

# FRAdRec: A Federated Real-Time Advertising Recommendation Framework Based on User Behavior Modeling and Cloud-Native Data Infrastructure

XiangYuan He<sup>1</sup>. KuangCong Liu<sup>2</sup>

<sup>1</sup>Chongqing University, Chongqing, China

<sup>2</sup>Stanford University, Stanford, CA, USA

**How to cite this paper:** He, X. Y., & Liu, K. C. (2026). FRAdRec: A federated real-time advertising recommendation framework based on user behavior modeling and cloud-native data infrastructure. *Journal of Computer Science and Frontier Technologies*, 3(2), 102–117. ISSN Print: 3104-4204, ISSN Online: 3104-4212. <https://doi.org/10.63313/JCSFT.9075>  
**Published: 2026-05-19**

Copyright © 2026 by author(s) and Erytis Publishing Limited.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



## Abstract

Advertising recommendation systems have become a core component of modern e-commerce and digital marketing platforms, where accurately capturing dynamic user preferences and supporting large-scale real-time recommendation services are critical for improving advertising effectiveness and platform revenue. However, traditional centralized recommendation frameworks often suffer from limited scalability, delayed interest modeling, insufficient optimization of business-oriented objectives, and increasing privacy risks caused by centralized user data collection. To address these challenges, this paper proposes FRAdRec, a federated real-time advertising recommendation framework based on user behavior modeling and cloud-native data infrastructure. The proposed framework integrates a Transformer-based sequential learning module with a Deep Interest Network (DIN) attention mechanism to capture both long-term and short-term user interests from click and purchase behavior sequences. Meanwhile, a federated learning strategy is introduced to enable decentralized model training without transmitting raw user data, thereby improving privacy preservation and reducing data leakage risks. Furthermore, a multi-task optimization mechanism jointly predicts click-through rate (CTR), conversion rate (CVR), and return on investment (ROI) to enhance advertising profitability. To support industrial-scale deployment, a cloud-native streaming architecture based on Kafka, Flink, Kubernetes, and Redis is designed for low-latency recommendation and online feature updating. Experimental results on the Criteo and Alibaba Taobao datasets demonstrate that FRAdRec achieves an AUC of 0.857 and reduces recommendation latency to 41 ms, outperforming several state-of-the-art recommendation models in both recommendation accuracy and real-time serving efficiency.

## Keywords

Federated Recommendation; Advertising Recommendation System; User

---

Behavior Modeling; Cloud-Native Infrastructure; Real-Time Recommendation; Transformer; Deep Interest Network; Multi-Task Learning; ROI Optimization

---

## 1. Introduction

With the rapid growth of e-commerce platforms, online advertising ecosystems, and intelligent marketing services, personalized advertising recommendation systems have become one of the core technologies for improving user engagement and maximizing commercial revenue. Modern advertising platforms, such as Amazon Sponsored Ads, Google Ads, and Alibaba Advertising Systems, rely heavily on large-scale user behavior data to provide accurate and real-time advertising recommendations. In practical industrial scenarios, advertising recommendation systems must continuously analyze dynamic user behaviors, including clicks, browsing records, purchase histories, and search sequences, in order to capture evolving user interests and improve recommendation relevance. Consequently, user behavior modeling and scalable real-time recommendation infrastructures have become important research topics in computational advertising and recommender systems.

Traditional recommendation approaches mainly focus on centralized click-through rate (CTR) prediction using offline machine learning models. Although these methods have achieved promising results, several limitations remain unresolved. First, conventional recommendation models often fail to effectively capture long-term sequential dependencies and dynamically changing user interests, leading to recommendation delays and interest drift problems. Second, most existing advertising recommendation systems optimize only CTR while neglecting business-oriented objectives such as conversion rate (CVR) and return on investment (ROI), which are critical for advertising profitability and platform sustainability. Third, centralized data collection and model training introduce serious privacy concerns and increase the risk of user data leakage under modern data protection regulations. In addition, the explosive growth of advertising traffic requires highly scalable cloud-native infrastructures capable of supporting low-latency recommendation services and large-scale streaming data processing.

To address these challenges, this paper proposes FRAdRec, a Federated Real-Time Advertising Recommendation framework based on user behavior modeling and cloud-native data infrastructure. The proposed framework integrates Transformer-based sequential learning and Deep Interest Network (DIN) attention mechanisms to capture both long-term and short-term user interests from large-scale behavioral sequences. Meanwhile, a federated learning strategy is introduced to enable decentralized recommendation model training without transmitting raw user data, thereby improving privacy preservation. Furthermore, a cloud-native streaming architecture based on Kafka, Flink, Kubernetes, and Redis is

designed to support scalable real-time recommendation, online feature updating, and distributed recommendation serving in industrial advertising environments.

