Mixture of Sequence: Theme-Aware Mixture-of-Experts for Long-Sequence Recommendation

Xiao Lin
UIUC
Urbana, IL, USA
xiaol13@illinois.edu

Zhicheng Tang
Meta
Sunnyvale, CA, USA
roberttang@meta.com

Weilin Cong, Mengyue Hang
Meta
Sunnyvale, CA, USA

Kai Wang
Meta
Sunnyvale, CA, USA
wangkai@meta.com

Yajuan Wang
Meta
Sunnyvale, CA, USA
yajuanwang@meta.com

Zhichen Zeng, Ting-Wei Li, Hyunsik Yoo
UIUC
Urbana, IL, USA

Zhining Liu, Xuying Ning, Ruizhong Qiu
UIUC
Urbana, IL, USA

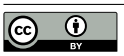
Wen-Yen Chen, Shuo Chang, Rong Jin
Meta
Sunnyvale, CA, USA

Huayu Li
Meta
Sunnyvale, CA, USA
huayuli@meta.com

Hanghang Tong
UIUC
Urbana, IL, USA
htong@illinois.edu

Abstract

Sequential recommendation has emerged as a rapidly growing research area in click-through rate prediction due to its ability to capture dynamic user interests from historical interaction sequences. A key challenge, however, lies in modeling long sequences, where users often exhibit pronounced interest shifts, thereby introducing substantial irrelevant or even misleading information into the prediction process. Our empirical analysis corroborates this challenge and further uncovers a recurring behavioral pattern in long sequences, which we term the *session hopping* phenomenon: while user interests remain stable within a short temporal span, referred to as a *session*, they often exhibit drastic shifts across sessions and may reappear after multiple sessions. To address this challenge, we propose the Mixture of Sequence (MoS) framework, a model-agnostic MoE approach that achieves accurate predictions by extracting theme-specific and multi-scale subsequences from noisy raw user sequences. First, MoS employs a theme-aware routing mechanism to adaptively learn the latent themes of user sequences and organizes these sequences into multiple coherent subsequences. Each subsequence contains only sessions aligned with a specific theme, thereby effectively filtering out irrelevant or even misleading information introduced by user interest shifts in session hopping. In addition, to alleviate potential information loss caused by subsequence extraction, we introduce a multi-scale fusion mechanism, which leverages three types of experts to capture global sequence characteristics, short-term user behaviors, and theme-specific semantic patterns. Together, these two mechanisms endow MoS with the ability to deliver accurate recommendations from multi-faceted and multi-scale perspectives. Experimental results demonstrate that MoS consistently improves the performance of long-sequence recommendation models while introducing fewer FLOPs compared with other MoE counterparts, providing strong evidence of its excellent balance between utility and efficiency. The code is available at https://github.com/xiaolin-cs/MoS.



WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792483

CCS Concepts

• **Information systems** → **Personalization**.

Keywords

Recommendation, Mixture of Experts, Sequence Modeling

ACM Reference Format:
Xiao Lin, Zhicheng Tang, Weilin Cong, Mengyue Hang, Kai Wang, Yajuan Wang, Zhichen Zeng, Ting-Wei Li, Hyunsik Yoo, Zhining Liu, Xuying Ning, Ruizhong Qiu, Wen-Yen Chen, Shuo Chang, Rong Jin, Huayu Li, and Hanghang Tong. 2026. Mixture of Sequence: Theme-Aware Mixture-of-Experts for Long-Sequence Recommendation. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792483

1 Introduction

Click-through rate (CTR) prediction [102, 105, 106] serves as a cornerstone in online advertising [77] and recommender systems [102, 105, 106], enabling users to efficiently discover items of interest from an ever-growing pool of products [25], videos [20, 104], and applications [17]. In recent years, modeling user behavior sequences has proven to be a highly effective strategy for enhancing CTR performance, as such sequences capture rich temporal dynamics that reflect the evolution of user interests. This advantage has spurred substantial research interest in sequence recommendation [9, 14, 34, 69]. Furthermore, extending the sequence length is particularly appealing, as it allows models to exploit long-term behavioral patterns. Motivated by this, recent studies [5, 7, 10, 85] have begun to investigate the feasibility of modeling very long sequences and have reported promising improvements.

Despite the advantages of leveraging long interaction sequences for CTR prediction, they introduce significant challenges. As user interests often undergo substantial shifts within long sequences [39, 41, 42], many interactions become irrelevant to the current prediction or even inject misleading signals. For instance, on an e-commerce platform [93], a user may simultaneously exhibit diverse interests, such as browsing both electronic devices and athletic equipment. An illustration is provided in Figure. 1. Ideally, when recommending an electronic product, the model should primarily focus on the subsequence associated with electronics while down-weighting interactions w.r.t. unrelated interests, such as athletic equipment. However, if the model indiscriminately incorporates

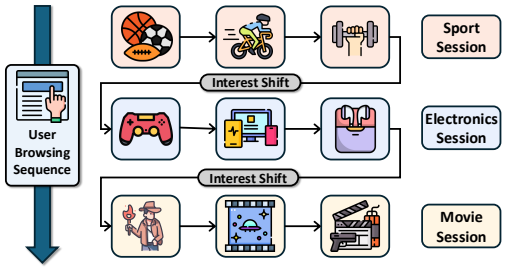


Figure 1: An illustration of user interests. The user undergoes two interest shifts, resulting in three clearly distinct sessions. Within each session, the browsing history reflects a consistent user interest.

the entire sequence, misleading interactions can dominate the representation and ultimately lead to erroneous predictions [12, 47]. This issue motivates us to raise a critical question for long-sequence modeling.

How could a recommendation model, by focusing on strategically chosen subsets of a long sequence, gain complementary perspectives that lead to more accurate predictions?

To address this question, we investigate the distinctive behavioral patterns embedded in long sequential data. Through a comprehensive empirical analysis, we identify a phenomenon we term *session hopping*: user interests remain highly consistent within a short temporal span (a "*session*"), yet may shift abruptly across adjacent sessions and reappear after several sessions. This phenomenon clearly demonstrates the sudden and recurring transitions in user preferences, highlighting the substantial challenges it poses for long-sequence recommendation. Existing methods often rely on attention [5, 10, 85], selective copying [53, 68] or similar mechanisms [75] to capture long-range dependencies, but these mechanisms are designed as continuous function approximators for smooth dynamics. As such, they struggle to fit the discontinuous signals from abrupt interest shifts and hence continue to exploit information from sessions prior to the shift. As illustrated in Figure 1, once an interest shift occurs, the items chosen by the user differ substantially from those in earlier sessions. Consequently, prior session information not only fails to provide useful guidance but may also mislead the prediction.

To address the challenges introduced by the session hopping phenomenon, we propose the Mixture of Sequence (MoS), a model-agnostic MoE framework that leverages subsequence extraction to selectively utilize informative sessions for accurate prediction. Intuitively, since similar sessions tend to reappear, these sessions can be grouped into subsequences that reflect specific themes of user behaviors (e.g., sports). By assigning a dedicated expert to each subsequence, MoS enables every expert to fully exploit informative sessions in the theme-aware subsequence while effectively filtering out misleading signals from other themes, thereby mitigating the negative effects of abrupt interest shifts. As a result, MoS serves as a versatile plug-in design that empowers existing sequential models to overcome the challenges posed by session hopping. Specifically, MoS consists of two key components: theme-aware routing and a multi-scale fusion mechanism. (1) The theme-aware routing mechanism introduces a router equipped with a codebook that defines user-behavior themes, allowing MoS to adaptively extract highly coherent, theme-specific subsequences from long interaction histories. (2) The multi-scale fusion mechanism mitigates the potential disruption of semantic and temporal continuity caused by subsequence extraction. It incorporates three types of experts that respectively capture global characteristics of the sequence, theme-specific patterns within subsequences, and short-term contiguous user behaviors. Together, these two components enable MoS to model user dynamics from multi-thematic and multi-scale perspectives, leading to more accurate sequential recommendations.

In summary, our main contributions are as follows:

- **Observation.** We empirically demonstrate the existence of the *session hopping* phenomenon on real-world data, which manifests as both abrupt shifts and reappearance of user interests, thereby highlighting the inherent challenges of modeling long-sequence recommendation.
- **Algorithm.** As a model-agnostic MoE method, MoS introduces a router that perform adaptive theme-aware subsequence extraction, and leverages three types of experts to jointly capture global, short-term, and semantic characteristics. These designs enable accurate recommendations even in the presence of highly noisy long-sequence data.
- **Experiment.** We conduct extensive experiments on three real-world datasets. The results show that MoS consistently improves performance across four sequence recommendation backbones, yielding an average gain of 0.68% in AUC and 0.72% in GAUC. MoS surpasses all other MoE methods while requiring fewer FLOPs, demonstrating an excellent balance between utility and efficiency.

2 Preliminary

2.1 Sequential Recommendation

Let $\mathcal{U}$ and $\mathcal{V}$ denote the sets of users and items, respectively, with $u \in \mathcal{U}$ and $v \in \mathcal{V}$ representing a specific user and item. We use $|\mathcal{U}|$ and $|\mathcal{V}|$ to denote the sizes of these sets. For each user u, we define the interaction history as a chronologically ordered sequence $\mathcal{S}_u = (v_{1,u}, v_{2,u}, \ldots, v_{t,u})$ where $v_{i,u}$ is the i-th interacted item and t is the sequence length for user u. For notational simplicity, we use v_i to denote $v_{i,u}$ and $\mathcal{S}$ to denote $\mathcal{S}_u$ in the following. Given a user interaction sequence $\mathcal{S}$, the objective of sequential recommendation is to predict the next item the user will interact with at time step $t+1$. Formally, the task can be expressed as:

$$v_{t+1}^{*} = \arg\max_{v \in \mathcal{V}} P(v|\mathcal{S}; \Theta) \tag{1}$$

where Θ denotes the model parameters.

