Review

A Systematic Survey on Federated Sequential Recommendation

Yichen Li , Qiyu Qin , Gaoyang Zhu , Wenchao Xu , Haozhao Wang , Yuhua Li , Rui Zhang [*] , Ruixuan Li [*]

Article

A Systematic Survey on Federated Sequential Recommendation

Yichen Li [1], Qiyu Qin [1,†], Gaoyang Zhu [1,†], Wenchao Xu [2], Haozhao Wang [1], Yuhua Li [1], Rui Zhang [1,*] and Ruixuan Li [1,*]

[1] School of Computer Science and Technology, Huazhong University of Science and Technology; ycli0204@hust.edu.cn (Y.L.); m202474246@hust.edu.cn (Q.Q.); zgy75602@hust.edu.cn (G.Z.); hz_wang@hust.edu.cn (H.W.); idcliyuhua@hust.edu.cn (Y.L.)

[2] Department of Computing, The Hong Kong Polytechnic University; wenchao.xu@polyu.edu.hk

[*] Correspondence: rayteam@yeah.net (R.Z.); rxli@hust.edu.cn (R.L.)

[†] These authors contributed equally to this work.

Abstract: Sequential recommendation is an advanced recommendation technique that utilizes the sequence of user behaviors to generate personalized suggestions by modeling the temporal dependencies and patterns in user preferences. However, it requires a server to centrally collect users' data, which poses a threat to the data privacy of different users. In recent years, federated learning has emerged as a distributed architecture that allows participants to train a global model while keeping their private data locally. This survey pioneers Federated Sequential Recommendation (FedSR), where each user joins as a participant in federated training to achieve a recommendation service that balances data privacy and model performance. We begin with an introduction to the background and unique challenges of FedSR. Then, we review existing solutions from two levels, each of which includes two specific techniques. Additionally, we discuss the critical challenges and future research directions in FedSR.

Keywords: federated learning; sequential recommendation; federated sequential recommendation

---

1. Introduction

Sequential recommendation has emerged as a pivotal area of research in the field of recommender systems [1,2]. Unlike traditional recommendation approaches that treat user interactions as independent events, sequential recommendation acknowledges the inherent order and temporal dependencies in user behavior [3]. By modeling the sequence of actions or items that users engage with over time, these systems aim to capture the dynamic nature of user preferences and generate more accurate and personalized recommendations. However, in such cases, data collection in central entities, referred to as centralized training, is often infeasible or impractical due to data privacy concerns or country regulations [4]. As a potential remedy to this dilemma, federated learning emerges as a promising approach, enforcing data localization and enabling the distributed training of a globally shared model [5,6]. This framework has achieved remarkable success and has been applied to various fields, such as recommendation systems [7] and smart healthcare [8].
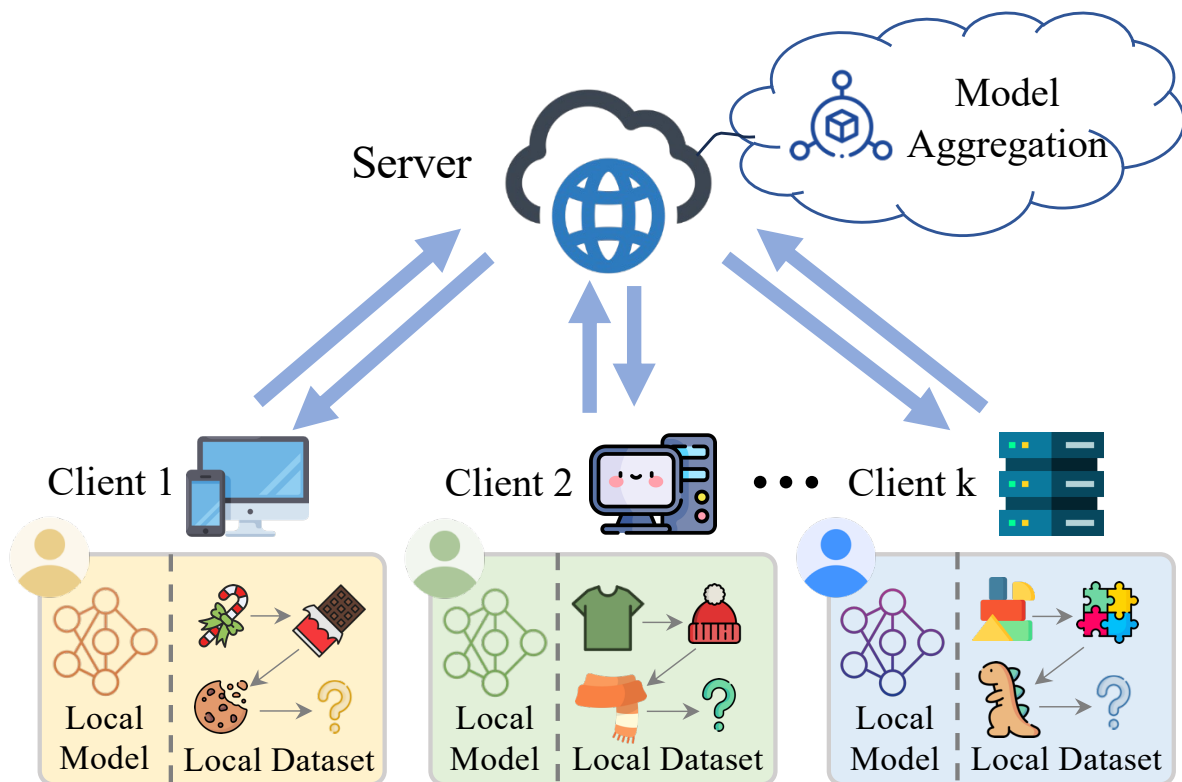
Figure 1. The framework of FedSR. In FedSR, each user participates in the federated training with the local model and dataset. Firstly, the user trains the local model with the private dataset locally and then uploads the model parameters to the server. The server aggregates the local models and broadcasts the global model to the users participating in the next communication round.

In recent years, Federated Sequential Recommendation (FedSR) has been proposed to allow users to collaboratively train a shared model locally without breaching their privacy [9]. Moreover, it effectively reduces the demand for computing resources on the server while efficiently utilizing the local resources of distributed devices [7]. Given the aforementioned advantages, FedSR has gained success in many applications. For instance, [10] proposes using pseudo-labeling and secret-sharing techniques to protect privacy while ensuring the performance of federated recommendation systems. [11] utilizes Word2Vec to efficiently learn embeddings for users and items, combined with privacy protection mechanisms to achieve cross-device federated recommendation. FedNRM is proposed in [12] to design a personalized news recommendation model that combines federated learning and privacy-preserving technologies to provide high-quality recommendation services while protecting privacy. The framework of FedSR is illustrated in Figure **??**.

