Length-Adaptive Interest Network for Balancing Long and Short Sequence Modeling in CTR Prediction

Zhicheng Zhang[1*], Zhaocheng Du[2*], Jieming Zhu[2†], Jiwei Tang[1], Fengyuan Lu[3], Wang Jiaheng[4], Song-Li Wu[1], Qianhui Zhu[1], Jingyu Li[5], Hai-Tao Zheng[1,6†], Zhenhua Dong[2]

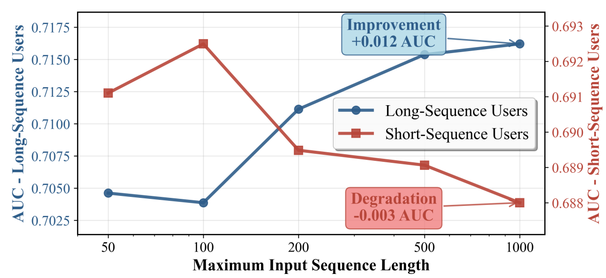
[1]Shenzhen International Graduate School, Tsinghua University
[2]Huawei Noah's Ark Lab
[3]School of Artificial Intelligence and Science, Nanjing University
[4]Hong Kong University of Science and Technology
[5]School of Cyber Science and Technology, Sun Yat-sen University
[6]Peng Cheng Laboratory
zhang-zc24@mails.tsinghua.edu.cn, zhaochengdu@huawei.com, jiemingzhu@ieee.org, zheng.haitao@sz.tsinghua.edu.cn

Abstract

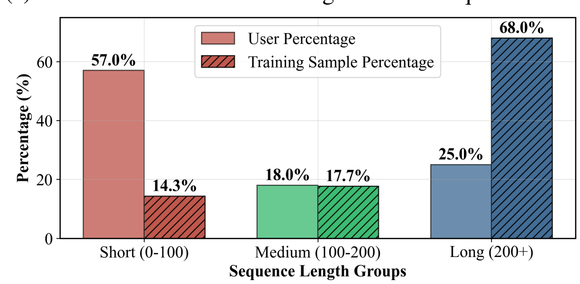
User behavior sequences in modern recommendation systems exhibit significant length heterogeneity, ranging from sparse short-term interactions to rich long-term histories. While longer sequences provide more context, we observe that increasing the maximum input sequence length in existing CTR models paradoxically degrades performance for short-sequence users due to attention polarization and length imbalance in training data. To address this, we propose **LAIN** (Length-Adaptive Interest Network), a plug-and-play framework that explicitly incorporates sequence length as a conditioning signal to balance long- and short-sequence modeling. LAIN consists of three lightweight components: a *Spectral Length Encoder* that maps length into continuous representations, *Length-Conditioned Prompting* that injects global contextual cues into both long- and short-term behavior branches, and *Length-Modulated Attention* that adaptively adjusts attention sharpness based on sequence length. Extensive experiments on three real-world benchmarks across five strong CTR backbones show that LAIN consistently improves overall performance, achieving up to 1.15% AUC gain and 2.25% log loss reduction. Notably, our method significantly improves accuracy for short-sequence users without sacrificing long-sequence effectiveness. Our work offers a general, efficient, and deployable solution to mitigate length-induced bias in sequential recommendation.

Introduction

Click-through rate (CTR) prediction is a core task in personalized recommendation and online advertising systems. Accurately capturing user preferences from historical behavior sequences has shown to significantly improve prediction performance (Zhou et al. 2018a, 2019; Yang et al. 2023; Lu et al. 2025a). As user engagement grows and digital content consumption increases, user behavior sequences have become longer and more diverse. Consequently, a growing body of work has focused on long-sequence modeling frameworks (Pi et al. 2020; Chang et al. 2023; Si et al. 2024), which typically adopt a two-stage architecture: a general search unit (GSU) that selects representative behaviors relevant to the target item from the full history, and an exact



(a) Performance Imbalance : Long- vs Short-Sequence Users



(b) Data Imbalance : Users vs Training Samples

Figure 1: (a) The effect of varying the maximum input sequence length on average AUC across five CTR models on the *EBNeRD-small* dataset. Longer sequences improve performance for users with extensive behavior histories ($>$ 200), while consistently degrading it for users with fewer interactions ($<$ 100). (b) The user behavior length distribution reveals a skewed pattern: short-sequence users form the majority, yet long-sequence users dominate the training sample.

[*]These authors contributed equally.
[†]Corresponding Authors: Jieming Zhu and Hai-Tao Zheng.

search unit (ESU) that applies target-aware attention over selected behaviors to derive a refined long-sequence representation. In parallel, a short-sequence branch often extracts the most recent interactions for high-resolution modeling of recent interests. The fused representation is then used for CTR prediction. This framework has been widely deployed in industrial systems, demonstrating its ability to leverage rich contextual information from long behavior sequences.

Despite its success, we identify a counterintuitive phenomenon in deployment: *as the maximum input sequence length increases, the prediction performance for long-sequence users improves, but the performance for short-sequence users degrades notably*. As shown in Figure 1a, users with long histories benefit from longer inputs, whereas users with fewer behaviors suffer from decreased AUC. This phenomenon is particularly concerning since short-sequence users often constitute the majority. As illustrated in Figure 1b, 57% of users contribute only 14.3% of training samples, while just 25% of users with long sequences dominate over 68% of training data.

