. Although increasing the layers yields some improvement, excessive layers can lead to over-smoothing, which diminishes the overall



(a) Graph Layer L_c (b) GRU Layer L_p

(c) Embedding Dimension d (d) Graph Mask Ratio

Figure 2: Hyperparameters analysis results.

effectiveness. Therefore, we recommend setting the number of graph layers to 1 and 2 for good results.

The Effect of GRU Layer L_p. The effect of GRU layer L_p determines the number of graph layer in dynamic personalized diagnosis module. As shown in Figure 2(b), FACD is not particularly sensitive to the number of GRU layers. However, when the number of layers is set to 3, it provides a well-balanced performance. Thus, we recommend setting the number of GRU layer as 3.

The Effect of Embedding Dimension d. The embedding dimension d determines the dimension of representation H_t. As shown in Figure 2(c), the performance achieves the highest point at 32 or 64 in most cases, so it is recommended to set d as 32 or 64.

The Effect of Graph Mask Ratio. The mask is used for graph layer for the purpose of the robustness of the graph in dynamic collaborative diagnosis module. As shown in Figure 2(d), employing the mask in the dynamic graph show effectiveness. FACD delivers better performance in early step with a small ratio, while a relatively big mask ratio results in improved performance in later step. So it is recommended to set the mask ratio between 0.1 and 0.2, if we want to get a fast adaptability in early step.

References

[Bi *et al.*, 2020] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. Quality meets diversity: A model-agnostic framework for computerized adaptive testing. In *Proceedings of the 20th IEEE International Conference on Data Mining*, pages 42–51, Sorrento, Italy, 2020.

[Bi *et al.*, 2023] Haoyang Bi, Enhong Chen, Weidong He, Han Wu, Weihao Zhao, Shijin Wang, and Jinze Wu. Beta-cd: A bayesian meta-learned cognitive diagnosis framework for personalized learning. In *Proceedings of the 37th AAAI conference on artificial intelligence*, volume 37, pages 110–118, Washington, DC, 2023.

[Brody *et al.*, 2021] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[Chang and Ying, 1996] Hua-Hua Chang and Zhiliang Ying. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229, 1996.

[De La Torre, 2009] Jimmy De La Torre. Dina model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1):115–130, 2009.

[DeCarlo, 2011] Lawrence T DeCarlo. On the analysis of fraction subtraction data: The dina model, classification, latent class sizes, and the q-matrix. *Applied Psychological Measurement*, 35(1):8–26, 2011.

[Embretson and Reise, 2013] Susan E Embretson and Steven P Reise. *Item Response Theory*. Psychology Press, 2013.

[Ethan Prihar, 2023] NHeffernan Ethan Prihar. Edm cup 2023, 2023.

[Gao *et al.*, 2021] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 501–510, Virtual Event, 2021.

[Gao *et al.*, 2023] Weibo Gao, Hao Wang, Qi Liu, Fei Wang, Xin Lin, Linan Yue, Zheng Zhang, Rui Lv, and Shijin Wang. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 983–992, Taiwan, China, 2023.

[Gao *et al.*, 2024] Weibo Gao, Qi Liu, Hao Wang, Linan Yue, Haoyang Bi, Yin Gu, Fangzhou Yao, Zheng Zhang, Xin Li, and Yuanjing He. Zero-1-to-3: Domain-level zero-shot cognitive diagnosis via one batch of early-bird students towards three diagnostic objectives. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 8417–8426, Vancouver,Canada, 2024.

[Ghosh and Lan, 2021] Aritra Ghosh and Andrew Lan. Bobcat: Bilevel optimization-based computerized adaptive testing. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, Virtual Event, 2021.

[He *et al.*, 2020] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, Virtual Event, 2020.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Li *et al.*, 2022] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. HierCDF: A Bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 904–913, Virtual Event, 2022.

[Liu *et al.*, 2024] Zirui Liu, Yan Zhuang, Qi Liu, Jiatong Li, Yuren Zhang, Zhenya Huang, Jinze Wu, and Shijin Wang. Computerized adaptive testing via collaborative ranking. *Advances in Neural Information Processing Systems 37*, pages 95488–95514, 2024.

