Federated Adaptation for Foundation Model-based Recommendations

Chunxu Zhang[1,2]* , **Guodong Long**[3] , **Hongkuan Guo**[4] , **Xiao Fang**[4] , **Yang Song**[4] , **Zhaojie Liu**[4] , **Guorui Zhou**[4] , **Zijian Zhang**[1,2] , **Yang Liu**[5]† , **Bo Yang**[1,2]†

[1]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[2]College of Computer Science and Technology, Jilin University, China
[3]Australian Artificial Intelligence Institute, FEIT, University of Technology Sydney
[4]Kuaishou Technology
[5]Institute for AI Industry Research, Tsinghua University
{cxzhang19, zhangzj2114}@mails.jlu.edu.cn, guodong.long@uts.edu.au, {guohongkuan, zhouguorui}@kuaishou.com, {ustcfx0727, liuzj03}@gmail.com, ys@sonyis.me, liuy03@air.tsinghua.edu.cn, ybo@jlu.edu.cn

Abstract

With the recent success of large language models, particularly foundation models with generalization abilities, applying foundation models for recommendations becomes a new paradigm to improve existing recommendation systems. It becomes a new open challenge to enable the foundation model to capture user preference changes in a timely manner with reasonable communication and computation costs while preserving privacy. This paper proposes a novel federated adaptation mechanism to enhance the foundation model-based recommendation system in a privacy-preserving manner. Specifically, each client will learn a lightweight personalized adapter using its private data. The adapter then collaborates with pre-trained foundation models to provide recommendation service efficiently with fine-grained manners. Importantly, users' private behavioral data remains secure as it is not shared with the server. This data localization-based privacy preservation is embodied via the federated learning framework. The model can ensure that shared knowledge is incorporated into all adapters while simultaneously preserving each user's personal preferences. Experimental results on four benchmark datasets demonstrate our method's superior performance. The code is available.

1 Introduction

Recently, the Foundation Models (FMs) [Radford *et al.*, 2019; Bommasani *et al.*, 2021; Achiam *et al.*, 2023] emerge rapidly and have made breakthroughs in various AI applications, ranging from language [Alayrac *et al.*, 2022], vision [Saharia *et al.*, 2022], reasoning [Kojima *et al.*, 2022] and recommendation [Geng *et al.*, 2022]. FMs are typically trained on extensive data sources, allowing them to capture and utilize inherent common knowledge. This capability empowers FMs to achieve outstanding performance in various downstream tasks. Applying foundation models to recommendation systems is considered a highly promising direction, which has significantly propelled the state-of-the-art in recommendation system studies [Harte *et al.*, 2023; Liu *et al.*, 2023a; Lin *et al.*, 2023].

Two new open challenges have been encountered when we introduce the foundation models into a practical recommendation system. First, given the fast changes in user preference, how to timely update the foundation models-based recommendation system with reasonable cost on communication and computation. An on-device parameter-efficient fine-tuning mechanism is desired to be incorporated with the foundation models. The second challenge is how to tackle the privacy-sensitive data of users that is needed to train or fine-tune the foundation model in a recommendation system.

Federated learning trains the global model by iterative model parameters transmission between server and clients without accessing private client data [McMahan *et al.*, 2017; Miao *et al.*, 2023; Zhong *et al.*, 2023; Liu *et al.*, 2024]. Due to its excellent privacy-preserving properties, federated learning has become a popular privacy protection enhanced scheme for recommendation, named federated recommendation systems [Chai *et al.*, 2020; Yang *et al.*, 2020; Wu *et al.*, 2022; Sun *et al.*, 2022; Zhang *et al.*, 2023a; Li *et al.*, 2023]. Trained on extensive data, the foundation model's attribute embedding and prediction function retain valuable general domain knowledge and user decision logic. Integrating federated learning into foundation model-based recommendation systems can benefit from this shared knowledge while ensuring privacy preservation. However, effectively addressing this integration remains unsettled.

Several challenges must be addressed urgently to adapt foundation model-based recommendations with federated learning frameworks. ***Challenge 1: efficient user personalization modeling.*** Given the substantial preference differences across isolated clients, the uppermost goal for this recommendation system is effective personalization modeling of diverse users under the privacy protection limitation.

At the same time, there are usually concerns about efficiency when using large-scale foundation models, which is more urgent in federated learning that requires end-device optimization. ***Challenge 2: common knowledge and user personalization fusion.*** Common knowledge maintained in the pre-trained foundation model can incorporate insights from collective user behavior. By integrating it with user personalization, the federated recommendation system can leverage the collective intelligence of the user community, leading to improved decision-making. However, the importance of common knowledge and user personalization varies across users, and ineffective knowledge fusion can lead to information confusion and misleading recommendations. Hence, balancing general knowledge and individual user preference emerges as a main challenge.