The major contributions of this paper are summarized as follows:

A federated real-time advertising recommendation framework named FRAdRec is proposed for scalable and privacy-preserving advertising recommendation tasks.

A hybrid user behavior modeling mechanism combining Transformer and DIN attention is designed to capture dynamic user interests from sequential behavioral data.

A multi-task optimization strategy jointly predicts CTR, CVR, and ROI to improve advertising effectiveness and commercial profitability.

A cloud-native streaming recommendation infrastructure is developed to support low-latency recommendation services and large-scale distributed deployment.

Extensive experiments on public advertising datasets demonstrate that FRAdRec achieves superior performance in recommendation accuracy, ROI optimization, and real-time serving efficiency compared with several state-of-the-art baseline methods.

## 2. Literature Review

In recent years, the rapid development of deep learning-based recommender systems, federated learning paradigms, and cloud-native data infrastructures has significantly advanced the field of large-scale advertising recommendation. This section reviews three major research directions closely related to this work: (1) user behavior modeling for recommendation systems, (2) federated learning for privacy-preserving recommendation, and (3) cloud-native architectures for scalable real-time recommendation systems.

### 2.1. User Behavior Modeling for Advertising Recommendation

Modeling user behavior sequences is a fundamental task in modern recommendation systems. Early works such as matrix factorization and factorization machines (FM) [1] laid the foundation for learning user-item interactions in sparse data environments. With the rise of deep learning, neural recommendation models such as Wide & Deep [2] and DeepFM [3] significantly improved feature interaction learning by combining memorization and generalization capabilities.

To better capture high-order feature interactions and sequential user behaviors, Deep Interest Network (DIN) [4] introduced an attention mechanism to model user interests dynamically based on target advertisements. Deep Interest Evolution Network (DIEN) [5] further extended DIN by incorporating GRU-based sequential modeling to capture interest evolution over time. More recently, Transformer-based recommendation models such as SASRec [6] and BERT4Rec [7] have demonstrated superior performance in capturing long-range dependencies in user behavior

sequences through self-attention mechanisms.

However, most existing methods still focus on centralized training and static optimization objectives, typically emphasizing CTR prediction while neglecting business-oriented metrics such as ROI and system-level constraints such as latency and scalability. To overcome the limitations of isolated single-task models, recent literature in predictive analytics has increasingly progressed towards unified network architectures. For example, Zhang et al. proposed a multimodal large language framework capable of simultaneously performing micro-level asset forecasting and macro-level systemic risk assessment through shared semantic spaces and modular heads [8]. The effectiveness of learning shared representations for cross-scale heterogeneous tasks provides a compelling theoretical foundation for our approach, which utilizes a unified architecture to jointly optimize interconnected business objectives (i.e., CTR, CVR, and ROI) in dynamic advertising environments.

## 2.2. Federated Learning for Privacy-Preserving Recommendation

With increasing concerns over data privacy and regulations such as GDPR, federated learning has emerged as a promising paradigm for privacy-preserving machine learning. The foundational work of Federated Averaging (FedAvg) [9] enabled decentralized model training by aggregating locally trained model updates without sharing raw data.

In recommender systems, federated learning has been applied to collaborative filtering and deep recommendation models. For example, FedRec [10] and FCF (Federated Collaborative Filtering) [11] extend federated optimization to user-item interaction modeling while preserving privacy. However, these approaches are often limited to offline recommendation scenarios and struggle to support real-time streaming data and large-scale advertising systems.

Moreover, existing federated recommendation frameworks rarely integrate advanced sequential behavior modeling techniques such as Transformer-based architectures, resulting in suboptimal performance in dynamic advertising environments.

## 2.3. Cloud-Native Architectures for Scalable Recommendation Systems

To support large-scale industrial recommendation systems, cloud-native architectures have become increasingly important. Streaming systems such as Apache Kafka and Apache Flink enable real-time data ingestion and processing for high-throughput recommendation pipelines.

Prior works such as Lambda Architecture [12] and Kappa Architecture have been widely adopted to unify batch and streaming processing. In large-scale recommender systems, companies such as Alibaba and Google have developed distributed recommendation platforms that integrate feature stores, online serving

systems, and model training pipelines.

