Constructing a Personalized Learning Path Recommendation Model Based on Federated Learning

Jing Zhang
School of Engineering, Changchun Normal University, Changchun 130032, Jilin, China
Corresponding author: Jing Zhang (e-mail: fuleiwazzg@163.com)

ABSTRACT To address the conflict between model generalization and personalized recommendation in distributed educational environments, this paper constructs a personalized learning path recommendation model based on federated learning. Such privacy-preserving collaborative learning frameworks are also valuable for intelligent information processing and distributed decision-making in modern electromagnetic communication and edge computing systems, where data sharing is often restricted. The proposed model employs a Graph Neural Network (GNN)-Transformer hybrid encoder that combines knowledge graphs with learning behavior sequences to accurately capture knowledge transfer relationships. A dynamic knowledge distillation aggregation strategy is introduced to generate soft labels from the global model for guiding local optimization, thereby preserving personalized characteristics while improving semantic consistency. Furthermore, an adaptive aggregation mechanism based on Kullback-Leibler (KL) divergence dynamically adjusts client weights to enhance robustness under heterogeneous data distributions. Experimental results demonstrate that the proposed method achieves excellent recommendation accuracy (average Hit@5 of 0.676 $\pm$ 0.009), sequence consistency (average NDCG@10 of 0.712 $\pm$ 0.008), and personalized responsiveness (average personalized score difference rate of 0. 38). The framework effectively balances global generalization and local adaptation while maintaining privacy protection, providing a feasible solution for secure and intelligent recommendation in distributed learning environments and offering technical insights for collaborative intelligence in privacy-sensitive electromagnetic information systems.

INDEX TERMS personalized learning path recommendation, federated learning, graph neural network-transformer, distributed intelligent systems, electromagnetic information processing

I. INTRODUCTION

WITH the rapid development of online education and intelligent learning systems, personalized learning path recommendation has become a key technology for improving teaching efficiency and student experience, a concept increasingly relevant to the industry for training new workers and upskilling existing employees in advanced manufacturing techniques [1,2]. However, learning data is widely distributed across different institutions, platforms, and devices. Due to privacy restrictions and data silos, traditional centralized modeling approaches struggle to achieve cross-domain collaborative optimization. Federated learning addresses this dilemma by providing a new paradigm for distributed training, enabling joint modeling without sharing raw data [3-5]. However, in real-world educational scenarios, student groups vary significantly across participants, and data exhibits non-independent and identically distributed (non-IID) characteristics, leading to a serious conflict between the generalization capabilities of global models and local personalized needs [6,7]. Therefore, how to balance model consistency and individual adaptability while ensuring privacy becomes a core challenge in building efficient federated recommendation systems. Refined modeling of learning behavior is key to achieving personalized learning path recommendations. Existing research mainly improves recommendation accuracy by fusing multi-source information. Jiang L [8] constructed a Massive Open Online Course (MOOC) knowledge graph to capture the logical dependencies between knowledge points, providing a structural basis for generating interpretable learning paths. Tzeng JW [9] used long short-term memory networks to model learning sequences and address students' personalized learning needs, whereas Gu R [10] proposed an algorithmic framework that combines deep reinforcement learning with a graph attention mechanism to customize learning paths for individual learners. Amin S [11] built an interactive recommendation system based on the Actor-Critic model of deep reinforcement learning, which can provide top-N course recommendations and personalized learning paths. Hussain T [12] used a multi-layer topic model to annotate unlabeled

student feedback, implemented the Feld-Silverman learning style model to automatically identify learning styles, and used latent Dirichlet allocation to analyze text feedback for sentiment analysis to further improve the learning experience. These works collectively demonstrate the effectiveness of combining knowledge priors with behavioral dynamics for accurate recommendation, providing important insights for constructing local encoders in this paper. However, most of these methods assume centralized data and fail to address the challenges of privacy protection and collaborative modeling in cross-institutional data silos, limiting their application in distributed educational environments. To address the core challenges posed by non-IID data in federated learning, researchers have proposed various improvement strategies. At the algorithm level, Nemati AA [13] evaluated the performance of federated learning algorithms such as Federated Averaging (FedAvg) and Federated Proximal (FedProx) in IID and non-IID scenarios, and pointed out that FedAvg is an ideal choice for scenarios containing IID data. Xu Y [14] alleviated data drift at the device edge through hierarchical aggregation and sub-model selection. In terms of personalization mechanism, Yu E [15] used a federated recommendation algorithm based on user clustering and meta-learning to improve both communication efficiency and recommendation personalization. Tan AZ [16] proposed a unique taxonomy of personalized federated learning technologies to address the fundamental challenges faced by federated learning on heterogeneous data. Singh N [17] explored a self-knowledge distillation framework to facilitate knowledge transfer on edge devices. While these methods have made some progress, they still have significant shortcomings. First, they generally lack effective integration of core semantic structures in the educational field, making it difficult for the models to deeply understand the laws of cognitive evolution. Second, their aggregation mechanisms are often static or simply weighted, unable to adaptively adjust to the client's real-time knowledge point distribution shifts, thus limiting the generalization and stability of the global model. However, in federated learning environments, models relying solely on behavioral sequences struggle to capture universal cognitive patterns under highly heterogeneous data and privacy-constrained conditions, while pure knowledge graph methods lack adaptability to individual learning dynamics. Therefore, there is an urgent need for a hybrid modeling mechanism that integrates structural prior knowledge with behavioral dynamics to achieve knowledge consistency across clients and unified personalized path generation while ensuring privacy protection. The GNN-Transformer hybrid architecture is specifically designed for this purpose: the GNN encodes knowledge logic, while the Transformer captures individual learning evolution. Together, they provide a representation foundation for federated recommendation that combines semantic stability with behavioral sensitivity. This paper proposes a federated personalized learning path recommendation method that combines dynamic knowledge distillation and adaptive aggregation, a method highly applicable to the industry for distributing specialized training knowledge across global factory networks while customizing skill development for individual employees or departments. First, each client constructs a GNN-Transformer hybrid encoder based on the knowledge graph and student behavior sequence to achieve joint modeling of cognitive state and knowledge structure. Second, a dynamic knowledge distillation mechanism is introduced in the global communication round. The server generates soft labels using public calibration data and sends them to the client, guiding the local model to align semantic representations without sharing data. A cosine annealing strategy is used to dynamically weight the distillation loss and local supervision loss. The paper further designs an adaptive aggregation mechanism based on KL divergence, dynamically adjusting the model update weights according to the similarity between the client's knowledge point distribution and the global distribution and suppressing the bias caused by non-IID and IID data. Finally, the paper integrated differential privacy and secure aggregation technology to perform gradient cropping and Gaussian noise injection before uploading to protect user privacy. The innovations of this method include:

(1) constructing a GNN-Transformer hybrid encoder to integrate knowledge graphs and learning behavior sequences to accurately model the transfer relationship of knowledge points;

(2) introducing a dynamic knowledge distillation mechanism to guide local training through global soft labels to improve model consistency;

(3) designing an adaptive aggregation mechanism based on KL divergence to dynamically adjust client weights and enhance robustness in heterogeneous environments;

(4) combining differential privacy with a periodic federated collaborative training framework to achieve efficient and secure model aggregation while ensuring (E, δ)- differential privacy.

Among them, GNN-Transformer and dynamic knowledge distillation synergistically constrain the semantic consistency of local models from the two levels of representation learning and optimization objectives, fundamentally alleviating client drift caused by non-IID and IID data, rather than relying solely on posterior corrections in the aggregation stage.

II. ALGORITHM DESIGN

CONSTRUCTION OF A LOCAL ENCODER BASED ON GNN-TRANSFORMER

In the federated learning framework, each client independently maintains and updates a local personalized learning path encoder to model the student's dynamic cognitive state in the knowledge space. Let $S_i = \{(k_\tau, r_\tau)\}_{\tau=1}^{T}$ be the historical learning behavior sequence of a student on client i, let k_τ denote the knowledge point interacted with at step τ, and let $r_\tau \in \{0, 1\}$ be the response result (0 for an incorrect answer, 1 for a correct answer). T is the sequence length.

