RESEARCH-ARTICLE

Beyond Immediate Click: Engagement-Aware and MoE-Enhanced Transformers for Sequential Movie Recommendation

HAOTIAN JIANG, Amazon.com, Inc., Seattle, WA, United States

SIBENDU PAUL, Amazon.com, Inc., Seattle, WA, United States

HAIYANG ZHANG, Amazon.com, Inc., Seattle, WA, United States

CAREN CHEN, Amazon.com, Inc., Seattle, WA, United States

Open Access Support provided by:

Amazon.com, Inc.

Published: 22 September 2025

Citation in BibTeX format

RecSys '25: Nineteenth ACM Conference on Recommender Systems
September 22 - 26, 2025
Prague, Czech Republic

Conference Sponsors:
SIGCHI

Beyond Immediate Click: Engagement-Aware and MoE-Enhanced Transformers for Sequential Movie Recommendation

Haotian Jiang
Amazon Prime Video
Sunnyvale, USA
haotij@amazon.com

Sibendu Paul
Amazon Prime Video
Seattle, USA
sibendu@amazon.com

Haiyang Zhang
Amazon Prime Video
Sunnyvale, USA
hhaiz@amazon.com

Caren Chen
Amazon Prime Video
Seattle, USA
carechen@amazon.com

Abstract

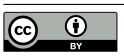
Modern video streaming services heavily rely on recommender systems. Although there are many methods for content personalization and recommendation, sequential recommendation models stand out due to their ability to summarize user behavior over time. We propose a novel sequential recommendation framework to address the following key issues: suboptimal negative sampling strategies, fixed user-history context lengths, and single-task optimization objectives, insufficient engagement-aware learning, and short-sighted prediction horizons, ultimately improving both immediate and multi-step next-title prediction for video streaming services. In this work, we propose a novel approach to capture patterns of interaction at different time scales. We also align long-term user happiness with instantaneous intent signals using multi-task learning with engagement-aware personalized loss. Finally, we extend traditional next-item prediction into a next-K forecasting task using a training strategy with soft positive label. Extensive experiments on large-scale streaming data validate the effectiveness of our approach. Our best model outperforms the baseline in NDCG@1 by up to 3.52% under realistic ranking scenarios showing the effectiveness of our engagement-aware and MoE-enhanced designs. Results also show that soft-label Multi-K training is a practical and scalable extension, and that a balanced personalized negative sampling strategy generalizes well. Our framework outperforms baselines across all ranking metrics, providing a robust solution for production-scale streaming recommendations.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Sequential Recommendation, Context-Adaptive Ranking, Multitask Learning, Engagement-aware Loss, Hard Negative Sampling, Next-K Title Prediction



RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748076

ACM Reference Format:
Haotian Jiang, Sibendu Paul, Haiyang Zhang, and Caren Chen. 2025. Beyond Immediate Click: Engagement-Aware and MoE-Enhanced Transformers for Sequential Movie Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3705328.3748076

1 Introduction

Modern video streaming services heavily rely on recommender systems, which enable users to rapidly discover relevant content inside large, constantly evolving video catalogs [12, 24, 25, 37]. Many streaming services like Prime Video, Netflix, Hulu, TikTok, and YouTube frequently refine their recommendation systems to keep users engaged and improve retention [5, 10, 31, 36]. Although there are many methods [19, 40] for content personalization and recommendation, sequential recommendation models stand out due to their ability to summarize user behavior over time. These models help services to summarize user viewing patterns and predict the next content for the user. Behavioral Sequential Transformers (BST) [4, 29, 43] play an important role in ranking systems, helping services analyze long-term user behavior while incorporating contextual signals to refine recommendations. Although BSTs offer many advantages, three key challenges remain in predicting the next title a user will watch. Firstly, rather than selecting user-specific, challenging negative examples, many studies choose negative samples randomly or according to popularity [22]. This reduces the model's ability to identify false positives, reducing its effectiveness in practical applications. Moreover, the fixed sequence lengths ignore changes in user engagement [27]. As a result, the model struggles to rank titles that satisfy both immediate needs and consistent interests. Furthermore, many ranking systems overly-prioritize click-through rate even if completion rate [42] is often a more reasonable indicator of user satisfaction and recommendation quality and these systems also narrowly focus on immediate next-item prediction [8, 21, 26, 32, 33]. Overcoming these issues will help to develop a ranking model that better aligns with user preferences and improves long-term engagement.

To address key limitations in sequential recommendation, such as generic negative sampling, fixed user-history lengths, narrow CTR-focused optimization, insufficient engagement-aware learning, and short-sighted prediction horizons, we introduce the unified

framework that jointly leverages engagement signals and context-adaptive architectures for personalized sequential recommendation. Our main contributions are summarized as follows:

(1) We introduce an adaptive context-aware Mixture of Experts (MoE) architecture that dynamically distributes user representation over several specialized experts focusing on various time scales.
(2) We propose a personalized hard negative sampling strategy that leverages user-specific completion rates to generate more informative contrastive examples.
(3) We design a multi-task learning framework with engagement-aware personalized loss that jointly optimizes CTR, ranking, and completion rate prediction.
(4) We extend traditional next-title prediction to a more realistic next-K title forecasting setup using soft positive label.

These contributions collectively improve personalization, robustness, and scalability in real-world streaming recommendation systems, outperforming the baseline in NDCG@1 by up to 3.52%. To our best knowledge, this is the first deployable system combining these techniques, with validated offline and real-world gains. We believe it offers practical value and can inspire future work on engagement-aligned ranking.

2 Related Work

Sequential recommendation models use historical data to project future interactions of users. Traditional approaches such Markov chains and recurrent neural networks (RNNs) [6, 34] are used to summarize sequential patterns in user historical behavior. However, they struggle with long-range dependencies and scalability. More recently, Behavioral Sequential Transformers [3, 39] and Transformer-based designs [16] have shown remarkable performance more recently by simulating long-term dependencies and integrating contextual signals. However, these models often rely on simple negative sampling techniques and fixed sequence durations, making them difficult to adapt to evolving user engagement patterns. Additionally, most of the previous works focus on ranking next immediate title, but many real-world applications require predicting multiple consecutive interactions (e.g., next-K title prediction) [11, 30]. In our work, we also discuss the next-K title prediction in a realistic scenario.

Training recommendation models to identify relevant from irrelevant objects requires negative sampling [7]. Popularity-based negativity or random sampling are common traditional approaches that might not give the model enough learning challenges. Therefore, hard negative mining methods have been further investigated [9, 15, 38]. However, most existing approaches do not personalize negative sampling based on user-specific engagement patterns, such as completion rates. Our approach introduces personalized hard negative-aware sampling, leveraging completion rates to encourage the model learns from more informative negatives, thus refining ranking precision.