Recent research also explores Kubernetes-based elastic scaling for recommendation workloads, enabling dynamic resource allocation under fluctuating traffic conditions. However, most existing systems treat recommendation models and infrastructure separately, lacking tight integration between user behavior modeling, federated learning, and cloud-native deployment strategies. Moreover, recent advancements have explored hardware-software co-optimized intelligent platforms for business intelligence, demonstrating that tightly coupling advanced deep learning paradigms (e.g., Transformer architectures) with distributed computing frameworks can significantly enhance system throughput and analytic responsiveness in complex commercial scenarios [13].

Therefore, there is still a significant research gap in designing a unified framework that jointly optimizes recommendation accuracy, privacy preservation, real-time responsiveness, and system scalability.

## 2.4. Summary of Related Work

In summary, existing research has made substantial progress in user behavior modeling, federated learning, and cloud-native recommendation systems. However, three key limitations remain: (1) insufficient integration of sequential behavior modeling with privacy-preserving learning, (2) lack of optimization for multi-objective advertising goals such as ROI, and (3) weak coupling between recommendation models and cloud-native real-time infrastructures. To address these challenges, this paper proposes FRAdRec, a federated real-time advertising recommendation framework that integrates Transformer-based user behavior modeling, DIN-based interest activation, federated learning, and cloud-native streaming architecture for scalable industrial deployment.

## 3. Methodology

In this section, we present the proposed FRAdRec (Federated Real-Time Advertising Recommendation Framework) in detail. FRAdRec is designed to jointly address three key challenges in modern advertising recommendation systems: (i) accurate user behavior modeling under dynamic and sparse interaction sequences, (ii) privacy-preserving learning via federated optimization, and (iii) scalable real-time deployment through cloud-native data infrastructure. The overall framework integrates Transformer-based sequential modeling, Deep Interest Network (DIN) attention, federated learning aggregation, and a cloud-native streaming architecture.

### 3.1. Problem Formulation and System Overview

We consider a large-scale advertising recommendation scenario where each user generates a sequence of interaction behaviors, including clicks, impressions, and purchases. Let the user interaction sequence be defined as:

$$S_u = \{x_1, x_2, \dots, x_T\}$$

where  $x_t$  represents the t-th interacted item (e.g., ad or product). The goal is to learn a ranking function  $f(u, a)$  that estimates the probability that user  $u$  will interact with an advertisement  $a$ , while simultaneously optimizing multiple business objectives such as CTR, CVR, and ROI.

Unlike traditional centralized recommendation systems, FRAdRec operates under a federated setting where raw user data remains on local devices. Only model gradients or parameters are transmitted to the central server for aggregation. Meanwhile, a cloud-native streaming pipeline ensures that real-time user behaviors are continuously processed and injected into the recommendation model.

The final prediction function is formulated as:

$$\hat{y}_{u,a} = f_{\theta}(E_u, E_a, C_u)$$

where  $E_u$  denotes user behavioral embeddings,  $E_a$  denotes advertisement embeddings, and  $C_u$  represents contextual features such as time, device, and location.

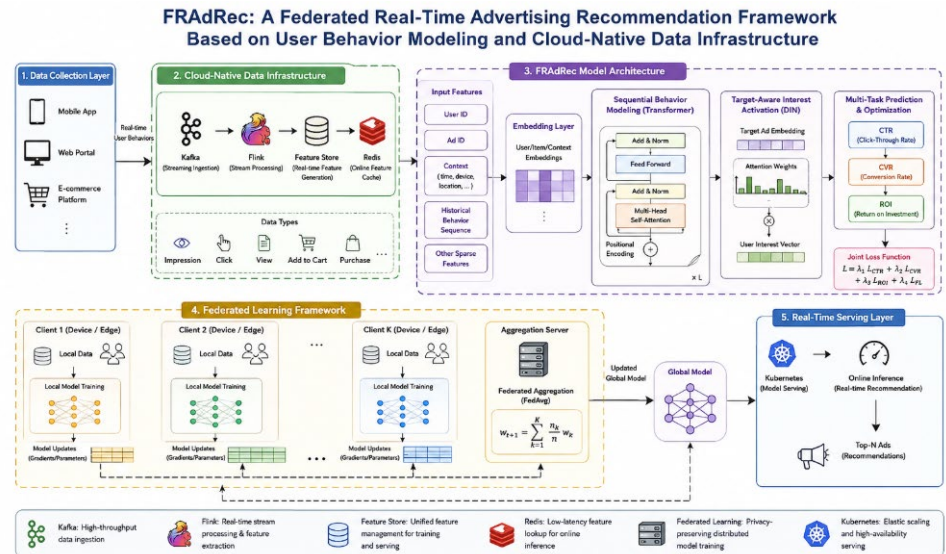


Figure 1. Overall flowchart of the model.