First, the paper defines a structured prior knowledge graph G=(V,E) on the knowledge point set K, where nodes $v_k \in V$ correspond to knowledge points k and edges $(v_a, v_b) \in E$ represent prior dependencies or semantic associations. A two-layer graph convolutional network (GCN) is used to encode the graph structure [18,19]:

$$H^{(0)} = E_k \in R^{|K| \times d} \tag{1}$$

$$H^{(l)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l-1)}W^{(l)}\right), \quad l = 1, 2 \tag{2}$$

In Formula (1), E_k is the learnable knowledge point embedding matrix. In Formula (2), $\tilde{A}$ is the adjacency matrix with self-loops added, $\tilde{D}$ is its degree matrix, $W^{(l)}$ is the trainable parameter of the l layer, and $\sigma(\cdot)$ is the ReLU activation function. The behavior sequence is mapped to a vector sequence. For each step τ, a joint input vector is constructed as follows:

$$x_\tau = [z_{k_\tau}; r_\tau \cdot w_r + (1 - r_\tau) \cdot w_w] \in R^{2d} \tag{3}$$

In Formula (3), w_r, w_w represent the learnable weight vectors of the mastered and unmastered states, respectively, and $[\cdot;\cdot]$ represents vector concatenation. x_τ is input into the Transformer encoder with positional encoding. The positional encoding uses the form of sine and cosine functions [20]:

$$\begin{aligned} PE(pos, 2j) &= \sin\left(\frac{pos}{10000^{2j/2d}}\right), \\ PE(pos, 2j+1) &= \cos\left(\frac{pos}{10000^{2j/2d}}\right) \end{aligned} \tag{4}$$

In Formula (4), pos is the time step position and j is the dimension index. The Transformer encoding layer calculates the context representation through the multi-head self-attention mechanism [21,22]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V \tag{5}$$

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{6}$$

Among them, $\text{head}_m = \text{Attention}(XW_m^Q, XW_m^K, XW_m^V)$, and W_m^Q, W_m^K, W_m^V, and W^O are learnable parameters. After l layers of stacking, the output sequence $\{s_\tau\}_{\tau=1}^T$ is obtained, and the final state $s_T \in R^{2d}$ is taken as the global representation of the current learning state. This state vector can serve as the input feature for subsequent recommendation tasks to predict the next knowledge point to be recommended. Figure 1 illustrates the overall architectural flow of the local personalized learning path encoder. The input is a student's historical learning behavior sequence. First, knowledge point embedding is combined with the knowledge graph structure, and a GCN is used to extract semantically relevant knowledge representations. Simultaneously, the behavior sequence is vectorized and positionally encoded before being fed into a Transformer encoder to capture temporal dependencies and long-term interaction patterns. The two features are combined in the fusion module to generate a joint representation that encompasses both cognitive state and structural priors. The final output, sT, serves as a global vector representation of the current learning state, which is used for subsequent personalized recommendation tasks. This design effectively integrates knowledge logic with individual behavioral dynamics, enabling accurate modeling of learning paths. Notably, the knowledge embeddings generated by GNN not only form part of the input features but also serve as initialization vectors for the Key and Value in the Transformer, enabling dynamic guidance of knowledge structure in modeling behavioral sequences.

DYNAMIC KNOWLEDGE DISTILLATION STRATEGY DESIGN

The term "dynamic" in this paper refers to the adaptive adjustment of distillation loss weights and the continuous updating of soft labels as the global model evolves, rather than changes in temperature parameters or distillation frequency. To alleviate model conflicts caused by non-IID data in federated learning, this paper introduces a dynamic knowledge distillation mechanism within a global-local collaborative training framework to achieve cross-client knowledge transfer and semantic alignment [23,24]. In each communication cycle $t \in \{1, 2, \ldots, T_{\text{comm}}\}$, the server maintains the current global model $f_{\theta_g}^{(t)}(\cdot)$ and holds a lightweight public calibration data $D_{\text{calib}} = \{(x_n, y_n)\}_{n=1}^N$ covering the distribution of major knowledge points. This set does not contain true labels for supervised training and is only used to generate soft targets. The server performs forward propagation on D_{calib} and obtains logits $z_n^{\text{global}} = f_{\theta_g}^{(t)}(x_n) \in R^{N_c}$ output produced by the global model, where N_c is the total number of candidate knowledge points. A smooth probability distribution (i.e., soft label) is generated through the temperature scaling function:

$$p_n^{\text{global}} = \text{Softmax}\left(\frac{z_n^{\text{global}}}{T}\right), \quad T > 1 \tag{7}$$

The temperature parameter T in Formula (7) controls the degree of distribution smoothness and preserves the relative relationship information between categories. This set of soft labels $\{p_n^{\text{global}}\}_{n=1}^N$ is sent to all clients participating in this round of training along with the global model. When each client i uses its private dataset D_i for local training, the knowledge distillation loss is introduced into the optimization objective. For each sample, the local model $f_{\theta_i}^{(t)}(\cdot)$ outputs logits z_n^{local} and the corresponding soft prediction:

$$q_n^{\text{local}} = \text{Softmax}\left(\frac{z_n^{\text{local}}}{T}\right) \tag{8}$$

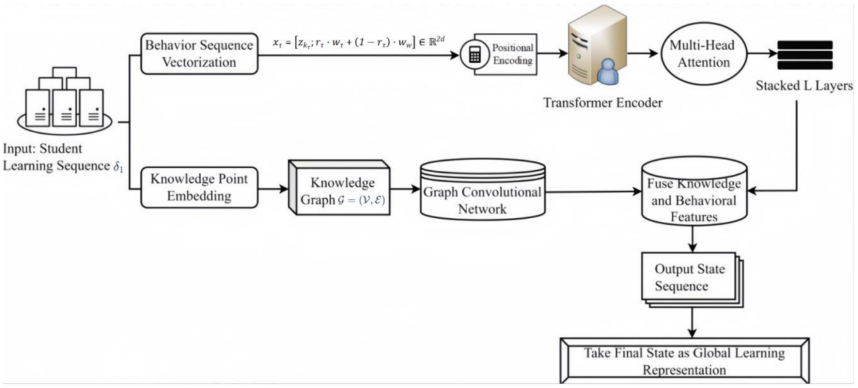

FIGURE 1. Structure Diagram

KL divergence is used to measure the semantic difference between the local model and the global model, and the distillation loss term is defined as:

$$\begin{aligned}\mathcal{L}_{\mathrm{kd}}^{(i)} &= \frac{1}{N}\sum_{n=1}^{N} D_{\mathrm{KL}}(p_n^{\mathrm{global}} \| q_n^{\mathrm{local}}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{N_c} p_n^{\mathrm{global}}(c)\log\frac{p_n^{\mathrm{global}}(c)}{q_n^{\mathrm{local}}(c)}\end{aligned} \tag{9}$$

At the same time, the standard cross entropy loss is retained to ensure the ability to fit local labels:

$$\mathcal{L}_{\mathrm{ce}}^{(i)} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{N_c}\mathbb{I}[y_n = c]\log \mathrm{Softmax}(z_n^{\mathrm{local}})(c) \tag{10}$$

Ultimately, the total training loss for client i is a dynamically weighted combination of:

$$\mathcal{L}^{(i)} = \alpha(t)\cdot\mathcal{L}_{\mathrm{ce}}^{(i)} + (1-\alpha(t))\cdot\mathcal{L}_{\mathrm{cd}}^{(i)} \tag{11}$$

The weight coefficient α(t) in Formula (11) is adjusted with the communication round according to the cosine annealing strategy [25]:

$$\alpha(t) = \frac{1}{2}\left(1+\cos\left(\pi\cdot\frac{t}{T_{\mathrm{comm}}}\right)\right) \tag{12}$$

In the initial phase, t→0 is α(t)→1 , emphasizing local supervised learning. As training progresses, α(t) gradually decreases, enhancing knowledge transfer and promoting model consistency. Gradient updates are based on total loss backpropagation. This mechanism effectively transfers global semantic knowledge without transmitting the original data, providing a high-quality local model foundation for subsequent adaptive aggregation.