Multi-task learning (MTL) techniques have been widely adopted in movie recommender systems to optimize multiple objectives together such as Click-through rate (CTR) and Conversion Rate [18, 20]. But conventional MTL models often inadequately balance short-term metrics such as CTR and long-term user satisfaction such as (completion rates or watch times), resulting in suboptimal trade-offs between instantaneous clicks and long-term user happiness. Recent works like Progressive Layered Extraction [28] and multi-gate mixture-of-experts [23] try to resolve this issue but lack explicit incorporation of personalized engagement signals. We propose an engagement-aware personalized loss function, dynamically weighting recommendation optimization objectives by individual user completion patterns, thereby better aligning recommendations with genuine user interests.

Exploration of Mixture-of-experts (MoE) architectures has been conducted to dynamically allocate various network segments to particular input patterns which improves representation learning [1, 35, 41]. Existing MoE-based recommendation models primarily use static gating mechanisms [2], where expert selection is fixed after training and does not dynamically adapt to user behavior. Some recent works investigate dynamic MoE gating in recommendation [1, 14], but they mostly focus on content-based routing or user intent modeling rather than temporal feature modeling [17]. Our approach introduces an adaptive MoE gating network that explicitly routes user interactions to different experts based on engagement history length (short-term, mid-term etc). This temporal-aware expert selection improves personalization, making the model more effective at capturing both immediate user intent and consistent viewing preferences.

3 Methodology

In this section, we introduce the details of our novel sequential recommendation framework. As illustrated in Fig 1, the proposed method integrates personalized hard negative sampling, a Mixture-of-Experts module, and multi-task learning with engagement-aware loss to improve the effectiveness and robustness of next-title prediction in streaming platforms.

3.1 Personalized Hard Negative-aware Sampling

Effective negative sampling is critical for training robust recommendation systems. Standard negative sampling approaches often overlook the varying degrees of relevance among negative titles to a given customer. To address this limitation, we propose a personalized hard negative sampling strategy leveraging user-specific viewing behavior, specifically using completion rates as an explicit measure of engagement. We first aggregate user interactions across sessions to calculate the total viewing time (seconds viewed) for each title. The completion rate for each user-title interaction is formally defined as:

$$\text{Completion Rate}_{u,i} = \frac{\text{Seconds Viewed}_{u,i}}{\text{Runtime}_i \times 60} \tag{1}$$

where u denotes the user and Runtime_i denotes the total minutes of title i.

user-specific hard negatives are titles that have low completion rates, which represent partial interest followed by abandonment. These titles act as difficult negative examples, allowing the model to more precisely separate truly preferred material from ostensibly related but finally boring titles. This therefore enables the model to better reflect complex variations in user preferences. We enrich our negative sampling technique by grouping negatives into three separate pools:

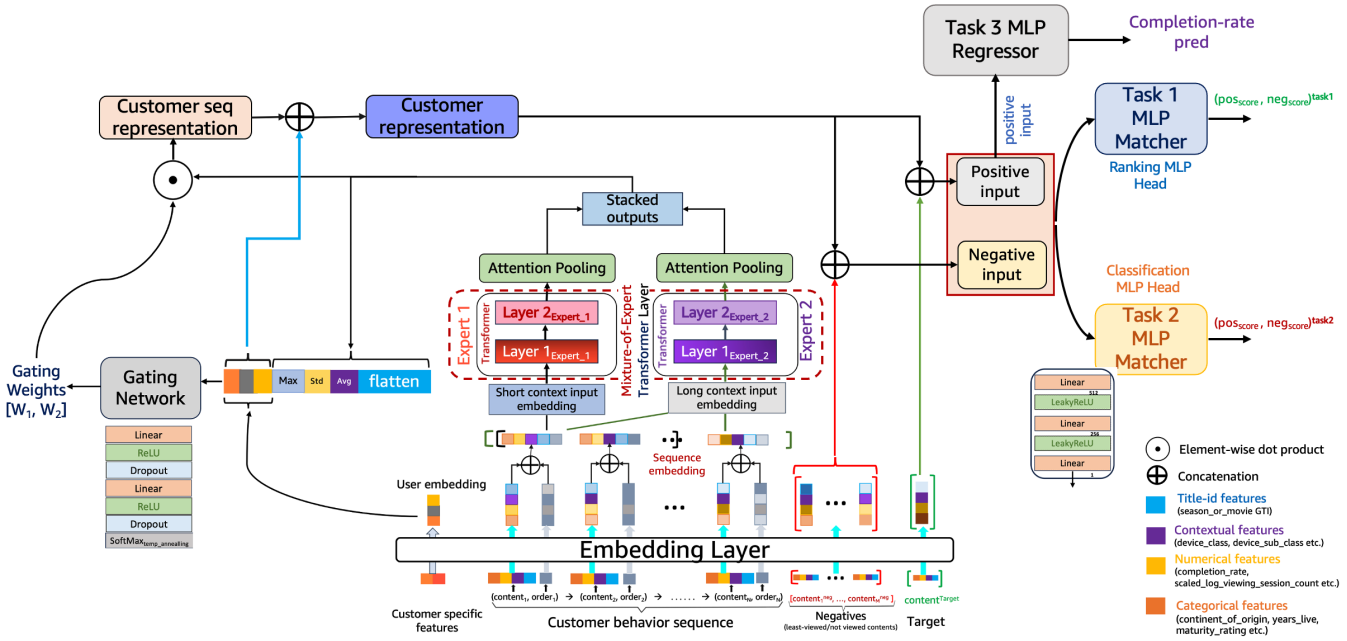


Figure 1: Overview of the proposed sequential recommendation framework. The embedding layer processes customer-specific features and behavior sequences, transforming them into customer and title embeddings. These are subsequently processed by a Behavioral Sequential Transformer to capture temporal dependencies. A MoE module dynamically assigns user interactions to specialized experts and each expert focuses on different time scales. A gating network route user interactions to the most appropriate expert by learning adaptive weights based on user engagement patterns. The final sequence representation is concatenated with customer-level features to form a comprehensive customer representation, which is fed into a multi-task learning framework for ranking and prediction tasks. Specifically, three task-specific heads jointy optimize click-through rate, ranking accuracy, and completion rate prediction using engagement-aware personalized loss, ensuring alignment between immediate engagement signals and long-term user satisfaction.

- **User-specific Hard Negatives**: Titles the user partly watched and abandoned.
- **Globally Trending Negatives**: Globally popular titles that the user has not interacted with.
- **Globally Tailing Negatives**: Globally less popular titles also not interacted with by the user.

These pools taken together offer a varied and useful collection of negative examples, therefore guaranteeing efficient model training.