In this paper, we present a novel method, named **Fed**erated recommendation with **P**ersonalized **A**dapter (**FedPA**), to explore the federated foundation model for recommendation. Our method leverages the pre-trained model as the foundation, allowing us to incorporate common knowledge and optimize the federated recommendation system from a well-established starting point. To capture user personalization efficiently, we propose a personalized adapter to deploy on the client, which can learn individual user preferences in a lightweight manner. Then, we learn these adapters with the pre-trained model in an adaptive fusion manner to balance the collaboration of common knowledge and user personalization in federated optimization.

Motivated by research about neural network optimization [Li *et al.*, 2018; Aghajanyan *et al.*, 2020], the model parameter solution for the target task usually resides within an intrinsic dimension. In the federated recommendation system, each user has task-specific intrinsic parameter space that can be learned from personal data. To this end, we design a low-rank adapter to learn user personalization in a lightweight manner. Particularly, we develop two levels of personalization to accommodate the recommendation scenario, including user-level and user-group-level. They cater to the unique preferences of individual users while also capturing and leveraging shared patterns and preferences within specific user groups. Furthermore, we design an adaptive gate learning mechanism that dynamically learns the weights for common knowledge and user personalization, enabling effective knowledge fusion. During federated optimization, our FedPA focuses on updating only the parameters relevant to user-specific modeling and the others are frozen and exempted from optimization, leading to a significant reduction in communication cost and achieving faster convergence.

We assess the performance of FedPA on four benchmark datasets and compare it with various advanced baselines. Experimental results consistently demonstrate that our method outperforms the baselines by a significant margin. In addition, we conduct comprehensive experiments to analyze FedPA's ability to capture user personalization and the impact of common knowledge in the pre-trained model on federated recommendation systems. Furthermore, we validate the model's feasibility in real-world applications. By distilling the pre-trained model into a smaller size, we address the computational and storage challenges of deploying the pre-trained model on edge devices. Additionally, we enhance privacy protection by leveraging the Local Differential Privacy technique. Experimental results demonstrate FedPA's stable performance with distilled smaller models and privacy preservation, affirming its practical applicability. To summarize, the **main contributions** are listed as follows,

- For the first time, we investigate the federated adaption paradigm for foundation model-based recommendation, named FedPA. It enables the integration of the rich knowledge encapsulated within pre-trained models while upholding privacy protection for users.

- We present a personalized low-rank adapter to learn user personalization from user-level and user-group-level in a lightweight manner. Furthermore, we design an adaptive gate learning mechanism to dynamically learn weights, allowing for the effective fusion of common knowledge with user personalization.

- Extensive experiments on four benchmark datasets demonstrate the superior performance of FedPA against advanced baselines. Additionally, FedPA also shows excellent feasibility in deploying on clients with limited computation capability and strengthening user privacy protection in federated recommendation systems.

2 Related Work

2.1 Foundation Models for Recommendation

Pre-training in Natural Language Processing (NLP) [Qiu *et al.*, 2020] has witnessed significant progress, with language models like GPT [Brown *et al.*, 2020] and BERT [Devlin *et al.*, 2018] achieving state-of-the-art results. The pre-training and fine-tuning paradigm allows for the extraction of valuable knowledge and eliminates the need to train new models from scratch. Given its remarkable benefits, increasing research has been on developing foundation models for recommendation systems [Liu *et al.*, 2023a; Wu *et al.*, 2023]. The essential learning objective of recommendation is to estimate the user's preferences for a certain item set. By incorporating the foundation models into the recommendation system, it can absorb valuable knowledge and improve the system's ability for characteristics extraction and user decision pattern learning. However, existing foundation models for recommendation rely on collecting personal data from users to optimize, which poses severe risks to protecting user privacy.