Although many research works investigate integrating FL algorithms with sequential recommendation, several challenges must be solved before FedSR can be scalable. Firstly, data heterogeneity across clients presents a significant challenge. Each user participates in federated training as a participant, and there are significant differences in preferences among individual users. Additionally, the wide variety of items results in sparse patterns for each user, necessitating robust aggregation strategies to ensure that the global model generalizes well. Secondly, communication efficiency is crucial. Compared to ResNet [13] backbones commonly used in other tasks like image classification, sequential recommendation models are often large in size, resulting in significant communication overhead. Meanwhile, in FL systems, there are often numerous clients with limited resources, requiring reduced communication without sacrificing performance. Thirdly, user privacy requires extra care. While FL keeps data localized, uploading the model updates may still leak sensitive information with malicious attacks, which may include the user's personal information.

While surveys on federated learning and sequential recommendation exist, the existing studies treat the two topics separately. The remaining surveys primarily focus on federated recommenda-

tions [14,15]. As an essential branch of recommendation systems with numerous research works, sequential recommendation merits separate analysis due to its value and significance. Moreover, most existing surveys still concentrate on the classification of federated learning based on heterogeneity and security, neglecting the technical classification of recommendation techniques. This paper proceeds from two levels and distills two mainstream techniques for each (parameter decomposition and LLM foundation for model-level, communication optimization, and aggregation balance for device-level).

Specific contributions of our survey are as follows: 1) We present the first comprehensive survey of recent advancements in federated sequential recommendation, backgrounds, related works and new insights. 2) Our systematic categorization of these methods into two subcategories based on their defining characteristics offers a thorough and structured overview. Based on this, we further elaborate on the existing methods by identifying and discussing two key techniques for each category. 3) We highlight the current challenges and potential future directions in federated sequential recommendations. We intend to shed light on under-researched aspects to spur possible paths within this field.

2. Preliminaries

The main objective of this section is to establish a solid foundation for the reader by providing the background and fundamentals of FedSR. We will first analyze the backgrounds of federated learning and sequential recommendation. Then, we discuss the new insights associated with FedRS.

2.1. Federated Learning

In recent years, due to growing concerns about privacy leakage, Federated Learning (FL) has been introduced for machine learning across distributed local clients. FL enables multiple users to collaboratively contribute to a global model by exchanging and aggregating model parameters [5,6]. This approach reduces communication overhead and effectively preserves data privacy, as only parameter transfers are involved. Google introduced the foundational FL algorithm, FedAvg, in 2017 [6]. Since then, numerous studies have aimed to optimize it further. For instance, FedProx [5] addresses the heterogeneity issue by adding a regularization term to penalize local model divergence from the global model. MOON [16] enhances local training of different clients by applying contrastive learning at the model level, comparing model representations.

Next, we will outline the federated learning process in three main steps: 1) The server initializes a global model w^0 and transmits it to all clients after determining the training tasks and specifying hyper-parameters; 2) In each communication round t, a random subset of clients is selected. These clients receive the current global model w^t and perform SGD on their local data. After local training, they upload the updated parameters to the server; 3) During the aggregation phase, the server combines the local models and distributes the updated global model w^{t+1} to the clients participating in the next round. The aggregation is performed as follows:

$$w^{t+1} = \sum_{k \in S_t} \frac{|D_k|}{|D|} w_k^{t+1}. \tag{1}$$

where $|D|$ is the total amount of data, $|D_k|$ is the amount of data on client k, and S_t is the set of clients participating in the t-th round.

2.2. Sequential Recommendation

In recent years, with the extensive application of user behavioral data, Sequential Recommendation (SR) has gradually become a key area of research in recommender systems. Its core objective is to predict the next item of interest based on the chronological order of a user's historical interaction behaviors. Early studies on sequential recommendation commonly relied on Markov chains to capture the sequential dependencies of user behaviors. For example, FPMC [17] combines a Markov chain with a matrix factorization model to simultaneously capture users' long-term interests and short-term

sequential transitions. Later, deep learning is extensively used to model user behavior sequences. Convolutional neural networks can be applied to sequential recommendation [18] by extracting sequential patterns from users' historical interactions through convolution operations.

We divide the main process of sequential recommendation into the following three stages: 1) Input Sequence Construction: Extract behavioral sequences from user interaction logs and represent them as item ID lists or embedding matrices in chronological order; 2) Sequential Pattern Modeling: Based on the backbone model, capture dependencies in user sequences and learn latent feature representations of these sequential patterns; 3) Next-Step Behavior Prediction: Utilize the modeled user sequence features to predict the user's next item of interest according to the current context or historical records, typically outputting a Top-N recommendation list. By continuously optimizing sequential modeling methods, this process effectively enhances the performance of recommender systems in dynamic environments, thereby providing significant value to personalized services and user experience.