Personalized hard negative sampling has a number of distinct advantages over simpler strategies, such as popularity-based or uniform random sampling, as a result of its capacity to identify nuanced differences in user preferences. While uniform random sampling introduces many negatives that are straightforward for the model to differentiate, popularity-based sampling ignores particular user behaviors, therefore perhaps introducing irrelevant or easily recognizable negatives. Personalized hard negative sampling, on the other hand, guarantees the negatives closely resemble titles of real user interest by using clear user involvement signals—such as partial watching and abandoning behaviors. Based on experimental settings and data availability, we balance sample weights among these groups. This sampling method helps the model to learn subtle differences in user preferences, leading to more accurate recommendations. Note that in this work 0.05 is empirically pre-defined as the low completion rate criterion.

3.2 Adaptive Context-aware MoE Architecture

As shown in Fig 1, we propose an Adaptive Context-Aware Mixture of Experts (MoE) architecture to dynamically distribute user representations over several specialized experts, hence improving tailored suggestions. Conventional single-model systems can find it difficult to distinguish between changing behavioral patterns across time, which results in less than ideal recommendations. Our approach overcomes this limitation by using a team of experts each focusing on a different length of user interactions. As a result, the model is better able to capture long-term user preferences, medium-term behavioral patterns, and brief involvement. Four central components define the architecture:

- **Embedding Layers**: Shared embedding layers for user features, categorical characteristics, contextual signals, numerical inputs, and title embeddings. These layers give all the experts a shared feature representation.
- **Behavioral Sequential Transformer**: A transformer model is used to track how users interact with titles over time, showing how their tastes and behaviors change over time.

Table 1: Example features used in this work.

Context	Title	Customer	Interaction
device	title_id	cust_seg1	viewing_sessions
time	genre	cust_seg2	completion_rate
...	...	...	...

- **Mixture-of-Experts Module**: A group of trained experts who can work on different length of user interactions. The model activates and weights various experts depending on user-specific interaction history, therefore guaranteeing more correct and tailored recommendations.
- **Adaptive Gating Network**: A dynamic routing system that identifies the most helpful expert for a given user. By consulting multiple experts, users get more accurate recommendations.

The proposed method let the model tell differences between various engagement patterns, so it can adapt to different user tastes. Learning to allocate behavioral data to the most qualified experts helps the model to increase the accuracy and granliness of its suggestions.

3.2.1 Embedding and Transformer Layer. Table 1 summarizes the example features used in this work. Our architecture uses an embedding layer followed by a Transformer-based encoder to efficiently model sequential user interactions, hence allowing effective representation learning for categorical features and sequential dependencies. An embedding function maps any title-level feature into a dense embedding space:

$$\mathbf{e}_i = E_i(f_i), \quad \mathbf{e}_i \in \mathbb{R}^{d_i} \tag{2}$$

where $E_i(.)$ is the embedding function for feature f_i, and d_i is the specific embedding dimension for that feature. The embeddings for all title-level features are then concatenated to get the unified title representation:

$$\mathbf{t} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_k], \quad \mathbf{t} \in \mathbb{R}^{\sum d_i} \tag{3}$$

where k is the number of embedded features per title. For a given user interaction sequence of length L, we concatenate sequence title features and positional encodings:

$$\mathbf{x}_t = [\mathbf{t}_t + \mathbf{p}_t], \quad \mathbf{x}_t \in \mathbb{R}^{\sum d_i} \tag{4}$$

where p_t is the positional encoding for timestamp t. The resulting sequence representation:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_L], \quad \mathbf{X} \in \mathbb{R}^{L \times \sum d_i} \tag{5}$$

is then fed into a Transformer encoder, producing contextualized sequence representations:

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{X}) \tag{6}$$

where H represents the learned user sequence embeddings. To obtain a final user sequence representation, we apply attention pooling, where attention scores are computed as:

$$\alpha_t = \frac{\exp(w^T \mathbf{h}t)}{\sum t' \exp(w^T \mathbf{h}_{t'})} \tag{7}$$

where w is a learnable attention weight vector and h_t is the hidden state at timestamp t. The final pooled sequence representation is obtained as:

$$\mathbf{h}_{\text{user}} = \sum_{t=1}^{L} \alpha_t \mathbf{h}_t, \quad \mathbf{h}_{\text{user}} \in \mathbb{R}^d \tag{8}$$

where higher weights are assigned to more relevant interactions.

3.2.2 Mixture-of-Experts Specialization. Every expert in our MoE system is meant to be specialized in handling sequences of user input at a certain temporal scale. This specialization helps the model to more precisely capture different behavioral patterns over many time periods, hence supporting a more customized and dynamic recommendation system. We characterize three primary experts:

- **Short-term Expert**: Pays more attention to recent encounters, which typically represent the user's immediate intent and short-term preferences.
- **Mid-term Expert**: Discovers the hidden interests of users over an extended period of time, such as variations in genre preferences and consistent interactions with related content.
- **Long-term Expert (optional)**: Learns more constant patterns like favored genres, consistent content choices, and persistent user behaviors over months or years.

Experts are heterogeneous. Each expert is implemented using separate transformer encoder layers, and created especially to replicate sequential dependencies within their allocated temporal range. These layers let the experts utilize self-attention techniques to highlight important historical exchanges and efficiently collect trends in user interactions.

3.2.3 Adaptive Gating Network for Expert Selection. We provide an adaptive gating network to properly route user interactions to the most appropriate expert(s). Dynamic determination of the weight each expert should get for a specific user sequence depends critically on the gating mechanism. The gating network is implemented as a multi-layer perceptron, mapping user embeddings and sequence representations into expert routing logits. In more details, the gating network follows this pattern:

(1) **Context-aware Routing**: The gating network receives user-level embeddings and expert-specific sequence representations as input and creates a set of routing weights for each expert.
(2) **Softmax-based Selection**: Softmax activation function with temperature annealing computes the expert selection probabilities:

$$\alpha_{u,e} = \frac{\exp\left((z_{u,e} + \varepsilon)/\tau\right)}{\sum_{j=1}^{E} \exp\left((z_{u,j} + \varepsilon)/\tau\right)}, \tag{9}$$

where $\alpha_{u,e}$ represents the gating weight for user u and expert e, denoting the gating logit output. ε denotes additive Gaussian noise scaled by a temperature parameter τ. By adding stochasticity to the expert selection, we encourage exploration and smoother expert assignment in early training stages. and is the softmax temperature parameter.
(3) **Entropy Regularization for Balanced Expert Usage**: To prevent expert collapse (i.e., a scenario where only one expert dominates), we introduce an entropy-based regularization term:

$$\mathcal{L}_{\text{entropy}} = \left(H(\alpha_u) - H^*\right)^2, \tag{10}$$

where $H(\alpha_u)$ is the entropy of the gating weight distribution, and H^* is a target entropy level encouraging diverse expert selection.

The final user representation is computed as a weighted sum of expert outputs:

$$\mathbf{u}_{\text{MoE}} = \sum_{e=1}^{E} \alpha_{u,e} \mathbf{h}_{u,e}, \tag{11}$$

where $\mathbf{h}_{u,e}$ is the output from expert e. This weighted representation $\mathbf{u}_{\text{MoE}}$ is then concatenated with customer-level features and fed into multitask learning framework for final ranking and prediction tasks.