2.2 Federated Recommendation System

Federated recommendation system is a rising service paradigm that can learn the model in a privacy-preserving manner [Lin *et al.*, 2020; Perifanis and Efraimidis, 2022; Qu *et al.*, 2023; Zhang *et al.*, 2023b]. Existing studies focus on developing federated recommendation models with mainstream recommendation architectures, *e.g.,* matrix factorization [Chai *et al.*, 2020] and neural collaborative filtering [Perifanis and Efraimidis, 2022], popular recommendation tasks, *e.g.,* POI prediction [Zhang *et al.*, 2023c] and multi-domain recommendation [Liu *et al.*, 2023c], and user privacy protection enhancement [Liu *et al.*, 2023b; Huang *et al.*, 2023]. In this paper, we investigate the integration of foundation model
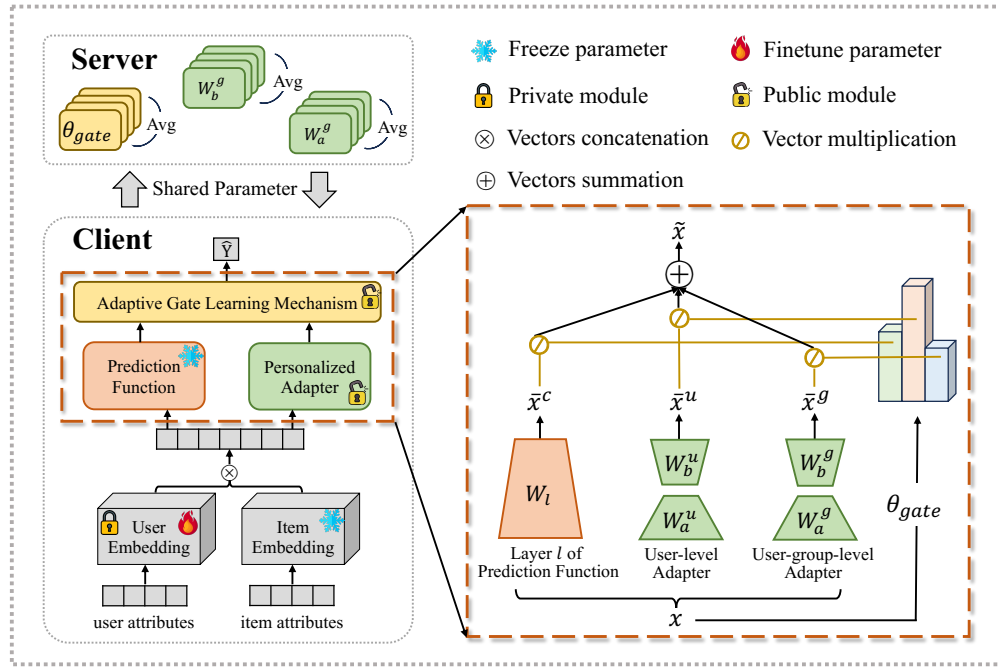
Figure 1: The framework of FedPA. The left part represents the workflow of our method. Each client learns the local recommendation model based on personal data, initializing it with parameters from the pre-trained model. During training, we only update the parameters about user personalization modeling and keep others frozen. The server is responsible for globally aggregating the shared parameters to transmit common information among clients. The right part illustrates the details of employing the adaptive gate learning mechanism to fuse the common information from each layer of the prediction function and user personalization from the personalized adapter at two granularities. It is worth noting that the user-level adapter is a private module and the user-group-level adapter is a public module.

into federated recommendation system, which can utilize the inherent common knowledge in the pre-trained model and prosper the powerful federated recommendation system.

3 Methodology

We present a novel federated adaption paradigm for foundation model-based recommendation, named **Fed**erated recommendation with **P**ersonalized **A**daptor (**FedPA**). In this section, we first provide an overview of the framework architecture. Then, we delve into the components to elucidate the details and summarize the workflow into an optimization algorithm. Following that, we engage in a discussion on the feasibility of deploying our FedPA in physical applications. Finally, we develop a privacy-protection enhanced FedPA that can strengthen the system's protection for user privacy.

3.1 Framework Overview

The pre-trained recommendation model is learned from a large amount of publicly available data, embodying rich knowledge. It can effectively characterize user and item attributes and possesses strong predictive capabilities for user decision patterns. Effective utilization of this common knowledge contributes to building a more powerful federated recommendation model. To achieve this, we take the pre-trained model as the foundation model for the federated recommendation system, enabling federated optimization from a favorable starting point.

To efficiently capture user-specific preferences, we propose a low-rank adapter that models user personalization from both the user-level and user-group-level perspectives. Additionally, we design an adaptive gate learning mechanism that effectively integrates common knowledge and personalized knowledge for better user modeling. The model architecture is illustrated in Figure 1. In this paper, we utilize an existing recommendation architecture as the base model, and it can be easily extended to other popular architectures.

3.2 Base Model

Given user attributes, item attributes, and user interaction records, we adopt a widely used two-tower recommendation model architecture. Specifically, the model consists of two input branches, one for learning user representations based on user attributes and the other for learning item representations based on item attributes. These representations are then fed into a prediction function to estimate user preferences for items. The user interaction records serve as supervision information to guide the model in updating its parameters.

User Embedding Module. There are certain attribute information available for the users, *e.g.,* user active degree. For each attribute i, we construct a learnable embedding table $E_i \in \mathbb{R}^{p \times d}$, where p is the total attribute categories and d is the embedding dimension. Then, for each user u with attributes A_u, we retrieve the embedding vectors from all the attribute embedding tables based on attribute values and ob-

tain the user representation u_r by concatenating them,

$$u_r = \text{Concat}(E_i(A_u^i)_{i=1}^{|A_u|}) \tag{1}$$

where $|A_u|$ is the total attribute number of user u.

Item Embedding Module. For each item v, we adopt a similar approach as the user embedding module to construct the item embedding tables and obtain item representation v_r,

$$v_r = \text{Concat}(E_i(A_v^i)_{i=1}^{|A_v|}) \tag{2}$$

and $|A_v|$ is the total attribute number of item v.