ADAPTIVE AGGREGATION MECHANISM BASED ON KL DIVERGENCE

This paper proposes an adaptive weight allocation mechanism based on distribution similarity. This mechanism dynamically adjusts the contribution weights of each participating client during each round of global aggregation, giving nodes with more representative data distributions a higher priority in model updates, thereby improving the generalization ability and training stability of the global model [26,27]. Assume that there are M clients C(t)={c1,c2,...,cM} participating in the training in round t of communication. Each client i∈C(t) locally calculates the empirical distribution of the knowledge point categories in its training samples:

$$q_i(k) = \frac{1}{|D_i|}\sum_{(k_\tau, r_\tau)\in D_i}\mathbb{I}[k_\tau = k], \quad \forall k \in K \tag{13}$$

In Formula (13), $\mathbb{I}[\cdot]$ is the indicator function, and K is the set of all knowledge points. The server collects the knowledge point distribution $\{q_i(k)\}_{i=1}^{M}$ of all participating clients and calculates the global average distribution:

$$\bar{q}(k) = \frac{1}{M}\sum_{i=1}^{M} q_i(k), \quad k \in K \tag{14}$$

Then the KL divergence between the local distribution $q_i(k)$ of client i and the global distribution $\bar{q}(k)$ is calculated as a quantitative indicator of its data deviation:

$$D_{\mathrm{KL}}(q_i \| \bar{q}) = \sum_{k\in K} q_i(k)\log\frac{q_i(k)}{\bar{q}(k)+\epsilon} \tag{15}$$

In Formula (15), ϵ is a smoothing term to prevent logarithmic zero value overflow. The KL divergence used in this paper measures the difference between the local knowledge experience distribution of the client and the global average distribution, rather than the divergence of the model output or parameters. The distribution deviation is converted into aggregation weight, and the negative exponential weighted mapping function is introduced:

$$w_i' = \exp\left(-\beta\cdot D_{\mathrm{KL}}(q_i \| \bar{q})\right) \tag{16}$$

The hyperparameter β>0 in Formula (16) controls the weight decay rate and adjusts the system's sensitivity to distribution differences. The weights are further normalized to ensure the numerical stability and interpretability of the aggregation operation:

$$w_i(t) = \frac{w_i'}{\sum_{j=1}^{M} w_j'} = \frac{\exp\left(-\beta \cdot D_{\mathrm{KL}}(q_i \| \bar{q})\right)}{\sum_{j=1}^{M} \exp\left(-\beta \cdot D_{\mathrm{KL}}(q_j \| \bar{q})\right)} \quad (17)$$

In the model aggregation phase, the server performs weighted fusion of the model parameter increments $\Delta\theta_i(t) = \theta_i(t) - \theta_g(t-1)$ uploaded by each client:

$$\Delta\theta_g(t) = \sum_{i=1}^{M} w_i(t)\Delta\theta_i(t) \quad (18)$$

Update global model parameters:

$$\theta_g{}^{(t)} = \theta_g{}^{(t-1)} + \eta_g \cdot \Delta\theta_g{}^{(t)} \quad (19)$$

In Formula (19), ηg is the global learning rate. This adaptive weight mechanism effectively suppresses the misleading influence of extreme non-IID clients on the global model, enhances the robustness and semantic consistency of the aggregation process, and provides a more reliable basic model support for subsequent personalized recommendation tasks. The weight distribution heatmap is shown in Figure 2.

Figure 2 shows the dynamic evolution of client weights during training under the adaptive aggregation mechanism. The horizontal axis represents five clients (A–E), and the vertical axis represents communication rounds (t=1, 5, 10, 15, 20). Initially (t=1), all clients have a weight of 0.20, demonstrating fair aggregation. As training progresses, the system dynamically adjusts the contribution ratio of each client based on its data distribution and the global KL divergence. Clients A, B, and C, whose distributions are close to the overall network, see their weights gradually increase to 0.26–0.29. D has a slight shift, with its weight slowly decreasing to 0.15. Due to the highly heterogeneous data, E's weight rapidly decays from 0.20 to 0.02, significantly reducing its impact on the global model. This heatmap intuitively reflects the core concept of the proposed approach, which implements representative data-driven aggregation through an adaptive mechanism, thereby ensuring model consistency and enhancing robustness to non-IID data.

INTEGRATION OF DIFFERENTIAL PRIVACY MECHANISMS

To ensure the privacy of student behavior data during federated learning, this paper integrates a Gaussian-mechanism-based differential privacy (DP) strategy into the model update process and combines it with a secure aggregation protocol to achieve end-to-end privacy preservation [28,29]. First, the gradient is clipped to limit its sensitivity. The L2-norm gradient clipping operation is used:

$$\widetilde{\Delta\theta}_i^t = \Delta\theta_i^t \cdot \min\left(1, \frac{B}{\|\Delta\theta_i^t\|_2}\right) \quad (20)$$

In Formula (20), B is the preset clipping threshold, which ensures that the maximum impact of a single client update is controlled within a fixed range. After clipping, the global sensitivity Δf of the gradient satisfies:

$$\Delta f = \max_{D_i \sim D_i'} \left\| \widetilde{\Delta\theta}_i^t(D_i) - \widetilde{\Delta\theta}_i^t(D_i') \right\|_2 \le B \quad (21)$$

Then, random noise from a Gaussian distribution is added to the clipped gradient to achieve (ϵ,δ)-differential privacy:

$$\widehat{\Delta\theta}_i(t) = \widetilde{\Delta\theta}_i(t) + n_i(t), \quad n_i(t) \sim \mathcal{N}(0, \sigma^2 B^2 I_d) \quad (22)$$

The noise standard deviation is controlled by the privacy parameter σ. All participating clients upload the noisy model update $\widehat{\Delta\theta}_i(t)$ through a secure aggregation protocol. This protocol is based on secret sharing and homomorphic encryption technology to ensure that the server can only obtain the aggregated results. To accurately monitor the consumption of the privacy budget, this paper adopts the Rényi Differential Privacy (RDP) framework for cumulative analysis [30]. Let T_{dp} represent a noisy Gaussian mechanism operation. The Rényi divergence under the order $\lambda \in (1, \infty)$ is defined as:

$$D_\lambda(T_{\mathrm{dp}}(D_i) \| T_{\mathrm{dp}}(D_i')) = \frac{\lambda}{2\sigma^2} + \frac{1}{\lambda - 1} \log\left[\Phi\left(\frac{\sigma - \frac{\lambda}{C}}{\sqrt{\sigma^2}}\right) - e^{\lambda\epsilon}\Phi\left(\frac{-\sigma - \frac{\lambda-1}{B}}{\sqrt{\sigma^2}}\right)\right] \quad (23)$$

In Formula (23), $\Phi(\cdot)$ is the standard normal cumulative distribution function. After federated training, the total privacy overhead is calculated using the RDP combination theorem, and the corresponding (ϵ,δ)-DP bound is obtained through transformation. When the cumulative privacy loss exceeds the preset threshold, the system automatically terminates the training process. This mechanism achieves strict privacy protection for the learner's sensitive behavior data while minimizing utility loss.