3.3 Multitask Learning for Joint Optimization with Engagement-aware Personalized Loss

Optimizing just the click-through rate can lead to clickbait bias, in which case the model prioritizes titles that gets clicks without ensuring that the users are engaged in a meaningful manner. Conversely, optimizing ranking loss simply ignores click likelihood, therefore restricting coverage and sometimes burying relevant material. To balance clickability and engagement, as shown in Fig 1, we propose to use a Multi-task Learning Framework, where the MoE-based sequence representation is jointly optimized for several learning objectives. Moreover, conventional recommendation loss models mostly assume that all good interactions have equal contribution, therefore neglecting to differentiate between strong and weak user involvement. At the same time, engagement behavior is quite individual; some users stream a specific title completely while others give it early abandon. Therefore, we propose a novel Engagement-aware Personalized Loss where the completion rate acts as a user-specific weighting element, dynamically changing the optimization process to reflect individual engagement preferences, hence including these behavioral variations.

3.3.1 Multi-task Learning Framework. In particular, we optimize the subsequent tasks:

- **CTR prediction**: A binary classification task that optimizes the likelihood of a user streaming a recommended title by adjusting the BCE loss based on the completion rate.
- **Ranking optimization**: A contrastive loss-based objective that guarantees that products with high engagement and relevance are prioritized over those with low engagement or irrelevant value.
- **Completion Rate Prediction (Optional)**: A regression task that assists the model in learning fine-grained engagement patterns by estimating how much of the content that a user is likely to view.

Sharing the same bottom layer, each task is optimized using a dedicated multi-layer perceptron (MLP) head:

- **CTR MLP Matcher:** Optimized with a completion-aware BCE loss to improve click prediction accuracy while taking engagement into account.
- **Ranking MLP Matcher:** Trained with completion-aware contrastive loss to encourage that highly-engaging titles are favored in the ranking process.
- **Completion Rate MLP Regressor (Optional):** Predicts the engagement-level for a given <user, title> as a continuous value, which can be used as an auxiliary signal to improve the training process of CTR or ranking.

The proposed multi-task learning is supposed to improving personalization, exploration and less overfitting to misleading behaviors.

3.3.2 Engagement-aware Personalized Loss. Unlike conventional loss functions, which treat all positive interactions equally across customers, the proposed engagement-aware personalized loss adapts per user by incorporating completion rate as an engagement signal. This guarantees that recommendations are optimized for meaningful streaming in addition to click likelihood, thereby reinforcing engagement-driven ranking.

To optimize ranking in a personalized engagement-aware manner, we introduce completion rate-based adaptive weighting to modify the contrastive loss. For each user u, the completion rate $c_{u,i}$ of a positive title i serves as a personalized scaling factor:

$$\mathcal{L}_{\text{ranking}} = \mathbb{E}\big[\max(0, 1 - (\text{pos\_score} - \text{neg\_score})) \cdot W_{u,i}\big] \tag{12}$$

where $W_{u,i}$ is the engagement-aware scaling factor defined as:

$$W_{u,i} = 1 + \alpha(c_{u,i} - c_{\text{threshold}}) \tag{13}$$

where $c_{\text{threshold}}$ is a global completion baseline ensuring weighting remains centered around an expected engagement level, α is a hyperparameter controlling the degree of personalization in ranking loss. By integrating $c_{u,i}$, the new loss dynamically adjusts to the unique behaviors of individual users, penalizing low-engagement recommendations and reinforcing high-engagement items.

Following the similar manner, we further incorporate engagement-awareness into CTR prediction by adjusting the binary cross-entropy (BCE) loss, thereby assuring that click probability predictions are influenced by both click likelihood and expected engagement,

$$\mathcal{L}_{\text{CTR}} = \mathbb{E}\big[-y\log(\hat{y}) - (1-y)\log(1-\hat{y})\big] \cdot W_{u,i} \tag{14}$$

where $W_{u,i}$ is the same completion-aware scaling factor as in ranking loss. This adjustment guarantees that items with higher engagement levels receive more robust reinforcement in CTR modeling, while simultaneously preventing overfitting to low-engagement but high-CTR items.

The final loss is given as below:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CTR}} + \lambda_2 \mathcal{L}_{\text{ranking}} + \lambda_3 \mathcal{L}_{\text{Reg}} + \lambda_4 \mathcal{L}_{\text{entropy}} \tag{15}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are dynamically-adjusted task-specific weights, $\mathcal{L}_{\text{Reg}} = \mathbb{E}\big[(c_{u,i} - \hat{c}_{u,i})^2\big]$ is optional completion rate regression loss used as an auxiliary task for training and it is not personalized. Note that the regression head will be discarded for inference. By incorporating a completion-aware loss scaling factor, our model guarantees that recommendations distinguish between superficial engagement and genuine user interest, resulting in recommendations that are engagement-driven and of higher quality.

Table 2: Detailed negative sampling strategies for each setting including training, validation and testing.

	PHNS	Globally Trending Negatives	Globally Tailing Negatives	Random Negatives
Training Negative Sampling Strategy				
Hard	50%	30%	20%	0%
Medium	30%	40%	30%	0%
Easy	20%	30%	50%	0%
No PHNS	0%	50%	30%	20%
Validation/Testing Negative Sampling Strategy				
PHNS-enabled	30%	30%	20%	20%
PHNS-disabled	0%	50%	30%	20%

4 Experiment

4.1 Experimental Setting

We utilize BST as our primary baseline in our research because of its real-world deployment in large-scale ranking systems, scalability, and capacity of self-attention mechanisms to integrate contextual embeddings at the user and item levels. Although other models including GRU4Rec [13] and SASRec [16] have been extensively investigated, they lack fundamental adjustments required for production-scale streaming recommendations. Though it models session-based interactions effectively, GRU4Rec suffers with long-range dependencies, is fundamentally sequential in computation, and does not scale effectively for millions of users in real-time production environments. While using self-attention to improve sequence modeling, SASRec does not specifically combine rich metadata signals—such as viewing device, content features, and temporal engagement patterns—which are absolutely essential for contextual personalizing in streaming recommendations.