2.2 Sparse Mixture of Experts

For a sparse MoE model with n experts $\{F^{(1)}, \ldots, F^{(n)}\}$, a learnable sparse router $G(\cdot)$ is leveraged to decide which experts are activated for a given item embedding $\boldsymbol{x}$. The model prediction is a weighted combination of expert outputs, i.e., $\boldsymbol{y} = \sum_{i=1}^{n} G(\boldsymbol{x})[i] F^{(n)}(\boldsymbol{x})$ where $G(\boldsymbol{x})[i]$ denotes the i-th element of $G(\boldsymbol{x})$, i.e., the routing weight assigned to expert F_i. In practice, $G(\cdot)$ is parameterized by a scoring network $H(\cdot)$ that produces relevance scores of experts. To enforce sparsity, only the top-k experts with the highest scores are selected:

$$G(\boldsymbol{x}) = \text{softmax}(\text{KeepTopK}(H(\boldsymbol{x}), k)) \tag{2}$$

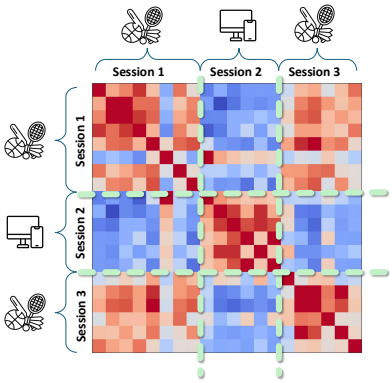


Figure 2: Heatmap of the self-similarity matrix for a representative user transaction history. Red indicates high similarity, blue indicates low similarity, and green lines denote session boundaries.

where KeepTopK$(\boldsymbol{v}, k)$ preserves the k largest elements of $\boldsymbol{v}$ and replaces the rest with $-\infty$.

3 Session Hopping

To better understand the challenges of long-sequence modeling, we conduct a case study on real-world datasets. Specifically, we compute the self-similarity matrix of user interaction history using cosine similarity. A representative heatmap of user behavior is shown in Figure 2, with more examples provided in Appendix A. From Figure 2, we can clearly identify a distinctive user behavior pattern, which we term the *session hopping* phenomenon. This phenomenon is characterized by three properties: (1) **Stability.** User interests remain highly consistent and stable within a short temporal span (a session), as indicated by the red area of Figure 2; (2) **Discontinuity.** User interests undergo abrupt and discontinuous shifts across sessions, reflected in the sharp red–blue boundary in Figure 2; (3) **Reappearance.** User interests may reappear after several sessions, as indicated by the strong similarity between the first and third sessions, both highlighted in red. This pattern closely aligns with everyday user behavior. For instance, when purchasing a computer, users often buy related accessories such as a mouse or keyboard within a short period of time. Once these needs are fulfilled, however, their interest in electronics typically drops, and attention shifts abruptly to unrelated categories. After some time, users may once again develop interest in electronics, such as upgrading to a better mouse. A clear illustration is provided in Figure 1. In summary, we formalize this phenomenon as follows:

Phenomenon 1. ***(Session Hopping)*** *User interests remain highly consistent within a session but differ substantially across adjacent sessions. Moreover, similar user interests may reappear across non-adjacent sessions.*

This phenomenon illustrates both the opportunities and challenges of long-sequence recommendation. On the one hand, the reappearance property of session hopping suggests that certain user behavior patterns tend to recur over time. These recurring patterns provide valuable cues for predicting the next item, and compared with short-sequence settings, long-sequence data offer richer opportunities to observe such recurring sessions, thereby holding substantial potential for improving predictive accuracy. On the other hand, the discontinuity inherent in session hopping introduces significant modeling challenges. Mainstream sequential recommendation models, such as RNN-based [75], Mamba-based [53, 68], and Transformer-based architectures [5, 10, 85], rely on continuous functions during forward propagation and thus lack mechanisms explicitly designed to capture abrupt, non-continuous shifts in user interests. When fitting such discontinuous signals with models designed smooth dynamics [6, 60], approximation errors naturally arise. In recommendation scenarios, when a sudden shift in user interest occurs, the content of preceding sessions may differ drastically from the current intent. As a result, earlier sessions not only fail to provide useful information but may even mislead the prediction. Taking Transformer-based models as an example, although attention mechanisms can partially down-weight irrelevant interactions, it is practically impossible for them to assign zero attention weights to all unrelated sessions preceding a shift. Consequently, the model is highly likely to attend to misleading contexts, thereby incorporating harmful information into recommendation decisions. In summary, current sequential recommendation models face a fundamental trade-off: leveraging as much informative historical context as possible while mitigating the adverse influence of misleading information in long sequences.

4 Methodology

To address the problem introduced by session hopping, we propose MoS framework, which adaptively selects highly coherent subsequences at multiple scales to provide diverse perspectives for accurate prediction. MoS is built upon two key mechanisms: the theme-aware routing mechanism in Section 4.1 and the multi-scale fusion mechanism in Section 4.2. Furthermore, we discuss the training paradigm of MoS in Section 4.3 and analyze its computational efficiency in Section 4.4. The pipeline of MoS is shown in Figure 3.

4.1 Theme-aware Routing

The session hopping phenomenon introduces substantial difficulties for long-sequence recommendation, and existing approaches struggle to precisely capture such user behavior patterns. However, we propose that a simple yet elegant solution is *subsequence extraction*. Since user interests remain relatively stable within a session and similar sessions often reappear, these sessions reflect a particular user interest or a combination of closely related interests. If we define such interests as a special "*theme*" (e.g., sports), then these sessions can be regarded as a subsequence aligned with that theme. Furthermore, if a long user sequence can be ideally decomposed into multiple subsequences according to themes, then temporally adjacent but semantically irrelevant sessions are automatically separated into different sequences, thereby effectively reducing the disruptive impact of sudden interest shifts.

The idea of subsequence extraction naturally aligns with the MoE framework. By assigning individual experts to model each subsequence, every expert is able to capture long-range dependencies across sessions in a theme-aware subsequence while effectively ignoring irrelevant information.

Model architecture. Based on the above intuition, we employ a router to adaptively extract multiple subsequences, each of which is then processed by a dedicated expert. Formally, given a router $G(\cdot)$ and a user sequence $\mathcal{S}$, the subsequence assigned to expert i is defined as $\mathcal{S}^{(i)} := (v|G(\boldsymbol{x}_v)[i] > 0, v \in \mathcal{S})$ with $\boldsymbol{x}_v$ being the embedding of the item v. This dispatch process needs to satisfy two key properties: (1) **Sparsity.** To avoid the influence of misleading

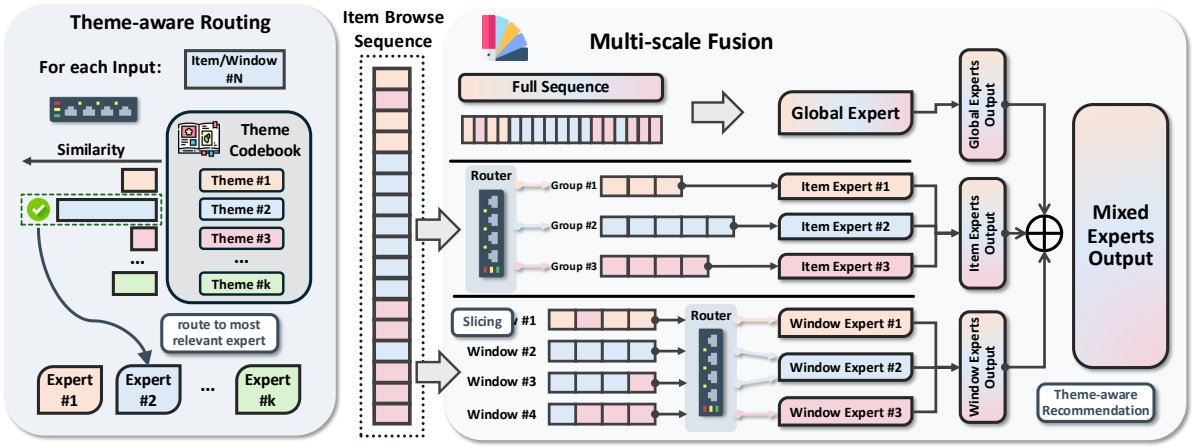


Figure 3: The pipeline of MoS. The left panel illustrates theme-aware routing, which assigns inputs to experts according to theme vectors in the codebook. The right panel shows multi-scale fusion, which models user behaviors by extracting subsequences at different granularities from the full sequence.

sessions, each expert should only have access to the corresponding subsequence instead of the entire sequence. Hence, a sparse MoE is leveraged by MoS rather than a dense MoE. (2) **Cohesion.** The subsequence assigned to an expert should be semantically coherent and aligned with a consistent theme, which imposes strong requirements on the router design. However, empirical evidence shows that conventional routers (e.g., MLPs or linear layers) fail to fulfill this requirement, as they tend to distribute items uniformly across experts rather than making theme-aware assignments. To address this limitation, we propose the theme-aware routing mechanism, which maintains a theme codebook $\boldsymbol{W} \in \mathbb{R}^{n \times D}$ with D being the dimension of the codebook and n being the number of experts. Here, the i-th row of the codebook describes the theme feature associated with the i-th expert. Given an item embedding $\boldsymbol{x}$, MoS first projects $\boldsymbol{x}$ into the theme space via a learnable MLP $h(\cdot)$, and then computes its cosine similarity with each entry of the codebook to obtain the relevance scores. Formally, the scoring network $H(\cdot)$ of the theme-aware router can be expressed as:

$$H(\boldsymbol{x}) = \left(\|h(\boldsymbol{x})\|_2 \sqrt{\mathrm{diag}(\boldsymbol{W}\boldsymbol{W}^T)} \right)^{-1} \boldsymbol{W} h(\boldsymbol{x}) \tag{3}$$

where $\mathrm{diag}(\cdot)$ preserves only diagonal elements while setting other elements to zero. Mathematically, equation 3 could easily ensure the following proposition with its proof provided in Appendix B.