2.3. New Insights in Sequential Recommendation by Federated Learning

Compared to training the sequential recommendation model by collecting data centrally, utilizing the FL paradigm brings three major characteristics. Firstly, FL is inherently designed for privacy preservation, as it trains a global model by allowing raw data to remain on the participants' local devices. Compared to other privacy-preserving technologies, FL does not require complex statistical computations for data encryption or the use of perturbations that sacrifice model accuracy for security. FL has been theoretically proven to achieve the performance upper bound of centralized training. Then, Although the performance of FL can be affected by non-IID data, the data heterogeneity in FedSR differs from traditional FL in that it primarily stems from variations in user information among different participants. This variation persists even in traditional sequential recommendations, where the model still needs to integrate and analyze information from different users' raw data. In FedSR, the participants are users, and during the training process, the server also needs to integrate and analyze information from updates uploaded by different users. Last but not least, introducing FL can enhance the system's update efficiency, enabling each user to analyze and upgrade their local model using local resources dynamically. It also allows each user to customize their information security, thereby providing a better recommendation service for users.
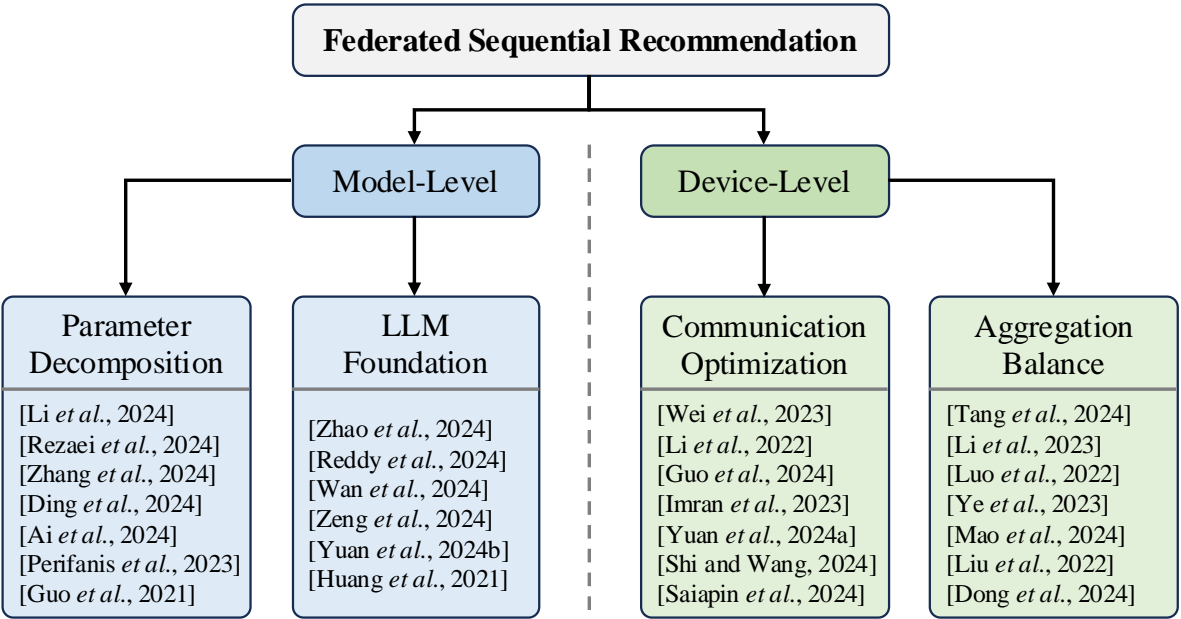

Figure 2. Taxonomy of FedSR. We classify them into two subcategories and four primary techniques, i.e., model-level (parameter decomposition and LLM foundation) and device-level (communication optimization and aggregation balance). Different colors indicate categories, and we list representative works in the boxes.

3. Solutions for Federated Sequential Recommendation

This section outlines a series of targeted solutions designed for FedSR. Considering the research methods of both FL and SR, we categorize the existing studies into model-level and device-level based on the level targeted by the technique. Then, for each level, we separately investigate two major techniques. We illustrate the outline of this survey in Figure **??**.

3.1. Model-Level

Model-level approaches aim to optimize model architectures and training strategies to balance the trade-offs between privacy preservation, personalized recommendations, and recommendation effectiveness. Specifically, parameter decomposition methods divide model parameters into global and local components, enabling collaborative learning across diverse data sources. This approach allows for the sharing of common information across users while maintaining adaptability to individual behaviors. On the other hand, LLM foundation methods leverage the powerful sequential modeling capabilities of LLMs, combined with the federated learning framework, to enhance the understanding of complex user behavior patterns. These methods improve recommendation accuracy and diversity while preserving user privacy.

3.1.1. Parameter Decomposition

Parameter decomposition has become an effective strategy for addressing the trade-off between model performance and privacy preservation in federated sequential recommendation tasks. Traditional federated learning methods often face challenges when dealing with Non-IID data, heterogeneous models, and computationally intensive tasks. Introducing decomposition in model representation enables efficient collaboration across diverse data sources, tasks, and model components. Through the adoption of global parameters to capture universal patterns across users while leveraging local parameters to adapt to individual user behaviors, this approach ensures privacy protection while optimizing recommendation performance. In sequential recommendation tasks, it guarantees model personalization and enhances robustness.

Several studies aim to balance generalization and personalization. For example, [19] proposes a semi-global modeling framework that dynamically optimizes personalization and generalization through alternating local training and global aggregation. On the client side, the model captures personalized features using local data; on the server side, aggregated model information updates the semi-global model, preserving global knowledge. This EM-like optimization strategy groups user behaviors to generate sub-models that align with different user group preferences. Experimental results demonstrate that during the personalized training phase, the proposed method achieves up to a 21.9% performance improvement on the ML-latest dataset using its two-stage training framework. This improvement is particularly significant in datasets with highly heterogeneous user preferences. The decomposition strategy between semi-global and personalized components significantly enhances the model's performance in sequential recommendation tasks. Another approach to addressing the generalization-personalization trade-off leverages knowledge distillation techniques. FedDist-POIRec [20] extracts knowledge from locally trained models in the form of soft labels, which are softened representations of model parameters. These soft labels are then uploaded to the server for aggregation, avoiding the direct parameter synchronization common in traditional federated learning. This method further improves the model's adaptability to heterogeneous data distributions.

Parameter decomposition has also demonstrated strong applicability in multi-scenario and cross-domain recommendation tasks. FedDCSR [21] proposes a cross-domain recommendation framework based on disentangled representation learning to address the challenges of privacy preservation and cross-domain knowledge sharing in cross-domain recommendation tasks. Disentangling domain-shared and domain-specific representations effectively captures user behavior characteristics across different domains. On the client side, local models focus on capturing domain-specific interests, while on the server side, the aggregation of domain-shared representations enhances the global