We evaluate our model using a dataset collected from November 2024 on the Prime Video service, covering 1 million randomly selected users. In total, we sample 7,244,982 sequences. **We guarantee that all customer data utilized in this work is completely anonymized and de-identified prior to analysis, in accordance with strict privacy and data protection principles. There is no personally identifiable information (PII), and all interactions are aggregated at a level that ensures customer confidentiality while still allowing insightful analysis of recommendation efficacy.**

To facilitate clarity and reproducibility, we summarize our detailed experimental setup in Table 2 (sampling strategy) and Table 3 (model configuration). These experimental settings remain consistent across all experiments, including ablation studies, to ensure fair and controlled comparisions. Our evaluation setup follows standard sequential recommendation practice, where each test query has one relevant (positive) item and multiple sampled negatives. Under this setup, F1 is not meaningful and Precision@K is equivalent to Recall@K, as there is only one ground-truth item per query. Thus, we report Recall@5 to represent both. We include NDCG@1 to assess top-1 accuracy, which is critical in high-precision applications. While NDCG@1 reduces to a binary indicator of top-1 correctness in this setup, it remains consistent with NDCG@K when K > 1,

Table 3: This table summarizes the core architectural and experimental settings. The setup is consistent for all variants unless othereise noted, making sure controlled ablation and fair comparisons.

Category	Parameter	Value
Framework	Pytorch	
	CUDA	
General	Batch size	8192
	Learning rate	0.001
	Epochs	20
	Optimizer	Adam
	Weight decay	0.01
	Scheduler	ReduceLROnPlateau
Sampling	Negative sampling	20 per positive
	Soft negative label	0.1
Gating	Initial temperature	2
	Noise	N(0, 0.1)
	Target Entropy	0.52
	MLP hidden dims	[128, 64]
	Short-term seq length	7
	Mid-term seq length	20
	Initial Dropout	0.2
Expert 1	Transformer layers	2
	Feedforward multiplier	1 x hidden_dim
	Attention heads	12
	Dropout	0.2
Expert 2	Transformer layers	2
	Feedforward multiplier	2 x hidden_dim
	Attention heads	12
	Dropout	0.2
Multi-task loss	Ranking	Contrastive loss
	CTR	BCE loss
	Completion rate regression	MSE loss
Evaluation	Metrics	NDCG@1, NDCG@5, Recall@5, MRR@5
	Statistical tests	Paired t-test, Wilcoxon

and allows fair comparison across other metrics such as NDCG@5 and MRR@5, which we also report. For every user sequence, the next immediate watched title from the user's history is selected as the target, and 20 negative samples are generated according to the strategies outlined in Table 2. We randomly split users into training (80%), validation (10%), and testing (10%) groups so as to guarantee a fair evaluation without overlap across these sets. As illustrated in Table 2, we examine four different negative sampling strategies with various difficulty level during training: Hard, Medium, Easy, and No PHNS (Personalized Hard Negative Sampling). For validation and testing, we assess models under both PHNS-enabled and PHNS-disabled conditions. We compare Single Task Learning (STL), which

optimizes CTR as the sole objective, against Multi-Task Learning (MTL), which jointly optimizes CTR, Ranking, and Completion Rate Regression with a weighting of [1.0, 1.0, 0.2], respectively. We also study Personalized Loss (PL), which dynamically scales CTR and ranking loss by including user-specific engagement signals, and Personalized Hard Negative Sampling (PHNS), which chooses negative samples depending on user-specific interaction patterns. Additionally, we assess both Specialized MoE (S-MoE) and Vanilla MoE (V-MoE). S-MoE consists of two experts specialized for different temporal windows. One expert focuses on short-term sequences, while the other processes mid-term sequences. On the other hand, V-MoE also has two experts but lacks specialization as both experts receive same sequences. To regulate expert selection diversity, we incorporate Target Entropy (TE), a regularization technique that encourages balanced expert utilization. Unless otherwise specified, all results presented in the following experiments are based on the medium training negative sampling strategy, as shown in Table 2, with TE set to 0.52.

4.2 Results

As shown in Table 4, S-MoE-BST (PHNS+MTL+PL) has the best ranking performance among all model variants when PHNS is disabled for testing, with NDCG@1 = 0.7235 and Recall@5 = 0.9589. Note that, incorporating multi-task learning and personalized loss improves ranking effectiveness over the single-task counterpart, confirming that jointly optimizing CTR, ranking, and completion rate regression leads to more reliable recommendations. When PHNS is enabled for testing, all model variants have a drop in absolute ranking performance, as PHNS introduces more challenging negative samples, making ranking optimization more difficult. However, S-MoE-BST (PHNS+MTL+PL) still maintains a leading performance (NDCG@1 = 0.6065, Recall@5 = 0.9356), demonstrating the robustness of S-MoE-BST with engagement-aware multi-task optimization in handling harder negatives. Furthermore, the performance gap between S-MoE and V-MoE confirms the advantage of temporal-aware expert specialization in MoE-based architectures. The results denote that static MoE models may not generalize effectively under difficult ranking conditions, whereas dynamic expert selection makes the model more capable to identify behavioral variations. The results also show that medium PHNS consistently give robust performance under both PHNS-enabled and PHNS-disabled testing scenarios, proving the importance of using a balanced PHNS approach. In addition to baseline-BST, we include two classical sequential recommenders—SASRec and GRU4Rec—as new baselines. Both models perform worse than not only our S-MoE variants, but also the baseline-BST model, particularly under PHNS-enabled testing (e.g., SASRec NDCG@1 = 0.4626 vs. 0.5862 for Baseline-BST; GRU4Rec NDCG@1 = 0.3074), indicating that conventional sequence models are less effective in engagement-aware ranking tasks. To verify statistical significance of our results, we conducted per-user significance testing on NDCG@1 values (726,520 samples, batch size = 1) comparing our best model variant S-MoE-BST (PHNS+MTL+PL) to the baseline-BST, using both paired t-test and Wilcoxon signed-rank test. Our best model variant yields a t-statistic of 21.68 ($p < 1e-8$) and Wilcoxon test statistic of $4.72 \times 1e9$ ($p < 1e-4$). These results strongly indicate that the observed improvements are statistically significant and unlikely due to chance. Overall, the results indicate that adaptive MoE design, personalized hard negative sampling, and multitask learning with engagement-aware loss jointly contribute to a better recommendation system that generalizes well to difficult ranking scenarios.

Table 4: Offline NDCG, Recall, MRR of model variants.

