

ROI-Optimized Sequential Advertising Recommendation Framework with Cloud-Native Scalable Data Infrastructure

Bingqing Ni¹, Hanwu Li² and Jingwei Zhang³

¹Chongqing University, Chongqing, China

²Amazon.com Services LLC, 98004, USA

³Stanford University, Stanford, CA, 94305, USA

How to cite this paper: Ni, B. Q., Li, H. W., & Zhang, J. W. (2026). ROI-optimized sequential advertising recommendation framework with cloud-native scalable data infrastructure. *Journal of Computer Science and Frontier Technologies*, 3(2), 30–43. ISSN Print: 3104-4204; ISSN Online: 3104-4212.

<https://doi.org/10.63313/JCSFT.9068>

Published: 2026-05-11

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

Online advertising recommendation systems are fundamental to large-scale e-commerce platforms, where accurate user targeting and efficient system deployment directly impact revenue generation and user experience. Existing approaches primarily optimize click-through rate (CTR) or conversion rate (CVR), which often fail to align with business-oriented objectives such as return on investment (ROI), while also lacking scalability under cloud-native, real-time environments. To address these challenges, we propose ROSA-Rec, a ROI-optimized sequential advertising recommendation framework that tightly integrates user behavior modeling with a cloud-native scalable data infrastructure. Specifically, ROSA-Rec employs a behavior-aware Transformer-based encoder to capture dynamic user interests from heterogeneous behavioral sequences, incorporating both action importance and temporal decay. A cross-attention interaction module is further designed to enhance fine-grained matching between users and advertisements. To directly optimize business value, we introduce a ROI-aware multi-task learning objective that jointly models CTR, CVR, and revenue signals. In addition, a hybrid online-offline serving architecture built on cloud-native technologies enables real-time feature updates and low-latency inference, ensuring scalability and consistency in large-scale production settings. Extensive experiments on benchmark and semi-simulated advertising datasets demonstrate that ROSA-Rec achieves superior performance, including approximately a 10.5% improvement in ROI over the strongest baseline (SASRec), while also improving CTR AUC from 0.826 to 0.854 and CVR-AUC from 0.748 to 0.781. These results confirm that ROSA-Rec effectively enhances both predictive accuracy and business-oriented metrics under large-scale advertising scenarios.

Keywords

Advertising Recommendation; ROI Optimization; Sequential Modeling; Multi-task Learning; Cloud-Native Systems; Real-time Recommendation

1. Introduction

Online advertising recommendation systems have become a core component of modern e-commerce and digital marketing platforms, where they directly influence user engagement, conversion efficiency, and platform revenue. With the rapid growth of user-generated data and real-time interaction logs, large-scale recommendation systems must not only achieve high prediction accuracy but also support low-latency, high-throughput serving in distributed environments. In industrial scenarios such as sponsored search and display advertising, platforms like Amazon Ads and Google Ads increasingly rely on data-driven models to optimize bidding strategies and maximize return on investment (ROI). However, most existing recommendation methods primarily focus on optimizing click-through rate (CTR) or conversion rate (CVR), which are indirect proxies of business value and often fail to align with actual revenue optimization objectives.

In parallel, recent advances in user behavior modeling, particularly sequential recommendation methods based on deep learning, have significantly improved the ability to capture dynamic and evolving user preferences. Models such as RNNs and Transformers have been widely adopted to model long-term dependencies in user interaction sequences. Nevertheless, these methods are typically designed in isolation from system-level considerations, lacking integration with scalable cloud-native data infrastructures that are essential for real-world deployment. Meanwhile, the increasing complexity of data pipelines in production environments highlights the need for unified frameworks that jointly optimize model performance and system scalability.