FEDERATED COLLABORATIVE TRAINING PROCESS

This paper constructs a periodic, privacy-safe federated collaborative training framework to achieve efficient collaboration between the global model and multiple local learning path recommenders. The entire process integrates dynamic knowledge distillation, adaptive aggregation, and differential privacy mechanisms to form a closed-loop optimization structure, while ensuring that data remains within the domain. The training process is iterated in communication rounds, and the specific steps include client selection and model delivery. At the beginning of each round, the server randomly selects M_t clients from all N registered clients, where $M_t \le M_{\mathrm{sel}}$, to participate in the training process. The following condition is satisfied:

$$P(c_i \in C(t)) = p_i, \quad \forall i \in \{1, \ldots, N\} \quad (24)$$

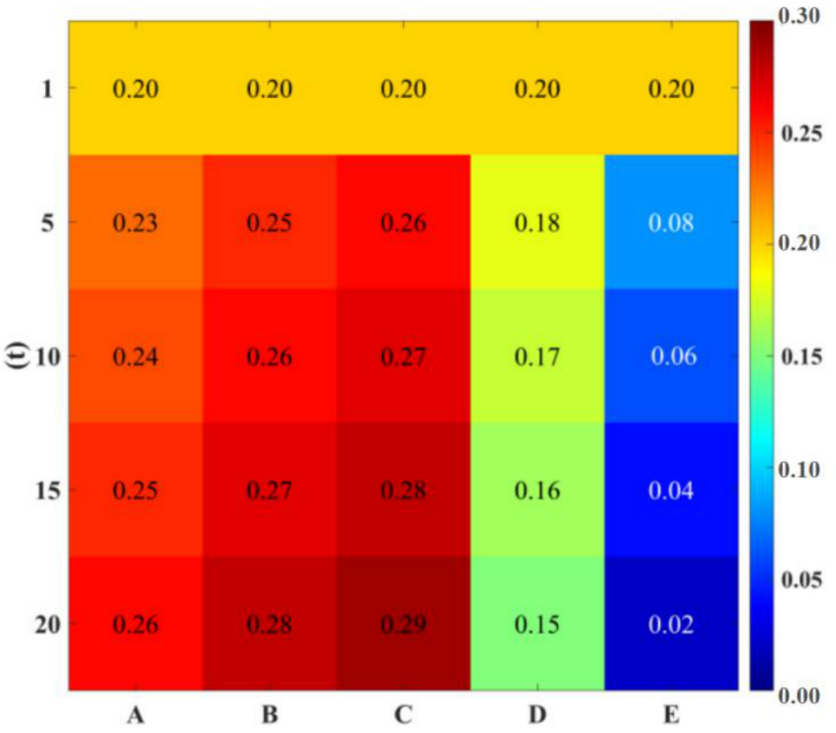

FIGURE 2. Weight Distribution Heatmap

In Formula (24), C(t) is the t-round participant set, and pi can be weighted sampled based on device status, data volume, or historical response rate. After selection, the server encrypts and sends the current global model parameters to each selected client as the initial weight for local training. Local knowledge-aware training and noise-added uploading are then performed. After each client receives the model, it performs E rounds of local iterative optimization on its private dataset Di . The e -th update uses stochastic gradient descent:

$$\theta_i^{(e)} = \theta_i^{(e-1)} - \eta_l \nabla_\theta \mathcal{L}^{(i)}(\theta_i^{(e-1)}; \xi_e) \tag{25}$$

In Formula (25), ηl is the local learning rate, ξe is the small batch sample, and the loss function L(i) contains the cross-entropy term and the knowledge distillation term. After completing the local training, the model update amount is calculated:

$$\Delta\theta_i(t) = \theta_i^{(E)} - \theta_g(t-1) \tag{26}$$

Gradient clipping and Gaussian noise injection are then performed. The client uploads the model update after gradient clipping and Gaussian noise injection to the server via a Transport Layer Security (TLS) encrypted channel, and records the Rényi privacy cost of this contribution. The server then accumulates the total privacy expenditure for the current round through privacy budget checks and adaptive aggregation:

$$\rho^{(t)}(\alpha) = \sum_{i \in C(t)} \rho_i^{(t)}(\alpha) = M_t \cdot \frac{\alpha}{2\sigma^2} \tag{27}$$

And update the cumulative RDP budget:

$$\rho_{\text{cum}}^{(t)}(\alpha) = \sum_{r=1}^{t} \rho^{(r)}(\alpha) \tag{28}$$

If $\epsilon(\delta) = \inf_{\alpha>1} \left(\rho_{\text{cum}}^{(t)}(\alpha) + \frac{\log(1/\delta)}{\alpha-1} \right) \geq \epsilon_0$ is achieved, training is terminated to ensure (ϵ_0, δ)-DP compliance. Otherwise, adaptive weighted aggregation based on KL divergence is performed. Finally, loop control is executed and the termination conditions are checked. The above process is repeated until any termination condition is met: reaching the maximum number of communication rounds, global loss convergence, or privacy budget is exceeded. This collaborative training mechanism achieves the unified optimization of security, personalization, and generalization, providing a sustainable modeling paradigm for intelligent recommendation in distributed education environments.

III. EXPERIMENTS AND VALIDATION

EXPERIMENTAL SETUP

This experiment established a cross-institutional federated learning environment, including interaction logs from the exam preparation platform and practice records of knowledge points in online courses. Five client nodes were selected to simulate a real-world non-IID scenario: the knowledge point coverage, difficulty distribution, and learning

path patterns varied significantly across the clients. Each client's data was split into a training set and a test set at an 8:2 ratio. The test set was used to evaluate the performance of personalized recommendations. This study was approved by the Ethics Committee of Changchun Normal University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. The experimental setup is shown in Table 1.

TABLE 1. Experimental Setup

Parameter name	Value/Setting
Number of Clients	5
Train/Test Ratio	8:2
Batch Size	32
Local Learning Rate	0.001
Global Learning Rate	1
Communication Rounds	30
Noise Scale	1.2
Clipping Threshold	1
Distillation Temperature	4
KL Weighting Hyperparameter	0.5

RECOMMENDATION ACCURACY RESULTS AND ANALYSIS

The top-5 recommendation hit rate (Hit@5) is used to measure the accuracy of the model's prediction of the next knowledge point. For each test example, the model outputs the top five recommended knowledge points. If the knowledge point is actually mastered, it is counted as a hit. The calculation formula is:

$$\text{Hit@5} = \frac{1}{T}\sum_{\tau=1}^{T}\mathbb{I}[k_{\tau+1} \in R_{\tau+1}] \tag{29}$$

Figure 3 shows the Hit@5 comparison results of the proposed method against FedAvg, FedProx, and SCAFFOLD (Federated Learning via Stochastic Controlled Averaging) on five clients under varying degrees of Non-IID data. The horizontal axis represents clients A through E, and the vertical axis represents the Hit@5 performance metric. The proposed method achieves optimal performance across all types of non-IID data and across all clients. Under mild non-IID conditions, the proposed method achieves an average of 69.5 $\pm$ 0.8%, significantly higher than FedAvg (0.621 $\pm$ 0.001), FedProx (0.634 $\pm$ 0.001), and SCAFFOLD (0.642 $\pm$ 0.001). In moderate and severe scenarios, the proposed method maintains a stable advantage, achieving average Hit@5 scores of 0.681 $\pm$ 0.009 and 0.651 $\pm$ 0.011, respectively, significantly higher than other methods, with an overall average of 0.676 $\pm$ 0.009. On client D, the proposed method maintains a hit rate of 0.607 $\pm$ 0.012 even under severe non-IID conditions, far exceeding FedAvg's 0.510 $\pm$ 0.019, demonstrating greater robustness and generalization.

Traditional aggregation methods (such as FedAvg) treat model parameters as homogeneous vectors and simply average them, ignoring the differences in the mapping of the cognitive semantics encoded by these parameters across the knowledge spaces of different clients. The dynamic knowledge distillation mechanism proposed in this paper uses soft labels generated from server-side public calibration data to construct a cross-domain, unified cognitive coordinate system. By fitting this global coordinate system, the local models of each client achieve nonlinear alignment of their internal knowledge representations. This ensures that when the models of client A and client B both predict a certain knowledge point, the meaning of the logits they output is comparable and consistent, laying a solid semantic foundation for subsequent aggregation. At the same time, adaptive aggregation based on KL divergence goes beyond simply suppressing noise or outliers. From an information-theoretic perspective, it treats each client's contribution as an observation of the global knowledge distribution and assigns trust weights based on the representativeness of their observations.