model's generalization capability. Experimental results show that FedDCSR significantly outperforms all baselines across various metrics, highlighting the critical role of disentangled representation learning and contrastive information strategies in capturing both intra-domain and inter-domain user preferences. This provides an effective solution to the problem of cross-domain knowledge transfer. For multi-task scenarios, [22] proposes a personalized federated learning framework. Through a multi-task aggregation strategy, the framework balances multiple scenarios within the global model. Its core lies in introducing personalized modules for each task, where local models adapt to user needs in specific scenarios while sharing global knowledge for inter-task collaboration.
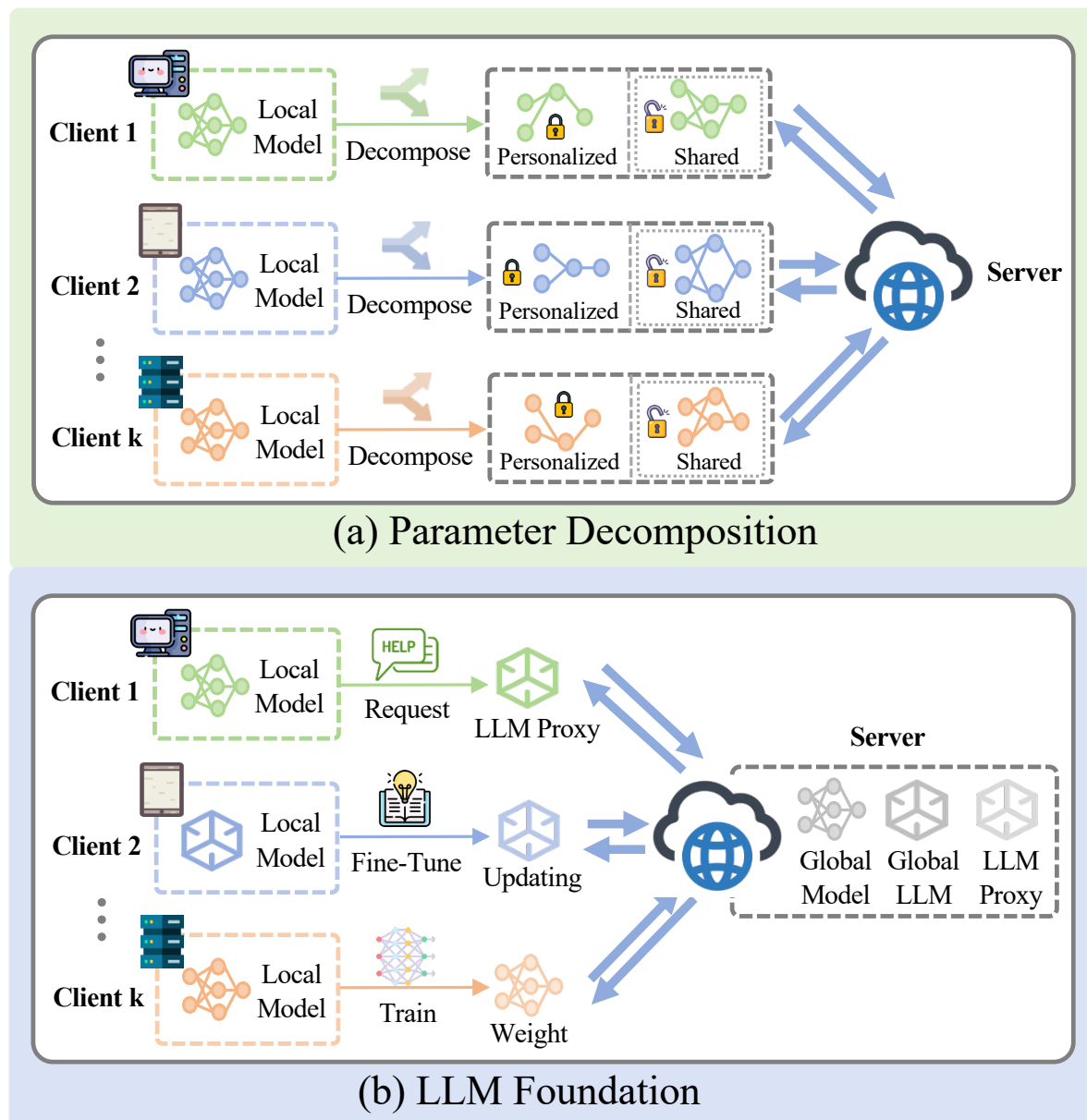

Figure 3. Model-level methods, including parameter decomposition and LLM foundation. Parameter decomposition captures universal patterns across clients while leveraging local parameters to adapt to local dataset, and LLM foundation guides knowledge retention and parameter updates by integrating knowledge from LLM.

In addition to multi-scenario and cross-domain research, personalization optimization and dynamic adaptation have emerged as critical directions for parameter decomposition methods. FML-Rec [23] introduces a meta-learning framework that captures shared characteristics among users through global meta-training while utilizing fast fine-tuning to adapt to local user behavior data. This

method performs exceptionally well in cold-start scenarios. The dynamic adaptation capability of meta-learning demonstrates greater flexibility and robustness in highly heterogeneous data scenarios.

Parameter decomposition methods have also shown unique value in specialized tasks. Fed-POIRec [24] incorporates social influence modeling to capture social relationships and user interaction behaviors, further enhancing the model's ability to personalize point-of-interest recommendations. Regarding sequential modeling capabilities, PREFER [25] focuses on point-of-interest recommendations in edge scenarios. Lightweight model designs and efficient parameter update strategies significantly optimize recommendation performance. Experimental results indicate that these methods not only improve recommendation metrics but also exhibit strong adaptability in heterogeneous data scenarios.

3.1.2. LLM Foundation

Beyond parameter decomposition-based approaches, a series of methods leveraging Large Language Models (LLM) have been developed to address the challenges of federated sequential recommendation. These methods combine the powerful sequential modeling capabilities of LLMs with the privacy-preserving mechanisms of federated learning, providing innovative solutions for recommendation systems. By leveraging LLMs' deep semantic modeling of user behavior sequences, item features, and contextual information, they significantly enhance recommendation performance. Meanwhile, the federated learning framework ensures privacy protection by avoiding the direct transmission of sensitive data, achieving a balance between privacy preservation and personalized recommendations. In recent years, researchers have proposed various innovative techniques focusing on deep modeling, retrieval-augmented generation, reinforcement learning, and knowledge enhancement, driving the practical application of federated sequential recommendation in complex tasks.

The introduction of LLM in federated sequential recommendation research has significantly enhanced the ability to capture complex patterns in user behavior and generate more accurate personalized recommendations. PPLR is proposed in [26] to combine the powerful semantic representation capabilities of LLMs with the privacy-preserving mechanisms of federated learning, achieving dual optimization of recommendation performance and user privacy protection. This method leverages pre-trained LLMs to generate high-dimensional semantic representations of user behaviors, item descriptions, and contextual information, enabling richer feature expression. Furthermore, it enhances the generalization capability of global models and the personalization of local models through parameter aggregation under the federated learning framework and personalized fine-tuning. Experimental results demonstrate that this method outperforms other privacy-preserving federated recommendation methods across three datasets: Games, MicroLens, and Book while achieving performance close to centralized LLM-based approaches. This establishes a groundbreaking solution for the future development of federated recommendation systems. In addition to LLMs, Transformer and lightweight deep learning architectures have shown significant potential in federated sequential recommendation.