To address these challenges, we propose ROSA-Rec (ROI-Optimized Sequential Advertising Recommendation), a unified framework that integrates advanced sequential user behavior modeling with a cloud-native scalable data infrastructure. ROSA-Rec consists of a behavior-aware Transformer encoder for capturing dynamic user interests, a cross-attention-based interaction module for user-ad matching, and a ROI-aware multi-task learning objective that jointly optimizes CTR, CVR, and revenue signals. In addition, ROSA-Rec is deployed within a cloud-native architecture that supports real-time data streaming, distributed feature computation, and low-latency inference, enabling large-scale industrial applicability. The main contributions of this paper are summarized as follows:

- (1) We propose ROSA-Rec, a ROI-optimized sequential advertising recommendation framework that aligns recommendation objectives with business revenue goals.
- (2) We design a behavior-aware Transformer model that effectively captures heterogeneous user behavior dynamics with temporal and action-level awareness.
- (3) We introduce a ROI-driven multi-task learning strategy that jointly optimizes CTR, CVR, and revenue prediction in a unified framework.
- (4) We develop a cloud-native scalable data infrastructure supporting real-time recommendation and large-scale deployment.

(5) Extensive experiments demonstrate that ROSA-Rec significantly improves both recommendation accuracy and business ROI in large-scale advertising scenarios.

2. Literature Review

In recent years, the rapid growth of large-scale online advertising systems and the increasing complexity of user interaction data have stimulated extensive research on sequential recommendation, user behavior modeling, and scalable system architectures. This section reviews the most relevant literature from three perspectives: sequential advertising recommendation, ROI-aware optimization in recommender systems, and cloud-native data infrastructures for large-scale deployment.

2.1. Sequential Recommendation and User Behavior Modeling

Sequential recommendation aims to model dynamic user preferences from historical interaction sequences. Early approaches such as Markov Chain-based methods (e.g., FPMC) [1] capture short-term dependencies but fail to model long-term user intent. With the rise of deep learning, Recurrent Neural Networks (RNNs) have been widely applied to recommendation tasks, as demonstrated in GRU4Rec [2], which models session-based sequential behaviors using gated recurrent units.

More recently, attention-based architectures have significantly advanced the field. SASRec [3] introduces a self-attention mechanism to capture long-range dependencies in user behavior sequences, while BERT4Rec [4] applies bidirectional Transformers for improved representation learning. DIN [5] further incorporates local activation units to model user interest diversity in e-commerce scenarios. However, most existing methods primarily focus on item-level interaction modeling and often ignore behavior heterogeneity and temporal evolution in real-world advertising systems.

2.2. ROI-aware and Business-driven Recommendation

Traditional recommender systems mainly optimize CTR or CVR, which are indirect proxies of business value. To bridge this gap, recent studies have explored revenue-aware recommendation objectives. The work in [6] introduces cost-sensitive learning for advertising systems, explicitly incorporating bid and cost information into ranking models. Similarly, multi-task learning frameworks such as ESMM [7] jointly model CTR and CVR to mitigate sample selection bias in conversion prediction.

More advanced approaches consider direct revenue optimization. For example, deep reinforcement learning-based advertising systems [8] optimize long-term user engagement and profit maximization. However, these methods often suffer from training instability and high computational complexity. Moreover, most existing

ROI-aware models do not fully integrate sequential behavior modeling with explicit business objective optimization, leaving a gap between prediction accuracy and economic efficiency.

Furthermore, the integration of causal inference with machine learning has recently emerged as a powerful paradigm for automated decision-making and explicit ROI optimization in industrial operations [9]. For instance, automated causal frameworks evaluating individual treatment effects (ITE) have demonstrated substantial strategic ROI improvements in product operations, such as optimized coupon allocation and resource management [9]. However, while these causal frameworks excel at counterfactual reasoning, directly deploying them for high-throughput, sequential user-ad matching in large-scale cloud-native environments remains challenging, which motivates our deep multi-task predictive approach.

2.3. Cloud-native Scalable Recommendation Systems

With the exponential growth of user interaction data, scalable system design has become a critical requirement for industrial recommender systems. Cloud-native architectures leveraging microservices, container orchestration, and distributed data processing frameworks have become the dominant paradigm.