RECOMMENDATION SEQUENCE CONSISTENCY RESULTS AND ANALYSIS

The normalized discounted cumulative gain (NDCG@10) is used to evaluate the rationality of the recommendation list order. Ideal Discounted Cumulative Gain (IDCG) and DCG are defined as follows:

$$\begin{aligned}\text{DCG@10} &= \sum_{j=1}^{10}\frac{2^{rel_j}-1}{\log_2(j+1)},\\ \text{IDCG@10} &= \sum_{j=1}^{10}\frac{2^{rel_j^*}-1}{\log_2(j+1)}\end{aligned} \tag{30}$$

In Formula (30), $rel_j \in \{0,1\}$ indicates whether the j-th recommendation matches the real target, and rel_j^* is its ideal ranking score. The final indicator is:

$$\text{NDCG@10} = \frac{\text{DCG@10}}{\text{IDCG@10}} \tag{31}$$

Figure 4 shows the sequence prediction consistency (NDCG@10) of the proposed method compared with FedAvg, FedProx, and SCAFFOLD on five clients under mild, moderate, and severe Non-IID data conditions. As data heterogeneity increases, the NDCG@10 values of all methods show a downward trend, reflecting that the quality of the recommendation order is significantly affected by distribution shift. FedAvg's average NDCG@10 under mild Non-IID is 0.660 $\pm$ 0.011, dropping to 0.589 $\pm$ 0.017 under severe Non-IID, showing significant performance fluctuations. The proposed method achieves average scores of 0.732 $\pm$ 0.007, 0.715 $\pm$ 0.008, and 0.689 $\pm$ 0.010 in mild, moderate, and severe scenarios, respectively, with an overall average of 0.712 $\pm$ 0.008, consistently significantly outperforming other methods. On client D, the proposed approach maintained a high score of 0.647 $\pm$ 0.011 under severe non-IID conditions, significantly exceeding FedAvg's score of 0.543 $\pm$ 0.018 and demonstrating greater robustness

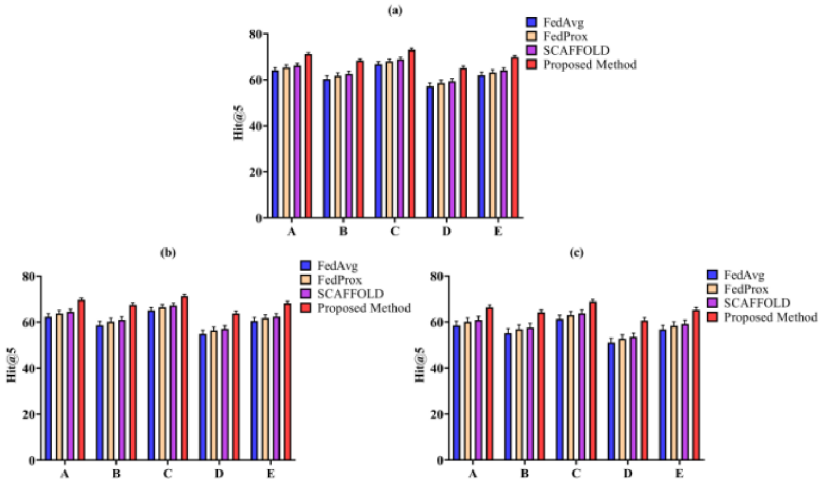

FIGURE 3. Learning Path Recommendation Accuracy; Figure 3(a). Mild Non-IID; Figure 3(b). Moderate Non-IID; Figure 3(c). Severe

and consistency in learning path ranking tasks. The GNN-Transformer architecture does not simply concatenate graph structure and sequence information, but rather constructs a knowledge-driven attention guidance system. The knowledge point embedding vectors learned by the GNN layer are used as the initialization basis for the keys and values of the multi-head self-attention in the Transformer. This mechanism of dynamically injecting static knowledge structures into the sequential decision-making process ensures that the recommendation results possess inherent and explainable rationality. Simultaneously, the soft labels generated by the server using public calibration data, with their inherent ranking probability distribution, provide all clients with a global consensus on the relative importance and dependency strength of knowledge points. The high NDCG@10 score of the proposed method stems not only from its accurate fit to historical behavior but also from the model's successful integration of decentralized local experience with a unified domain knowledge system.

MODEL CONVERGENCE AND STABILITY ANALYSIS

The Variance Index (VI) is defined to quantify the volatility of the convergence process. After each round of communication, the mean L2 distance between each client's uploaded gradient $\Delta\theta_i^t$ and the average gradient $\bar{\Delta}^t$ is calculated:

$$VI(t) = \frac{1}{M_t}\sum_{i=1}^{M_t} \|\Delta\theta_i^t - \bar{\Delta}^t\|_2, \qquad \bar{\Delta}^t = \frac{1}{M_t}\sum_{j=1}^{M_t} \Delta\theta_j^t \tag{32}$$

Figure 5a, 5b, 5c shows the variance index comparison results of the proposed method against FedAvg, FedProx, and SCAFFOLD at different communication rounds under mild, moderate, and severe Non-IID data distributions, thereby assessing the consistency of model updates during federated aggregation. The horizontal axis represents communication rounds, and the vertical axis represents the VI performance metric. As the degree of Non-IID increases, the VI values of all methods show an upward trend, indicating that client gradient divergence increases with increasing data heterogeneity. Under severe scenarios, FedAvg's average VI reaches 1.00, significantly higher than the 0.76 under mild scenarios. Under severe conditions, the average VI of the proposed method is only 0.60, far lower than FedAvg's 1.00, FedProx's 0.92, and SCAFFOLD's 0.86.

The proposed approach builds an inherently stable collaborative framework through dual mechanisms. First, dynamic knowledge distillation establishes an implicit contract between all clients regarding knowledge point relationships and cognitive logic through server-generated soft labels. This process forcibly pulls decentralized optimization paths toward a shared cognitive subspace, reducing fundamental conflicts between clients' update directions. Second, adaptive aggregation based on KL divergence acts as a dynamic reputation system. It does not pre-set any client weights, but rather dynamically allocates voice based on the fit between each client's current data distribution and the current global knowledge system. Even in high-noise or highly heterogeneous environments, the global model can evolve smoothly and consistently, guided by a majority of reliable nodes. This achieves an extremely low variance index and ensures high stability during the training process. While the GNN-Transformer encoder and dynamic knowledge distillation introduce additional computational overhead, these operations are executed exclusively on the client side without increasing parameter upload volume per round. Furthermore, their faster convergence reduces total communication rounds, resulting in communication costs comparable to FedAvg.