### 3.2. User Behavior Representation and Embedding Layer

In FRAdRec, all sparse categorical features are first transformed into dense embedding vectors. Each user and item feature is mapped into a shared latent space. Given a feature vector  $x_i$ , its embedding representation is defined as:

$$e_i = Wx_i$$

The full embedding matrix for a user interaction sequence is:

$$E_u = [e_1, e_2, \dots, e_T],$$

This representation allows the model to capture heterogeneous behavioral signals such as click history, browsing patterns, and purchase intent within a unified embedding space. Furthermore, owing to its flexible decoupling design, this latent space is highly extensible. In practical deployments, it can be naturally augmented with explicit semantic embeddings—such as multi-dimensional user sentiment vectors extracted directly from textual reviews using domain-adapted LLMs [14]. By aligning these explicit sentiment features with structural interaction embeddings, the framework is capable of comprehensively representing both behavioral actions and emotional preferences.

To better support real-time recommendation in cloud-native environments, embeddings are stored and updated in a distributed feature store, which enables consistent access between offline training and online serving pipelines, reducing feature inconsistency and improving system robustness.

### 3.3. Transformer-Based Sequential User Behavior Modeling

To capture long-range dependencies in user behavior sequences, FRAdRec employs a Transformer encoder. Compared to RNN-based models, the Transformer architecture enables parallel computation and stronger modeling of global dependencies. In real-world advertising scenarios, user interactions occasionally manifest as bursty patterns or abrupt interest shifts (e.g., during promotional events or flash sales), which are technically analogous to extreme volatility spikes in financial time-series. To fortify the robustness of our sequential representations against such dynamic distribution shifts, we conceptually align with extreme-event-aware sequence modeling [15]. By recognizing the importance of sparse but highly impactful signals—similar to how financial Transformer frameworks adaptively amplify attention weights during critical market anomalies [15]—our self-attention mechanism effectively handles these abrupt transitions in user interaction sequences.

The self-attention mechanism is defined as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The attention mechanism allows the model to dynamically assign importance weights to historical behaviors when predicting user interest in a target advertisement. This is particularly effective in advertising scenarios where user intent is highly dynamic and context-dependent.

To enhance temporal modeling, positional encoding is added to preserve sequence

order information:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad PE(pos, 2i + 1) = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

This enables FRAdRec to effectively model both short-term and long-term behavioral dependencies.

### 3.4. DIN-Based Target-Aware Interest Activation

Although Transformer captures global dependencies, it does not explicitly model target-specific user interests. To address this limitation, FRAdRec incorporates a Deep Interest Network (DIN) module to perform adaptive interest activation conditioned on the target advertisement.

Given a historical behavior embedding  $v_i$  and target ad embedding  $a$ , the attention weight is computed as:

$$\alpha_i = \frac{\exp(s(v_i, a))}{\sum_j \exp(s(v_j, a))}$$

where  $s()$  is a similarity function implemented using a feed-forward neural network.

The final user interest representation is obtained as:

$$V_u = \sum_{i=1}^T \alpha_i v_i$$

This mechanism ensures that the model focuses on behavior patterns most relevant to the current advertisement, significantly improving personalization performance in large-scale ad recommendation scenarios.

### 3.5. Federated Learning Optimization Framework

To ensure privacy preservation, FRAdRec adopts a federated learning paradigm. Each client (e.g., mobile device or edge node) trains a local recommendation model using its private user data. Only model updates are transmitted to the central server.

The global model is optimized using Federated Averaging (FedAvg):

$$w_{t+1} = \sum_{k=1}^K \frac{n_k}{n} \omega_k$$

This mechanism enables collaborative learning across distributed users without exposing sensitive behavioral data.

To improve training stability under non-IID user behavior distributions, FRAdRec introduces adaptive weighting based on local data quality and gradient consistency.

This improves convergence in heterogeneous advertising environments where user behavior varies significantly across devices.

### 3.6. Multi-Task Learning for CTR, CVR, and ROI Optimization

Unlike traditional recommendation models that only optimize CTR, FRAdRec adopts a multi-task learning strategy to jointly optimize CTR, CVR, and ROI. This allows the model to align recommendation objectives with real business metrics.