The study in [27] introduces the transformer architecture into federated recommendation systems. By leveraging multi-head attention mechanisms, the model effectively captures complex patterns in user behavior sequences. At the same time, its lightweight Transformer design makes it well-suited for federated learning scenarios with limited communication resources, enhancing global user behavior modeling capabilities while maintaining privacy protection. Another study, Fed-AttGRU [28], combines attention mechanisms with GRU, focusing on the most important parts of user behavior sequences and modeling both short-term preferences and long-term interests. Experimental results indicate that Fed-AttGRU performs exceptionally well in long-sequence modeling and highly heterogeneous data scenarios, highlighting its value in federated recommendation research.

For cold-start and sparse data scenarios, strategies combining retrieval and generation have provided practical solutions for federated recommendations. GPT-FedRec [29] proposes a retrieval-augmented generation framework, which generates more contextually relevant recommendations through a hybrid retrieval and generation strategy. The federated learning framework protects user data during local training while integrating retrieval and generation modules, which improves system performance in sparse data and cold-start scenarios. GPT-FedRec demonstrated remarkable

improvements across six datasets and four evaluation metrics, achieving average increases of 36.12% in Hit@5, 29.88% in NDCG@5, 45.44% in Hit@10, and 36.56% in NDCG@10. his study highlights the potential of combining retrieval and generation techniques in the federated recommendation and further advances the application of such systems to address cold-start challenges.

Regarding knowledge enhancement, FELLAS [30] proposes a framework that leverages LLMs as external services to enhance federated sequential recommendations. By interacting with LLMs, the framework provides local models with rich semantic information and contextual features, enabling the recommendation model to capture better user behavior characteristics in sparse data and long-tail scenarios. By optimizing the frequency of LLM calls and communication strategies, FELLAS effectively reduces local resource consumption and avoids the challenges of directly deploying complex LLM models, showcasing the potential of external knowledge enhancement in FedSR.

Furthermore, LLM technologies have been widely applied in dynamically optimized recommendation tasks. [31] introduces deep reinforcement learning into federated learning. By locally training reinforcement learning agents, this method models the dynamics and long-term preferences of users' schedules. Meanwhile, global parameter sharing integrates knowledge across multiple users. This combination of reinforcement learning and federated learning demonstrates advantages in dynamic recommendation tasks and further validates the applicability of federated learning frameworks in complex recommendation scenarios.

3.2. Device-Level

Device-level approaches aim to address the heterogeneity challenges in the federated sequential recommendation by enhancing the performance and adaptability of recommendation systems through the optimization of model aggregation strategies. To better capture user behavior characteristics, optimizing device selection and device clustering mechanisms becomes crucial. These methods effectively improve both model training efficiency and recommendation performance by focusing on optimizations at the device level. In addition to optimizing device selection and clustering mechanisms to capture user behavior features better, reducing communication overhead is also a vital component. Optimizing communication strategies can effectively alleviate the communication burden between clients and servers, thereby accelerating the global model update speed and reducing latency and resource consumption during the training process.

3.2.1. Communication Optimization

Communication optimization primarily focuses on reducing the communication overhead between clients and servers, enhancing system operational efficiency, and decreasing overall communication costs. In addition to reducing communication overhead through techniques such as sparsification, which involves transmitting only critical parameter updates and compressing model updates, some communication optimization strategies also balance the reduction of overhead with the protection of user privacy, which is especially important in distributed environments. Firstly, in the KG-FedTrans4Rec framework [32], the authors optimize communication efficiency through an edge device aggregation mechanism. In this model, edge devices not only perform local training but also undertake information aggregation tasks, thereby reducing the frequency of model updates and the volume of data transmission.

In [33], the authors propose a method to effectively reduce the amount of communication data by utilizing low-rank tensor projections and device clustering strategies. This approach models users' long-term preferences by projecting user behavior data into a low-rank space and transmitting only smaller updates during each communication round. Similarly, researchers in [34] address communication challenges in horizontal federated recommendation systems by applying model compression techniques. By quantizing and sparsifying model parameters, they significantly decrease the volume of transmitted gradient data, thereby improving communication efficiency. Furthermore, the proposed strategy, based on gradient optimization for model compression, further reduces the size of transmitted data by sharing compressed model updates, optimizing overall communication efficiency.
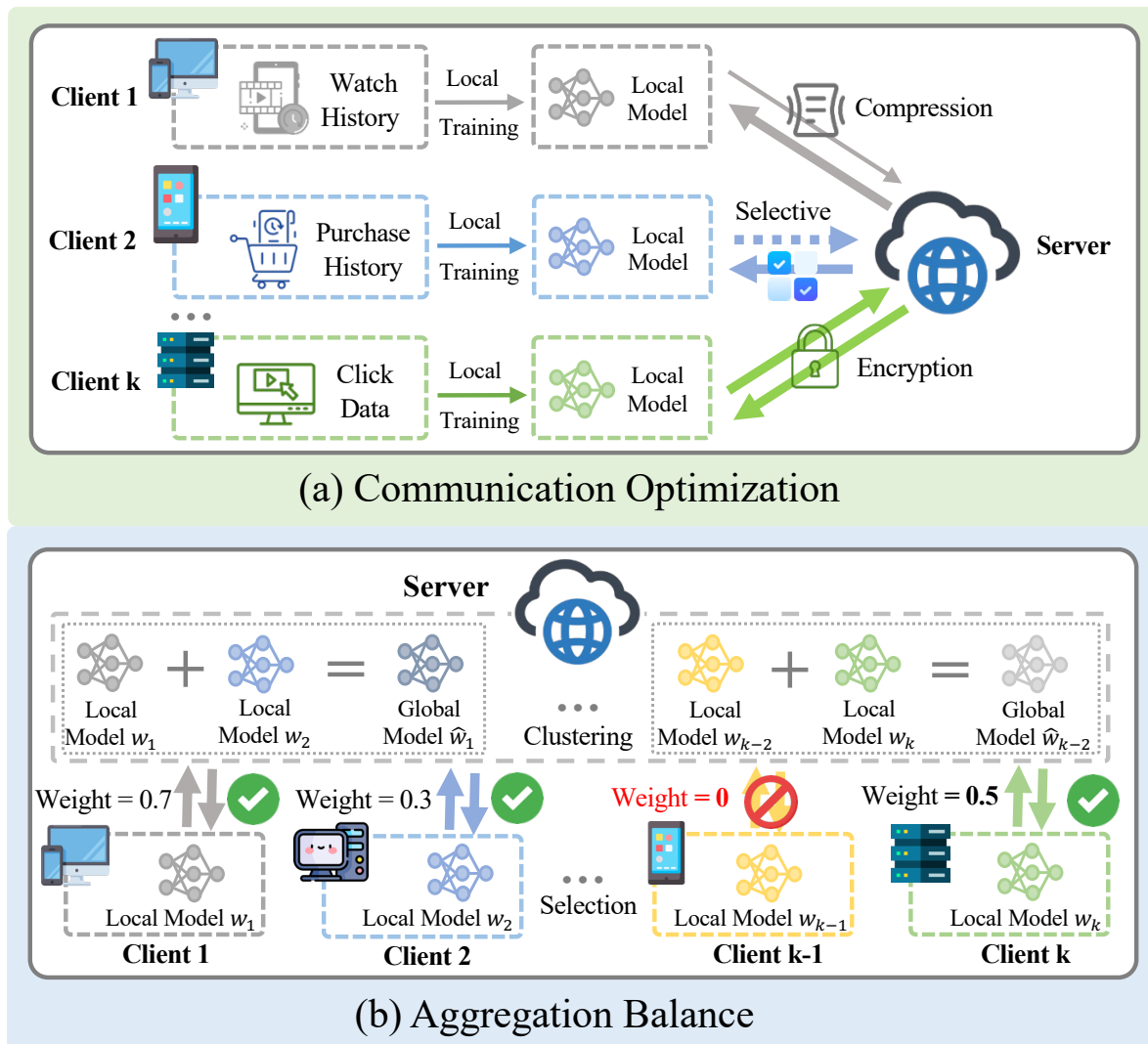
Figure 4. Device-level methods, including communication optimization and aggregation balance. Communication optimization focuses on reducing the communication cost between clients and the server, and the aggregation balance often strategically aggregates models to maximize the optimization of the global model.