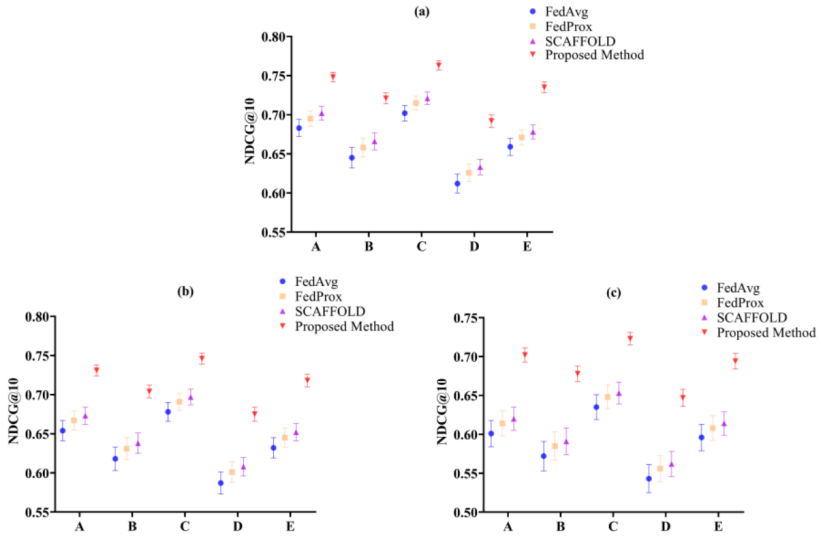

FIGURE 4. Sequence Prediction Consistency; Figure 4(a). Mild Non-IID; Figure 4(b). Moderate Non-IID; Figure 4(c). Severe Non-IID

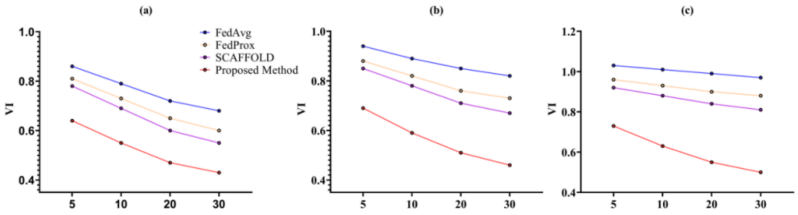

FIGURE 5. Model Convergence Stability; Figure 5(a). Mild Non-IID; Figure 5(b). Moderate Non-IID; Figure 5(c). Severe Non-IID

PERSONALIZED ADAPTABILITY ANALYSIS

The paper introduced the personalized score difference rate to evaluate the responsiveness of local fine-tuning. The metric is defined as: for each user, the average absolute change rate between the recommendation score vectors of all candidate knowledge points before and after local training is calculated, and then averaged across all users. For each user, the paper calculated the rate of change in their recommendation score vector before and after local training:

$$\mathrm{PSDR}_u = \frac{\|s_u^{\mathrm{post}} - s_u^{\mathrm{pre}}\|_2}{\|s_u^{\mathrm{pre}}\|_2 + \epsilon} \tag{33}$$

Figure 6 shows the difference in personalized scores between the proposed method and FedAvg, FedProx, and SCAFFOLD across five clients under mild, moderate, and severe Non-IID scenarios. These results are used to evaluate the model's responsiveness to locally learned features. As data heterogeneity increases, the variance of personalized scores across all methods generally decreases. FedAvg drops from an average of 0.20 under mild conditions to 0.15 under severe conditions, demonstrating its difficulty in effectively capturing individual differences in highly dispersed data environments. The proposed method achieves average personalized score variances of 0.40, 0.38, and 0.35 under mild, moderate, and severe Non-IID conditions, respectively, with an overall average of 0.38. This consistently outperforms other methods, reaching a maximum of 0.43 (mild) on client D and maintaining 0.35 under the most adverse conditions, demonstrating strong personalized adaptability. This demonstrates that the proposed method, after federated aggregation, can still fully respond to individual student behavior patterns through local fine-tuning, achieving precisely customized recommendations.

The knowledge distillation mechanism, guided by the global model, preserves the learning freedom of the local model. This allows soft labels to provide semantic guidance rather than mandatory constraints, leaving room for personalized updates. The GNN-Transformer architecture fuses the common logic of the knowledge graph with the individual trajectories of student behavior sequences to construct a state representation that is both generalizable and discriminative. The adaptive aggregation mechanism prevents the global model from being dominated by a small number of distributions and tends to over-average, ensuring the effectiveness of local updates. These mechanisms collectively enable the model to absorb global knowledge while deeply adapting to local features without sacrificing privacy, demonstrating greater individual sensitivity and adaptability in the rate of change of recommendation scores.

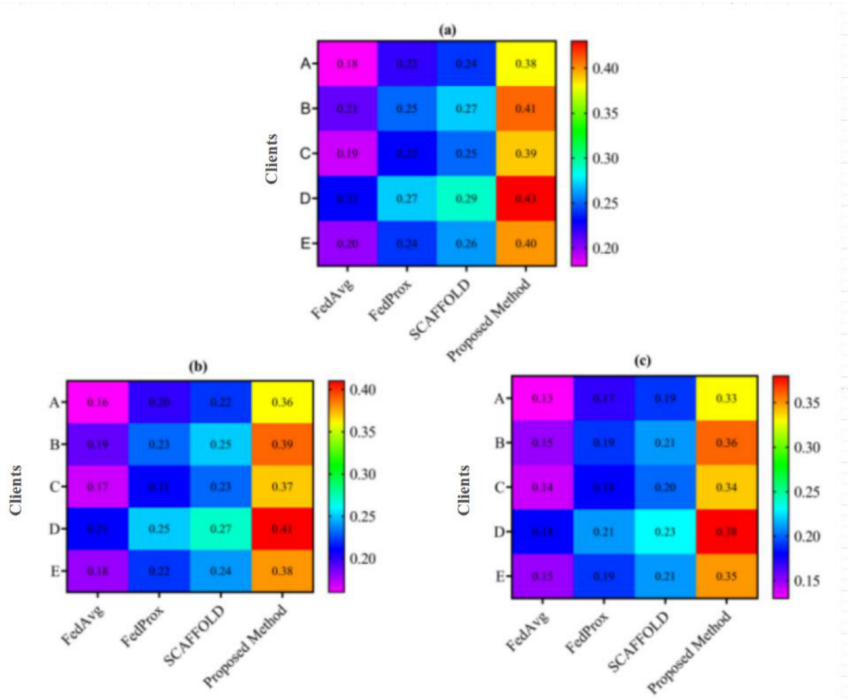

FIGURE 6. Personalized Adaptation Capability; Figure 6(a). Mild Non-IID; Figure 6(b). Moderate Non-IID; Figure 6(c). Severe Non-IID

GLOBAL GENERALIZATION PERFORMANCE

Cross-domain F1-score is used to evaluate the model's transferability. Data from external institutions that did not participate in training are selected as the test set, and comprehensive indicators under the classification task are calculated:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \\ \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{34}$$

In Formula (34), TP, FP and FN represent the number of correct/false positive/missing recommendations, respectively. This indicator reflects the adaptability of the global model to the new environment. Figure 7 compares the cross-domain F1-score results of the proposed method with those of FedAvg, FedProx, and SCAFFOLD on four external test sets not used in training, thereby evaluating the model's global generalization ability. All methods exhibit some performance fluctuation across target domains, reflecting the impact of data distribution differences on transfer effectiveness. FedAvg's F1-score on test domain D' is only 0.576 $\pm$ 0.017, significantly lower than its performance on C' (0.631 $\pm$ 0.013), indicating its strong dependence on the source domain. The proposed method maintains a high level of performance across all target domains: 0.687 $\pm$ 0.008 on A', 0.673 $\pm$ 0.009 on B', 0.701 $\pm$ 0.007 on C', and 0.654 $\pm$ 0.009 on D', for an average of 0.679 $\pm$ 0.008, an improvement of 0.051 over SCAFFOLD, with a smaller standard deviation. This demonstrates not only high accuracy but also strong cross-environmental stability and good external adaptability.

This exceptional generalization performance is achieved through the coordinated optimization of dynamic knowledge distillation and an adaptive aggregation mechanism. The global model uses soft labels to convey the semantic relationships and cognitive logic between knowledge points, rather than simply fitting local statistical patterns, making the learned representations more transferable. Furthermore, a weight adjustment mechanism based on KL divergence effectively suppresses model bias caused by client data drift, forcing global parameters to focus on common knowledge structures. The GNN-Transformer encoder integrates the prior logic of the knowledge graph into local modeling, enhancing its deep understanding of knowledge point dependencies. This allows the model to make reasonable inferences based on common cognitive pathways even when faced with unprecedented learning behavior patterns in new organizations. These designs collectively enhance the model's cross-domain reasoning capabilities under privacy isolation and data heterogeneity conditions, enabling a transition from local fitting to universal cognitive modeling.