Prediction Function Module. Given user representation u_r and item representation v_r, we use a simple MLP (Multi-Layer Perceptron) as the prediction function to estimate user preferences for items,

$$\hat{Y}_{uv} = \text{MLP}(\text{Concat}(u_r, v_r)) \tag{3}$$

Loss Function. To update the model parameters, we construct a loss function that encourages the model's predictions to be as close as possible to the true labels. For common implicit feedback recommendation tasks, where the label value is $Y_{uv} = 1$ when a user u interacts with an item v and $Y_{uv} = 0$ otherwise, we use binary cross-entropy as the loss function,

$$L(\theta_{base}) = -\frac{1}{|D|} \sum_{Y_{u,v} \in D} Y_{uv} \log \hat{Y}_{u,v} + (1 - Y_{uv}) \log(1 - \hat{Y}_{u,v}) \tag{4}$$

where θ_{base} is the model parameter, including user embedding tables θ_{ue}, item embedding tables θ_{ie} and the MLP parameters θ_{mlp}. D is the user-item interaction record set, and $|D|$ is the total interactions number.

3.3 Personalized Adapter

Existing federated recommendation methods typically model user personalization by preserving partial model parameters locally. However, it limits direct access from other clients and potentially hinders the collaborative context utilization. Moreover, the personal data on each client is generally limited, which can introduce biases and compromise model performance. To overcome the challenge, we propose to learn user personalization from two perspectives, *i.e.,* user-level and user-group-level. Particularly, we devise a personalized adapter applied to the prediction function module due to its crucial role in predicting user preferences.

Drawing inspiration from research on neural network optimization, model parameters can be embedded within an intrinsic dimension [Li *et al.*, 2018; Aghajanyan *et al.*, 2020]. To this end, we propose a low-rank adapter that leverages low-rank matrices to model user-specific knowledge on each client. This approach offers two prominent advantages: First, it can learn user personalization from user-specific and user groups with similar characteristics, which enhances the system's ability to model individual user preferences. Besides, the low-rank matrices introduce only a small number of parameters, making it a parameter-efficient solution.

Low-Rank adapter. For each layer l of MLP in the prediction function module, it maps the input $x \in \mathbb{R}^d$ into a new space $\overline{x} \in \mathbb{R}^k$ with the weight matrix $W_l \in \mathbb{R}^{k \times d}$. We intensify the personalization learning by adding a low-rank decomposition matrix $W_{lr} = W_a W_b$, where $W_a \in \mathbb{R}^{k \times r}$ and

$W_b \in \mathbb{R}^{r \times d}$, and the rank $r \ll min(d, k)$. Then, the forward pass of each layer can be modified as follows,

$$\overline{x} = W_l x + W_a W_b x \tag{5}$$

where W_{lr} is responsible for learning user personalization. In the context of recommendation tasks, we develop two levels of personalization: user-level personalization and user-group-level personalization. They cater to individual user preferences as well as capturing patterns and preferences shared within specific user groups.

User-Level Personalization. For the user-level personalization, we aim to learn a specific low-rank adapter for each user so the parameters W_a^u and W_b^u would be preserved locally and not be shared globally. For the user u, we formulate the user-specific low-rank adapter as follows,

$$\overline{x}^u = W_a^u W_b^u x \tag{6}$$

User-Group-Level Personalization. For the user-group-level personalization, we aim to learn the same low-rank adapter for users in a specific group. In recommendation systems, users with similar characteristics tend to share similar preferences. To fully leverage this information, we learn multiple groups of low-rank adapters. For each user group g, we formulate the group-specific low-rank adapter as follows,

$$\overline{x}^g = W_a^g W_b^g x \tag{7}$$

Users within the same group share parameter W_a^g and W_b^g. It is worth noting that users in the system can be grouped in multiple ways, meaning each user u can belong to multiple groups $\{g_i\}_{i=1}^{total}$. For example, a user can belong to the "young adults" group while also to the "highly active degree" group. By categorizing users from multiple orthometric perspectives, our model can learn more detailed personalized parameters and enhance the user preference capture.

3.4 Adaptive Gate Learning Mechanism

By incorporating the low-rank adapter, the prediction function module can learn both the shared decision patterns among users and the personalized decision logic at two granularities, *i.e.,* user-level and user-group-level. To effectively combine common knowledge and personalized knowledge, we propose an adaptive gate learning mechanism that dynamically assigns weights to each decision branch. Specifically, we utilize a two-layer non-linear mapping to learn the weights for the branches based on the input. The fusion process can be formulated as follows,

$$\widetilde{x} = \text{Sum}(softmax(W_2 Relu(W_1 x)) \odot \\ [\overline{x}^c, \overline{x}^u, \{\overline{x}^{g_i}\}_{i=1}^{total}]) \tag{8}$$

where W_1 and W_2 are the parameters of the adaptive gate learning mechanism, $\overline{x}^c$ is the output of the MLP layer in the prediction function module. $\odot$ and $\text{Sum}(\cdot)$ denote element-wise multiplication and summation calculation, respectively. The parameters of the adaptive gate learning mechanism can be updated based on gradients along with other model parameters, eliminating the cumbersome manual setting of hyperparameters. This approach enhances the flexibility of multi-branch fusion, allowing for adaptive adjustment of the weights for all branches during the model's training stages.