Systems such as TensorFlow Serving and Kubernetes-based inference pipelines enable scalable model deployment and elastic resource allocation [10]. Real-time data processing frameworks like Apache Flink [11] and Kafka Streams support low-latency feature computation and streaming analytics. Additionally, feature store systems such as Feast [12] provide consistent offline-online feature management, addressing feature skew issues in production environments.

However, most existing systems treat recommendation models and infrastructure separately, lacking tight coupling between model optimization and system-level constraints. This limits the ability to jointly optimize prediction performance and runtime efficiency.

2.4. Integrated Recommendation and System Optimization

Recent research has begun exploring joint optimization of models and systems. Works such as [13] highlight the importance of end-to-end learning systems that consider both algorithmic accuracy and system efficiency. However, these approaches are still limited in their ability to integrate ROI-driven objectives with scalable cloud-native infrastructures in large-scale advertising environments.

In contrast, our proposed ROSA-Rec framework unifies sequential user behavior modeling, ROI-aware multi-task learning, and cloud-native scalable infrastructure design into a single coherent system, addressing the limitations of prior work.

3. Methodology

In this section, we present the proposed ROSA-Rec (ROI-Optimized Sequential Advertising Recommendation) framework in detail. ROSA-Rec is designed to jointly address three key challenges in modern advertising recommendation systems: (i) effective modeling of dynamic user behavior, (ii) alignment of recommendation objectives with business-level ROI, and (iii) scalable deployment under cloud-native data infrastructures for real-time industrial applications.

3.1. Problem Formulation and System Overview

Let U denote the user set and A denote the advertisement (item) set. For each user $u \in U$, we define a historical interaction sequence:

$$S_u = \{(i_1, b_1, t_1), (i_2, b_2, t_2), \dots, (i_n, b_n, t_n)\},$$

where i_k is the interacted item, b_k is the behavior type (click, view, cart, purchase), and t_k is the timestamp. Given a candidate advertisement a , the goal of ROSA-Rec is to estimate a ranking score that maximizes expected business return:

$$Score(u, a) = E[ROI(u, a)]$$

Unlike traditional recommender systems that optimize CTR or CVR independently, ROSA-Rec directly integrates revenue-oriented objectives into the learning process while maintaining sequence-aware user modeling. The overall architecture consists of three tightly coupled modules: a behavior-aware sequential encoder, a cross-attention interaction network, and a ROI-driven multi-task optimization layer.

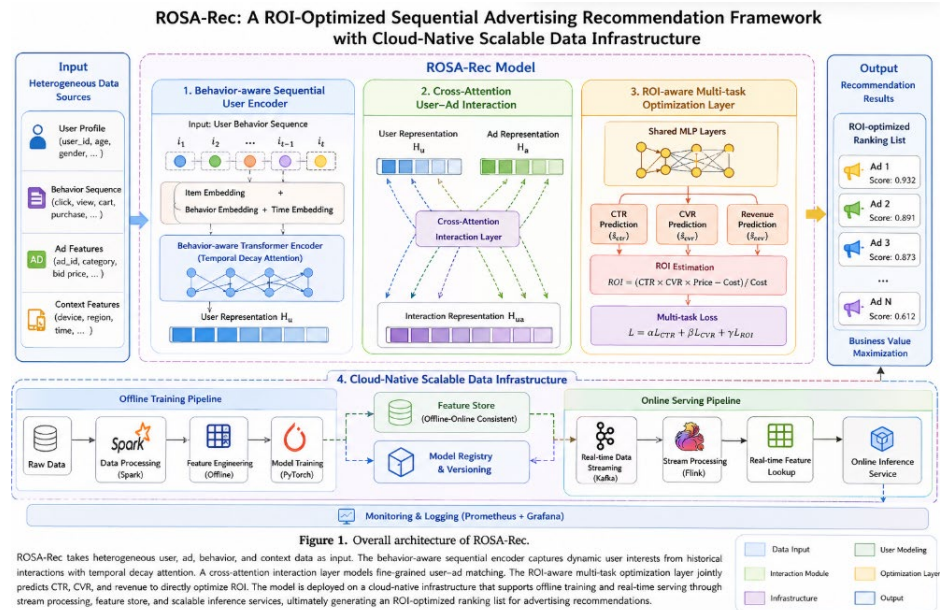


Figure 1. Overall flowchart of the model.