Medium PHNS	NDCG@1	NDCG@5	Recall@5	MRR@5
Personalized Hard Negative-aware Sampling Disabled				
S-MoE-BST (PHNS+MTL +PL)	**0.7235**	**0.8547**	**0.9589**	**0.8195**
S-MoE-BST (PHNS+MTL)	0.7205	0.8527	0.9577	0.8172
S-MoE-BST (PHNS+STL)	0.7108	0.8470	0.9552	0.8104
V-MoE-BST (PHNS+STL)	0.7085	0.8462	0.9557	0.8092
Baseline-BST (PHNS+STL)	0.7094	0.8463	0.9551	0.8095
SASRec (PHNS+STL)	0.5886	0.7647	0.9120	0.7154
GRU4Rec (PHNS+STL)	0.4383	0.6468	0.8299	0.5858
Personalized Hard Negative-aware Sampling Enabled				
S-MoE-BST (PHNS+MTL +PL)	**0.6065**	**0.7864**	**0.9356**	**0.7363**
S-MoE-BST (PHNS+MTL)	0.5978	0.7813	0.9337	0.7302
S-MoE-BST (PHNS+STL)	0.5885	0.7749	0.9303	0.7227
V-MoE-BST (PHNS+STL)	0.5846	0.7724	0.9292	0.7198
Baseline-BST (PHNS+STL)	0.5862	0.7738	0.9300	0.7213
SASRec (PHNS+STL)	0.4626	0.6743	0.8605	0.6124
GRU4Rec (PHNS+STL)	0.3074	0.5449	0.7639	0.4726

To make our work more comprehensive, we extend our research to next-K title prediction as a more realistic scenario in streaming recommendations. Formally, given a user interaction history $S = \{t_1, t_2, ..., t_n\}$, instead of only predicting the immediate next title t_{n+1}, we aim to predict the next K titles $\{t_{n+1}, ..., t_{n+K}\}$, where $K = 3$ in this work. We propose a **Soft-Label Multi-K Training** framework that integrates learning signals from multiple future interactions. Rather than treating all K steps equally, we assign larger label strengths to earlier steps and smaller label strengths to later ones. Specifically, the model uses soft labels to encode diminishing confidence in farther future interactions—e.g., the next title t_{n+1} has a positive label of 1.0, the second-next title t_{n+2} has 0.6 and the third-next title t_{n+3} has 0.3. This soft integration allows the

model to remain focused on short-term prediction accuracy while still learning from longer-term engagements. For comparison, we also experiment the setting where all K steps are equal-weighted, namely, no soft label is used.

Table 5: Next-K Title Prediction with Soft Positive Label.

S-MoE-BST (PHNS+MTL+PL)	NDCG@1	NDCG@5	Recall@5	MRR@5
Personalized Hard Negative-aware Sampling Disabled (K = 1)				
K=1	0.7235	0.8547	0.9589	0.8195
K=3, equal	0.6935	0.8404	0.9567	0.8011
K=3, weighted	**0.7272**	**0.8575**	**0.9605**	**0.8227**
Personalized Hard Negative-aware Sampling Disabled (K = 3)				
K=1	0.6696	0.8242	0.9482	0.7824
K=3, equal	0.6586	0.8205	0.9499	0.7768
K=3, weighted	**0.6770**	**0.8302**	**0.9521**	**0.7890**
Personalized Hard Negative-aware Sampling Enabled (K = 1)				
K=1	0.6065	0.7864	0.9356	0.7363
K=3, equal	0.5601	0.7618	0.9292	0.7056
K=3, weighted	**0.6068**	**0.7870**	**0.9363**	**0.7368**
Personalized Hard Negative-aware Sampling Enabled (K = 3)				
K=1	0.5290	0.7388	0.9162	0.6794
K=3, equal	0.5093	0.7293	0.9152	0.6670
K=3, weighted	**0.5304**	**0.7418**	**0.9198**	**0.6821**

Notably, when extending testing from $K = 1$ to $K = 3$, the test set size increases by 2.6× (from 726,520 to 1,894,994 sequences). This increases task complexity significantly due to more diverse targets and shifted intents. However, the gradual decrease in NDCG@5 and Recall@5 suggests that the model still retains strong predictive power and generalizes well beyond immediate next-title prediction. As shown in Table 5, our proposed soft-label Multi-K training consistently delivers good performance. Under PHNS-disabled conditions, the soft-label strategy outperforms both single-step prediction and equal-weighted Multi-K training. For instance, at $K = 1$ testing, the soft-label multi-K training achieves an NDCG@1 of 0.7272 versus 0.7235 (standard next-title-only) and 0.6935 (equal-K). Similarly, for $K = 3$ testing, the soft-label multi-K training improves NDCG@1 to 0.6770 compared to 0.6696 (standard) and 0.6586 (equal-K). These gains appear under PHNS-enabled settings as well, indicating that the model focuses mainly on immediate next-title prediction but still learns from multi-K interactions. The key insight is that weighting future steps encourages the model to prioritize short-term accuracy and prevent overfitting to far future noise while softly considering longer-term patterns. Equal weighting, by contrast, dilutes the learning signal and slightly reduces precision, particularly for the immediate next title. Overall, the results demonstrate that soft-label Multi-K training is a practical and scalable extension to next-title recommendation.

4.3 Ablation Study

To get more insights about the effectiveness of each module, we conduct ablation experiments under PHNS-enabled and PHNS-disabled conditions. First, we evaluate the impact of different training negative sampling strategies on the ranking performance. As shown in Table 6, with S-MoE-BST (STL) model, when testing is PHNS-disabled, the standard training negative sampling without PHNS achieves highest ranking performance (NDCG@1 = 0.7568, Recall@5 = 0.9673). However, the performance drops significantly (NDCG@1 = 0.3694, Recall@5 = 0.8834) when testing with PHNS-enabled. That means the standard negative sampling is easier to be optimized but potentially overfits to simplistic patterns. Interestingly, when training includes PHNS, particularly at medium level, the ranking performance is quite balanced. In details, when testing is PHNS-enabled, training with medium PHNS significantly improves performance compared to the standard training negative sampling (NDCG@1 = 0.5885, Recall@5 = 0.9303). Even if testing is PHNS-disabled, training with medium PHNS can still achieve reasonable performance (NDCG@1 = 0.7108, Recall@5 = 0.9552). The results indicate that PHNS successfully help the model in differentiating between titles that appear relevant but fail to hold interest and those that people truly enjoy, leading to better recommendation. Overall, the study shows that a balanced negative sampling strategy, such as medium PHNS, can be generalized well and optimally improves ranking effectiveness while avoiding overfitting to simple or complicated negative titles.

Table 6: Ablation study for training negative sampling strategies.

S-MoE-BST (STL)	NDCG@1	NDCG@5	Recall@5	MRR@5
Personalized Hard Negative-aware Sampling Disabled				
PHNS Disabled	**0.7568**	**0.8753**	**0.9673**	**0.8441**
Hard PHNS	0.6932	0.8362	0.9510	0.7975
Medium PHNS	**0.7108**	**0.8470**	**0.9552**	**0.8104**
Easy PHNS	0.7100	0.8470	0.9558	0.8102
Personalized Hard Negative-aware Sampling Enabled				
PHNS Disabled	**0.3694**	**0.6416**	**0.8834**	**0.5612**
Hard PHNS	0.5756	0.7647	0.9236	0.7114
Medium PHNS	**0.5885**	**0.7749**	**0.9303**	**0.7227**
Easy PHNS	0.5850	0.7737	0.9308	0.7210

Second, we study how multitask learning with engagement-aware personalized loss affects model performance under both PHNS-enabled and PHNS-disabled conditions. As shown in Table 7, optimizing ranking and regression targets in addition to CTR consistently improves model performance across all metrics. Specifically, with joint optimization (CTR+Ranking+Regression), NDCG@1 is increased from 0.7108 to 0.7205 (PHNS-disabled testing) and 0.5885 to 0.5978 (PHNS-enabled testing). Furthermore, including engagement-aware personalized loss (CTR+Ranking+Regression+Places) improves model capability especially under PHNS-enabled conditions, where NDCG@1 get improved from 0.5978 to 0.6065 and achieves the highest scores across all metrics (NDCG@5 = 0.7864, Recall@5 = 0.9456, MRR@5 = 0.7363). Also, we observe that CTR+Ranking+PL performs better than CTR+Ranking+Regression, implying that personalization plays an important role in ranking optimization. However, the best results come from considering both regression and engagement-aware personalized loss (CTR+Ranking+Regression+PL). This indicates that completion rate regression helps capture additional engagement signals, and personalized loss further adjusts the

Table 7: Ablation study for multitask learning with engagement-aware personalized loss.