This imbalance poses a practical dilemma. Models are implicitly biased toward optimizing for high-frequency long-sequence users, whose patterns dominate the training data. Yet in reality, short-sequence users are more prevalent(Yin et al. 2020). Standard CTR models do not explicitly differentiate between users of varying sequence lengths. Instead, they apply a one-size-fits-all architecture across the entire spectrum of sequence lengths, assuming that user behaviors are homogeneously structured (Mao et al. 2023; Liu et al. 2023). As a consequence, new or inactive users—a critical cohort for platform growth—may suffer degraded recommendation quality, limiting system-wide engagement and retention.

To address this issue, we conduct an in-depth diagnosis (Section Empirical Analysis) and identify two key factors. The first is *attention polarization*: softmax-based attention tends to over-concentrate on a few salient behaviors in long sequences, which benefits interest extraction but amplifies noise in short sequences with limited context. The second is *length signal deficiency*: existing models treat behavior sequences as homogeneous event sets, failing to leverage sequence length as a prior that reflects user status. Together, these effects impair short-sequence modeling and hinder balanced performance across user groups.

Motivated by these findings, we propose that *sequence length should be explicitly modeled as a conditional signal* to guide interest extraction and attention behavior adaptively. To this end, we introduce the **Length-Adaptive Interest Network (LAIN)**, a lightweight, plug-and-play framework that injects length-awareness into CTR prediction. LAIN comprises three key components: a *Spectral Length Encoder* that maps raw length values into rich embeddings via continuous Fourier bases; a *Length-Conditioned Prompting* module that generates prompt tokens conditioned on sequence length to modulate user representations; and a *Length-Modulated Attention* mechanism that adaptively adjusts attention sharpness by modulating the temperature of the softmax function and augmenting key-query representations based on length priors.

By design, LAIN is compatible with mainstream CTR models and can be seamlessly integrated into existing architectures with negligible inference overhead. Our approach enables a shared model to dynamically adapt its attention and representation strategies to user sequences of varying lengths. Extensive experiments on multiple large-scale public datasets demonstrate that LAIN consistently improves both overall CTR performance and, crucially, significantly boosts prediction quality for short-sequence users. We believe this study highlights an important yet under-explored dimension of sequence modeling in recommendation systems and provides a generalizable paradigm for length-adaptive user modeling.

Our contributions are summarized as follows:

- We identify and analyze the underexplored challenge of length imbalance in CTR modeling and characterize its effects as *attention polarization* and *length signal deficiency*.
- We propose LAIN, a novel, lightweight and modular length-aware enhancement framework with three components that explicitly encode, inject, and modulate sequence length in user modeling.
- We conduct extensive experiments demonstrating that LAIN substantially improves both long- and short-sequence modeling, and provides a generalizable solution to mitigating length-induced bias in CTR prediction.

Related Work

Long-sequence CTR models. Modern recommender systems often handle very long user behavior sequences using a two-stage architecture: a General Search Unit (GSU) retrieves a subset of relevant behaviors, and an Exact Search Unit (ESU) applies target-aware attention for final modeling. Prominent models following this paradigm include SIM (Pi et al. 2020), ETA (Chen et al. 2022), SDIM (Cao et al. 2022), and TWIN (Chang et al. 2023). The more recent TWIN-V2 further introduces hierarchical clustering to compress life-cycle behaviors and allows modeling of ultra-long sequences in a scalable manner (Si et al. 2024). However, such frameworks focus primarily on expanding sequence length and search efficiency while largely overlooking the performance imbalance between long- and short-sequence users. Recent work DARE (Feng et al. 2025) critiques this by decoupling attention and representation embeddings to mitigate interference, but still does not condition on sequence length during modeling.

Length imbalance and fairness in recommendation. Despite the widespread presence of long-tail behavior distributions (Luo et al. 2023), few prior works explicitly address the modeling imbalance between short- and long-history users. Traditional CTR models optimize aggregate metrics like AUC or log loss, implicitly favoring data-rich long-history users, while potentially under-serving the majority short-history cohort (Zhu et al. 2021). Although fairness and cold-start literature address new or infrequent users (Lee et al. 2019; Jiang et al. 2024), to our knowledge no existing work systematically diagnoses the phenomenon of length imbalance or proposes conditioning the model on sequence length to balance across user groups.

Prompting and conditional encoding in recommendation. Recent research has explored soft prompt and prefix mechanisms to inject auxiliary context into sequential recommendation (Wang et al. 2025). For example, Personalized Prompt for Sequential Recommendation (PPR) generates soft prefix prompts from user profiles to improve cold-start performance (Wu et al. 2024). Other studies propose custom prompt tuning atop pre-trained recommendation models to enable efficient adaptation (Lu et al. 2025b; Han et al. 2024). Unlike these methods, our Length-Conditioned Prompting dynamically generates prompts conditioned on behavior sequence length—not on user profile or item metadata—allowing the model to adapt its attention strategy according to implicit user state.

Empirical Analysis

In this section, we conduct an in-depth empirical investigation to reveal why the widely-adopted long-short sequence modeling paradigm may systematically underperform for short-sequence users, despite improving the overall performance. Our diagnosis identifies two key factors: *Attention Polarization* and *Length Signal Deficiency*. These factors indicate that user sequence length is not just a passive statistic, but an active condition that influences model behavior, and must be explicitly modeled.

Attention Polarization Transformer-based CTR models rely on attention mechanisms—particularly target-aware attention—to extract user preferences from historical behavior sequences. However, the use of softmax attention inherently involves a normalization constraint over all tokens in the sequence. As the input sequence grows longer, the softmax distribution becomes more *polarized*, concentrating attention weights on a few "salient" items and suppressing the rest. This effect, visualized via Gini coefficient or entropy measurements, becomes more pronounced with longer input lengths (see Figure 2).