Proposition 1. *(**Theme-aware Dispatch**) A subsequence with high internal similarity will be consistently routed to the same expert.*

Optimization. Although Proposition 1 theoretically guarantees the feasibility of the router, in practice the codebook $\boldsymbol{W}$ cannot be trained by gradient updates under joint optimization. This is because the codebook and the learnable MLP tend to collapse into a single MLP due to linearity during joint training, thereby undermining the intended role of the codebook. To address this problem, inspired by VQ-VAE [73], we update the codebook using an exponential moving average (EMA):

$$\boldsymbol{W}_i^{(t)} = \gamma \boldsymbol{W}_i^{(t-1)} + (1-\gamma) \frac{\sum_{\boldsymbol{x} \in \mathcal{B}} \mathbb{1}\left(G(\boldsymbol{x})[i] > 0\right) h(\boldsymbol{x})}{\sum_{\boldsymbol{x} \in \mathcal{B}} \mathbb{1}\left(G(\boldsymbol{x})[i] > 0\right)} \tag{4}$$

where $\mathcal{B}$ is a batch sampled from the dataset. In this way, although the item embeddings evolve throughout training, each entry of the codebook adaptively tracks the centroid of item embeddings under its corresponding theme, thereby maintaining stable and meaningful theme representations.

Initialization. A notable limitation of EMA is that if two entries are initialized too closely, they tend to collapse toward each other during the updating process, thereby reducing the diversity of the codebook. To mitigate this issue, it is essential to adopt an initialization strategy that ensures sufficient differentiation among themes. Specifically, we first estimate the distribution of item embeddings using a pretrained embedding layer, and then apply k-means clustering to obtain k cluster centroids, which are subsequently used as the initial values of rows in the codebook.

4.2 Multi-scale Fusion

A notable limitation of theme-aware routing is that subsequence extraction disrupts temporal dependencies. As a result, information about users' evolving preferences is largely ignored, and the model becomes biased toward recommending items that shows high semantic similarity rather than capturing dynamic behavioral shifts. This issue arises because the extraction process focuses solely on semantic coherence while neglecting temporal continuity. To remedy this, we introduce a multi-scale fusion mechanism that leverages a global expert, item experts, and window experts to capture temporal information at different granularities.

Global Expert. The objective of the global expert is to effectively extract global features from the sequence, thereby capturing users' globally drifted interests and modeling their evolving preferences. To this end, the global expert directly takes the entire historical sequence as input, without relying on subsequence extraction. When only the global expert is activated, the model naturally degenerates into the non-MoE backbone. Concretely, given any model with a repeated block structure, if we treat one block $F(\cdot)$ as an expert, the output of the global expert $\boldsymbol{y}_g$ can be expressed as:

$$\boldsymbol{y}_G = F\left(\mathcal{X}\left(\mathcal{S}\right)\right) \tag{5}$$

where $\mathcal{S} := (v_1, \ldots, v_t)$ denotes a user historical sequence and

$\mathcal{X}(\mathcal{S}) := (\boldsymbol{x}_{v_1}, \ldots, \boldsymbol{x}_{v_t})$ denotes the corresponding embedding sequence with $\boldsymbol{x}_{v_i}$ being the embedding of the item v_i.

Item Experts. The goal of the item experts is to deliver fine-grained and accurate recommendations by decomposing user behaviors into multiple latent themes and analyzing the theme-aware behavioral patterns. To this end, leveraging theme-aware routing mechanism, the item router $G_I(\cdot)$ first organizes the historical sequence $\mathcal{S}$ into a set of thematically coherent subsequences $\mathcal{S}_I^{(i)} := (v | G_I(\boldsymbol{x}_v)[i] > 0, v \in \mathcal{S})$. Then each item expert $F_I^{(i)}(\cdot)$ independently analyzes its subsequence to capture user behaviors from a specific theme. Finally, the outputs of all item experts are aggregated to form multi-faceted recommendation results. Mathematically, this process could be expressed as:

$$\boldsymbol{y}_I = \sum_{i=1}^{n} G_I\left(\boldsymbol{x}_{v_t}\right)[i] \cdot F_I^{(i)}\left(\mathcal{X}\left(\mathcal{S}_I^{(i)}\right)\right) \tag{6}$$

where $\boldsymbol{x}_{v_t}$ represents the embedding of the last item interacted with by the user u and $\boldsymbol{y}_I$ is the aggregated output by item experts.

Window Experts. Positioned between the global expert and the item experts, window experts jointly account for both semantic coherence and temporal continuity. These experts aim to capture short-term dynamics of user preferences while preserving theme-specific behavioral patterns. Concretely, we first transform the user's historical item sequence into a historical window sequence using a sliding window of size L and stride s. Mathematically, each window is defined as $w_m = (v_{s \cdot m+1}, \ldots, v_{s \cdot m+L})$, and the corresponding window embeddings is then computed as the average of the embeddings of the items it contains, i.e., $\boldsymbol{x}_{w_m} = \frac{1}{L} \sum_{j=s \cdot m+1}^{s \cdot m+L} \boldsymbol{x}_{v_j}$. Then, given this transformed window sequence $\mathcal{S}_W = (w_1, \ldots, w_{\tilde{t}})$, similarly, a window router $G^W(\cdot)$ applies the theme-aware routing mechanism to extract several theme-aware windows subsequences, $\mathcal{S}_W^{(i)} := (w | G_W(\boldsymbol{x}_w)[i] > 0, w \in \mathcal{S}_W)$, with each subsequence $\mathcal{S}_W^{(i)}$ assigned to a window expert $F_W^{(i)}(\cdot)$. Finally, their outputs are aggregated to obtain the final prediction:

$$\boldsymbol{y}_W = \sum_{i=1}^{n} G_W\left(\boldsymbol{x}_{w_{\tilde{t}}}\right)[i] \cdot F_W^{(i)}\left(\mathcal{X}\left(\mathcal{S}_W^{(i)}\right)\right) \tag{7}$$

Multi-scale Fusion. By integrating the outputs of the global, item, and window experts, MoS employs a multi-scale fusion strategy that effectively balances semantic coherence and temporal continuity in recommendation. When applied to sequential data exhibiting the session hopping phenomenon, MoS provides an elegant solution that enables accurate recommendations across multiple thematic perspectives and temporal granularities. Concretely, the final output of MoS is obtained as a weighted combination of the three expert groups:

$$\boldsymbol{y} = (1 - \alpha_I - \alpha_W)\boldsymbol{y}_G + \alpha_I \boldsymbol{y}_I + \alpha_W \boldsymbol{y}_W, \tag{8}$$

where α_I and α_W are two hyperparameters that control the relative importance of item and window experts in the final prediction.

4.3 Training Paradigm

Beyond model architecture design, a dedicated training paradigm is essential to ensure stable and effective optimization, particularly given the high training complexity of MoE models. To this end, we adopt a staged progressive training paradigm, which consists of three phases: backbone warm-up, theme-aware expert warm-up, and joint optimization.

Backbone Warm-up. An insightful observation in our model design is that when only the global expert is activated, the entire architecture degenerates into the backbone model without MoE. This implies that training solely the global expert substantially reduces the optimization difficulty, while guaranteeing performance at least comparable to the backbone. Motivated by this, in the backbone warm-up stage, MoS trains only the backbone components, including the global expert, embedding layer, and classifier head. Meanwhile, all the output from theme-aware experts are ignored, i.e., $\alpha_I = \alpha_W = 0$.

Theme-aware Expert Warm-up. In this stage, we update only the parameters of the item experts and window experts, while freezing the embedding layer and classifier head, and ignoring the output of the global expert, i.e., $\alpha_I = \alpha_W = 0.5$. Since the embedding layer and classifier head have already been well warmed up, this stage forces the item and window experts to align their outputs with those of the global expert, hence providing a strong initialization for the theme-aware experts.

Joint Optimization. In the final stage, we jointly fine-tune the entire model, updating all parameters based on the strong initialization provided by the previous phases. Here, α_I and α_W are treated as tunable hyperparameters to balance the contributions of item and window experts. This stage ensures global consistency across all components and maximizes the overall performance of the model.

4.4 Efficiency Analysis

As an MoE-based method, MoS not only achieves strong effectiveness but also introduces only a marginal increase in computational overhead, thereby ensuring efficiency. Given the prevalence of Transformer architectures, we adopt Transformer blocks as an example. Suppose MoS activates the top-k_1 experts out of n_1 item experts and the top-k_2 experts out of n_2 window experts. In this case, each item expert is responsible for approximately $O(\frac{N}{n_1})$ data points, while each window expert processes about $O(\frac{N}{s \cdot n_2})$ data points due to the slicing window, where N is the sequence length and s is the stride. Consequently, compared with the backbone, the additional time complexity introduced by MoS is $O(\frac{k_1^2 N^2}{n_1} + \frac{k_2^2 N^2}{s^2 \cdot n_2})$. Under the same conditions, this complexity is substantially lower than that of other MoE methods with a shared expert [21, 28, 44, 107], whose complexity is $O(\frac{(k_1+k_2)^2 N^2}{(n_1+n_2)})$. For instance, in our experimental setup with $k_1 = k_2$, $n_1 = n_2$, and $s = 4$, the theoretical time complexity of MoS is only about 53% of that of conventional MoE counterparts, highlighting its high efficiency.