Beyond data compression, selective data transmission can also be employed. ReFRS [35] significantly reduces communication data volume by transmitting only the encoder parameters of the client-embedded model to the server. The server side employs asynchronous dynamic clustering methods and semantic samplers to extract low-dimensional embedding representations and group similar client models, thereby avoiding sensitivity to Non-IID data and reducing communication overhead. ReFRS demonstrates superior memory and computational efficiency, with embeddings using only 270KB compared to FedAvg and FedFast, which require 2.43MB-3.78MB, and batch computation time reduced from 45 minutes for FedAvg and FedFast to 3 seconds. Furthermore, in PTF-FSR [36], the authors take it a step further by proposing a parameter-free federated learning framework that replaces model parameters with user-generated sequential data for knowledge transfer, completely eliminating the high communication overhead associated with parameter transmission. This approach is suitable for the efficient operation of large-scale complex models, with PTF-FSR's communication cost ranging from only 1.2KB to 2.4KB, significantly lower than traditional federated methods such as Fed-SASRec, which range from 1.6MB to 14.8MB, reducing communication overhead by more than an order of magnitude.

Additionally, some communication optimization methods offer the dual benefits of reducing communication overhead and enhancing privacy protection. For example, the FedSeqRec framework [37] employs local differential privacy techniques to encrypt and compress model parameters,

thereby optimizing both user data privacy and communication efficiency. After local model training and parameter updates, clients encrypt the parameters using LDP techniques before sending them to the central server. This ensures the security and compressibility of the transmitted data, alleviating the communication burden. Similarly, the SeqMF model [38] incorporates the QHarmony mechanism, which significantly reduces communication overhead by transmitting sparse gradients with quantization perturbations. This mechanism selectively transmits partially quantized gradients and associated metadata, minimizing the amount of data transmitted while preserving training effectiveness and privacy.

3.2.2. Aggregation Balance

Aggregation balance addresses the challenges of uneven data distribution and heterogeneous device characteristics in federated learning by dynamically adjusting each device's contribution to the global model. This ensures that devices with more critical or representative data play a more significant role in the aggregation process. Specific strategies include assigning weights to local models based on data importance and clustering devices with similar data distributions. These methods effectively mitigate data sparsity issues and enhance the model's ability to provide personalized recommendations.

Some methods primarily optimize recommendation performance through client selection. For example, FedGST [39] optimizes recommendation performance through a device-level client selection mechanism. This method uses influence functions to assess each client's contribution to model training dynamically. In each training round, the server selects high-contribution clients based on their contribution values to participate in the next round of training, thereby improving model training efficiency and recommendation performance. DistVAE [40] optimizes recommendation performance through a contribution-based client selection method and clustering strategy. This framework evaluates each client's computational capability and training contribution, selects high-contribution clients to participate in training, and clusters clients based on device resources and behavioral characteristics to reduce communication overhead and enhance recommendation performance, making it suitable for large-scale distributed recommendation systems. CF-FedSR [41] combines adaptive client selection, client clustering-based sampling, fairness-aware model aggregation, and personalized recommendation modules to optimize recommendation performance. This framework optimizes participating clients through an adaptive selector, groups and proportionally samples similar clients based on clustering methods, employs weighted strategies for model aggregation, and integrates global models with locally fine-tuned models to achieve personalized recommendations, thereby significantly enhancing communication efficiency, fairness, and recommendation performance while also reducing communication consumption. For instance, on the Beauty dataset, CF-FedSR improves HR@10 by 8.90% and NDCG@10 by 15.46% compared to FedAvg while reducing communication rounds by approximately 10.67% compared to FedAvg.

Other methods primarily focus on enhancing recommendation performance through client clustering. For instance, CPF-POI [42] employs an adaptive clustering mechanism to optimize recommendations. This framework utilizes a gate control strategy based on Gumbel-Softmax sampling to dynamically adjust client clustering schemes, grouping similar clients together to facilitate knowledge sharing and prevent negative transfer. This balance enhances both knowledge sharing and personalized recommendations. In CPF-GCN [43], a cluster-driven approach is used to optimize recommendation performance. The server clusters users based on their embeddings and selects representative clients from each cluster proportionally for model updates. This strategy not only reduces communication costs but also improves the model's generalization and personalized recommendation capabilities.

Additionally, FCLUB [44] proposes a method that combines stage-wise clustering detection algorithms with asynchronous communication protocols to boost recommendation performance. Clients generate local clustering information based on users' historical interaction data, dynamically adjust local connection graphs to form multiple clusters, and upload this information to the global server for global clustering, maximizing collaborative effects. Furthermore, SFL [45] optimizes recommendation performance by incorporating semantic information and client clustering strategies. This framework

clusters clients based on user behavior's semantic features, ensuring similar users are grouped to promote knowledge sharing and reduce interference from unrelated data. It also reduces communication overhead by transmitting perturbed semantic information. On the NYC dataset, using the SASRec model, SFL improves NDCG@5 from 0.2789 to 0.4492 compared to FedAvg, a 60.8% increase. In single-round computations, server-side computation time decreases from 22.86 seconds with FedAvg to 9.28 seconds with SFL.