3.5 Optimization and Model Scalability

Optimization Objective

In the federated recommendation system, each user u acts as a client and trains the local recommendation model according to the personal dataset D_u. Let $L_u(\theta_{all}^u)$ denote the local model's loss function, where θ_{all}^u contains of base model parameter θ_{base}, low-rank adapter parameter θ_{lr}^u and $\{\theta_{lr}^{g_i}\}_{i=1}^{total}$ and adaptive gate learning mechanism parameter θ_{gate}. The overall optimization of the federated recommendation model can be formulated as follows,

$$\min_{\{\theta_{all}^1, \dots, \theta_{all}^n\}} \sum_{u=1}^{n} \frac{1}{n} \mathcal{L}_u(\theta_{all}^u) \qquad (9)$$

where n is the total number of clients in the federated recommendation system. Here, we employ a naive average aggregation approach to optimize system parameters. It is also possible to use a more flexible weighted aggregation approach further to enhance the optimization objective [McMahan *et al.*, 2017; Wang *et al.*, 2020].

Efficient Parameter Update

To leverage the knowledge inherent in the pre-trained model effectively, we propose to freeze the item embedding module and prediction function module during federated optimization. However, since the federated recommendation system operates on different user sets, we continue to update the user embedding module to adapt to the specific characteristics. To summarize, during the federated optimization process, we focus solely on updating the parameters related to personalized user modeling, *e.g.*, user embedding, low-rank adapter, and adaptive gate learning mechanism parameters. It effectively saves computational and communication costs.

Discussion about Model Scalability

The pre-trained recommendation models learned with abundant computational resources often have complex structures and large sizes in real-world scenarios. Deploying such models directly to the clients poses significant challenges in terms of limited storage and computational capability. To address this issue, we propose to leverage the knowledge distillation (KD) technique [Gou *et al.*, 2021] to distill the pre-trained model into a smaller-sized recommendation model that can be accommodated by client devices. We then utilize this distilled model to warm up the federated recommendation system. This approach effectively enhances the scalability of the proposed framework in real-world applications.

3.6 Privacy-Preserving Enhanced FedPA

The distributed training nature of federated learning avoids direct exposure to private user data. To further mitigate the risk of the server inferring user privacy through model parameter reverse engineering, we integrate the Local Differential Privacy technique [Choi *et al.*, 2018] into our method, whose basic insight is to add a zero-mean Laplacian noise to the shared model parameters before uploaded to the server. Our method's shared model parameters include user group-level low-rank adapter and the adaptive gate learning mechanism parameters. By adjusting the intensity of noise, we can control the privacy protection capability of the system, that is, increasing the intensity of the noise enhances the effectiveness of privacy protection.

4 Experiment

In this section, we conduct comprehensive experiments to verify the efficacy and illustrate a deep analysis of various aspects of our proposed FedPA. Implementation code is available to ease reproducibility[1].

4.1 Experimental Setup

Datasets

We evaluate FedPA on four practical industrial recommendation datasets collected from the short video platform Kuaishou [2], *i.e.*, KuaiRand [3] (KuaiRand-Pure and KuaiRand-small) and KuaiSAR [4] (KuaiSAR-S and KuaiSAR-R). For dataset split, we first divide each dataset into two subsets: one for training the foundation model and the other for training the federated recommendation system. The dataset for the federated recommendation system is further split into train, validation, and test sets for each user based on interaction timestamps, with a ratio of 6:2:2.

Evaluation Protocols

We evaluate the model performance by calculating the evaluation metrics on the test set of private users. Specifically, we take widely used AUC (Area Under Curve) and Precision as two evaluation metrics. All experimental results are in units of 1e-2 and the average values of five individual runs. For the user group level low-rank personalization, we group users based on their attribute values. It is important to note that the clients can update the corresponding user group low-rank adaptor parameters based on their own attribute values locally without exposing the attribute values to the server. Additionally, we have incorporated Local Differential Privacy technique to further protect user privacy, hence user privacy protection can be ensured in our method.

Baselines

This paper concentrates on developing a personalized federated recommendation system and investigating the assistance provided by the common knowledge contained in pre-trained models. To assess the feasibility and effectiveness of our proposed FedPA, we compare it with two types of baseline models: (1) Train and evaluate the model on the private user dataset without warm-staring from the pre-trained model, *i.e.*, **w/o Warm**. (2) Warm-start the federated recommendation models with the pre-trained model, *i.e.*, **w/ Warm**. Specifically, we select four representative personalized federated recommendation models, including FedNCF [Perifanis and Efraimidis, 2022], PFedNCF [Pillutla *et al.*, 2022], FedRecon [Singhal *et al.*, 2021] and PFedRec [Zhang *et al.*, 2023a], and the corresponding warm-starting variants as baselines. Besides, we remove the warm-starting strategy