5 Experiments

5.1 Experiment Settings

Dataset descriptions. We evaluate MoS on three real-world datasets: MicroVideo [13], KuaiVideo [49], and EBNeRD-Small [40]. The MicroVideo dataset focuses on short-video recommendation with multimodal image embeddings in the entertainment domain, KuaiVideo originates from the Kuaishou competition and models user–video interactions on online video platforms, and EBNeRD-Small targets personalized news recommendation in the digital publishing domain. Detailed dataset descriptions are provided in Appendix C.

Evaluation metrics. We evaluate model performance by ranking predictions over the entire item set without negative sampling. To measure utility, we adopt AUC and GAUC, while FLOPs are used to assess efficiency. For AUC and GAUC, higher values indicate better performance, whereas for FLOPs, lower values are preferred.

Models. We demonstrate the superiority of MoS from two perspectives: routing strategy and multi-scale fusion. From the routing perspective, we compare MoS with three MoE baselines, namely GShard [44], DSelect-k [28], and Expert Choice Routing [107], and leverage them on four different sequence recommendation models as backbones: Mamba4Rec [53], TransAct [85], TWIN [10], and SDIM [5]. These backbones cover both Transformer-based and Mamba-based architectures. From the multi-scale fusion perspective, we further compare MoS with long-sequence recommendation models that adopt similar fusion designs, including MIRRN [91], AttenMixer [97], and MiasRec [19].

Parameter Settings. Unless otherwise specified, we follow the default hyperparameter settings provided in the released code of the BARS Benchmark [108][1]. We adopt the Adam optimizer [37] to train all models, with the initial learning rate tuned from {1e-4, 5e-4, 1e-3}. For MoS, the codebook dimension is searched from $\{16, 32, 64\}$, and a 2-layer MLP is used as $h(\cdot)$. The EMA update weight γ is set to 0.999. For the window experts, we set the stride to 4 and the window size to 8. The multi-scale fusion weights are fixed at $\alpha_I = \alpha_W = 0.25$. For all MoE routers, the number of experts is set to 5, and the number of experts selected by the router k is tuned from $\{1, 2\}$. All experiments are conducted on NVIDIA A100 GPUs.

5.2 Experimental Results

Main results. The primary utility evaluation results are summarized in Table 1 and Table 2. (1) Table 1 compares MoS with other MoE methods, highlighting the superiority of its routing strategy. **Across diverse backbones, MoS consistently delivers overall performance gains on both AUC and GAUC**, achieving an average improvement of 0.68% in AUC and 0.72% in GAUC. Specifically, **MoS attains the top rank in AUC and GAUC across all datasets**, with average ranks of 1 / 1 / 1 / 1 for AUC and 1 / 1 / 1 / 1.33 for GAUC across the four backbones. In certain cases, the improvement is particularly pronounced. For instance, on the MicroVideo dataset, MoS improves the performance of Mamba4Rec by 2.91% in AUC and 0.85% in GAUC. (2) Table 2 further compares MoS with sequential recommendation models that adopt similar fusion designs. **Across all datasets, models enhanced with MoS consistently achieve both the first and second highest ranks, maintaining a clear performance margin over the remaining backbones.** On three datasets, the best-performing MoS variants surpass the strongest baseline by 2.39 (1.62), 1.68 (2.60), and 1.87 (1.70) in AUC (GAUC), respectively. These results collectively demonstrate the superiority of MoS in enhancing predictive performance. Importantly, this advantage is not tied to a specific architecture but generalizes robustly across heterogeneous backbones.

Efficiency Analysis. To evaluate the tradeoff between utility and efficiency of MoS, we compare it with all MoE methods using the same number of experts on the MicroVideo dataset across four backbones. The results are presented in Figure 4, where colors denote different backbones and shapes correspond to different MoE methods. As shown, MoS achieves a highly competitive tradeoff, with all of its cross markers positioned near the upper-left corner across all backbones. More concretely, **MoS consistently delivers the best utility (both AUC and GAUC)**, outperforming all baselines on every backbone. At the same time, **compared with all MoE methods, MoS achieves the lowest FLOPs**, indicating that its subsequence extraction mechanism is highly efficient. These findings collectively demonstrate the superiority of MoS in balancing performance and efficiency.

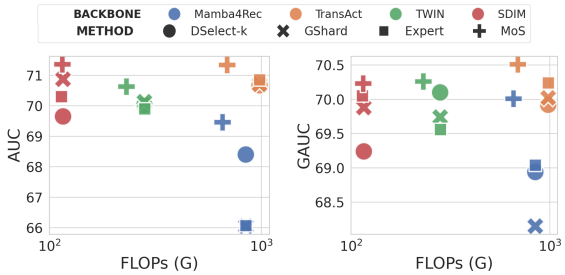


Figure 4: Trade-off between utility and efficiency. MoS achieves superior utility–efficiency balance, appearing closer to the upper-left region for every backbone (color).

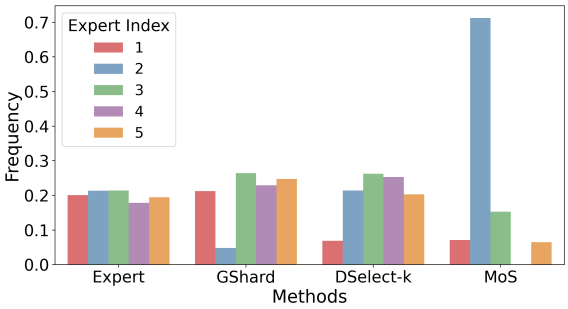


Figure 5: Dispatch behavior of routers for the most popular item. MoS consistently routes the popular item to a fixed expert, indicating strong consistency of theme-aware routing and clear expert specialization compared to other routers.

Routing Behavior Analysis. We analyze the dispatch behavior of routers from all MoE variants to understand why MoS consistently achieves superior performance on recommendation datasets. On the MicroVideo dataset, we select the most popular item, i.e., the item most frequently interacted with by all users, and then track how different MoE variants assign it to experts. Using the TransAct backbone, we record the expert selected for this item each time it appears and compute the dispatch frequency for each expert, as shown in Figure 5. The results reveal that MoS consistently routes the popular item to a fixed expert (here, expert #2), demonstrating strong **theme-aware consistency and clear expert specialization**. In contrast, other MoE variants tend to distribute the same item almost uniformly across all experts. This uniformity indicates insufficient differentiation among experts, which in turn leads to degraded recommendation performance. More detailed explanation and analyses of routing behaviors are provided in Appendix D.

[1] https://github.com/reczoo/LongCTR

Table 1: Main evaluation of model utility on sequential recommendation. Higher AUC and GAUC values indicate better performance. Utility performance as well as the average ranking are reported on four sequential recommendation backbones, comparing MoS with three MoE baselines and the vanilla backbone (without MoE enhancement). Red font highlights the best performance, while blue font denotes the second best. Numbers in parentheses indicate the performance gain brought by MoE.

Dataset		MicroVideo		KuaiVideo		Ebnerd		Rank	
Backbone	Method	AUC	GAUC	AUC	GAUC	AUC	GAUC	AUC	GAUC
Mamba4Rec	Vanilla	66.55	69.16	66.27	66.26	63.08	62.68	4.00	2.67
	DSelect-k	68.40(+1.85)	68.94(-0.22)	66.48(+0.21)	66.22(-0.04)	63.59(+0.51)	62.50(-0.18)	2.00	4.00
	GShard	66.05(-0.50)	68.15(-1.01)	65.89(-0.38)	65.74(-0.52)	63.26(+0.18)	63.79(+1.11)	4.67	4.33
	Expert	66.07(-0.48)	69.04(-0.12)	66.29(+0.02)	66.06(-0.20)	63.56(+0.48)	64.18(+1.50)	3.33	3.00
	MoS	69.46(+2.91)	70.01(+0.85)	66.67(+0.40)	66.67(+0.41)	64.07(+0.99)	64.54(+1.86)	**1.00**	**1.00**
TransAct	Vanilla	70.58	69.90	67.67	65.92	70.89	70.25	3.33	3.67
	DSelect-k	70.67(+0.09)	69.92(+0.02)	67.87(+0.20)	65.38(-0.54)	70.68(-0.21)	70.30(+0.05)	3.67	3.33
	GShard	70.70(+0.12)	70.02(+0.12)	67.62(-0.05)	65.42(-0.50)	70.69(-0.20)	70.08(-0.17)	3.67	3.67
	Expert	70.85(+0.27)	70.24(+0.34)	67.52(-0.15)	65.28(-0.64)	70.87(-0.02)	70.27(+0.02)	3.33	3.33
	MoS	71.34(+0.76)	70.51(+0.61)	67.90(+0.23)	68.10(+2.18)	71.06(+0.17)	70.71(+0.46)	**1.00**	**1.00**
TWIN	Vanilla	70.34	69.67	69.36	66.60	69.85	69.54	2.33	3.67
	DSelect-k	70.02(-0.32)	70.10(+0.43)	68.96(-0.40)	66.64(+0.04)	69.40(-0.45)	68.99(-0.55)	4.00	3.00
	GShard	70.13(-0.21)	69.74(+0.07)	68.84(-0.52)	66.62(+0.02)	70.10(+0.25)	69.58(+0.04)	3.33	2.67
	Expert	69.90(-0.44)	69.56(-0.11)	68.96(-0.40)	66.34(-0.26)	69.80(-0.05)	69.33(-0.21)	4.00	4.67
	MoS	71.11(+0.77)	70.26(+0.59)	69.62(+0.26)	66.89(+0.29)	70.30(+0.45)	69.81(+0.27)	**1.00**	**1.00**
SDIM	Vanilla	70.50	69.62	68.56	66.61	69.63	69.03	3.00	4.00
	DSelect-k	69.65(-0.85)	69.24(-0.38)	68.78(+0.22)	66.48(-0.13)	69.47(-0.16)	69.07(+0.04)	4.00	4.33
	GShard	70.87(+0.37)	69.88(+0.26)	68.90(+0.34)	66.52(-0.09)	69.47(-0.16)	69.07(+0.04)	2.33	3.33
	Expert	70.30(-0.20)	70.05(+0.43)	68.85(+0.29)	66.66(+0.05)	69.40(-0.23)	69.13(+0.10)	4.00	2.00
	MoS	71.36(+0.86)	70.23(+0.61)	69.38(+0.82)	66.75(+0.14)	69.49(-0.14)	69.40(+0.37)	**1.33**	**1.00**

Table 2: Performance comparison between MoS and baselines employing similar multi-scale fusion architectures.