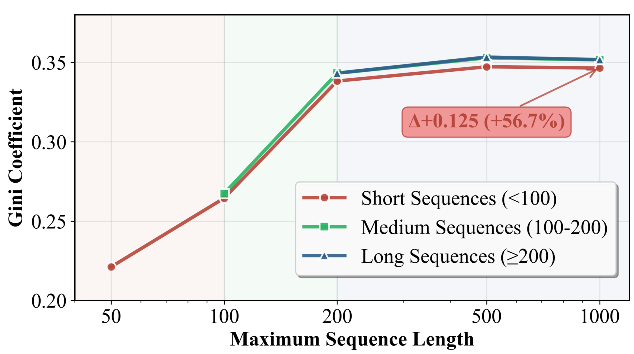


Figure 2: Attention polarization trend in baseline CTR models. Gini coefficient increases dramatically with sequence length, demonstrating progressive attention concentration that particularly affects short-sequence users.

While this selective focus may benefit long-history users by isolating strong preferences, it proves detrimental for short-history users. When only a few behaviors are available (e.g., < 10 items), softmax often places overwhelming weight on one or two items, effectively ignoring others. This leads to underutilization of limited signals, resulting in unstable and brittle interest representations.

Such behavior aligns with findings in other domains, such as language modeling and vision, where attention entropy is negatively correlated with input length (Tang et al. 2025a,b). In our CTR setting, this "attention collapse" disproportionately harms short-sequence users and necessitates strategies that regulate attention sharpness as a function of sequence length.

Metric	Short (< 100)	Medium (100–200)	Long (≥ 200)
User Distribution			
User Count	10,735	3,379	4,713
Percentage (%)	57.0	18.0	25.0
Sequence Characteristics			
Avg Length	37.1	144.2	401.4
Length Std	26.9	28.7	183.3
Behavioral Patterns			
Click Rate (%)	8.28	8.60	9.41
Click Rate Std (%)	4.19	3.04	2.52
Behavioral Stability	0.515	0.354	0.268
Content Engagement			
Avg Unique Types	2.5	3.6	4.6
Avg Sessions/User	3.1	7.5	15.7
Avg Impr./User	5.5	13.4	32.0

Table 1: Length Signal Deficiency: Empirical Evidence from User Behavioral Analysis. Behavioral Stability measured by Coefficient of Variation of click rates; lower values indicate more stable behavior. Data from EBNeRD-small.

Length Signal Deficiency Another critical issue is that most CTR models treat behavior sequences as homogeneous unordered event sets and omit sequence length from the modeling process. This design overlooks the fact that sequence length is a powerful proxy for a user's latent state.

Empirically, we observe strong correlations between sequence length and user behavioral characteristics: (1) **short-sequence users** tend to be new, cold-start, or low-activity users with unstable interests and lower click-through rates; (2) **long-sequence users** exhibit higher activity, greater interest diversity, and more stable long-term preferences.

Our comprehensive analysis on the Ebnerd dataset reveals significant disparities across sequence length groups. Short-sequence users constitute 57.0% of the user base but achieve only 8.28% click-through rate, while long-sequence users (200+ behaviors) represent 25.0% of users with 9.41% click-through rate—a substantial 1.13 percentage point improvement. More critically, short-sequence users exhibit high behavioral instability (coefficient of variation: 0.515) compared to long-sequence users (0.268), indicating that their preferences are indeed more volatile and harder to model consistently. Furthermore, content engagement patterns differ markedly: short-sequence users interact with an average of 2.5 article types, while long-sequence users engage with 4.6 types, demonstrating the diversity gap mentioned above.

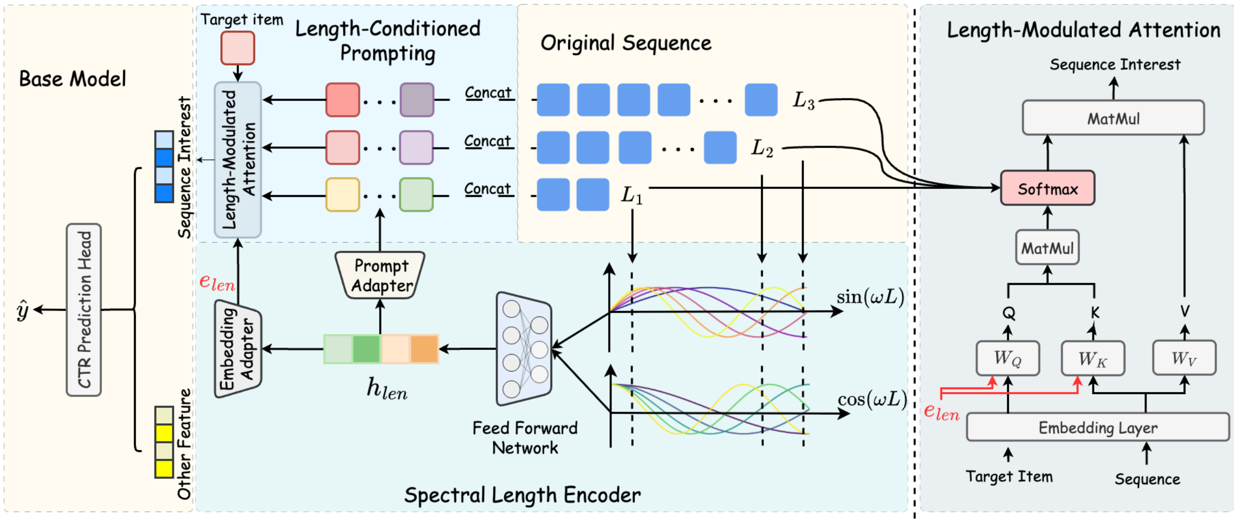


Figure 3: Overview of the Length-Adaptive Interest Network (LAIN) for CTR prediction. LAIN conditions on sequence length via a Spectral Length Encoder (SLE), which generates a continuous embedding $\mathbf{h}_{\text{len}}$ from the raw length L. This embedding modulates two components: Length-Conditioned Prompting (LCP), which prepends length-aware prompt tokens to the behavior sequence, and Length-Modulated Attention (LMA), which adjusts attention via query/key conditioning and dynamic temperature scaling. The resulting sequence representation is fused with other features for final CTR prediction.