---

[1]Code: https://github.com/Zhangcx19/IJCAI-24-FedPA

[2]https://www.kuaishou.com/cn

[3]https://kuairand.com/

[4]https://kuaisar.github.io/

Method	Dataset	KuaiRand-Pure		KuaiRand-small		KuaiSAR-S		KuaiSAR-R	
	Metric	AUC	Precision	AUC	Precision	AUC	Precision	AUC	Precision
w/o Warm	FedNCF	68.17	73.33	69.35	65.69	55.19	**81.82**	70.83	65.45
	PFedNCF	62.99	72.09	63.73	62.23	55.31	79.44	67.69	63.81
	FedRecon	65.21	61.72	68.81	65.32	**58.56**	76.57	66.65	62.68
	PFedRec	59.48	70.80	61.09	61.33	56.71	79.42	68.32	63.83
	FedPA w/o Warm	**68.44***	**73.75***	**69.65***	**66.69***	57.58	81.04	**71.30***	**65.61***
w/ Warm	Warm_FedNCF	69.90	74.15	70.54	65.83	59.95	79.03	71.47	66.38
	Warm_PFedNCF	62.78	71.75	62.58	61.59	57.86	79.24	66.82	63.43
	Warm_FedRecon	70.02	74.22	70.75	66.56	55.68	78.54	65.49	61.31
	Warm_PFedRec	62.71	71.73	63.96	61.83	59.74	79.30	67.50	63.81
	FedPA	**70.28***	**75.12***	**71.14***	**66.86**	**61.99***	**86.98***	**72.21***	**66.70***

Table 1: Experimental results of baselines and our method on four datasets. "**w/o Warm**" ("**w/ Warm**") denotes training the federated recommendation system without (with) a pre-trained model. The best results are bold. "＊" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

from our method, named **FedPA w/o Warm**, to assess the contribution of our personalization modeling insight for the federated recommendation system.

4.2 Comparison with Baselines

Table 1 shows the performance of AUC and Precision on four datasets. We provide a summary of the experimental results and discuss noteworthy observations as follows,

(1) **Compared to the naive FedNCF, existing user personalization modeling techniques provide modest performance improvements and, in some cases, can even degrade the model performance.** Existing federated recommendation models focus on modeling user preferences based on user (item) ID attributes. The key idea behind these personalization techniques is to learn user-specific model parameters (*e.g.,* prediction function module) via personal data. However, when the system can leverage additional attributes, it provides mode auxiliary information for model optimization and increases learning difficulty. Consequently, selecting specific model parameters exclusively for modeling user personalization loses effectiveness. In contrast, our method introduces low-rank matrices to modeling personalization from both user-level and user-group-level, which ensures common knowledge sharing and enhances model performance.

(2) **Warm-starting the federated recommendation systems with pre-trained models can improve their performances in most cases.** For instance, when incorporating the warm-start strategy, FedNCF achieves 2.54% and 8.62% performance gains on KuaiRand-Pure and KuaiSAR-S, respectively. The pre-trained model has already learned general user preferences. By transferring the common knowledge to the federated recommendation system, it is able to supplement the beneficial information and alleviate the performance bottleneck caused by distributed data storage. In our method, we devise an adaptive gate learning mechanism. It can dynamically learn the weights for an effective blend of common knowledge and personalized knowledge and hence achieves state-of-the-art performance.

(3) **Our FedPA is a communication-efficient federated foundation model for recommendation.** In the federated optimization phase, our method only updates the parameters relevant to user personalization modeling, leading to significant savings in communication overhead. For example, on the KuaiRand dataset, the number of trainable parameters of baseline models is 13,649, while only 8,189 for our method, resulting in a 40% reduction. In federated recommendation systems, the number of clients is typically extensive, and frequent parameter communication between clients and servers presents a substantial challenge to system optimization. Our method effectively reduces communication overhead, making it suitable for deployment in real-world environments.

4.3 Low-Rank Personalization Analysis

Our FedPA enables the learning of low-rank personalization at both the user and user-group levels, providing comprehensive modeling of user preferences from multiple perspectives. To make an in-depth analysis of the efficacy of the two forms of personalization, we conduct two model variants: one focusing on user-level personalization and the other on user-group-level personalization. Specifically, we take FedNCF as the baseline and incorporate two forms of personalization, denoted as **w/ UP** (with user-level) and **w/ GP** (with user-group-level), respectively. Given the user grouping based on their attributes, we conduct the experiments according to multiple user attributes to make a comprehensive analysis. As shown in Table 2, incorporating either user-level or user-group-level low-rank personalization into FedNCF can improve its performance. Therefore, combining these two forms of personalization allows them to complement each other and achieve superior performance.

4.4 Effect of Common Knowledge on Federated Optimization

In the recommendation model, different modules have specific roles. User embedding and item embedding focus on attribute information learning, while the prediction function captures user decision patterns. To further explore the impact of the common knowledge inherent in the pre-trained model on federated optimization, we assess the specific effects of each module in the recommendation model on model per-

Model	Warm_FedNCF	w/ UP	w/ GP				FedPA
			user active degree	register days range	onehot feat10	follow user num range	
AUC	69.90	70.01	70.05	70.17	70.02	70.04	**70.28**
Precision	74.15	74.26	74.18	75.00	75.04	74.98	**75.12**

Table 2: Experimental results for user-level (**w/ UP**) and user-group-level (**w/ GP**) low-rank personalization analysis on KuaiRand-Pure. Specifically, we select multiple user attributes, *e.g.,* user active degree, to guide the user grouping in user-group-level personalization.