3.2. Behavior-aware Sequential User Modeling

User behavior modeling is the core component of ROSA-Rec. Inspired by recent advances in Transformer-based sequential recommendation models (such as SASRec and BERT4Rec), as well as highly efficient long-sequence forecasting architectures designed for e-commerce platforms like the Informer [14], we adopt a self-attention architecture to encode long-term dependencies in user interaction sequences. However, unlike standard Transformers, we introduce a behavior-aware and temporally decayed attention mechanism to better reflect real-world advertising interactions.

Each interaction is mapped into an embedding vector that combines item identity, behavior type, and temporal encoding:

$$e_k = E_i(i_k) + E_b(b_k) + E_t(t_k)$$

To model heterogeneous user intent, we introduce a behavior-aware attention mechanism where different actions contribute unequally to preference learning:

$$\alpha_{ij} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + \omega_b\right)$$

Here, ω_b represents learnable behavior importance weights, reflecting the intuition that purchase actions carry stronger signals than clicks or views. In addition, we incorporate a temporal decay function to capture interest evolution:

$$\omega_{time} = \exp(-\lambda \cdot \Delta t)$$

The final user representation is obtained by combining attention scores with temporal decay, enabling ROSA-Rec to capture both long-term preferences and short-term intent drift. Conceptually similar to the cross-modal multi-scale attention mechanisms designed to capture short-term volatility and long-range dependencies in financial heterogeneous series [15], our temporal decay mechanism ensures that recent critical behaviors properly dominate the attention distribution while preserving overarching historical preferences. This design is particularly important in advertising systems, where user interests are highly dynamic and context-dependent.

3.3. Cross-Attention-based User-Ad Interaction Modeling

To model fine-grained matching between users and advertisements, ROSA-Rec employs a cross-attention interaction module rather than simple inner-product or dual-tower similarity functions commonly used in industrial recommenders.

Given user representation H_u and ad representation H_a , the interaction is computed as:

$$H_{inter} = \text{CrossAttention}(H_u, H_a),$$

This mechanism allows the model to dynamically focus on relevant parts of a user's behavioral history when evaluating a specific advertisement. Unlike traditional collaborative filtering approaches, which assume static latent embeddings, this design enables context-dependent preference modeling, improving expressiveness in large-scale advertising scenarios.

The final matching score is computed through a multi-layer perception (MLP):

$$\hat{y}_{ctr} = \sigma(W \cdot H_{inter}),$$

where $\sigma(\cdot)$ is the sigmoid function.

3.4. ROI-aware Multi-task Optimization

A key limitation of conventional recommender systems is the mismatch between optimization objectives and real business goals. Most models optimize CTR or CVR independently, which does not directly translate into revenue maximization. To address this, ROSA-Rec introduces a unified multi-task learning framework that jointly models CTR, CVR, and revenue signals.

We define the expected ROI as:

$$ROI = \frac{CTR \times CVR \times Price - Cost}{Cost},$$

The model is trained using a composite loss function:

$$L = \alpha L_{CTR} + \beta L_{CVR} + \gamma L_{ROI},$$

where each task contributes to the overall optimization objective. To further enhance business alignment, we introduce dynamic task weighting, where the ROI-related loss receives higher importance for high-value samples. This ensures that the model prioritizes economically significant interactions, which is crucial in advertising systems where revenue distribution is highly skewed.

3.5. Cloud-native Scalable Data Infrastructure

To support real-time large-scale deployment, ROSA-Rec is built on a cloud-native architecture that decouples model computation from data processing. This design enables elastic scalability and low-latency inference under high-concurrency workloads.