	Method	MicroVideo		KuaiVideo		Ebnerd	
		AUC	GAUC	AUC	GAUC	AUC	GAUC
Baseline	MIRRN	67.17	68.74	67.94	65.5	69.19	69.01
	MiasRec	68.67	67.83	65.05	64.70	55.95	55.45
	AttenMixer	68.97	68.89	66.95	64.18	62.37	62.29
MoS	TWIN	70.83	70.26	69.62	66.89	66.89	69.81
	SDIM	71.36	70.23	69.38	66.75	69.49	69.49
	TransAct	71.34	70.51	67.90	68.10	71.06	70.71
	Mamba4Rec	69.46	70.01	66.67	66.67	64.07	64.54

5.3 Ablation Studies

Impact of multi-scale fusion. To examine the contributions of global, item, and window experts to prediction performance, we conduct an ablation study by varying α_I and α_W over 0, 0.2, 0.4, 0.6, 0.8, 1. The corresponding AUC and GAUC results are shown in Figure 7. From these results, two clear conclusions can be drawn:

(1) The multi-scale fusion mechanism effectively enhances recommendation performance. As shown in Figure 7, the heatmap corresponds to an upper-triangular region constrained by $\alpha_I + \alpha_W \leq 1$, where the vertices represent the use of single type of experts, the edges correspond to pairwise fusion of any two expert types, and the interior region denotes full fusion of all experts. It is evident that **full fusion consistently achieves the best results, while pairwise fusion outperforms using the single expert**. This observation suggests that although the global expert alone already provides strong representations for recommendations, the three types of experts capture complementary behavioral patterns and offer diverse analytical perspectives, thus leading to consistent performance gains when fused.

(2) The multi-scale fusion mechanism exhibits strong robustness. When all experts are fused, corresponding to the interior region of the triangle, the model performance remains consistently high without noticeable fluctuations. The AUC (GAUC) varies only within a narrow range from 70.94% (70.20%) to 71.20% (70.27%). The insensitivity of performance to α_I and α_W demonstrates the strong robustness of the multi-scale fusion mechanism.

Scaling on the sequence length. To assess the effectiveness of MoS in long-sequence recommendation, we vary the input sequence length on the MicroVideo dataset from 100 to 400 and set the number of experts as 5. As shown in Figure 6, both AUC and GAUC increase steadily with longer sequences across all four backbones. This trend confirms that MoS's strategy of decomposing long sequences into shorter, theme-coherent subsequences effectively alleviates the impact of irrelevant information and training difficulty, resulting in superior and stable performance.

Scaling on the number of experts. To examine the impact of the number of experts in MoS, we gradually increase the number of experts from 0 to 8 on the MicroVideo dataset with the sequence length

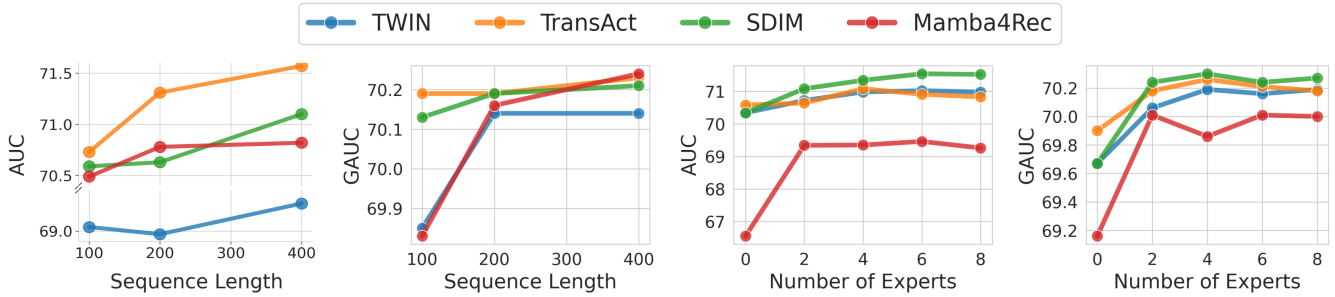
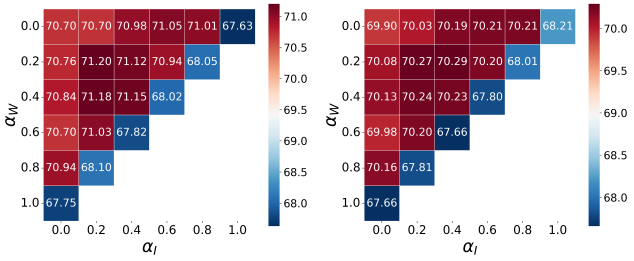


Figure 6: Scaling study of MoS. MoS consistently enhances AUC/GAUC with increased sequence length and number of experts.



(a) Impact of α_I and α_W on AUC. (b) Impact of α_I and α_W on GAUC.

Figure 7: Impact of α_I and α_W on model utility.

of 200. As shown in Figure 6, model performance, including both AUC and GAUC, improves as the number of experts increases, but the improvement trend slows markedly once the number exceeds four. This phenomenon may be related to the underlying number of latent themes. When the number of experts is small, adding more experts effectively introduces additional themes, which helps to separate unrelated items or sessions and thereby enhances model performance. However, when the number of experts becomes sufficiently large, further increases lead to overly fine-grained distinctions between themes, causing many informative items to be split across different experts. As a result, the effective information within subsequences diminishes, producing performance saturation while increasing computational cost. These findings suggest that an appropriate number of experts yields beneficial effects for MoS.

6 Related Work

6.1 Click-Through Rate Prediction

CTR [16, 26, 90] prediction is a fundamental task in online advertising, aiming to estimate the probability that a user clicks on a given item or ad. A dominant paradigm is feature-interaction-based modeling, which typically follows the embedding–interaction–prediction pipeline. Early works [65] capture low-order interactions, while Wide&Deep [16] and DeepFM [26] extend to higher-order with deep neural networks. More advanced designs include cross networks [46], graph-based [57] and attention-based models [67, 90].

6.2 Sequential Recommendation

Sequential recommendation [30, 35] aims to predict the next item a user will interact with based on their historical behavior sequence. Advances in deep learning have driven the development of diverse models that capture sequential dependencies and uncover latent user interest from historical interactions. Representative approaches include Convolutional Neural Networks (CNNs) [72], Recurrent Neural Networks (RNNs) [31, 96], Transformer-based models [36, 70, 100], and Graph Neural Networks (GNNs) [27, 84, 92], all of which have been widely adopted to improve recommendation performance. In addition, self-supervised learning has become a key paradigm in sequential recommendation [15, 64, 86, 99], where contrastive objectives are applied to improve representation learning from unlabeled sequences. More recently, research on long-term user behavior modeling [10, 53, 54, 85] has focused on improving efficiency and scalability under strict latency constraints. Representative directions include sampling-based hashing methods like SDIM [5], architecture innovations such as Mamba4Rec [53], industrial hybrid solutions like TransAct [85], and life-long user modeling approaches such as TWIN [10], which advance the balance between effectiveness and efficiency in long-term sequential modeling. Our method provides a plug-and-play solution that can be integrated into existing long-sequence recommendation models.

7 Conclusion

In this paper, we identify the session hopping phenomenon, which characterizes the stability, discontinuity, and recurrence of user interests—revealing the intrinsic challenges of long-sequence modeling. To address this issue, we propose MoS, a framework that extracts multi-scale, theme-consistent subsequences for accurate prediction. MoS introduces a *theme-aware router* that learns latent user-interest themes via a special-designed codebook and extracts theme-specific subsequences accordingly. It further employs a *multi-scale fusion mechanism* with three types of experts to capture global, short-term, and semantic representations for improved prediction. Extensive experiments show that MoS consistently enhances existing long-sequence recommendation models, outperforming other MoE approaches while requiring fewer FLOPs, demonstrating an excellent balance between utility and efficiency.

Acknowledgment

This work is supported by NSF (2433308) and IBM-Illinois Discovery Accelerator Institute. The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. We would like to thank Zhichen Tang, Weilin Cong, Mengyue Hang, Kai Wang, Yajuan Wang, Wen-Yen Chen, Shuo Chang, Rong Jin, and Huayu Li for their management support.

References

[1] Mengting Ai, Tianxin Wei, Yifan Chen, Zhichen Zeng, et al. 2025. ResMoE: Space-efficient Compression of Mixture of Experts LLMs via Residual Restoration. In *Proceedings of the 31st ACM SIGKDD Conference*. 1–12.

[2] Yuanchen Bei, Weizhi Zhang, Siwen Wang, Weizhi Chen, Sheng Zhou, Hao Chen, Yong Li, et al. 2025. Graphs Meet AI Agents: Taxonomy, Progress, and Future Opportunities. *arXiv preprint arXiv:2506.18019* (2025).