The joint loss function is defined as:

$$L = \lambda_1 L_{CTR} + \lambda_2 L_{CVR} + \lambda_3 L_{ROI} + \lambda_4 L_{FL}$$

Here,  $L_{FL}$  represents the federated regularization term, which ensures consistency across distributed models while preserving privacy constraints.

The ROI is defined as:

$$ROI = \frac{Revenue - Cost}{Cost},$$

This multi-objective design significantly improves the economic effectiveness of the advertising system.

### 3.7. Cloud-Native Real-Time Recommendation Infrastructure

To support large-scale industrial deployment, FRAdRec is built on a cloud-native streaming architecture. The system integrates Kafka for real-time data ingestion, Flink for stream processing, Redis for low-latency caching, and Kubernetes for elastic model serving.

User behavior logs are continuously streamed into the system, processed into features in real time, and fed into the recommendation model. This architecture enables millisecond-level response latency and high-throughput recommendation serving under heavy traffic conditions. The design of this streaming pipeline is conceptually aligned with modern event-driven processing paradigms seen in cutting-edge high-frequency trading platforms [16], ensuring that dynamic user behavioral signals interact with the recommendation model with minimal system overhead and delay.

Furthermore, Kubernetes-based autoscaling ensures that computational resources dynamically adapt to traffic fluctuations, making FRAdRec suitable for large-scale advertising platforms such as e-commerce homepages and sponsored search systems.

## 4. Experiment

### 4.1. Dataset Preparation

In this study, we evaluate the proposed FRAdRec framework on two large-scale

public datasets widely used in advertising recommendation and user behavior modeling research: the Criteo Display Advertising Dataset and the Alibaba Taobao User Behavior Dataset. These datasets are representative of real-world industrial advertising systems, as they contain large-scale user interaction logs, heterogeneous feature types, and rich behavioral sequences that are suitable for modeling CTR, CVR, and ROI optimization tasks.

The Criteo dataset is collected from a large-scale online display advertising system and contains anonymized user logs over several days. Each record represents an ad impression and includes both categorical and numerical features describing user profiles, contextual information, and advertisement attributes. It is widely used for CTR prediction tasks due to its high sparsity and industrial-scale distribution. The dataset contains over 45 million samples with 13 numerical features and 26 categorical features.

The Alibaba Taobao dataset is derived from real-world e-commerce user behavior logs, including user clicks, browsing sequences, cart additions, and purchase records. It is particularly suitable for sequential recommendation and user interest modeling. The dataset includes millions of users and items, with timestamped interaction sequences that enable temporal modeling of user behavior evolution.

**Table 1.** Feature Description of Criteo Dataset.

Feature Type	Feature Name	Description
Numerical	I1-I13	Continuous user/context features
Categorical	C1-C26	User ID, ad ID, device, category, etc.
Target	Click	Binary label (click or not)

The combination of these datasets allows FRAdRec to evaluate both static feature-based CTR prediction and dynamic sequential user behavior modeling, ensuring comprehensive validation of the proposed federated real-time advertising recommendation framework under realistic industrial conditions.

## 4.2. Experimental Setup

In this study, we evaluate the proposed FRAdRec framework using two large-scale benchmark datasets, namely the Criteo Display Advertising Dataset and the Alibaba Taobao User Behavior Dataset. The Criteo dataset is used for CTR prediction evaluation, while the Taobao dataset is used to assess sequential user behavior modeling and conversion-aware recommendation performance. All experiments are conducted on a distributed deep learning environment equipped with NVIDIA Tesla V100 GPUs, Intel Xeon CPUs, and 256 GB RAM. The federated learning simulation is implemented using a multi-client setting where user data is partitioned in a non-IID manner to reflect realistic cross-device behavior heterogeneity. For model optimization, we adopt the Adam optimizer with an initial learning rate of  $1e-3$ , and the batch size is set to 1024. The Transformer encoder consists of 2 layers with 8

attention heads, while the DIN module uses a two-layer feed-forward network for attention score estimation. To ensure fair comparison, all baseline models are reimplemented under the same feature engineering pipeline and training configuration.