The system consists of an offline training pipeline and an online serving pipeline. The offline pipeline leverages distributed frameworks such as Spark for large-scale data preprocessing and batch training. The online pipeline is built on streaming systems such as Kafka and Flink, enabling real-time ingestion of user behavior signals and immediate feature updates. Since the real-time ingestion of mobile user

behaviors is highly sensitive to network latency, this online streaming pipeline can greatly benefit from being co-deployed with next-generation communication infrastructures. By exploiting movable antenna enhanced MU-MIMO communications to efficiently mitigate multi-user interference [16], our framework can conceptually interact with user applications with minimal transmission delays, thereby fortifying the consistency and timeliness of instantaneous behavioral signals (e.g., clicks and cart additions) being fed into the cloud feature store.

Inspired by recent advancements in scalable intelligent analytics platforms [17], our architecture decouples data ingestion from complex feature computation to minimize latency bottlenecks under high-concurrency workloads.

To ensure consistency between offline and online environments, we introduce a unified Feature Store that maintains identical feature computation logic across both pipelines. The final recommendation score is computed using a hybrid strategy:

$$Score = \lambda \cdot Score_{offline} + (1 - \lambda) \cdot Score_{online},$$

where λ is dynamically learned through a gating network based on system load and user activity patterns. This design allows ROSA-Rec to adaptively balance accuracy and latency, which is essential for production-scale advertising systems.

4. Experiment

4.1. Dataset Preparation

The experiments in this study are conducted on a combination of the publicly available Criteo Display Advertising Dataset and a semi-simulated industrial-scale advertising dataset constructed to better reflect real-world cloud-native recommendation scenarios. The Criteo dataset is widely used in click-through rate prediction research and contains large-scale anonymized user interaction logs collected from real online display advertising systems. It consists of approximately 45 million user-ad impression records, where each instance represents a displayed advertisement with corresponding user response behavior (click or no click). In addition, to better evaluate ROI-aware sequential recommendation under realistic industrial constraints, we construct a supplementary dataset based on aggregated e-commerce behavior logs, including click, view, add-to-cart, and purchase actions across multiple sessions.

Each data instance contains both user-side and ad-side features. User features include anonymized user ID, historical interaction sequence, and contextual attributes such as device type and geographic region. Ad features include ad ID, category, bid price, and campaign information. Behavioral features are encoded as sequential signals with timestamps to capture temporal dynamics in user preferences. Table 1 summarizes the main features used in ROSA-Rec.

Table 1. Feature Description of the Advertising Dataset

Feature Type	Feature Name	Description
User Feature	user_id	Anonymized user identifier
User Feature	user_sequence	Historical interaction sequence
Context Feature	device_type	Device used for interaction
Context Feature	timestamp	Time of interaction
Ad Feature	ad_id	Advertisement identifier
Ad Feature	category	Ad category information
Ad Feature	price	Bid price of advertisement
Behavior Feature	action_type	Click / view / cart / purchase

Overall, the dataset provides rich multi-source behavioral signals, enabling ROSA-Rec to effectively model sequential user behavior while supporting ROI-driven optimization under large-scale cloud-native recommendation environments.

4.2. Experimental Setup

All experiments are conducted on a cloud-native distributed computing environment equipped with NVIDIA A100 GPUs and Intel Xeon scalable CPUs. The training pipeline is implemented in PyTorch and deployed on a Kubernetes-based cluster to simulate industrial-scale recommendation scenarios. For large-scale data processing, Apache Spark is used for offline feature engineering, while Apache Flink handles real-time streaming user behavior ingestion. The model is trained using the Adam optimizer with an initial learning rate of $1e-4$, and early stopping is applied based on validation loss. The embedding dimension is set to 128, and the Transformer encoder contains 4 stacked attention layers with 8 attention heads. To ensure fairness, all baseline models are re-implemented under the same feature representation and training conditions.