PRIVACY-UTILITY TRADEOFF ANALYSIS

Define the privacy cost ratio to quantify the performance loss efficiency of the DP mechanism:

$$\text{PCR} = \frac{\sigma}{\Delta\text{Acc}} \tag{35}$$

In Formula (35), σ is the noise intensity, and ΔAcc is the accuracy drop after the introduction of DP. Figure 8 compares the privacy-utility tradeoff performance of

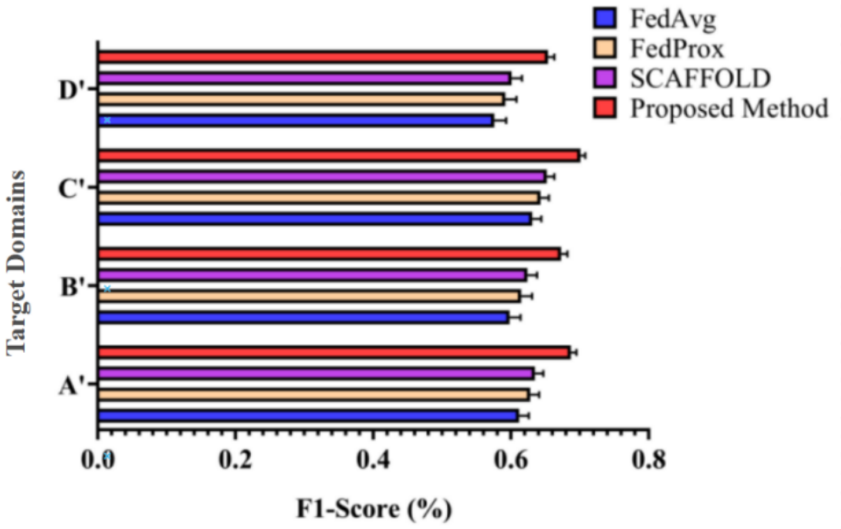

FIGURE 7. Global Generalization Performance

the proposed method with that of FedAvg, FedProx, and SCAFFOLD under mild, moderate, and severe non-IID data distributions. As data heterogeneity increases, the privacy-cost ratios of all methods decrease. FedAvg dropped from 0.351 ± 0.012 in the mild scenario to 0.321 ± 0.018 in the severe scenario, indicating that in high-bias scenarios, due to exacerbated model update conflicts, DP noise has a more significant impact on performance. The privacy-cost ratio of the proposed method in the three scenarios is 0.588 ± 0.007, 0.579 ± 0.008, and 0.552 ± 0.010, respectively, which is consistently higher than that of the other methods and is associated with smaller standard deviations. This demonstrates that the proposed method minimizes accuracy loss at the same privacy cost and optimally balances privacy and recommendation utility. Even under severe non-IID conditions, the proposed method maintains a privacy cost ratio exceeding 0.55, demonstrating its high efficiency and robustness in complex distributed environments.

The dynamic knowledge distillation mechanism provides a stable, global navigation signal for all clients. This signal, composed of soft labels, defines the macro-level direction and semantic boundaries of model optimization. When DP noise disrupts local gradients, the knowledge distillation loss term acts as a corrective force, continuously pulling the deviated optimization path back to the pre-set cognitive track, preventing noise accumulation from causing model divergence. The adaptive aggregation mechanism based on KL divergence acts as a noise filter. It identifies and reduces the weight of clients that experience significant fluctuations due to highly heterogeneous data. Because these clients' updates inherently contain large variance, their uncertainty increases exponentially when noise is added. By reducing their aggregate weights, the system effectively mitigates the destructive effects of the combined uncertainty of high noise and high bias. This approach goes beyond simply tolerating noise. Through knowledge guidance and intelligent weighting, it proactively offsets the negative effects of noise at the architectural level. This approach maintains efficient model convergence and excellent recommendation performance while maintaining strong privacy protection.

IV. CONCLUSION

Building upon the urgent need for tailored workforce upskilling in high-tech manufacturing, particularly across the geographically dispersed factories of the modern industry, this paper proposes a federated personalized learning path recommendation method that combines dynamic knowledge distillation and adaptive aggregation to leverage decentralized training data while ensuring skill alignment across all sites. By constructing a GNN-Transformer local encoder, the method jointly models knowledge graph structure and student behavior sequences. Public calibration data are introduced to achieve dynamic knowledge distillation and improve cross-client semantic consistency. The GNN-Transformer encoder provides each client with a unified representation foundation that combines knowledge structure and individual behavior. Dynamic knowledge distillation guides local models to align with global semantics on this basis, while adaptive aggregation dynamically updates based on the representativeness of the representation distribution. These three components form a closed loop of representation–guidance–aggregation, jointly improving personalization and generalization. In addition, an adaptive aggregation mechanism based on KL divergence is designed to effectively alleviate the bias caused by Non-IID data, and differential privacy is integrated to ensure user security. Experiments show that this method outperforms FedAvg, FedProx, and SCAFFOLD in terms of recommendation accuracy (average Hit@5: $67.6 \pm 0.9\%$), sequence consistency (average NDCG@10: 0.712 ± 0.008), and model stability. It achieves the coordinated optimization of privacy protection, global generalization, and personalized adapta-

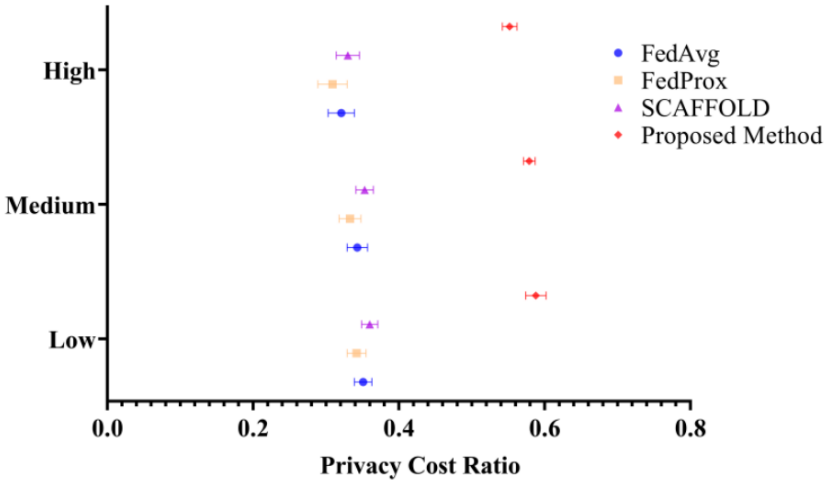

FIGURE 8. Privacy-Utility Tradeoff

tion, thereby providing an efficient and feasible solution for intelligent recommendation in distributed education environments, which is directly applicable to reducing the critical skills gap and boosting production quality within the modern manufacturing sector.

AUTHOR CONTRIBUTIONS

Jing Zhang designed, collected and analyzed the data, and drafted the manuscript. Jing Zhang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jing Zhang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Changchun Normal University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] D. Gm, R. H. Goudar, A. A. Kulkarni, V. N. Rathod, and G. S. Hukkeri, "A digital recommendation system for personalized learning to enhance online education: A review," IEEE Access, vol. 12, pp. 34019-34041, 2024, doi: 10.1109/ACCESS.2024.3369901.

[2] A. Ellikkal and S. Rajamohan, "AI-enabled personalized learning: Empowering management students for improving engagement and academic performance," Vilakshan-XIMB Journal of Management, vol. 22, no. 1, pp. 28-44, 2025, doi: 10.1108/XJM-02-2024-0023.

[3] P. K. Myakala, A. K. Jonnalagadda, and C. Bura, "Federated learning and data privacy: A review of challenges and opportunities," International Journal of Research Publication and Reviews, vol. 5, no. 12, Art. no. 10.55248, 2024, [Online]. Available: https://ssrn.com/abstract=5086425.