Model	FZ_UE	FZ_IE	FZ_PF	FedPA
AUC	65.55	**70.31**	70.28	70.28
Precision	72.67	75.16	**75.27**	75.12

Table 3: Effect of freezing different modules of the pre-trained model on federated optimization on KuaiRand-Pure. "**FZ_UE**", "**FZ_IE**" and "**FZ_PF**" denote freezing user embedding, item embedding and prediction function, respectively.

Model	8-(32, 8, 1)	8-(8, 1)	4-(32, 8, 1)	4-(8, 1)
AUC	70.28	**71.41**	71.12	70.61
Precision	75.12	75.47	**77.49**	76.81

Table 4: Experimental results of warm-starting the federated recommendation system with different-sized models by knowledge distillation on KuaiRand-Pure.

4.6 Privacy-Enhanced FedPA

We conduct experiments to evaluate the privacy-protection enhanced FedPA by integrating the Local Differential Privacy technique. Particularly, we set the Laplacian noise with different intensities, *e.g.,* from 0.1 to 0.5 with an interval of 0.1, to observe the effect, and the results are summarized in Table 5. As the noise intensity increases, the model performance deteriorates, while the decline is slight when the noise is not excessively large. Therefore, a moderate noise intensity, *e.g.,* 0.2 is desirable to strike a balance between model performance and system privacy protection capability.

Intensity	0	0.1	0.2	0.3	0.4	0.5
AUC	**70.28**	70.06	69.93	69.92	69.80	69.89
Precision	**75.12**	74.17	74.35	74.28	74.20	74.16

Table 5: Experimental results of privacy-protection enhanced FedPA with various noise intensity on KuaiRand-Pure.

formance. Specifically, we conduct warm-start experiments using pre-trained models, freezing user embedding, item embedding, and the prediction function separately, and analyze the experimental results.

From Table 3, we can summarize two conclusions: **(1) Fixing user embedding during the optimization process of federated recommendation system leads to a significant decline in model performance.** This is because the pre-trained model and the federated recommendation system are trained on different user sets, and the variations between user sets make it challenging for the pre-trained user embedding to adapt well to the federated recommendation system, and consequently resulting in a degradation of performance. **(2) Updating item embedding or the prediction function in the federated recommendation system leads to further performance improvement.** This finding aligns with our expectations, as updating a larger number of parameters facilitates the model in learning intricate user preferences. Our FedPA, which simultaneously fixes item embedding and the prediction function, greatly alleviates communication overhead, presenting an effective compromise to balance the performance and cost of federated recommendation systems.

4.5 Lightweight FedPA with KD

In the physical setting, the service provider generally learns a large-scale pre-trained model, which poses challenges in terms of computational and storage capabilities when deployed directly to clients. To fill the gap, we develop a lightweight FedPA with knowledge distillation technique. Specifically, we first distill a small-scale model from the pre-trained large-scale model and then deploy it on each client as the base model. For a comprehensive investigation, we distill three different sizes, *i.e.,* 8-(8, 1), 4-(32, 8, 1) and 4-(8, 1), from the original model whose size is 8-(32, 8, 1) by adjusting the embedding dimension or the prediction function architecture. As shown in Table 4, employing the distilled small-scale models for warm-starting the federated recommendation system not only preserves the model's performance but also yields performance improvements. This finding further strengthens our method's viability in real-world applications.

5 Conclusion

In this paper, we develop FedPA, the first federated adaption paradigm for foundation model-based recommendation. It optimizes from a solid starting point, leveraging a pre-trained model as the backbone. Our method involves learning personalized adapters for each client based on local data, focusing on lightweight low-rank adapters at user and user-group levels to learn detailed personalization with complementary perspectives. In addition, we design an adaptive gate learning mechanism to effectively blend common knowledge and user personalization with dynamic weights. Our method significantly reduces computational and communication costs by updating only user-specific parameters during federated optimization. Extensive experiments demonstrate superior performance compared to advanced baselines. We address the challenge of deploying models on resource-constrained devices by distilling a compact model from the original pre-trained model. Additionally, we enhance privacy protection in FedPA by incorporating Local Differential Privacy, achieving a solid balance between recommendation performance and privacy preservation.

Acknowledgments

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Aghajanyan *et al.*, 2020] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Chai *et al.*, 2020] Di Chai, Leye Wang, Kai Chen, and Qiang Yang. Secure federated matrix factorization. *IEEE Intelligent Systems*, 36(5):11–20, 2020.

[Choi *et al.*, 2018] Woo-Seok Choi, Matthew Tomei, Jose Rodrigo Sanchez Vicarte, Pavan Kumar Hanumolu, and Rakesh Kumar. Guaranteeing local differential privacy on ultra-low-power systems. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, pages 561–574. IEEE, 2018.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Geng *et al.*, 2022] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 299–315, 2022.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[Harte *et al.*, 2023] Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1096–1102, 2023.