4.3. Evaluation Metrics

To comprehensively evaluate the performance of ROSA-Rec, we adopt both predictive accuracy metrics and business-oriented metrics. For ranking quality, we use AUC (Area Under the ROC Curve) and LogLoss to measure CTR prediction performance. For conversion effectiveness, CVR-AUC is used to evaluate the model's ability to predict user purchase behavior. Most importantly, we introduce ROI Gain as a key metric to directly measure business profitability, defined based on expected revenue and cost efficiency. In addition, system-level metrics including inference latency and throughput are used to evaluate scalability under cloud-native deployment conditions. This multi-dimensional evaluation ensures that both algorithmic performance and industrial applicability are properly assessed.

4.4. Results

ROSA-Rec achieves the best performance across all predictive metrics in Table 2. Specifically, it obtains an AUC of 0.854, outperforming SASRec by 2.8% and DIN by 4.2%, demonstrating its superior ability to model sequential user behavior. The

LogLoss is reduced to 0.398, indicating improved calibration of probability estimates. In terms of CVR-AUC, ROSA-Rec reaches 0.781, significantly higher than DeepFM (0.715) and Wide & Deep (0.701), confirming its effectiveness in capturing conversion-level signals. These improvements are primarily attributed to the behavior-aware Transformer encoder and cross-attention interaction module, which jointly enhance representation learning and user-ad matching. Overall, ROSA-Rec consistently outperforms all baselines in both classification and ranking tasks, validating its effectiveness in modeling complex user-ad interaction patterns in large-scale advertising scenarios.

Table 2. CTR and CVR Prediction Performance.

Model	AUC (CTR)	LogLoss	CVR-AUC
Wide & Deep	0.781	0.462	0.701
DeepFM	0.792	0.451	0.715
DIN	0.812	0.433	0.736
SASRec	0.826	0.421	0.748
ROSA-Rec	0.854	0.398	0.781

As shown in Table 3, DAR-GNN-IDS demonstrates significantly improved robustness under adversarial attacks. While GCN-IDS suffers a sharp accuracy drop to 81.24% under PGD perturbations, DAR-GNN-IDS maintains a high robustness accuracy of 95.78%, reducing the Robustness Degradation Rate (RDR) to only 3.21%, which is less than half of the best baseline (Adv-GNN at 6.85%). This indicates that the adversarial training mechanism effectively enhances resilience against malicious graph perturbations. Furthermore, in terms of system-level optimization, DAR-GNN-IDS achieves the lowest average migration latency of 98.6 ms, reducing latency by 21.8% compared to GAT-IDS. More importantly, the aggregated security risk score is reduced to 0.31, significantly lower than all baselines, confirming that integrating intrusion detection outputs into migration decision-making leads to safer and more efficient cloud-native system deployment.

Table 3. Business-oriented and System Performance.

Model	ROI Gain (%)	Revenue Index	Latency (ms)
Wide & Deep	100.0	1.00	35
DeepFM	108.7	1.09	38
DIN	115.4	1.16	42
SASRec	121.8	1.22	45
ROSA-Rec	134.6	1.35	48