[Lord, 2012] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 2012.

[Ma *et al.*, 2022] Haiping Ma, Manwei Li, Le Wu, Haifeng Zhang, Yunbo Cao, Xingyi Zhang, and Xuemin Zhao. Knowledge-sensed cognitive diagnosis for intelligent education platforms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1451–1460, Atlanta, GA, 2022.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035, British Columbia, Canada, 2019.

[Pei *et al.*, 2022] Xiaohuan Pei, Shuo Yang, Jiajun Huang, and Chang Xu. Self-attention gated cognitive diagnosis for faster adaptive educational assessments. In *Proceedings of the 22nd IEEE International Conference on Data Mining*, pages 408–417, Orlando, FL, 2022.

[Qian *et al.*, 2024] Hong Qian, Shuo Liu, Mingjia Li, Bingdong Li, Zhi Liu, and Aimin Zhou. Orcdf: An oversmoothing-resistant cognitive diagnosis framework for student learning in online education systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2455–2466, Barcelona, Spain, 2024.

[Shen *et al.*, 2024a] Junhao Shen, Hong Qian, Shuo Liu, Wei Zhang, Bo Jiang, and Aimin Zhou. Capturing homogeneous influence among students: Hypergraph cognitive diagnosis for intelligent education systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2628–2639, Barcelona, Spain, 2024.

[Shen *et al.*, 2024b] Junhao Shen, Hong Qian, Wei Zhang, and Aimin Zhou. Symbolic cognitive diagnosis via hybrid optimization for intelligent education systems. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 14928–14936, Vancouver, Canada, 2024.

[Sympson, 1978] James B Sympson. A model for testing with multidimensional items. In *Proceedings of the 1977 Computerized Adaptive Testing Conference*, Minneapolis, MN, 1978.

[Tatsuoka, 1984] Kikumi K Tatsuoka. Analysis of errors in fraction addition and subtraction problems. final report. 1984.

[Wang *et al.*, 2020a] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020.

[Wang *et al.*, 2020b] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061*, 2020.

[Wang *et al.*, 2023a] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. Neuralcd: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8), 2023.

[Wang *et al.*, 2023b] Hangyu Wang, Ting Long, Liang Yin, Weinan Zhang, Wei Xia, Qichen Hong, Dingyin Xia, Ruiming Tang, and Yong Yu. Gmocat: A graph-enhanced multi-objective method for computerized adaptive testing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2279–2289, Long Beach, CA, 2023.

[Wang *et al.*, 2023c] Shanshan Wang, Zhen Zeng, Xun Yang, and Xingyi Zhang. Self-supervised graph learning for long-tailed cognitive diagnosis. In *Proceedings of the 37th AAAI conference on artificial intelligence*, volume 37, pages 110–118, Washington, DC, 2023.

[Wu *et al.*, 2021] Zhanghao Wu, Paras Jain, Matthew A. Wright, Azalia Mirhoseini, Joseph E. Gonzalez, and Ion Stoica. Representing long-range context for graph neural networks with global attention. In *Advances in Neural Information Processing Systems 34*, pages 13266–13279, Virtual Event, 2021.

[Zhuang *et al.*, 2022a] Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Shuanghong Shen, and Haiping Ma. Fully adaptive framework: Neural computerized adaptive testing for online education. In *Proceeddings of the 36th AAAI Conference on Artificial Intelligence*, pages 4734–4742, Virtual Event, 2022.

[Zhuang *et al.*, 2022b] Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Shuanghong Shen, and Haiping Ma. Fully adaptive framework: Neural computerized adaptive testing for online education. In *Proceedings of the 36th AAAI conference on artificial intelligence*, volume 36, pages 4734–4742, 2022.

[Zhuang *et al.*, 2023] Yan Zhuang, Qi Liu, GuanHao Zhao, Zhenya Huang, Weizhe Huang, Zachary Pardos, Enhong Chen, Jinze Wu, and Xin Li. A bounded ability estimation for computerized adaptive testing. In *Advances in Neural Information Processing Systems 36*, New Orleans, LA, 2023.