Despite this, current models apply the same architecture and attention logic regardless of sequence length, lacking any conditional adaptation. This implicitly assumes that a user with 3 interactions and a user with 300 interactions can be modeled with identical inductive biases. As a result, the model overfits to patterns seen in data-rich long-history users (who dominate the training sample size) and underperforms on the majority of short-history users.

We argue that this is a manifestation of **length signal deficiency**: failing to condition on sequence length prevents the model from distinguishing between user states at different stages of the behavioral lifecycle. As a result, model capacity is inefficiently allocated, and generalization suffers.

Together, attention polarization and length signal deficiency point to the need for *length-aware modeling*. Models should not only **observe** sequences but also **understand how long** they are—and what that implies about the user.

Method: Length-Adaptive Interest Network

Preliminaries CTR prediction estimates the probability of user-item interactions in recommendation systems, serving as a cornerstone for enhancing user engagement through personalized content delivery and optimizing business outcomes via targeted audience matching. Formulated as a binary classification task, the goal is to learn a predictor $f : \mathbb{R}^d \to \mathbb{R}$ from training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ concatenates user, item, and contextual features, and $y_i \in \{0, 1\}$ indicates click events. The predicted CTR is:

$$\hat{y}_i = \sigma(\phi(\mathbf{x}_i)), \tag{1}$$

with $\sigma(\cdot)$ as the sigmoid function. Training minimizes the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right], \tag{2}$$

The primary challenge for *long-sequence CTR* lies in designing $\phi(\cdot)$ to capture long-range temporal dependencies in user behavior while ensuring computationally efficient training and inference on large-scale sequences.

The architecture for CTR prediction with long sequence modeling follows a structured approach encompassing four core components: **(1)** feature embedding that maps categorical features to dense vectors; **(2)** feature interaction modeling to capture cross-field dependencies; **(3)** short-term interest modeling that extracts recent behavioral patterns; and **(4)** long-term interest modeling that leverages extended user history through two-stage architectures (GSU + ESU). LAIN enhances this framework by injecting length-awareness into both short-term and long-term interest modeling components.

CTR models typically share a single parameter set θ to process all sequences regardless of length. However, users with different behavior lengths may exhibit fundamentally different modeling needs. Specifically, short-sequence users benefit from smooth, distributed attention to maximize limited signal utilization, whereas long-sequence users require sparse, concentrated attention to focus on key interests. This induces conflicting gradient signals during training:

$$\nabla_\theta \mathcal{L}_{\text{short}} \propto -\alpha \cdot \text{smoothness gradient}, \tag{3}$$

$$\nabla_\theta \mathcal{L}_{\text{long}} \propto +\beta \cdot \text{sparsity gradient}, \quad \alpha, \beta > 0, \tag{4}$$

leading to negative inner products $\langle \nabla \mathcal{L}_{\text{short}}, \nabla \mathcal{L}_{\text{long}} \rangle < 0$ and degraded generalization.

LAIN mitigates this by explicitly conditioning the model on sequence length L and decoupling the optimization space via dynamic, length-dependent parameters. Specifically, we decompose model parameters into shared and adaptive components:

$$\theta = \{\theta_{\text{shared}}, \theta_{\text{prompt}}(L)\}, \tag{5}$$

where $\theta_{\text{prompt}}(L)$ are generated on-the-fly via a learnable function of L and injected as prompts into the model. The loss becomes:

$$\mathcal{L} = \sum_{L} \sum_{(u,v_t,y)\in\mathcal{D}_L} \ell(f_{\theta_{\text{shared}}}([\mathbf{S}_u; \mathbf{P}(L)], v_t), y), \tag{6}$$

where $\mathbf{P}(L)$ denotes the prompt tokens conditioned on L. This decouples gradient flows for users with varying sequence lengths and enables length-specific inductive biases.

Spectral Length Encoder (SLE) We first encode the sequence length L into a dense vector via trainable Fourier projections:

$$\mathbf{f}_{\text{fourier}} = [\sin(L \cdot \boldsymbol{\omega}); \cos(L \cdot \boldsymbol{\omega})] \in \mathbb{R}^{2d_f}, \tag{7}$$

where $\boldsymbol{\omega} \in \mathbb{R}^{d_f}$ are learnable frequency parameters. This representation avoids discretization and captures continuous length semantics (Tancik et al. 2020). It is projected via an MLP to produce a shared embedding:

$$\mathbf{h}_{\text{len}} = \text{MLP}(\text{LayerNorm}(\text{Linear}(\mathbf{f}_{\text{fourier}}))) \in \mathbb{R}^{d}. \tag{8}$$

Length-Conditioned Prompting (LCP) We generate k learnable prompt tokens from $\mathbf{h}_{\text{len}}$:

$$\mathbf{P}(L) = \text{reshape}(\text{MLP}_{\text{prompt}}(\mathbf{h}_{\text{len}})) \in \mathbb{R}^{k\times d}. \tag{9}$$

These tokens are prepended to both short-term and long-term behavior sequences:

$$\mathbf{S}'_{\text{short}} = [\mathbf{P}; \mathbf{S}_{\text{short}}], \quad \mathbf{S}'_{\text{long}} = [\mathbf{P}; \mathbf{S}_{\text{long}}], \tag{10}$$

and then passed to the attention encoder. LCP effectively injects global user state and expands the parameterization space, thus avoiding interference across sequences of varying length.