[3] Shuqing Bian, Xingyu Pan, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, and Ji-Rong Wen. 2023. Multi-modal mixture of experts represetation learning for sequential recommendation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 110–119.

[4] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. 2023. LightGCL: Simple yet effective graph contrastive learning for recommendation. *arXiv preprint arXiv:2302.08191* (2023).

[5] Yue Cao, Xiaojiang Zhou, Jiaqi Feng, Peihao Huang, Yao Xiao, Dayao Chen, and Sheng Chen. 2022. Sampling is all you need on modeling long-term user behaviors for CTR prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2974–2983.

[6] Valérie Castin, Pierre Ablin, and Gabriel Peyré. 2023. How smooth is attention? *arXiv preprint arXiv:2312.14820* (2023).

[7] Zheng Chai, Qin Ren, Xijun Xiao, Huizhi Yang, et al. 2025. Longer: Scaling up long sequence modeling in industrial recommenders. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 247–256.

[8] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Bundle recommendation with graph convolutional networks. In *Proceedings of the 43rd international ACM SIGIR conference*. 1673–1676.

[9] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference*. 378–387.

[10] Jianxin Chang, Chenbin Zhang, Zhiyi Fu, Xiaoxue Zang, et al. 2023. TWIN: TWo-stage interest network for lifelong user behavior modeling in CTR prediction at kuaishou. In *Proceedings of the 29th ACM SIGIR conferenceACM SIGKDD Conference*. 3785–3794.

[11] Hao Chen, Yuanchen Bei, Qijie Shen, Yue Xu, Sheng Zhou, et al. 2024. Macro graph neural networks for online billion-scale recommender systems. In *Proceedings of the ACM web conference 2024*. 3598–3608.

[12] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, et al. 2021. Curriculum disentangled recommendation with noisy multi-feedback. *NeurIPS* 34 (2021), 26924–26936.

[13] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. 2018. Temporal hierarchical attention at category-and item-level for micro-video click-through prediction. In *Proceedings of the 26th ACM international conference on Multimedia*. 1146–1153.

[14] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Sequential recommendation with user memory networks. In *Proceedings of the eleventh ACM WSDM conference*. 108–116.

[15] Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. 2022. Intent contrastive learning for sequential recommendation. In *Proceedings of the ACM Web Conference 2022*. 2172–2182.

[16] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.

[17] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.

[18] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, et al. 2022. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems* 35 (2022), 34600–34613.

[19] Minjin Choi, Hye-young Kim, Hyunsouk Cho, and Jongwuk Lee. 2024. Multi-intent-aware Session-based Recommendation. *arXiv preprint arXiv:2405.00986* (2024).

[20] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.

[21] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066* (2024).

[22] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*. PMLR, 5547–5569.

[23] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The world wide web conference*. 417–426.

[24] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

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

[26] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[27] Jiayan Guo, Yaming Yang, Xiangchen Song, Yuan Zhang, Yujing Wang, Jing Bai, and Yan Zhang. 2022. Learning multi-granularity consecutive user intent unit for session-based recommendation. In *Proceedings of the fifteenth ACM WSDM conference*. 343–352.

[28] Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, et al. 2021. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *NeurIPS* 34 (2021), 29335–29347.

[29] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference*. 639–648.

[30] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[31] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[32] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGIR conferenceACM SIGKDD Conference*. 585–593.

[33] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.

[34] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[35] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[36] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[37] Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[38] TN Kipf. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[39] Ivan Koychev and Ingo Schwab. 2000. Adaptation to drifting user's interests. In *Proceedings of ECML2000 workshop: machine learning in new information age*. 39–46.

[40] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, et al. 2024. EB-NeRD a large-scale dataset for news recommendation. In *Proceedings of the Recommender Systems Challenge 2024*. 1–11.

[41] Wai Lam and Javed Mostafa. 2001. Modeling user interest shift using a bayesian approach. *Journal of the American society for Information Science and Technology* 52, 5 (2001), 416–429.

[42] Wai Lam, Snehasis Mukhopadhyay, Javed Mostafa, and Mathew Palakal. 1996. Detection of shifts in user interests for personalized information filtering. In *Proceedings of the 19th annual international ACM SIGIR conference*. 317–325.

[43] Gyuseok Lee, Yaokun Liu, Yifan Liu, Susik Yoon, et al. 2025. Session-Based Recommendation with Validated and Enriched LLM Intents. *arXiv preprint arXiv:2508.00570* (2025).

[44] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668* (2020).

[45] Dingcheng Li, Xu Li, Jun Wang, and Ping Li. 2020. Video recommendation with multi-gate mixture of experts soft actor critic. In *Proceedings of the 43rd International ACM SIGIR conference*. 1553–1556.

[46] Honghao Li, Yiwen Zhang, Yi Zhang, Hanwei Li, Lei Sang, and Jieming Zhu. 2024. FCN: Fusing Exponential and Linear Cross Network for Click-Through Rate Prediction. *arXiv preprint arXiv:2407.13349* (2024).

[47] Haoxuan Li, Chunyuan Zheng, Wenjie Wang, Hao Wang, Fuli Feng, and Xiao-Hua Zhou. 2024. Debiased recommendation with noisy feedback. In *Proceedings of the 30th ACM SIGIR conferenceACM SIGKDD Conference*. 1576–1586.

[48] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1419–1428.

[49] Yongqi Li, Meng Liu, Jianhua Yin, Chaoran Cui, Xin-Shun Xu, and Liqiang Nie. 2019. Routing micro-videos via a temporal graph-guided recommendation system. In *Proceedings of the 27th ACM international conference on multimedia*. 1464–1472.

[50] Xiao Lin, Jian Kang, Weilin Cong, and Hanghang Tong. 2024. Bemap: Balanced message passing for fair graph neural network. In *Learning on Graphs Conference*. PMLR, 37–1.

[51] Xiao Lin, Zhining Liu, Dongqi Fu, Ruizhong Qiu, and Hanghang Tong. 2024. Backtime: Backdoor attacks on multivariate time series forecasting. *Advances in Neural Information Processing Systems* 37 (2024), 131344–131368.

[52] Xiao Lin, Zhichen Zeng, Tianxin Wei, Zhining Liu, et al. 2025. Cats: Mitigating correlation shift for multivariate time series classification. *arXiv preprint arXiv:2504.04283* (2025).

[53] Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. 2024. Mamba4rec: Towards efficient sequential recommendation with selective state space models. *arXiv preprint arXiv:2403.03900* (2024).

[54] Xin Liu, Zheng Li, Yifan Gao, Jingfeng Yang, Tianyu Cao, et al. 2024. Enhancing User Intent Capture in Session-Based Recommendation with Attribute Patterns. *Advances in Neural Information Processing Systems* 36 (2024).

[55] Jiaqi Ma, Zhe Zhao, Jilin Chen, Ang Li, Lichan Hong, and Ed H Chi. 2019. Snr: Sub-network routing for flexible parameter sharing in multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 216–223.

[56] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference*. 1930–1939.

[57] Xuan Ma, Hao Peng, Jia Duan, Zhanhao Ye, Langlang Ye, Zehua Zhang, Jie He, Changping Peng, and Zhangang Lin. 2025. Graph Isomorphism Network-Based Cohort Modeling In Click-Through Rate Prediction. In *Proceedings of the 48th International ACM SIGIR conference*. 4219–4223.

[58] Kelong Mao, Jieming Zhu, Xi Xiao, Biao Lu, Zhaowei Wang, and Xiuqiang He. 2021. UltraGCN: ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 1253–1262.

[59] Xuying Ning, Dongqi Fu, Tianxin Wei, Wujiang Xu, and Jingrui He. 2025. Graph4MM: Weaving multimodal learning with structural information. *arXiv preprint arXiv:2510.16990* (2025).

[60] Naoki Nishikawa and Taiji Suzuki. 2024. State Space Models are Comparable to Transformers in Estimating Functions with Dynamic Smoothness. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

[61] Xinglin Pan, Wenxiang Lin, Lin Zhang, Shaohuai Shi, et al. 2025. Fsmoe: A flexible and scalable training system for sparse mixture-of-experts models. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*. 524–539.

[62] Shutong Qiao, Wei Zhou, Junhao Wen, Chen Gao, et al. 2025. Multi-view Intent Learning and Alignment with Large Language Models for Session-based Recommendation. *ACM Transactions on Information Systems* 43, 4 (2025), 1–25.

[63] Zhen Qin, Yicheng Cheng, Zhe Zhao, Zhe Chen, Donald Metzler, and Jingzheng Qin. 2020. Multitask mixture of sequential experts for user activity streams. In *Proceedings of the 26th ACM SIGKDD international conference*. 3083–3091.

[64] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM WSDM conference*. 813–823.

[65] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.

[66] Jinghan Shi, Houye Ji, Chuan Shi, Xiao Wang, Zhiqiang Zhang, and Jun Zhou. 2020. Heterogeneous graph neural network for recommendation. *arXiv preprint arXiv:2009.00799* (2020).

[67] Qingquan Song, Dehua Cheng, Hanning Zhou, Jiyan Yang, Yuandong Tian, and Xia Hu. 2020. Towards automated neural interaction discovery for click-through rate prediction. In *Proceedings of the 26th ACM SIGKDD international conference*. 945–955.

[68] Jinzhao Su and Zhenhua Huang. 2024. Mlsa4rec: Mamba combined with low-rank decomposed self-attention for sequential recommendation. *arXiv preprint arXiv:2407.13135* (2024).

[69] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[70] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[71] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM conference on recommender systems*. 269–278.

[72] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM WSDM conference*. 565–573.

[73] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *NeurIPS* 30 (2017).

[74] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, et al. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[75] Kiewan Villatel, Elena Smirnova, Jérémie Mary, and Philippe Preux. 2018. Recurrent neural networks for long and short-term sequential recommendation. *arXiv preprint arXiv:1807.09142* (2018).