[Huang *et al.*, 2023] Xinyi Huang, Yuchuan Luo, Lin Liu, Wentao Zhao, and Shaojing Fu. Randomization is all you need: A privacy-preserving federated learning framework for news recommendation. *Information Sciences*, 637:118943, 2023.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[Li *et al.*, 2018] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

[Li *et al.*, 2023] Zhiwei Li, Guodong Long, and Tianyi Zhou. Federated recommendation with additive personalization. *arXiv preprint arXiv:2301.09109*, 2023.

[Lin *et al.*, 2020] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Dongxiao Yu, Jun Ma, Maarten de Rijke, and Xiuzhen Cheng. Meta matrix factorization for federated rating predictions. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 981–990, 2020.

[Lin *et al.*, 2023] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, et al. How can recommender systems benefit from large language models: A survey. *arXiv preprint arXiv:2306.05817*, 2023.

[Liu *et al.*, 2023a] Peng Liu, Lemei Zhang, and Jon Atle Gulla. Pre-train, prompt and recommendation: A comprehensive survey of language modelling paradigm adaptations in recommender systems. *arXiv preprint arXiv:2302.03735*, 2023.

[Liu *et al.*, 2023b] Ruixuan Liu, Yang Cao, Yanlin Wang, Lingjuan Lyu, Yun Chen, and Hong Chen. Privaterec: Differentially private model training and online serving for federated news recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4539–4548, 2023.

[Liu *et al.*, 2023c] Weiming Liu, Chaochao Chen, Xinting Liao, Mengling Hu, Jianwei Yin, Yanchao Tan, and Longfei Zheng. Federated probabilistic preference distribution modelling with compactness co-clustering for privacy-preserving multi-domain recommendation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2206–2214, 2023.

[Liu *et al.*, 2024] Ziqiao Liu, Hao Miao, Yan Zhao, Chenxi Liu, Kai Zheng, and Huan Li. Lighttr: A lightweight framework for federated trajectory recovery. In *ICDE*, 2024.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Miao *et al.*, 2023] Hao Miao, Xiaolong Zhong, Jiaxin Liu, Yan Zhao, Xiangyu Zhao, Weizhu Qian, Kai Zheng, and Christian S Jensen. Task assignment with efficient federated preference learning in spatial crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[Perifanis and Efraimidis, 2022] Vasileios Perifanis and Pavlos S Efraimidis. Federated neural collaborative filtering. *Knowledge-Based Systems*, 242:108441, 2022.

[Pillutla *et al.*, 2022] Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated learning with partial model personalization. In *International Conference on Machine Learning*, pages 17716–17758. PMLR, 2022.

[Qiu *et al.*, 2020] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.

[Qu *et al.*, 2023] Liang Qu, Ningzhi Tang, Ruiqi Zheng, Quoc Viet Hung Nguyen, Zi Huang, Yuhui Shi, and Hongzhi Yin. Semi-decentralized federated ego graph learning for recommendation. In *Proceedings of the ACM Web Conference*, page 339–348, 2023.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[Singhal *et al.*, 2021] Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, and Sushant Prakash. Federated reconstruction: Partially local federated learning. *Advances in Neural Information Processing Systems*, 34:11220–11232, 2021.

[Sun *et al.*, 2022] Zehua Sun, Yonghui Xu, Yong Liu, Wei He, Lanju Kong, Fangzhao Wu, Yali Jiang, and Lizhen Cui. A survey on federated recommendation systems. *arXiv preprint arXiv:2301.00767*, 2022.

[Wang *et al.*, 2020] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[Wu *et al.*, 2022] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Tao Qi, Yongfeng Huang, and Xing Xie. A federated graph neural network framework for privacy-preserving personalization. *Nature Communications*, 13(1):3091, 2022.

[Wu *et al.*, 2023] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*, 2023.

[Yang *et al.*, 2020] Liu Yang, Ben Tan, Vincent W Zheng, Kai Chen, and Qiang Yang. Federated recommendation systems. *Federated Learning: Privacy and Incentive*, pages 225–239, 2020.

[Zhang *et al.*, 2023a] Chunxu Zhang, Guodong Long, Tianyi Zhou, Peng Yan, Zijian Zhang, Chengqi Zhang, and Bo Yang. Dual personalization on federated recommendation. *arXiv preprint arXiv:2301.08143*, 2023.

[Zhang *et al.*, 2023b] Chunxu Zhang, Guodong Long, Tianyi Zhou, Xiangyu Zhao, Zijian Zhang, and Bo Yang. When federated recommendation meets cold-start problem: Separating item attributes and user interactions. *arXiv preprint arXiv:2305.12650*, 2023.

[Zhang *et al.*, 2023c] Xiao Zhang, Ziming Ye, Jianfeng Lu, Fuzhen Zhuang, Yanwei Zheng, and Dongxiao Yu. Fine-grained preference-aware personalized federated poi recommendation with data sparsity. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 413–422, 2023.

[Zhong *et al.*, 2023] Xiaolong Zhong, Hao Miao, Dazhuo Qiu, Yan Zhao, and Kai Zheng. Personalized location-preference learning for federated task assignment in spatial crowdsourcing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3534–3543, 2023.