Length-Modulated Attention (LMA) To further improve adaptivity, we modulate the attention computation using $\mathbf{h}_{\text{len}}$.

Query-Key Conditioning. We concatenate $\mathbf{h}_{\text{len}}$ to each query and key:

$$\mathbf{Q}' = W_q([\mathbf{Q}; \mathbf{e}_{\text{len}}]), \tag{11}$$

$$\mathbf{K}' = W_k([\mathbf{K}; \mathbf{e}_{\text{len}}]), \tag{12}$$

where $\mathbf{e}_{\text{len}} = \text{MLP}_{\text{emb}}(\mathbf{h}_{\text{len}})$ is a length embedding generated by a small MLP.

Softmax Temperature Scaling. We define a length-aware temperature:

$$\tau = 1 + \text{sigmoid}(-\beta(L - L_0)) \cdot \gamma, \tag{13}$$

where L_0 is the mean sequence length, and γ and β are learnable parameters. The attention weights are computed as:

$$\alpha_{ij} = \frac{\exp\left(\frac{\mathbf{Q}'_i \cdot \mathbf{K}'_j}{\sqrt{d}\cdot\tau}\right)}{\sum_j \exp\left(\frac{\mathbf{Q}'_i \cdot \mathbf{K}'_j}{\sqrt{d}\cdot\tau}\right)}, \tag{14}$$

yielding output:

$$\mathbf{O}_i = \sum_j \alpha_{ij} \cdot \mathbf{V}_j. \tag{15}$$

When L is small, τ increases and smooths the attention distribution, reducing over-polarization. When L is large, τ shrinks, promoting sharper focus on salient items.

Training and Complexity Analysis LAIN is trained end-to-end with minimal overhead. The Spectral Length Encoder (SLE) introduces $\mathcal{O}(d_f d + d^2)$ parameters. Length-Conditioned Prompting (LCP) adds $k \cdot d + \mathcal{O}(kd^2)$ parameters and increases the sequence length from L to $L + k$. Length-Modulated Attention (LMA) contributes $\mathcal{O}(d^2)$ parameters for the conditioned projections and MLP, plus scalar parameters for temperature scaling.

The overall attention complexity is $\mathcal{O}((L + k)^2 d)$. Since k is a small constant ($k = 2 \sim 4$), this simplifies to $\mathcal{O}(L^2 d)$, with only negligible linear terms added. In total, LAIN introduces less than 1.5% additional parameters and incurs about 2.3% inference time overhead compared to the base transformer model.

Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposed *Length-Conditioned Prompting* framework.

Datasets. We evaluate our method on three real-world sequential CTR datasets: **KuaiVideo** (Chen et al. 2018), **MicroVideo1.7M** (Chen et al. 2018), and **EBNeRD-small** (Kruse et al. 2024). These datasets span diverse domains—short videos and news—and exhibit long-tailed user sequence distributions, making them suitable for evaluating length-aware recommendation methods. The datasets statistics are shown in Table 2.

Dataset	Users	Items	Interactions	Avg-Len
MicroVideo1.7M	10,951	1,704,880	4,287,008	148
KuaiVideo	10,000	3,239,534	4,664,549	278
EBNeRD-small	18,828	20,739	5,514,689	147

Table 2: Datasets statistics.

Experimental Setup

Baselines. We compare LAIN-enhanced models against five representative long-sequence CTR baselines that cover the spectrum of attention-based user modeling approaches: **(1)** DIN (Zhou et al. 2018b) uses target-aware attention for interest extraction; **(2)** DIEN (Zhou et al. 2019) models interest evolution with GRU and attention; **(3)** SIM (Pi et al. 2020) introduces two-stage architecture with general and exact search units; **(4)** SDIM (Cao et al. 2022) enhances SIM with multi-head attention; and **(5)** TWIN (Chang et al. 2023) employs twin towers for long-sequence modeling. All baselines are implemented in the FuxiCTR[1] framework with identical feature engineering and hyperparameter settings for fair comparison (Zhu et al. 2022).