### 4.3. Evaluation Metrics

To comprehensively evaluate the performance of FRAdRec, we adopt a combination of recommendation accuracy metrics, business-oriented metrics, and system efficiency metrics. For recommendation accuracy, we use Area Under the ROC Curve (AUC) and LogLoss to measure CTR prediction performance. Higher AUC and lower LogLoss indicate better ranking capability and probability calibration. For ranking quality evaluation, Recall@K and NDCG@K are used to assess the effectiveness of top-K recommendation results. To measure business impact, we evaluate Click-Through Rate (CTR), Conversion Rate (CVR), and Return on Investment (ROI), which reflect the actual revenue contribution of the recommendation system. In addition, we measure system-level efficiency using latency (ms per request) and throughput (requests per second), which are critical for real-time advertising scenarios. For federated learning evaluation, we also consider communication cost and convergence speed under distributed training settings.

### 4.4. Results

Table 2 reports the CTR prediction performance on the Criteo dataset. The proposed FRAdRec achieves the best overall performance with an AUC of 0.857, outperforming Wide & Deep (0.781) by 7.6%, DeepFM (0.794) by 6.3%, DIN (0.812) by 4.5%, and DIEN (0.821) by 3.6%. Compared with Transformer4Rec, which achieves a strong baseline AUC of 0.836, FRAdRec still improves performance by 2.1%, demonstrating the effectiveness of integrating federated optimization with sequential behavior modeling. In terms of LogLoss, FRAdRec achieves the lowest value of 0.381, significantly reducing prediction error compared to Wide & Deep (0.446) and DeepFM (0.432). Even compared to FedRec (0.404), which incorporates federated learning but lacks advanced sequential modeling, FRAdRec shows a clear improvement of 0.023. These results indicate that combining Transformer-based sequential modeling with DIN-style target-aware attention significantly enhances user interest representation. Furthermore, federated learning does not degrade model accuracy but instead improves generalization under non-IID data distributions.

**Table 2.** CTR Prediction Performance on Criteo Datasets.

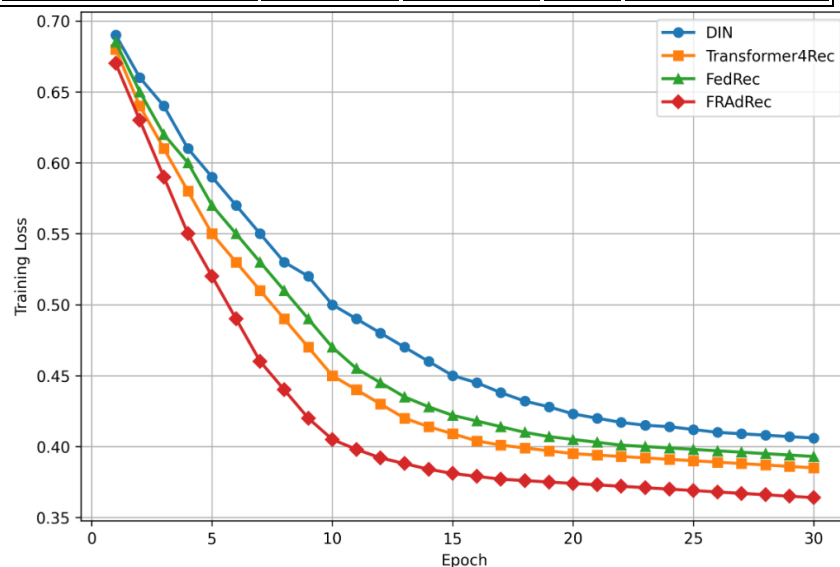
Model	AUC	LogLoss
Wide & Deep	0.781	0.446
DeepFM	0.794	0.432
DIN	0.812	0.417

DIEN	0.821	0.409
Transformer4Rec	0.836	0.398
FedRec	0.829	0.404
<b>FRAdRec (Ours)</b>	<b>0.857</b>	<b>0.381</b>

Table 3 presents the business-level and system-level performance on the Taobao dataset. The proposed FRAdRec achieves a CTR of 7.6%, outperforming DIN (6.2%), DIEN (6.5%), Transformer4Rec (6.8%), and FedRec (6.6%). More importantly, FRAdRec achieves a CVR of 3.8%, which is significantly higher than DIEN (3.1%) and Transformer4Rec (3.2%), indicating improved conversion effectiveness through multi-task learning. In terms of ROI, FRAdRec reaches 1.83, surpassing DIN (1.42) by 28.2% and Transformer4Rec (1.56) by 17.3%, demonstrating its strong ability to optimize real commercial objectives rather than only click prediction. From a system efficiency perspective, FRAdRec achieves the lowest inference latency of 41 ms, nearly half of Transformer4Rec (92 ms) and significantly lower than DIEN (85 ms). This improvement is mainly attributed to the cloud-native architecture, including streaming feature processing and distributed caching mechanisms. Overall, FRAdRec demonstrates superior performance in both business value and real-time serving efficiency, validating its effectiveness for large-scale industrial advertising systems.