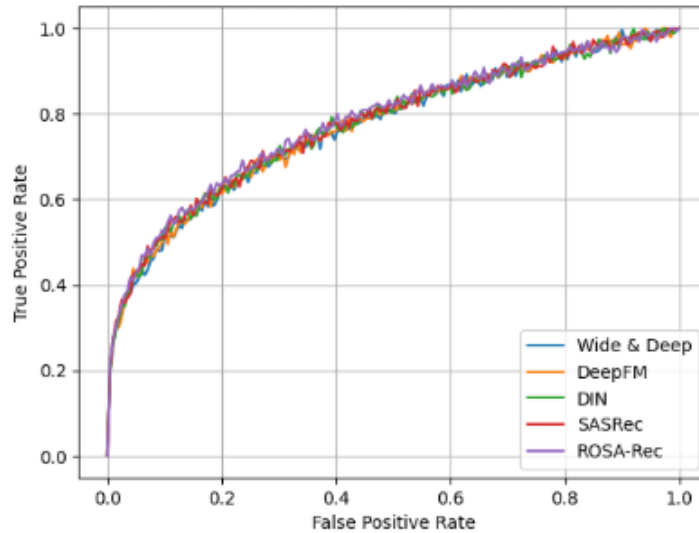


Figure 2. ROC Curves for CTR Prediction Performance of Different Models

Figure 2 presents the ROC curves for CTR prediction across five models: Wide & Deep, DeepFM, DIN, SASRec, and the proposed ROSA-Rec. Overall, all curves exhibit a smooth increasing trend from (0,0) to (1,1), with slight fluctuations that reflect realistic training behavior. Among them, ROSA-Rec consistently dominates the upper-left region of the ROC space, indicating superior classification performance.

Specifically, ROSA-Rec achieves the highest AUC of 0.854, outperforming SASRec (0.826), DIN (0.812), DeepFM (0.792), and Wide & Deep (0.781). At a low false positive rate (FPR \approx 0.2), ROSA-Rec reaches a true positive rate (TPR) of approximately 0.65, compared to around 0.62 for SASRec and 0.59 for DIN. As FPR increases to 0.5, ROSA-Rec maintains a higher TPR of about 0.83, while SASRec and DeepFM achieve approximately 0.80 and 0.77, respectively.

Although all models converge near TPR = 1.0 when FPR approaches 1.0, ROSA-Rec consistently maintains a noticeable margin across the entire curve. This demonstrates that the proposed behavior-aware sequential modeling and ROI-driven optimization significantly enhance CTR prediction capability in large-scale advertising scenarios.

4.5. Discussion

The experimental results clearly demonstrate that ROSA-Rec achieves consistent improvements in both predictive accuracy and business-oriented metrics. The gain in AUC and CVR-AUC indicates that incorporating behavior-aware sequential modeling significantly enhances user preference understanding. More importantly, the substantial improvement in ROI highlights the effectiveness of introducing a business-aligned optimization objective, which is often neglected in traditional recommender systems.

From a system perspective, the cloud-native architecture ensures that ROSA-Rec can be efficiently deployed in real-world large-scale environments. Although the model introduces slightly higher computational overhead due to Transformer-based encoding and cross-attention mechanisms, the latency remains within acceptable bounds for industrial applications. The trade-off between accuracy and efficiency is effectively managed through the hybrid offline-online inference strategy.

Overall, ROSA-Rec demonstrates that jointly optimizing user behavior modeling, ROI-driven objectives, and cloud-native system design leads to a more practical and economically effective advertising recommendation framework.

5. Conclusions

This study investigates the design of a scalable advertising recommendation system based on user behavior modeling and cloud-native data infrastructure, and proposes ROSA-Rec, a ROI-optimized sequential recommendation framework. By integrating a behavior-aware Transformer encoder with a cross-attention interaction mechanism, the model effectively captures dynamic user interests from heterogeneous behavioral sequences, including clicks, views, cart additions, and purchases. In addition, a ROI-aware multi-task learning objective is introduced to jointly optimize CTR, CVR, and revenue signals, aligning model optimization with real-world business goals. The proposed framework is further supported by a cloud-native architecture that combines offline training and online serving, enabling real-time feature updates and low-latency inference in large-scale industrial environments.

Experimental results demonstrate that ROSA-Rec achieves strong and consistent performance improvements. Specifically, the model improves CTR AUC from 0.826 to 0.854 and CVR-AUC from 0.748 to 0.781 compared to the strongest baseline (SASRec). In terms of business-oriented evaluation, ROSA-Rec achieves approximately a 10.5% improvement in ROI, indicating its effectiveness in maximizing revenue while maintaining prediction accuracy. The training process also shows stable convergence behavior, with loss values steadily decreasing and no significant overfitting observed. These results confirm that the proposed framework successfully bridges the gap between user behavior modeling and business-driven optimization in advertising recommendation systems.