[76] Jianling Wang, Kaize Ding, Ziwei Zhu, and James Caverlee. 2021. Session-based recommendation with hypergraph attention networks. In *Proceedings of the 2021 SIAM international conference on data mining (SDM)*. SIAM, 82–90.

[77] Xinfei Wang. 2020. A survey of online advertising click-through rate prediction models. In *2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, Vol. 1. IEEE, 516–521.

[78] Yifan Wang, Suyao Tang, Yuntong Lei, et al. 2020. Disenhan: Disentangled heterogeneous graph attention network for recommendation. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1605–1614.

[79] Tianxin Wei, Bowen Jin, Ruirui Li, Hansi Zeng, Zhengyang Wang, Jianhui Sun, Qingyu Yin, et al. [n. d.]. Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Generative Recommendation and Beyond. In *The Twelfth International Conference on Learning Representations*.

[80] Tianxin Wei, Xuying Ning, Xuxing Chen, et al. 2025. Coarse-to-Fine Tokenization for Generative Recommendation. *arXiv preprint arXiv:2511.22707* (2025).

[81] Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, et al. 2025. Evo-Memory: Benchmarking LLM Agent Test-time Learning with Self-Evolving Memory. *arXiv preprint arXiv:2511.20857* (2025).

[82] Tianxin Wei, Ziwei Wu, Ruirui Li, Ziniu Hu, et al. 2020. Fast adaptation for cold-start collaborative filtering with meta-learning. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 661–670.

[83] Tianxin Wei, Yuning You, Tianlong Chen, Yang Shen, Jingrui He, and Zhangyang Wang. 2022. Augmentations in hypergraph contrastive learning: Fabricated and generative. *Advances in neural information processing systems* 35 (2022), 1909–1922.

[84] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 346–353.

[85] Xue Xia, Pong Eksombatchai, Nikil Pancha, et al. 2023. Transact: Transformer-based realtime user action model for recommendation at pinterest. In *Proceedings of the 29th ACM SIGIR conferenceACM SIGKDD Conference*. 5249–5259.

[86] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, et al. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.

[87] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, et al. 2019. Graph contextualized self-attention network for session-based recommendation.. In *IJCAI*, Vol. 19. 3940–3946.

[88] Wujiang Xu, Xuying Ning, Wenfang Lin, Mingming Ha, et al. 2024. Towards open-world cross-domain sequential recommendation: A model-agnostic contrastive denoising approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 161–179.

[89] Wujiang Xu, Yunxiao Shi, Zujie Liang, et al. 2025. Instructagent: Building user controllable recommender via llm agent. *arXiv preprint arXiv:2502.14662* (2025).

[90] Xiang Xu, Hao Wang, Wei Guo, Luankang Zhang, et al. 2025. Multi-granularity interest retrieval and refinement network for long-term user behavior modeling in ctr prediction. In *Proceedings of the 31st ACM SIGIR conferenceACM SIGKDD Conference V. 1*. 2745–2755.

[91] Xiang Xu, Hao Wang, Wei Guo, Luankang Zhang, Wanshan Yang, et al. 2025. Multi-granularity interest retrieval and refinement network for long-term user behavior modeling in ctr prediction. In *Proceedings of the 31st ACM SIGIR conferenceACM SIGKDD Conference V. 1*. 2745–2755.

[92] Yuhao Yang, Chao Huang, Lianghao Xia, Chunzhen Huang, Da Luo, and Kangyi Lin. 2023. Debiased contrastive learning for sequential recommendation. In *Proceedings of the ACM web conference 2023*. 1063–1073.

[93] Dingyu Ye, Fufan Liu, Dongmin Cho, and Zhengzhi Jia. 2022. Investigating switching intention of e-commerce live streaming users. *Heliyon* 8, 10 (2022).

[94] Hyunsik Yoo, Ruizhong Qiu, Charlie Xu, Fei Wang, and Hanghang Tong. 2025. Generalizable Recommender System During Temporal Popularity Distribution Shifts. In *Proceedings of the 31st ACM SIGIR conferenceACM SIGKDD Conference*. 1833–1843.

[95] Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. 2023. XSimGCL: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering* 36, 2 (2023), 913–926.

[96] Zhenrui Yue, Yueqi Wang, Zhankui He, Huimin Zeng, Julian McAuley, and Dong Wang. 2024. Linear recurrent units for sequential recommendation. In *Proceedings of the 17th ACM WSDM conference*. 930–938.

[97] Peiyan Zhang, Jiayan Guo, Chaozhuo Li, Yueqi Xie, Jae Boum Kim, Yan Zhang, Xing Xie, Haohan Wang, and Sunghun Kim. 2023. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *Proceedings of the sixteenth ACM WSDM conference*. 168–176.

[98] Shengzhe Zhang, Liyi Chen, Dazhong Shen, Chao Wang, and Hui Xiong. 2025. Hierarchical Time-Aware Mixture of Experts for Multi-Modal Sequential Recommendation. In *Proceedings of the ACM on Web Conference 2025*. 3672–3682.
[99] Shengyu Zhang, Lingxiao Yang, Dong Yao, Yujie Lu, Fuli Feng, et al. 2022. Re4: Learning to re-contrast, re-attend, re-construct for multi-interest recommendation. In *Proceedings of the ACM Web Conference 2022*. 2216–2226.
[100] Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, et al. 2019. Feature-level deeper self-attention network for sequential recommendation.. In *IJCAI*. 4320–4326.
[101] Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, et al. 2025. Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap. *arXiv preprint arXiv:2501.01945* (2025).
[102] Weinan Zhang, Jiarui Qin, Wei Guo, Ruiming Tang, and Xiuqiang He. 2021. Deep learning for click-through rate estimation. *arXiv preprint arXiv:2104.10584* (2021).
[103] Yijie Zhang, Yuanchen Bei, Hao Chen, Qijie Shen, Zheng Yuan, Huan Gong, Senzhang Wang, Feiran Huang, and Xiao Huang. 2024. Multi-behavior collaborative filtering with partial order graph convolutional networks. In *Proceedings of the 30th ACM SIGIR conferenceACM SIGKDD Conference*. 6257–6268.
[104] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, et al. 2019. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM conference on recommender systems*. 43–51.
[105] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.
[106] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, et al. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference*. 1059–1068.
[107] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. 2022. Mixture-of-experts with expert choice routing. *NeurIPS* 35 (2022), 7103–7114.
[108] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. 2022. Bars: Towards open benchmarking for recommender systems. In *SIGIR*. 2912–2923.

Appendix

A Examples of Session Hopping

To analyze user behavior patterns, we first use a pretrained embedding layer to obtain the embedding of each item in a user's interaction history. The, we compute a self-similarity matrix of every user interaction sequence based on cosine similarity. A careful examination of these matrices reveals a widely recurring pattern, referred to in the main text as *session hopping*. Representative cases are shown in Figure 8. These examples consistently exhibit stability of interests within sessions, discontinuity across sessions, and the reappearance of interests over time, which together corroborate the prevalence of session hopping.

B Proof of Theme-aware Dispatch

Setup. Let $\boldsymbol{W} \in \mathbb{R}^{n\times d}$ with rows $\boldsymbol{w}_i^\top$ and define $\boldsymbol{a}_i := \boldsymbol{w}_i/\|\boldsymbol{w}_i\|_2$ for $i = 1,\dots,n$. For any input $\boldsymbol{x}$, suppose $h(\boldsymbol{x}) \neq \boldsymbol{0}$ and its direction $\boldsymbol{u}(\boldsymbol{x}) := h(\boldsymbol{x})/\|h(\boldsymbol{x})\|_2 \in \mathbb{S}^{d-1}$. Under the scoring in Eq. (1), each coordinate of $H(\boldsymbol{x})$ equals the cosine similarity

$$H_i(\boldsymbol{x}) = \frac{\boldsymbol{w}_i^\top h(\boldsymbol{x})}{\|\boldsymbol{w}_i\|_2\,\|h(\boldsymbol{x})\|_2} = \boldsymbol{a}_i^\top \boldsymbol{u}(\boldsymbol{x}), \qquad i = 1,\dots,E.$$

Let $H_{(1)}(\boldsymbol{x}) \geq \cdots \geq H_{(n)}(\boldsymbol{x})$ be the sorted scores and let $S_k(\boldsymbol{x})$ be the index set of the top-k entries of $H(\boldsymbol{x})$. Easy to prove, the support of $G(\boldsymbol{x})$ coincides with $S_k(\boldsymbol{x})$.

Theorem B.1 (Routing stability under cosine proximity). *Fix $k \in \{1,\dots,n\}$ and two inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$. Assume: (i) the embedding directions are close in cosine,*

$$\rho := \boldsymbol{u}(\boldsymbol{x}_1)^\top \boldsymbol{u}(\boldsymbol{x}_2) \;\geq\; 1-\delta \qquad \textit{for some } \delta \in (0,1);$$

(ii) there is a strictly positive top-k margin at $\boldsymbol{x}_1$,

$$\gamma_k(\boldsymbol{x}_1) := H_{(k)}(\boldsymbol{x}_1) - H_{(k+1)}(\boldsymbol{x}_1) > 0.$$

If $\delta < \gamma_k(\boldsymbol{x}_1)^2/8$, then the selected experts are identical, *e.g., $S_k(\boldsymbol{x}_2) = S_k(\boldsymbol{x}_1)$, , hence* $\mathrm{supp}G(\boldsymbol{x}_2) = \mathrm{supp}G(\boldsymbol{x}_1)$.