S-MoE-BST (PHNS)	NDCG@1	NDCG@5	Recall@5	MRR@5
Personalized Hard Negative-aware Sampling Disabled				
CTR	0.7108	0.8470	0.9552	0.8104
CTR+Ranking	0.7207	0.8526	0.9575	0.8172
CTR+Ranking+ PL	0.7219	0.8534	0.9580	0.8180
CTR+Ranking+ Regression	0.7205	0.8527	0.9577	0.8172
CTR+Ranking+ Regression+PL	**0.7235**	**0.8547**	**0.9589**	**0.8195**
Personalized Hard Negative-aware Sampling Enabled				
CTR	0.5885	0.7749	0.9303	0.7227
CTR+Ranking	0.5996	0.7819	0.9333	0.7311
CTR+Ranking+ PL	0.6052	0.7854	0.9348	0.7352
CTR+Ranking+ Regression	0.5978	0.7813	0.9337	0.7302
CTR+Ranking+ Regression+PL	**0.6065**	**0.7864**	**0.9356**	**0.7363**

Table 8: Ablation study for temporal-aware MoE with different target entropy.

S-MoE-BST (PHNS+STL)	NDCG@1	NDCG@5	Recall@5	MRR@5
Personalized Hard Negative-aware Sampling Disabled				
TE=0.48	0.7102	**0.8472**	**0.9561**	0.8103
TE=0.52	**0.7108**	0.8470	0.9552	**0.8104**
TE=0.56	0.7088	0.8454	0.9542	0.8087
TE=0.60	0.7062	0.8448	0.9551	0.8075
TE=0.64	0.7091	0.8460	0.9550	0.8092
Baseline-BST (PHNS+STL)	0.7094	0.8463	0.9551	0.8095
Personalized Hard Negative-aware Sampling Enabled				
TE=0.48	0.5842	0.7722	0.9293	0.7194
TE=0.52	**0.5885**	**0.7749**	**0.9303**	**0.7227**
TE=0.56	0.5872	0.7736	0.9291	0.7215
TE=0.60	0.5828	0.7711	0.9285	0.7183
TE=0.64	0.5862	0.7730	0.9288	0.7207
Baseline-BST (PHNS+STL)	0.5862	0.7738	0.9300	0.7213

ranking process by dynamically scaling loss contributions based on user-specific satisfaction. These results reinforce the idea that hard negative sampling makes the ranking task more challenging, where personalized loss plays a crucial role in improving robustness. The consistent improvements in Recall@5 and MRR@5 also confirm that multi-task learning combined with engagement-aware personalized loss maximizes both click likelihood and long-term user satisfaction. Overall, our findings emphasize the significance of optimizing for more than just CTR and the efficacy of engagement-aware personalized loss in enhancing ranking quality in real-world streaming recommendation systems.

Third, we explore the different settings of target entropy and evaluate the performance of S-MoE-BST (PHNS+PL) compared to Baseline-BST (PHNS+STL) under both PHNS-enabled and PHNS-disabled conditions. Table 8 shows that target entropy is an important factor in balancing expert use and has direct impact on model's capability of reasonably distributing user sequences to specialized experts. Under PHNS-disabled testing, the best performance is observed at TE = 0.48 and TE = 0.52. In details, TE = 0.48 achieves the highest NDCG@5 (0.8472) and Recall@5 (0.9561), while TE = 0.52 achieves the highest NDCG@1 (0.7108) and MRR@5 (0.8104). When TE increases over 0.52, we observe a gradual decline in ranking effectiveness, indicating that too much expert diversity could lower sequential modeling's strength and compromise specialization. Under PHNS-enabled testing, TE = 0.52 again shows the best performance across all metrics (NDCG@1 = 0.5885, NDCG@5=0.7749, Recall@5 = 0.9303, MRR@5=0.7227), confirming its robustness in handling challenging negative samples. Similarly, performance degrades slightly as TE increases over 0.52, indicating that ideal ranking requires a little degree of expert variability. Importantly, S-MoE consistently outperforms Baseline-BST under both PHNS-disabled and PHNS-enabled testing, demonstrating the effectiveness of temporal-aware expert specialization. Overall, these results show the importance of carefully tuning TE to find the reasonable balance between expert specialization and diversity. Our findings suggest that TE = 0.52 offers the best trade-off, leading to robust ranking performance in both standard and challenging ranking scenarios.

5 Conclusion

In summary, this research demonstrates that jointly optimizing CTR, ranking, and completion rate regression significantly enhances ranking effectiveness, with engagement-aware personalized loss further improving robustness, particularly under PHNS-enabled conditions. We verify that a balanced negative sampling strategy, such as medium PHNS, can be generalized well and optimally improves ranking effectiveness while avoiding overfitting to simple or complicated negative titles. Additionally, our studyies on MoE architectures show that Specialized MoE, which assigns different experts to short-term and mid-term interaction patterns, consistently outperforms Vanilla MoE and baseline models. We also extend traditional next-title prediction to a more realistic next-K forecasting scenario using a soft-label Multi-K training. Our findings highlight the need of engagement-aligned learning in improving both instantaneous ranking performance and long-term user happiness. These insights also directly inform the design of scalable and generalizable response prediction systems. In particular, our findings support a two-stage modeling paradigm for industrial recommender systems: using rich multi-task objectives and personalized engagement signals during pretraining to enhance generalization, followed by fine-tuning under serving distributions to optimize alignment with online behavior. Together, this work offers a practical framework for bridging offline learning objectives with online personalization needs, spurring follow-up research on engagement-aligned ranking.

References

[1] 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.

[2] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2024. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204* (2024).

[3] Ming Chen, Weike Pan, and Zhong Ming. 2024. Explicit and Implicit Modeling via Dual-Path Transformer for Behavior Set-informed Sequential Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 329–340.