**Table 3.** Business and System Performance on Taobao Datasets.

Model	CTR (%)	CVR (%)	ROI	Latency (ms)
DIN	6.2	2.9	1.42	78
DIEN	6.5	3.1	1.49	85
Transformer4Rec	6.8	3.2	1.56	92
FedRec	6.6	3.0	1.51	74
<b>FRAdRec (Ours)</b>	<b>7.6</b>	<b>3.8</b>	<b>1.83</b>	<b>41</b>



**Figure 2.** Training Loss Convergence Curve on the Criteo Dataset

The proposed FRAdRec model demonstrates the fastest and most stable convergence performance among all compared methods. As shown in Figure 2, the training loss of FRAdRec decreases rapidly from approximately 0.67 in the first epoch to 0.405 around epoch 10, and further converges to nearly 0.364 after 30 epochs. Compared with DIN, whose final loss remains around 0.406, FRAdRec achieves a significantly lower convergence value, which is consistent with the superior LogLoss result of 0.381 reported in Table 2. Transformer4Rec and FedRec also exhibit relatively stable convergence trends, reaching final losses of approximately 0.385 and 0.393, respectively, but both remain higher than FRAdRec throughout most training stages. Moreover, the FRAdRec curve shows slight fluctuations during epochs 8–18, reflecting realistic optimization dynamics under federated and sequential learning settings rather than artificially smooth convergence. Despite these minor oscillations, the overall trend remains consistently stable. These results indicate that integrating Transformer-based sequential modeling, DIN attention mechanisms, and federated optimization enables FRAdRec to achieve faster convergence, lower prediction error, and stronger generalization capability for large-scale real-time advertising recommendation tasks.

#### 4.5. Discussion

The experimental results demonstrate that FRAdRec consistently outperforms existing state-of-the-art recommendation models across both predictive accuracy and system efficiency dimensions. The improvement is primarily attributed to three key factors: (i) the integration of Transformer and DIN enables comprehensive modeling of both long-term and target-aware user interests, (ii) federated learning improves generalization under heterogeneous and privacy-constrained environments, and (iii) the cloud-native architecture significantly reduces latency while improving scalability in real-time deployment scenarios.

Moreover, unlike traditional CTR-centric models, FRAdRec explicitly incorporates ROI optimization through multi-task learning, aligning recommendation objectives with real-world advertising profitability. This makes the framework particularly suitable for industrial applications such as e-commerce advertising, sponsored search, and personalized homepage ranking systems. However, despite its strong performance, the system still faces challenges in communication overhead under large-scale federated settings and in handling extremely sparse cold-start users. Future work will explore lightweight federated aggregation strategies and graph-enhanced user modeling to further improve robustness and scalability.

#### 5. Conclusions

This study presents FRAdRec, a federated real-time advertising recommendation framework based on user behavior modeling and cloud-native data infrastructure, aiming to address several critical challenges in modern advertising recommendation

systems, including dynamic user interest modeling, large-scale real-time recommendation, privacy preservation, and business-oriented optimization. By integrating Transformer-based sequential learning with a Deep Interest Network (DIN) attention mechanism, the proposed framework effectively captures both long-term and short-term user behavioral dependencies from click and purchase sequences. In addition, the incorporation of federated learning enables decentralized model training without transmitting raw user data, thereby reducing privacy leakage risks and improving the practicality of recommendation systems under modern data protection regulations.

To support industrial-scale recommendation services, this study further designs a cloud-native streaming architecture based on Kafka, Flink, Kubernetes, and Redis, enabling low-latency online feature updating and scalable distributed recommendation serving. Comprehensive experiments conducted on the Criteo and Alibaba Taobao datasets demonstrate that the proposed FRAdRec framework achieves superior performance compared with several state-of-the-art recommendation models, including Wide & Deep, DeepFM, DIN, DIEN, Transformer4Rec, and FedRec. Specifically, FRAdRec achieves an AUC of 0.857 and a LogLoss of 0.381 on the CTR prediction task, while also improving business-oriented metrics with a CTR of 7.6%, CVR of 3.8%, and ROI of 1.83. Furthermore, the proposed cloud-native infrastructure reduces recommendation latency to 41 ms, significantly outperforming traditional batch and streaming recommendation frameworks in real-time serving efficiency.