Proof. Set $\boldsymbol{u}_1 := \boldsymbol{u}(\boldsymbol{x}_1)$ and $\boldsymbol{u}_2 := \boldsymbol{u}(\boldsymbol{x}_2)$. By unit-row normalization, each score is linear in $\boldsymbol{u}$ with unit coefficient norm: for all i, $H_i(\boldsymbol{x}) = \boldsymbol{a}_i^\top \boldsymbol{u}$ and $\|\boldsymbol{a}_i\|_2 = 1$. Thus the coordinates of H are 1-Lipschitz in $\boldsymbol{u}$:

$$\big|H_i(\boldsymbol{x}_1) - H_i(\boldsymbol{x}_2)\big| = \big|\boldsymbol{a}_i^\top(\boldsymbol{u}_1 - \boldsymbol{u}_2)\big| \leq \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2 =: \varepsilon \qquad \forall i.$$

The cosine condition gives a bound on ε:

$$\varepsilon = \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2 = \sqrt{2 - 2\,\boldsymbol{u}_1^\top \boldsymbol{u}_2} \leq \sqrt{2\delta}.$$

Pick any $i \in S_k(\boldsymbol{x}_1)$ and any $j \notin S_k(\boldsymbol{x}_1)$. By the above coordinate-wise Lipschitz bound,

$$H_i(\boldsymbol{x}_2) \;\geq\; H_i(\boldsymbol{x}_1) - \varepsilon, \qquad H_j(\boldsymbol{x}_2) \;\leq\; H_j(\boldsymbol{x}_1) + \varepsilon,$$

so $H_i(\boldsymbol{x}_2) - H_j(\boldsymbol{x}_2) \;\geq\; \big(H_i(\boldsymbol{x}_1) - H_j(\boldsymbol{x}_1)\big) - 2\varepsilon \;\geq\; \gamma_k(\boldsymbol{x}_1) - 2\varepsilon$. If $\gamma_k(\boldsymbol{x}_1) > 2\varepsilon$, then $H_i(\boldsymbol{x}_2) > H_j(\boldsymbol{x}_2)$ for every such pair (i,j), which preserves the entire top-k ordering relation and hence the top-k index set: $S_k(\boldsymbol{x}_2) = S_k(\boldsymbol{x}_1)$. Since $\varepsilon \leq \sqrt{2\delta}$, the sufficient condition $\gamma_k(\boldsymbol{x}_1) > 2\varepsilon$ holds whenever $\delta < \frac{\gamma_k(\boldsymbol{x}_1)^2}{8}$, which proves the claim. Finally, $\mathrm{supp}G(\boldsymbol{x}) = S_k(\boldsymbol{x})$ by construction of KeepTopK and softmax, hence the selected experts coincide as stated. □

Remarks. In MoS, $h(\cdot)$ is implemented as an MLP that satisfies Lipschitz continuity. The Lipschitz continuity of $h(\cdot)$ guarantees that small input perturbations induce small changes of the embedding direction and therefore of the cosine term in (ii). The stability threshold in Theorem B.1 is explicit: the larger the top-k margin $\gamma_k(\boldsymbol{x}_1)$, the larger the admissible δ. Moreover, if any two items within a sequence exhibit high similarity, then by Theorem B.1, all such items will be routed consistently to the same expert, thereby completing the proof of Proposition 1.

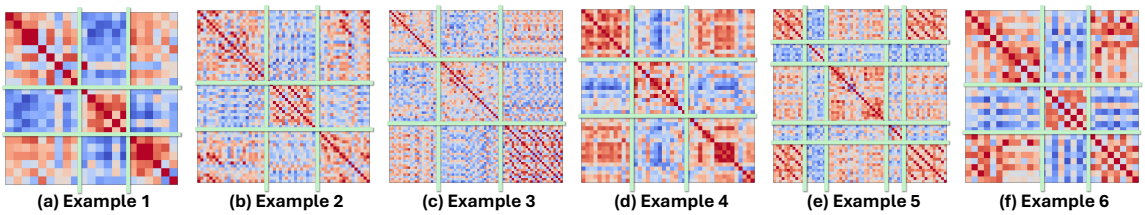
C Dataset Descriptions

We evaluate MoS on three public datasets:

- **MicroVideo** is released by the THACIL project and contains 12.7M user–item interactions from 10,986 users over 1.7M micro-videos. Available features include user and item IDs, categories, and visual embeddings.
- **KuaiVideo**, provided by the ChinaMM 2018 Kuaishou Challenge, focuses on CTR prediction for micro-videos and includes multiple interaction types such as click, non-click, and follow.
- **EBNeRD** (Ekstra Bladet News Recommendation Dataset) consists of over 37M impression logs from more than one million users, along with 125K Danish news articles with rich textual and categorical metadata.

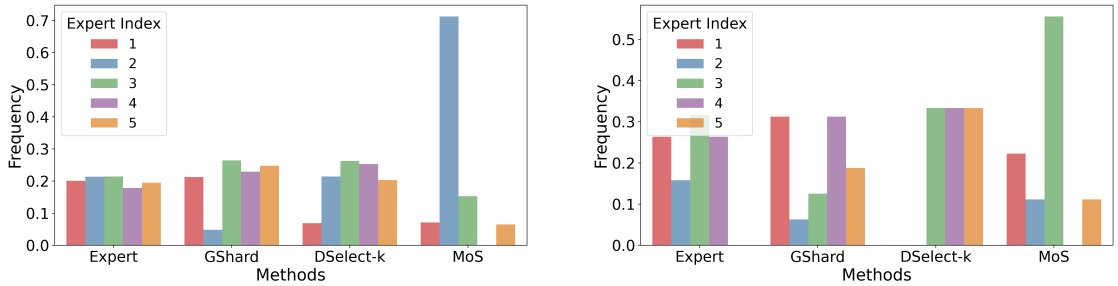
D Routing Behavior Analysis

To better understand why MoS outperforms other MoE methods, we conduct a fine-grained analysis of the router's dispatch behavior at the item level. Specifically, on the MicroVideo dataset, many items are repeatedly selected by different users. We first filter out items selected fewer than eight times, and then identify two representative targets: the most frequently selected item and the least frequently selected item. For each target item, we record the expert



(a) Example 1 (b) Example 2 (c) Example 3 (d) Example 4 (e) Example 5 (f) Example 6

Figure 8: Examples of Session Hopping



(a) Router dispatch behavior for the most popular item. **(b) Router dispatch behavior for the least popular item.**

Figure 9: Comparison between dispatch behavior of different MoE routers.

assigned by the router at every selection and compute the frequency with which each expert is chosen. All MoE variants are configured with five experts. Figure 9a and Figure 9b illustrate the dispatch behavior for the most popular and least popular items, respectively.

From these results, several observations can be made. (1) MoS consistently routes the same item to a fixed expert, effectively leveraging the theme-aware specialization of experts. In contrast, other MoE methods tend to distribute items uniformly across experts, reflecting weak expert differentiation. (2) When handling the most popular and the least popular items, MoS relies on distinct experts. Specifically, expert #2 is responsible for the most popular item, whereas expert #3 dominates for the least popular one. Based on these two observations, we safely draw the conclusion that **the consistency of dispatch behavior and the specialization of experts are central to the superior performance of MoS**.

E More Related Works

Mixture of Experts in Recommendation. The Mixture-of-Experts architecture [33] was originally introduced to improve supervised learning by combining multiple specialized models through a gating mechanism [1]. MoE has been widely adopted in recommendation systems [3, 45, 63, 98] particularly in multi-task learning scenarios. For instance, MMOE [56] introduces a multi-gate mechanism to balance multiple objectives, SNR [55] improves the flexibility of parameter sharing via sub-network routing, and PLE [71] proposes progressive layered extraction to mitigate the seesaw effect in multi-task recommendation. MoE has also been further extended to sequential recommendation [3] and cross-domain recommendation [32, 101]. Despite these advances, traditional MoE methods route inputs to all experts, leading to high computational cost. Sparse MoE (SMoE) addresses this by activating only a subset of experts [18, 22, 24, 61], significantly improving efficiency while preserving model capacity.

Structured User Behavior Modeling. In the era of big data, user recommendation data are collected and represented through a wide range of data modalities, including text [80, 81, 89], images [59, 79], time series [51, 52, 88, 94], and graph-structured data [11, 82, 103]. Consequently, user behaviors exhibit rich structural patterns, such as temporal dynamics and relational dependencies, making effective modeling of structured signals a fundamental challenge in recommendation.

Graph neural networks (GNNs) [2, 38, 50, 74] have been widely adopted to model such graph-structured user–item interactions. Homogeneous GNNs [8, 23, 29, 58] are commonly used for collaborative filtering due to their low-pass filtering property, which facilitates learning smooth and similar user–item representations. In contrast, heterogeneous GNNs [66, 78] incorporate diverse side information, such as social relations, item attributes, and contextual links, yielding expressive representations. Furthermore, graph contrastive learning [4, 83, 95] improves robustness by enforcing representation consistency under graph augmentations.

Another active line of research focuses on session-based recommendation [30, 43, 48, 62, 76, 87], which aims to model short-term user intent from sequential interactions. Early methods rely on recurrent neural networks [30] to capture within-session transitions. More recent works construct session graphs [76, 87] to explicitly model item-to-item transitions, enabling richer local dependency modeling beyond strict sequential order. In parallel, studies have explored disentangling short-term intent from contextual signals and adopting more expressive sequence encoders [43, 62] to enhance robustness under sparse or noisy sessions.

F Ethical Use of Data and Informed Consent

All experiments in this paper use only publicly available datasets from the LongCTR benchmark [2]. We did not collect new data, contact users, or conduct any interventions. We complied with the licenses and terms of use specified by the benchmark and its constituent datasets. To support transparency, we will release our code, configuration files, and evaluation scripts upon publication.

[2]https://github.com/reczoo/LongCTR