[4] J. Fan, H. Lian, and W. Liu, "Privacy-preserving AI analytics in cloud computing: A federated learning approach for cross-organizational data collaboration," Spectrum of Research, vol. 4, no. 2, pp. 1-21, 2024, [Online]. Available: http://spectrumofresearch.com/index.php/sr/article/view/15.

[5] T. Z. Sana, S. Abdulla, A. Das, A. Nag, M. M. Hassan, Z. Z. Fiza, et al., "Advancing Federated Learning: A Systematic Literature Review of Methods, Challenges, and Applications," IEEE Access, vol. 13, pp. 153817-153844, 2025, doi: 10.1109/ACCESS.2025.3605165.

[6] Z. Zhang, Y. Zhang, D. Guo, S. Zhao, and X. Zhu, "Communication-efficient federated continual learning for distributed learning system with Non-IID data," Science China Information Sciences, vol. 66, no. 2, Art. no. 122102, 2023, doi: 10.1007/s11432-020-3419-4.

[7] L. Yang, J. Huang, W. Lin, and J. Cao, "Personalized federated learning on non-IID data via group-based metalearning," ACM Transactions on Knowledge Discovery from Data, vol. 17, no. 4, pp. 1-20, 2023, doi: 10.1145/3558005.

[8] L. Jiang, K. Liu, Y. Wang, D. Wang, P. Wang, and Fu Y et al, "Reinforced explainable knowledge concept recommendation in MOOCs," ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 3, pp. 1-20, 2023, doi: 10.1145/3579991.

[9] J. W. Tzeng, N. F. Huang, Y. H. Chen, T. W. Huang, and Y. S. Su, "Personal learning material recommendation system for MOOCs based on the LSTM neural network," Educational Technology & Society, vol. 27, no. 2, pp. 25-42, 2024, [Online]. Available: https://www.jstor.org/stable/48766161.

[10] R. Gu, "Personalized learning path based on graph attention mechanism deep reinforcement learning research on recommender systems," Journal of Computational Methods in Sciences and Engineering, vol. 25, no. 3, pp. 2411-2426, 2025, doi: 10.1177/14727978241313260.

[11] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. O. Alzahrani, and H. A. Alzahrani, "An adaptable and personalized framework for top-N course recommendations in online learning," Scientific Reports, vol. 14, no. 1, Art. no. 10382, 2024, doi: 10.1038/s41598-024-56497-1.

[12] T. Hussain, L. Yu, M. Asim, A. Ahmed, and M. A. Wani, "Enhancing e-learning adaptability with automated learning style identification and sentiment analysis: A hybrid deep learning approach for smart education," Information, vol. 15, no. 5, pp. 277, 2024, doi: 10.3390/info15050277.

[13] A. A. Nemati, M. H. Sadreddini, and M. Mahdizade, "Comparative Federated Algorithms for Solving Non-IID Data Challenges," Risk Assessment and Management Decisions, vol. 1, no. 2, pp. 277-283, 2024, doi: 10.48314/ramd.v1i2.56.

[14] Y. Xu, Y. Zhu, Z. Wang, H. Xu, and Y. Liao, "Enhancing Federated Learning Through Layer-Wise Aggregation Over Non-IID Data," IEEE Transactions on Services Computing, vol. 18, no. 2, pp. 798-811, 2025, doi: 10.1109/TSC.2025.3536309.

[15] E. Yu, Z. Ye, Z. Zhang, L. Qian, and M. Xie, "A federated recommendation algorithm based on user clustering and meta-learning," Applied Soft Computing, vol. 158, Art. no. 111483, 2024, doi: 10.1016/j.asoc.2024.111483.

[16] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," IEEE transactions on neural networks and learning systems, vol. 34, no. 12, pp. 9587-9603, 2022, doi: 10.1109/TNNLS.2022.3160699.

[17] N. Singh, J. Rupchandani, and M. Adhikari, "Personalized federated learning for heterogeneous edge device: Self-knowledge distillation approach," IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 4625-4632, 2023, doi: 10.1109/TCE.2023.3327757.

[18] Y. Wang, S. Guo, D. Qiao, G. Liu, and M. Li, "Fedsg: A personalized subgraph federated learning framework on multiple non-iid graphs," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 8, no. 5, pp. 3678-3690, 2024, doi: 10.1109/TETCI.2024.3372381.

[19] G. Dai and J. Tang, "A Short-Term Traffic Flow Prediction Method Based on Personalized Lightweight Federated Learning," Sensors, vol. 25, no. 3, pp. 967, 2025, doi: 10.3390/s25030967.

[20] A. Belhadi, Y. Djenouri, F. A. de Alcantara Andrade, and G. Srivastava, "Federated constrastive learning and visual transformers for personal recommendation," Cognitive Computation, vol. 16, no. 5, pp. 2551-2565, 2024, doi: 10.1007/s12559-024-10286-0.

[21] Z. Li, Z. Zhong, P. Zuo, and H. Zhao, "A personalized federated learning method based on the residual multi-head attention mechanism," Journal of King Saud University-Computer and Information Sciences, vol. 36, no. 4, Art. no. 102043, 2024, doi: 10.1016/j.jksuci.2024.102043.

[22] S. Jiang, M. Lu, K. Hu, J. Wu, Y. Li, L. Weng, et al., "Personalized federated learning based on multi-head attention algorithm," International Journal of Machine Learning and Cybernetics, vol. 14, no. 11, pp. 3783-3798, 2023, doi: 10.1007/s13042-023-01864-z.

[23] Y. Jiang, X. Zhao, H. Li, Xue, and Y, "A Personalized Federated Learning Method Based on Knowledge Distillation and Differential Privacy," Electronics, vol. 13, no. 17, pp. 3538, 2024, doi: 10.3390/electronics13173538.

[24] Z. W. Tang, S. W. Xu, H. Jin, S. Liu, R. Zhai, and K. Lu, "Personalized federated learning via decoupling self-knowledge distillation and global adaptive aggregation," Multimedia Systems, vol. 31, no. 2, pp. 1-20, 2025, doi: 10.1007/s00530-025-01719-3.

[25] E. Ardic and Y. Genc, "Enhanced Privacy and Communication Efficiency in Non-IID Federated Learning With Adaptive Quantization and Differential Privacy," IEEE Access, vol. 13, pp. 54322-54337, 2025, doi: 10.1109/ACCESS.2025.3554138.

[26] Y. Liu, S. Li, W. Li, H. Qian, and H. Xia, "A Personalized Federated Learning Algorithm Based on Dynamic Weight Allocation," Electronics, vol. 14, no. 3, pp. 484, 2025, doi: 10.3390/electronics14030484.

[27] X. Han, Q. Zhang, Z. He, and Z. Cai, "Confidence-based similarity-aware personalized federated learning for autonomous IoT," IEEE Internet of Things Journal, vol. 11, no. 7, pp. 13070-13081, 2023, doi: 10.1109/JIOT.2023.3337520.

[28] Z. Tan, J. Le, F. Yang, M. Huang, T. Xiang, and X. Liao, "Secure and accurate personalized federated learning with similarity-based model aggregation," IEEE Transactions on Sustainable Computing, vol. 10, no. 1, pp. 132-145, 2024, doi: 10.1109/TSUSC.2024.3403427.

[29] X. Wu, L. Xu, and L. Zhu, "Local differential privacy-based federated learning under personalized settings," Applied Sciences, vol. 13, no. 7, pp. 4168, 2023, doi: 10.3390/app13074168.

[30] X. Yang, W. Huang, and M. Ye, "Dynamic personalized federated learning with adaptive differential privacy," Advances in Neural Information Processing Systems, vol. 36, pp. 72181-72192, 2023, [Online]. Available: https://github.com/xiyuanyang45/DynamicPFL.