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

Model	EBNeRD-small			KuaiVideo			MicroVideo1.7M		
	GAUC↑	AUC↑	logloss↓	GAUC↑	AUC↑	logloss↓	GAUC↑	AUC↑	logloss↓
DIN	0.7053	0.7080	0.2695	0.6716	0.6979	0.4498	0.7023	0.7093	0.4196
+LAIN	0.7096	0.7156	0.2678	0.6742	0.7043	0.4457	0.7019	0.7101	0.4179
Rel. Gain	**0.61%**	**1.07%**	**-0.65%**	**0.40%**	**0.93%**	**-0.92%**	-0.05%	**0.12%**	**-0.40%**
DIEN	0.7090	0.7099	0.2737	0.6678	0.6977	0.4509	0.7143	0.7185	0.4171
+LAIN	0.7078	0.7118	0.2676	0.6688	0.6991	0.4476	0.7146	0.7219	0.4161
Rel. Gain	-0.17%	**0.27%**	**-2.25%**	**0.14%**	**0.20%**	**-0.71%**	**0.04%**	**0.47%**	**-0.23%**
SIM	0.6960	0.6992	0.2720	0.6672	0.6875	0.4574	0.7017	0.7095	0.4168
+LAIN	0.6990	0.7037	0.2706	0.6678	0.6896	0.4566	0.7070	0.7131	0.4151
Rel. Gain	**0.43%**	**0.65%**	**-0.50%**	**0.08%**	**0.31%**	**-0.16%**	**0.75%**	**0.50%**	**-0.41%**
SDIM	0.7039	0.7099	0.2719	0.6729	0.6924	0.4536	0.6984	0.7161	0.4129
+LAIN	0.7099	0.7123	0.2704	0.6732	0.6949	0.4525	0.6993	0.7163	0.4121
Rel. Gain	**0.85%**	**0.34%**	**-0.56%**	**0.05%**	**0.35%**	**-0.23%**	**0.13%**	**0.03%**	**-0.18%**
TWIN	0.6930	0.6993	0.2718	0.6729	0.6918	0.4530	0.7060	0.7158	0.4164
+LAIN	0.7012	0.7074	0.2698	0.6748	0.6976	0.4508	0.7093	0.7233	0.4097
Rel. Gain	**1.19%**	**1.15%**	**-0.73%**	**0.29%**	**0.84%**	**-0.48%**	**0.47%**	**1.05%**	**-1.63%**

Table 3: Overall performance comparison on three real-world datasets. "+LAIN" denotes the model enhanced with our LAIN method, and "Rel. Gain" reports the relative improvement (%) over the base model. ↑ indicates higher is better; ↓ indicates lower is better.

Implementation. Following standard CTR prediction protocols, we configure all models with maximum sequence length of 1000 to handle long user histories. Feature embeddings use dimension 64, and models are optimized using Adam optimizer with learning rate 0.001 and early stopping based on validation AUC to prevent overfitting. The CTR prediction head employs a simple MLP with ReLU activation and dropout (0.2) for regularization. For the Spectral Length Encoder, we set Fourier dimension $d_f = 32$ and use a 2-layer MLP with hidden dimension 512 for length embedding projection. The Length-Conditioned Prompting module generates $k = 4$ prompt tokens with embedding dimension matching the backbone model (64). For Length-Modulated Attention, we initialize temperature scaling parameters $\gamma = 0.5$ and $\beta = 0.01$, with L_0 set as the dataset mean sequence length. The query-key conditioning uses a single linear layer to project concatenated features.

Metrics. We employ traditional CTR metrics to comprehensively evaluate LAIN's effectiveness. For overall CTR performance, we report: **(1)** AUC (Area Under ROC Curve) measuring ranking quality; **(2)** GAUC (Group AUC) computing per-user AUC then averaging to mitigate high-traffic user dominance; and **(3)** LogLoss measuring calibration quality for probabilistic predictions.

Overall Performance

Table 3 shows that integrating LAIN into each backbone model leads to consistent improvements on all datasets. Notably, GAUC increases by up to +1.2% and logloss drops by up to -1.6%, demonstrating that LAIN enhances sequence modeling capacity without sacrificing robustness. Variance arises from differing average sequence-length distributions and attention architectures; gains are larger for two-stage attention backbones (SIM/SDIM/TWIN) that expose this imbalance more clearly.

Length-Specific Evaluation

We further analyze TWIN's performance across short, medium, and long sequences. Table 4 shows that LAIN achieves consistent gains across all buckets, especially for short sequences (<100), where AUC improves by +1.08% and logloss reduces by -2.17%. This supports our hypothesis that prompt-based length conditioning better adapts to both cold-start and long-history users by dynamically adjusting attention focus.

Attention Polarization Mitigation

We quantify attention polarization by computing the Gini coefficient of attention scores across sequences of varying lengths. As demonstrated in our empirical analysis (Section 2), baseline models exhibit significant attention polarization that increases with maximum sequence length, particularly affecting short-sequence users.

Table 5 provides a comprehensive quantitative analysis of how LAIN mitigates this attention polarization problem. We compare the Gini coefficients across different sequence length groups for baseline models configured with varying maximum sequence lengths (200, 500, 1000) against our proposed LAIN model. These results verify that LAIN effectively mitigates over-concentration in softmax attention—an issue exacerbated by sequence length—while maintaining the model's capacity to focus on relevant behavioral signals.

Length	TWIN			+LAIN			Improvement		
	GAUC	AUC	Logloss	GAUC	AUC	Logloss	ΔGAUC	ΔAUC	ΔLogloss
0-100	0.6951	0.7108	0.4120	0.6986	0.7184	0.4031	+0.51%	+1.08%	-2.17%
100-200	0.7068	0.7212	0.4037	0.7100	0.7248	0.3986	+0.45%	+0.50%	-1.26%
200+	0.7133	0.7125	0.4395	0.7166	0.7238	0.4319	+0.46%	+1.58%	-1.74%
Overall	0.7060	0.7158	0.4164	0.7093	0.7233	0.4097	+0.47%	+1.05%	-1.63%

Table 4: Overall performance of TWIN(MicroVideo1.7M) with and without LAIN across different sequence lengths.

Model Configuration	**Short** (< 100)	**Medium** (100–200)	**Long** (≥ 200)
Baseline Models (Different Max Sequence Lengths)			
Baseline (L=200)	0.338	0.343	0.343
Baseline (L=500)	0.347	0.353	0.353
Baseline (L=1000)	0.346	0.351	0.352
Proposed LAIN Model			
LAIN	0.318	0.322	0.321
Improvement Analysis			
Variance Reduction	**50.6%** (vs. worst baseline)		
Range Reduction	**33.3%** (vs. worst baseline)		

Table 5: Attention Polarization Mitigation Analysis: LAIN vs. Baseline Models. Lower Gini scores indicate less attention polarization. LAIN exhibits a more uniform attention distribution across sequence lengths.