Moreover, spatial and temporal information can be leveraged for clustering. For example, SCFL [46] adopts a space-time consistency-based clustering strategy to optimize recommendation performance. This method enhances user collaboration through hierarchical aggregation strategies and edge device clustering, utilizes trajectory optimization modules to extract deep behavioral patterns, and facilitates information sharing and aggregation through edge devices. This approach improves personalized recommendation capabilities while reducing computational overhead. On the NYC dataset, using the SASRec model, SCFL achieves an NDCG@10 of 35.42, approximately 40% higher than FedProx's 25.26. It also significantly reduces server-side computation time, decreasing from 16.35 minutes with FedProx to 2.07 minutes with SCFL.

4. Future Directions

Although there is a lot of existing research, there are still challenging new research directions in the deployment of FedSR to be discussed as follows.

- *Personalized Strategy*: Because different users have varying preferences, it is inadequate to provide the same recommendation service to all users in FedSR. Federated sequential recommendation can investigate personalized strategies to optimize recommendation performance by analyzing user preferences and their individual understanding of items.

- *LLM Foundation*: Recently, LLM has gained significant attention due to its outstanding capabilities, leading researchers to explore its incorporation into the sequential recommendation. Nevertheless, it presents a considerable challenge in FedSR. In these environments, distributed devices must communicate frequently to exchange knowledge from ongoing tasks, while the server needs to efficiently extract new insights from the LLM. Problems concerning communication efficiency and training resources can hinder model convergence. Future research should investigate novel methods for LLM-based FedSR to address these challenges.

- *Multi-Modal Fusion*: In the research of recommendation systems, enhancing knowledge representation is commonly employed to improve model performance. In existing research on FedSR, the issue of multi-modality has been relatively under-considered. However, in addition to user-item interaction records, user reviews, user personal information, and item information can all serve as auxiliary information to enhance model performance. Future research could focus more on how to integrate information from different modalities while protecting the privacy therein to improve model performance.

- *Cold-Start Adaptation*: In the context of shopping recommendations, some items may have fewer interaction records and present a cold-start problem for FedSR systems. Traditional methods may struggle to provide accurate recommendations when there is limited or no historical interaction data. Future research should explore adaptive strategies to handle cold-start scenarios, such as leveraging transfer learning techniques to utilize knowledge from similar cases or incorporating expert knowledge to initialize recommendations.

- *Dynamic Preference*: User preferences are typically dynamic and ever-changing. Federated sequential recommendation needs to capture such dynamic changes to provide more accurate recommendations. Federated continual learning has attracted growing interest by enabling distributed devices to collaboratively learn novel concepts from streaming training data while avoiding knowledge forgetting of previous tasks. Future research could consider combining CL and FedSR to provide dynamic recommendation services.

5. Conclusion

We provide a comprehensive survey of the federated sequential recommendations. First, we begin with an introduction to the motivation for integrating the FL paradigm into the sequential recommendation and related works. Then, we categorize existing methods into two categories with four primary techniques. Finally, we point out the future directions of federated sequential recommendation. We expect this survey to provide an up-to-date summary of recent work and inspire new insights into the sequential recommendation field.

References

1. Kang, W.C.; McAuley, J. Self-attentive sequential recommendation. In Proceedings of the 2018 IEEE international conference on data mining (ICDM). IEEE, 2018, pp. 197–206.
2. Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; Jiang, P. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1441–1450.
3. Tang, J.; Wang, K. Personalized top-n sequential recommendation via convolutional sequence embedding. In Proceedings of the Proceedings of the eleventh ACM international conference on web search and data mining, 2018, pp. 565–573.
4. Voigt, P.; Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Pract. Guid. 1st Ed., Cham: Springer Int. Publ.* **2017**, *10*, 10–5555.
5. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
6. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial intelligence and statistics. PMLR, 2017, pp. 1273–1282.
7. Yuan, Z.; Wang, Y. FedMLP4SR: Federated MLP-Based Sequential Recommendation System. In Proceedings of the International Artificial Intelligence Conference. Springer, 2023, pp. 363–375.
8. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311.
9. Lin, Z.; Pan, W.; Yang, Q.; Ming, Z. A generic federated recommendation framework via fake marks and secret sharing. *ACM Trans. Inf. Syst.* **2022**, *41*, 1–37.
10. Lin, Z.; Pan, W.; Ming, Z. FR-FMSS: Federated recommendation via fake marks and secret sharing. In Proceedings of the Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 668–673.
11. Lee, T.H.; Kim, S.; Lee, J.; Jun, C.H. Word2Vec-based efficient privacy-preserving shared representation learning for federated recommendation system in a cross-device setting. *Inf. Sci.* **2023**, *651*, 119728.
12. Yu, S.; Jie, Z.; Wu, G.; Zhang, H.; Shen, S. FedNRM: A federal personalized news recommendation model achieving user privacy protection. *Intell. Autom. Soft Comput.* **2023**, *37*, 1729–1751.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
14. Sun, Z.; Xu, Y.; Liu, Y.; He, W.; Kong, L.; Wu, F.; Jiang, Y.; Cui, L. A survey on federated recommendation systems. *arXiv preprint arXiv:2301.00767* **2022**.
15. Asad, M.; Shaukat, S.; Javanmardi, E.; Nakazato, J.; Tsukada, M. A comprehensive survey on privacy-preserving techniques in federated recommendation systems. *Appl. Sci.* **2023**, *13*, 6201.
16. Li, Q.; He, B.; Song, D. Model-contrastive federated learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 10713–10722.
17. Rendle, S.; Freudenthaler, C.; Schmidt-Thieme, L. Factorizing personalized markov chains for next-basket recommendation. In Proceedings of the Proceedings of the 19th international conference on World wide web, 2010, pp. 811–820.

18. Tang, J.; Wang, K. Personalized top-n sequential recommendation via convolutional sequence embedding. In Proceedings of the Proceedings of the eleventh ACM international conference on web search and data mining, 2018, pp. 565–573.

19. Li, L.; Zhang, Z.; Huang, C.; Zhang, J. Semi-global sequential recommendation via EM-like federated training. *Expert Syst. Appl.* **2024**, *248*, 123460.

20. Rezaei, E.; Seyedmohammadi, S.J.; Abouei, J.; Plataniotis, K.N. FedDist-POIRec: Federated Distillation for Point-Of-Interest Recommendation in Human Mobility Prediction. In Proceedings of the 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS). IEEE, 2024, pp. 1–6.