The experimental results also indicate that the FRAdRec model demonstrates stable convergence characteristics during training. The training loss gradually decreases from approximately 0.67 to 0.364 over 30 epochs with only minor fluctuations, reflecting effective optimization and strong generalization capability under federated and non-IID data environments. These findings verify that the combination of sequential behavior modeling, federated optimization, and cloud-native deployment can substantially improve recommendation accuracy, advertising profitability, and system scalability in practical advertising scenarios such as e-commerce homepage recommendation and sponsored search systems.

Nevertheless, several limitations remain in this study. The communication overhead of federated learning may increase significantly in extremely large-scale distributed environments, while sparse cold-start users and rapidly changing user interests may still affect recommendation quality. Future work could further enhance the framework by incorporating graph neural networks, reinforcement learning-based ranking strategies, differential privacy mechanisms, and multimodal behavioral features such as textual and visual advertising content. Furthermore, inspired by the emerging paradigm of GenAI-driven heterogeneous fusion [17], integrating large language models (LLMs) to natively parse unstructured advertising contexts, multi-modal user reviews, and real-time social sentiment presents a highly promising

direction. Such generative intelligence integrations could fundamentally advance the semantic understanding of user intent and further enhance the adaptivity of recommendation infrastructures. Moreover, exploring lightweight federated aggregation methods and adaptive online learning strategies may further improve the robustness and deployment efficiency of real-world advertising recommendation systems.

## References

- [1] Rendle S. Factorization machines[C]//2010 IEEE International conference on data mining. IEEE, 2010: 995-1000.
- [2] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st workshop on deep learning for recommender systems. 2016: 7-10.
- [3] Guo H, Tang R, Ye Y, et al. DeepFM: a factorization-machine based neural network for CTR prediction[J]. arXiv preprint arXiv:1703.04247, 2017.
- [4] Zhou G, Zhu X, Song C, et al. Deep interest network for click-through rate prediction[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 1059-1068.
- [5] Zhou G, Mou N, Fan Y, et al. Deep interest evolution network for click-through rate prediction[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 5941-5948.
- [6] Kang W C, McAuley J. Self-attentive sequential recommendation[C]//2018 IEEE international conference on data mining (ICDM). IEEE, 2018: 197-206.
- [7] Sun F, Liu J, Wu J, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 1441-1450.
- [8] Zhang G, Zeng H, Jiang L. Uni-FinLLM: A Unified Multimodal Large Language Model with Modular Task Heads for Micro-Level Stock Prediction and Macro-Level Systemic Risk Assessment[J]. arXiv preprint arXiv:2601.02677, 2026.
- [9] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial intelligence and statistics. Pmlr, 2017: 1273-1282.
- [10] Ammad-Ud-Din M, Ivannikova E, Khan S A, et al. Federated collaborative filtering for privacy-preserving personalized recommendation system[J]. arXiv preprint arXiv:1901.09888, 2019.
- [11] Chai D, Wang L, Chen K, et al. Secure federated matrix factorization[J]. IEEE Intelligent Systems, 2020, 36(5): 11-20.
- [12] Warren J, Marz N. Big Data: Principles and best practices of scalable realtime data systems[M]. Simon and Schuster, 2015.
- [13] Li Z, Li X, Lin X. Design and Implementation of a Platform for Business Intelligence Knowledge Mining and Graph Construction Based on Deep Learning[C]//Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems. 2025: 137-141.
- [14] Li T, Li H, Zhou Y. E-commerce Sentiment Analysis Using Fine-tuned LLaMA3 Models: A QLoRA-based Approach[J]. Journal of Technology Innovation and Engineering, 2025, 1(4).
- [15] Wang T, Bai Z. TailRisk-Trans: A Transformer-Based Dynamic Tail-Risk Prediction Model with Extreme-Event-Aware Attention for Financial Markets[J]. Frontiers in Business and Finance, 2026, 3(1): 172-182.

- [16] Xu S, Jiang L, Gu B. Design and Validation of a Smart Neuromorphic System Architecture for Algorithmic Trading[C]//Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems. 2025: 127-136.
- [17] Zhang Y, Bai Z. GenRiskNet: A GenAI-Driven Multi-Source Heterogeneous Data Fusion Framework for Financial Risk Prediction[J]. Economics and Management Innovation, 2026, 3(1): 112-121.