Variant	GAUC	AUC	Logloss
LAIN (Full)	**0.7093**	**0.7233**	**0.4097**
w/o LCP	0.7083	0.7228	0.4107
w/o Query-Key Conditioning	0.7071	0.7195	0.4148
w/o Temperature Scaling	0.7082	0.7212	0.4111
w/o LMA	0.7076	0.7189	0.4157
w/o Short-term Branch	0.7077	0.7226	0.4137

Table 6: Ablation study on TWIN (MicroVideo1.7M).

Ablation Study

We conduct an ablation study on TWIN(MicroVideo1.7M) to examine the effect of each LAIN component. As shown in Table 6, removing any module degrades performance, confirming that all parts contribute to overall effectiveness. Excluding the Length-Modulated Attention (LMA), which couples temperature scaling with query-key conditioning, leads to the largest drop, indicating its importance in stabilizing attention. While temperature scaling alone slightly hurts performance, it shows strong synergy with length-conditioned prompting and query-key conditioning. Removing the short-term branch—where LAIN is applied to short behavior sequences—also reduces accuracy, highlighting the benefit of incorporating length awareness even in short contexts.

Parameter Sensitivity Analysis

We conducted a sensitivity study by jointly varying three major hyperparameters—Fourier feature dimension, hidden dimension, and the number of prompt tokens—across 12 configurations.

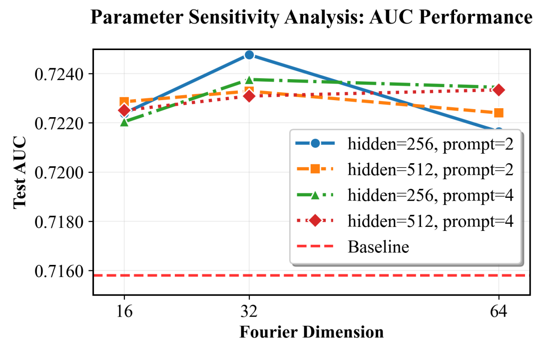


Figure 4: Parameter sensitivity analysis for AUC performance. All hyperparameter configurations consistently outperform the baseline (red dashed line), demonstrating the robustness of our approach across different parameter settings.

As shown in Figure 4, all 12 hyperparameter configurations achieve AUC scores ranging from 0.7216 to 0.7248, significantly outperforming the baseline (0.7158). This demonstrates the robustness of our method across different parameter choices. Our selected configuration (64, 512, 4) achieves competitive performance with 1.05% improvement over baseline, validating our hyperparameter selection.

Conclusion

In this paper, we present **LAIN**, a length-adaptive framework for CTR prediction that explicitly incorporates sequence length as a conditioning signal. By introducing three complementary components—Spectral Length Encoder, Length-Conditioned Prompting, and Length-Modulated Attention—LAIN enables adaptive modeling across users with diverse behavior histories. Our method addresses critical challenges such as attention polarization and gradient conflicts caused by length imbalance. Extensive experiments on multiple real-world datasets and strong baselines demonstrate that LAIN consistently improves overall performance, while significantly enhancing CTR prediction for short-sequence users without sacrificing long-sequence accuracy. We believe that LAIN provides a generalizable and practical solution for balanced interest modeling in large-scale recommendation systems.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province (Grant No. 2024TQ08X729 and 2023A1515012914), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033, JCYJ20240813112009013 and GJHZ20240218113603006), the Major Key Projectof PCL for Experiments and Applications (PCL2023A09). We also acknowledge partial support from MindSpore, a new deep learning computing framework (https://www.mindspore.cn).

References

Cao, Y.; Zhou, X.; Feng, J.; Huang, P.; Xiao, Y.; Chen, D.; and Chen, S. 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*, CIKM '22, 2974–2983. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392365.

Chang, J.; Zhang, C.; Fu, Z.; Zang, X.; Guan, L.; Lu, J.; Hui, Y.; Leng, D.; Niu, Y.; Song, Y.; and Gai, K. 2023. TWIN: TWo-stage Interest Network for Lifelong User Behavior Modeling in CTR Prediction at Kuaishou. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3785–3794. ACM. ISBN 979-8-4007-0103-0.

Chen, Q.; Xu, Y.; Pei, C.; Lv, S.; Zhuang, T.; and Ge, J. 2022. Efficient Long Sequential User Data Modeling for Click-Through Rate Prediction. arXiv:2209.12212.

Chen, X.; Liu, D.; Zha, Z.-J.; Zhou, W.; Xiong, Z.; and Li, Y. 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. ACM. ISBN 978-1-4503-5665-7.

Feng, N.; Pan, J.; Wu, J.; Chen, B.; Wang, X.; QianLi; Hu, X.; Jiang, J.; and Long, M. 2025. Long-Sequence Recommendation Models Need Decoupled Embeddings. In *The Thirteenth International Conference on Learning Representations*.

Han, Y.; Wang, H.; Wang, K.; Wu, L.; Li, Z.; Guo, W.; Liu, Y.; Lian, D.; and Chen, E. 2024. Efficient Noise-Decoupling for Multi-Behavior Sequential Recommendation. In *Proceedings of the ACM Web Conference 2024*, WWW '24, 3297–3306. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701719.