21. Zhang, H.; Zheng, D.; Yang, X.; Feng, J.; Liao, Q. FedDCSR: Federated cross-domain sequential recommendation via disentangled representation learning. In Proceedings of the Proceedings of the 2024 SIAM International Conference on Data Mining (SDM). SIAM, 2024, pp. 535–543.

22. Ding, Y.; Ji, Y.; Cai, X.; Xin, X.; Lu, Y.; Huang, S.; Liu, C.; Gao, X.; Murata, T.; Lu, H. Towards Personalized Federated Multi-Scenario Multi-Task Recommendation. *arXiv preprint arXiv:2406.18938* **2024**.

23. Ai, Z.; Wu, G.; Wan, X.; Qi, Z.; Wang, Y. Towards better personalization: A meta-learning approach for federated recommender systems. In Proceedings of the International Conference on Knowledge Science, Engineering and Management. Springer, 2022, pp. 520–533.

24. Perifanis, V.; Drosatos, G.; Stamatelatos, G.; Efraimidis, P.S. FedPOIRec: Privacy-preserving federated poi recommendation with social influence. *Inf. Sci.* **2023**, *623*, 767–790.

25. Guo, Y.; Liu, F.; Cai, Z.; Zeng, H.; Chen, L.; Zhou, T.; Xiao, N. PREFER: Point-of-interest REcommendation with efficiency and privacy-preservation via Federated Edge leaRning. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.* **2021**, *5*, 1–25.

26. Zhao, J.; Wang, W.; Xu, C.; Ren, Z.; Ng, S.K.; Chua, T.S. Llm-based federated recommendation. *arXiv preprint arXiv:2402.09959* **2024**.

27. Reddy, M.S.; Karnati, H.; Sundari, L.M. Transformer based Federated Learning models for Recommendation Systems. *IEEE Access* **2024**.

28. Wan, J.; Chi, C.; Yu, H.; Liu, Y.; Xu, X.; Lyu, H.; Wang, W. Fed-AttGRU Privacy-preserving Federated Interest Recommendation. In Proceedings of the Proceedings of the ACM Turing Award Celebration Conference-China 2024, 2024, pp. 138–143.

29. Zeng, H.; Yue, Z.; Jiang, Q.; Wang, D. Federated recommendation via hybrid retrieval augmented generation. *arXiv preprint arXiv:2403.04256* **2024**.

30. Yuan, W.; Yang, C.; Ye, G.; Chen, T.; Hung, N.Q.V.; Yin, H. FELLAS: Enhancing Federated Sequential Recommendation with LLM as External Services. *ACM Transactions on Information Systems* **2024**.

31. Huang, W.; Liu, J.; Li, T.; Huang, T.; Ji, S.; Wan, J. FedDSR: Daily schedule recommendation in a federated deep reinforcement learning framework. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3912–3924.

32. Wei, S.; Meng, S.; Li, Q.; Zhou, X.; Qi, L.; Xu, X. Edge-enabled federated sequential recommendation with knowledge-aware Transformer. *Future Gener. Comput. Syst.* **2023**, *148*, 610–622.

33. Li, L.; Lin, F.; Xiahou, J.; Lin, Y.; Wu, P.; Liu, Y. Federated low-rank tensor projections for sequential recommendation. *Knowl.-Based Syst.* **2022**, *255*, 109483.

34. Guo, K.; Xie, K.; Shi, Z.; Gao, R. Deep Leakage from Horizontal Federated Sequential Recommender Systems. *IEEE Access* **2024**.

35. Imran, M.; Yin, H.; Chen, T.; Nguyen, Q.V.H.; Zhou, A.; Zheng, K. ReFRS: Resource-efficient federated recommender system for dynamic and diversified user preferences. *ACM Trans. Inf. Syst.* **2023**, *41*, 1–30.

36. Yuan, W.; Yang, C.; Qu, L.; Hung, N.Q.V.; Ye, G.; Yin, H. PTF-FSR: A Parameter Transmission-Free Federated Sequential Recommender System. *ACM Transactions on Information Systems* **2024**.

37. Shi, Y.; Wang, B. A Privacy-Preserving Method for Sequential Recommendation in Vertical Federated Learning. In Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2024, pp. 2221–2226.

38. Saiapin, A.; Balitskiy, G.; Bershatsky, D.; Katrutsa, A.; Frolov, E.; Frolov, A.; Oseledets, I.; Kharin, V. Federated privacy-preserving collaborative filtering for on-device next app prediction. *User Modeling and User-Adapted Interaction* **2024**, pp. 1–30.

39. Tang, T.; Hou, M.; Yu, S.; Cai, Z.; Han, Z.; Oatley, G.; Saikrishna, V. Fedgst: An efficient federated graph neural network for spatio-temporal poi recommendation. *ACM Transactions on Sensor Networks* **2024**.

40. Li, L.; Xiahou, J.; Lin, F.; Su, S. Distvae: Distributed variational autoencoder for sequential recommendation. *Knowl.-Based Syst.* **2023**, *264*, 110313.

41. Luo, S.; Xiao, Y.; Liu, Y.; Li, C.; Song, L. Towards communication efficient and fair federated personalized sequential recommendation. In Proceedings of the 2022 5th International Conference on Information Communication and Signal Processing (ICICSP). IEEE, 2022, pp. 1–6.

42. Ye, Z.; Zhang, X.; Chen, X.; Xiong, H.; Yu, D. Adaptive clustering based personalized federated learning framework for next poi recommendation with location noise. *IEEE Transactions on Knowledge and Data Engineering* **2023**.

43. Mao, X.; Liu, Y.; Qi, L.; Duan, L.; Xu, X.; Zhang, X.; Dou, W.; Beheshti, A.; Zhou, X. Cluster-driven Personalized Federated Recommendation with Interest-aware Graph Convolution Network for Multimedia. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 5614–5622.

44. Liu, X.; Zhao, H.; Yu, T.; Li, S.; Lui, J.C. Federated online clustering of bandits. In Proceedings of the Uncertainty in Artificial Intelligence. PMLR, 2022, pp. 1221–1231.

45. Dong, X.; Zeng, J.; Wen, J.; Gao, M.; Zhou, W. SFL: A Semantic-based Federated Learning Method for POI Recommendation. *Information Sciences* **2024**, p. 121057.

46. Zhong, L.; Zeng, J.; Wang, Z.; Zhou, W.; Wen, J. SCFL: Spatio-temporal consistency federated learning for next POI recommendation. *Inf. Process. Manag.* **2024**, *61*, 103852.