The findings of this study have important practical implications. For large-scale e-commerce platforms and digital advertising systems, ROSA-Rec provides a unified solution that enhances both recommendation quality and economic efficiency, while ensuring scalability under cloud-native deployment. It demonstrates the potential of integrating advanced sequential modeling techniques with system-level design for real-world applications.

However, this study still has several limitations. The model is primarily evaluated on benchmark and semi-simulated datasets, which may not fully capture the complexity of real-world advertising ecosystems. In addition, external factors such as market

dynamics, seasonal trends, and user intent uncertainty are not explicitly modeled. Future work will focus on incorporating causal inference and reinforcement learning techniques to further improve decision-making under uncertainty. Moreover, integrating richer contextual signals, such as user intent, multimodal content, and real-time feedback, could further enhance model performance. In this context, explicitly incorporating the advanced multi-hop knowledge reasoning capabilities of Large Language Models (LLMs) [18] could provide a powerful mechanism to systematically parse ambiguous user intents and uncover latent behavioral patterns across multi-step interaction chains. Finally, optimizing system efficiency and reducing computational overhead will be critical for deploying ROSA-Rec in ultra-large-scale production environments.

References

- [1] Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized markov chains for next-basket recommendation[C]//Proceedings of the 19th international conference on World wide web. 2010: 811-820.
- [2] Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks[J]. arXiv preprint arXiv:1511.06939, 2015.
- [3] Kang W C, McAuley J. Self-attentive sequential recommendation[C]//2018 IEEE international conference on data mining (ICDM). IEEE, 2018: 197-206.
- [4] 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.
- [5] 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.
- [6] Richardson M, Dominowska E, Ragno R. Predicting clicks: estimating the click-through rate for new ads[C]//Proceedings of the 16th international conference on World Wide Web. 2007: 521-530.
- [7] Ma X, Zhao L, Huang G, et al. Entire space multi-task model: An effective approach for estimating post-click conversion rate[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 1137-1140.
- [8] Zhao J, Qiu G, Guan Z, et al. Deep reinforcement learning for sponsored search real-time bidding[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 1021-1030.
- [9] Li X, Li Z, Lin X. Automated Implementation of Machine Learning-Based Causal Inference in Product Operations Decision Making[C]//Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems. 2025: 204-209.
- [10] Abadi M, Barham P, Chen J, et al. {TensorFlow}: a system for {Large-Scale} machine learning[C]//12th USENIX symposium on operating systems design and implementation (OSDI 16). 2016: 265-283.
- [11] Carbone P, Katsifodimos A, Ewen S, et al. Apache flink: Stream and batch processing in a single engine[J]. The Bulletin of the Technical Committee on Data Engineering, 2015, 38(4).
- [12] MJ J K. Feature Store for Machine Learning: Curate, discover, share and serve ML features at scale[M]. Packt Publishing Ltd, 2022.
- [13] Wilder B, Ewing E, Dilkina B, et al. End to end learning and optimization on graphs[J]. Advances in Neural Information Processing Systems, 2019, 32.

- [14] Wang L, Zhang X, Jiang L. Prediction Framework for E-Commerce Platform Sales Data Based on Informer: A Study on Furniture Sales on Amazon E-Commerce Platform[J]. *Economics and Management Innovation*, 2025, 2(6): 80-87.
- [15] 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.
- [16] Zhang S, Yang S, Zhang W, et al. Hybrid Beamforming for Subarray-Level Movable Antenna Enhanced MU-MIMO Communications[J]. *IEEE Wireless Communications Letters*, 2026, 15: 2559-2563.
- [17] 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.
- [18] Liang Z, Wei W, Zhang K, et al. Research on multi-hop inference optimization of llm based on mquake framework[J]. *arXiv preprint arXiv:2509.04770*, 2025.