Jiang, Y.; Chen, G.; Zhang, W.; Wang, J.; Jiang, Y.; Zhang, Q.; Lin, J.; Jiang, P.; and Bian, K. 2024. Prompt Tuning for Item Cold-start Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, 411–421. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705052.

Kruse, J.; Lindskow, K.; Kalloori, S.; Polignano, M.; Pomo, C.; Srivastava, A.; Uppal, A.; Andersen, M. R.; and Frellsen, J. 2024. EB-NeRD a large-scale dataset for news recommendation. In *Proceedings of the Recommender Systems Challenge 2024*, RecSysChallenge '24, 1–11. New York, NY, USA: Association for Computing Machinery. ISBN 9798400711275.

Lee, H.; Im, J.; Jang, S.; Cho, H.; and Chung, S. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 1073–1082. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.

Liu, Q.; Yan, F.; Zhao, X.; Du, Z.; Guo, H.; Tang, R.; and Tian, F. 2023. Diffusion Augmentation for Sequential Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 1576–1586. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701245.

Lu, X.; Wang, J.; Zhu, J.; Zhang, Z.; Zou, D.; Zheng, H.-T.; Xia, S.-T.; and Zhang, R. 2025a. ROMA: Recommendation-Oriented Language Model Adaptation Using Multi-Modal Multi-Domain Item Sequences. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, 4670–4681. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714542.

Lu, Y.; Du, Z.; Li, X.; Jia, P.; Wang, Y.; Liu, W.; Wang, Y.; Guo, H.; Tang, R.; Dong, Z.; Duan, Y.; and Zhao, X. 2025b. Prompt Tuning as User Inherent Profile Inference Machine. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, 5898–5906. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720406.

Luo, S.; Ma, C.; Xiao, Y.; and Song, L. 2023. Improving Long-Tail Item Recommendation with Graph Augmentation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 1707–1716. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701245.

Mao, K.; Zhu, J.; Su, L.; Cai, G.; Li, Y.; and Dong, Z. 2023. FinalMLP: An Enhanced Two-Stream MLP Model for CTR Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4): 4552–4560.

Pi, Q.; Zhou, G.; Zhang, Y.; Wang, Z.; Ren, L.; Fan, Y.; Zhu, X.; and Gai, K. 2020. Search-Based User Interest Modeling with Lifelong Sequential Behavior Data for Click-Through Rate Prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2685–2692. ACM. ISBN 978-1-4503-6859-9.

Si, Z.; Guan, L.; Sun, Z.; Zang, X.; Lu, J.; Hui, Y.; Cao, X.; Yang, Z.; Zheng, Y.; Leng, D.; Zheng, K.; Zhang, C.; Niu, Y.; Song, Y.; and Gai, K. 2024. TWIN V2: Scaling Ultra-Long User Behavior Sequence Modeling for Enhanced CTR Prediction at Kuaishou. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4890–4897. ACM. ISBN 979-8-4007-0436-9.

Tancik, M.; Srinivasan, P. P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron,

J. T.; and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Tang, J.; Xu, J.; Lu, T.; Zhang, Z.; YimingZhao, Y.; LinHai, L.; and Zheng, H.-T. 2025a. Perception Compressor: A Training-Free Prompt Compression Framework in Long Context Scenarios. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, 4093–4108. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7.

Tang, J.; Zhang, Z.; Wu, S.; Ye, J.; Bai, L.; Wang, Z.; Lu, T.; Chen, J.; Hai, L.; Zheng, H.-T.; and Kim, H.-G. 2025b. GMSA: Enhancing Context Compression via Group Merging and Layer Semantic Alignment. arXiv:2505.12215.

Wang, X.; Du, Z.; Xu, H.; Yin, S.; Han, Y.; Zhu, J.; Zhang, K.; and Liu, Q. 2025. Personalized Visual Content Generation in Conversational Systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wu, Y.; Xie, R.; Zhu, Y.; Zhuang, F.; Zhang, X.; Lin, L.; and He, Q. 2024. Personalized Prompt for Sequential Recommendation. *IEEE Trans. on Knowl. and Data Eng.*, 36(7): 3376–3389.

Yang, B.; Gu, J.; Liu, K.; Xu, X.; Xu, R.; Sun, Q.; and Liu, H. 2023. Empowering General-purpose User Representation with Full-life Cycle Behavior Modeling. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, 2908–2917. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.

Yin, J.; Liu, C.; Wang, W.; Sun, J.; and Hoi, S. C. 2020. Learning Transferrable Parameters for Long-tailed Sequential User Behavior Modeling. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, 359–367. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379984.

Zhou, G.; Mou, N.; Fan, Y.; Pi, Q.; Bian, W.; Zhou, C.; Zhu, X.; and Gai, K. 2019. Deep Interest Evolution Network for Click-Through Rate Prediction. 33(01): 5941–5948.

Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; and Gai, K. 2018a. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1059–1068.

Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; and Gai, K. 2018b. Deep Interest Network for Click-Through Rate Prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1059–1068. ACM. ISBN 978-1-4503-5552-0.

Zhu, J.; Dai, Q.; Su, L.; Ma, R.; Liu, J.; Cai, G.; Xiao, X.; and Zhang, R. 2022. BARS: Towards Open Benchmarking for Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 2912–2923. New York, NY, USA: Association for Computing Machinery. ISBN 9781450387323.

Zhu, J.; Liu, J.; Yang, S.; Zhang, Q.; and He, X. 2021. Open Benchmarking for Click-Through Rate Prediction. In Demartini, G.; Zuccon, G.; Culpepper, J. S.; Huang, Z.; and Tong, H., eds., *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, 2759–2769. ACM.