[4] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior sequence transformer for e-commerce recommendation in alibaba. In *Proceedings of the 1st international workshop on deep learning practice for high-dimensional sparse data*. 1–4.

[5] 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.

[6] Qiang Cui, Shu Wu, Qiang Liu, Wen Zhong, and Liang Wang. 2018. MV-RNN: A multi-view recurrent neural network for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering* 32, 2 (2018), 317–331.

[7] Jingtao Ding, Yuhan Quan, Quanming Yao, Yong Li, and Depeng Jin. 2020. Simplify and robustify negative sampling for implicit collaborative filtering. *Advances in Neural Information Processing Systems* 33 (2020), 1094–1105.

[8] Shereen Elsayed, Ahmed Rashed, and Lars Schmidt-Thieme. 2024. Multi-Behavioral Sequential Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 902–906.

[9] Lu Fan, Jiashu Pu, Rongsheng Zhang, and Xiao-Ming Wu. 2023. Neighborhood-based hard negative mining for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2042–2046.

[10] Mattias Frey. 2021. *Netflix recommends: Algorithms, film choice, and the history of taste*. Univ of California Press.

[11] 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 International conference on web search and data mining*. 343–352.

[12] Mayor Inna Gurung, Md Monoarul Islam Bhuiyan, Ahmed Al-Taweel, and Nitin Agarwal. 2024. Decoding YouTube's Recommendation System: A Comparative Study of Metadata and GPT-4 Extracted Narratives. In *Companion Proceedings of the ACM Web Conference 2024*. 1468–1472.

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

[14] 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 SIGKDD Conference on Knowledge Discovery and Data Mining*. 585–593.

[15] Kaixi Hu, Lin Li, Qing Xie, Jianquan Liu, and Xiaohui Tao. 2021. What is next when sequential prediction meets implicitly hard interaction?. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 710–719.

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

[17] Jiamin Li, Qiang Su, Yitao Yang, Yimin Jiang, Cong Wang, and Hong Xu. 2023. Adaptive gating in mixture-of-experts based language models. *arXiv preprint arXiv:2310.07188* (2023).

[18] Yuanguo Lin, Yong Liu, Fan Lin, Lixin Zou, Pengcheng Wu, Wenhua Zeng, Huanhuan Chen, and Chunyan Miao. 2023. A survey on reinforcement learning for recommender systems. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[19] Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6566–6576.

[20] Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Why I like it: multi-task learning for recommendation and explanation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 4–12.

[21] Sichun Luo, Bowei He, Haohan Zhao, Wei Shao, Yanlin Qi, Yinya Huang, Aojun Zhou, Yuxuan Yao, Zongpeng Li, Yuanzhang Xiao, et al. 2024. Recranker: Instruction tuning large language model as ranker for top-k recommendation. *ACM Transactions on Information Systems* (2024).

[22] Haokai Ma, Ruobing Xie, Lei Meng, Fuli Feng, Xiaoyu Du, Xingwu Sun, Zhanhui Kang, and Xiangxu Meng. 2024. Negative Sampling in Recommendation: A Survey and Future Directions. *arXiv preprint arXiv:2409.07237* (2024).

[23] 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 on knowledge discovery & data mining*. 1930–1939.

[24] Kabir Nagrecha, Lingyi Liu, Pablo Delgado, and Prasanna Padmanabhan. 2023. Intune: Reinforcement learning-based data pipeline optimization for deep recommendation models. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 430–442.

[25] Efrat Nechushtai, Rodrigo Zamith, and Seth C Lewis. 2024. More of the same? Homogenization in news recommendations when users search on Google, YouTube, Facebook, and Twitter. *Mass Communication and Society* 27, 6 (2024), 1309–1335.

[26] Aleksandr V Petrov and Craig Macdonald. 2023. Generative sequential recommendation with gptrec. *arXiv preprint arXiv:2306.11114* (2023).

[27] Mostafa Rahmani, James Caverlee, and Fei Wang. 2023. Incorporating time in sequential recommendation models. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 784–790.

[28] 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.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[30] Jianling Wang, Kaize Ding, Liangjie Hong, Huan Liu, and James Caverlee. 2020. Next-item recommendation with sequential hypergraphs. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1101–1110.

[31] Pengda Wang. 2022. Recommendation algorithm in TikTok: Strengths, dilemmas, and possible directions. *Int'l J. Soc. Sci. Stud.* 10 (2022), 60.

[32] Yuhao Wang, Ha Tsz Lam, Yi Wong, Ziru Liu, Xiangyu Zhao, Yichao Wang, Bo Chen, Huifeng Guo, and Ruiming Tang. 2023. Multi-task deep recommender systems: A survey. *arXiv preprint arXiv:2302.03525* (2023).

[33] Lianghao Xia, Chao Huang, Yong Xu, and Jian Pei. 2022. Multi-behavior sequential recommendation with temporal graph transformer. *IEEE Transactions on Knowledge and Data Engineering* 35, 6 (2022), 6099–6112.

[34] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Jiajie Xu, Victor S Sheng S. Sheng, Zhiming Cui, Xiaofang Zhou, and Hui Xiong. 2019. Recurrent convolutional neural network for sequential recommendation. In *The world wide web conference*. 3398–3404.

[35] Jiahui Xu, Lu Sun, and Dengji Zhao. 2024. Mome: Mixture-of-masked-experts for efficient multi-task recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2527–2531.

[36] Xiaoran Xu, Laming Chen, Songpeng Zu, and Hanning Zhou. 2018. Hulu video recommendation: from relevance to reasoning. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 482–482.

[37] Yanwu Yang and Panyu Zhai. 2022. Click-through rate prediction in online advertising: A literature review. *Information Processing & Management* 59, 2 (2022), 102853.

[38] Zhen Yang, Ming Ding, Tinglin Huang, Yukuo Cen, Junshuai Song, Bin Xu, Yuxiao Dong, and Jie Tang. 2024. Does negative sampling matter? a review with insights into its theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[39] Enming Yuan, Wei Guo, Zhicheng He, Huifeng Guo, Chengkai Liu, and Ruiming Tang. 2022. Multi-behavior sequential transformer recommender. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1642–1652.

[40] Hengyu Zhang, Junwei Pan, Dapeng Liu, Jie Jiang, and Xiu Li. 2024. Deep Pattern Network for Click-Through Rate Prediction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1189–1199.

[41] Junjie Zhang, Ruobing Xie, Hongyu Lu, Wenqi Sun, Xin Zhao, Zhanhui Kang, et al. [n. d.]. Frequency-Augmented Mixture-of-Heterogeneous-Experts Framework for Sequential Recommendation. In *THE WEB CONFERENCE 2025*.

[42] Haiyuan Zhao, Lei Zhang, Jun Xu, Guohao Cai, Zhenhua Dong, and Ji-Rong Wen. 2023. Uncovering user interest from biased and noised watch time in video recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 528–539.

[43] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.