SIREN: Unified Multi-Granularity Semantic Interaction for Multi-Modal Lifelong User Interest Modeling

Yaqian Zhang[1,*], Ruyi Yu[1,*], Tianyi Li[2,*], Bohan Liu[1], Maoquan Ye[1], Ke Wang[1], Shifeng Wen[1,†], Junwei Pan[1], Lijie Wang[1], Qi Zhou[1], Yeshou Cai[1], Chengguo Yin[1], Lifeng Wang[1], Hui Li[2], Lei Xiao[1], Haijie Gu[1]
[1]Tencent Inc., China [2]School of Informatics, Xiamen University, China
[1]{yaqianzhang, sparkletyu, leobhliu, adamye, kirkkwang, romeowen, jonaspan, lijiewang, joeyqzhou, showcai, turingyin, fandywang, shawnxiao, jerrickgu}@tencent.com
[2]litianyi@stu.xmu.edu.cn, hui@xmu.edu.cn
[*]Equal contribution. [†]Corresponding author.

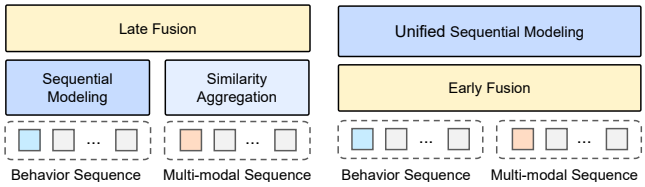
Abstract

Industrial recommender systems increasingly leverage lifelong user behavior histories and rich multi-modal content to capture evolving user preferences. However, effectively integrating multi-modal features into lifelong interest modeling remains challenging due to the inherent misalignment between multi-modal and collaborative spaces. Existing paradigms typically rely on *separate modeling* of multi-modal sequence and behavior sequence, and late fusion to alleviate the modality gap, which results in coarse-grained multi-modal representation and limited integration. In this paper, we propose SIREN, a *unified multi-granularity semantic interaction framework* for multi-modal lifelong user interest modeling. In the General Search Unit stage, we introduce two alternative retrieval strategies: multi-modal similarity-based soft retrieval for retrieval effectiveness, and Semantic ID (SemID)-based hard retrieval for efficient industrial serving. For the Exact Search Unit stage, we explicitly incorporate target-aware relevance via coarse similarity buckets and fine-grained prefix-encoded SemIDs, *enabling unified interaction* with collaborative ID features within the target-conditioned transformer architecture. Extensive experiments on the offline dataset demonstrate that SIREN achieves a state-of-the-art GAUC. Online A/B tests further demonstrate consistent GMV gains across multiple production scenarios, including +2.28% in Weixin Moments, +3.87% in Weixin Official Accounts, and +1.61% in Weixin Channels. From July 2025, SIREN has been fully launched for full-traffic serving in Tencent's advertising platform.

1 Introduction

Lifelong user behavior histories provide rich personalized signals and play a crucial role in industrial recommender systems, especially in advertising, feed, and e-commerce scenarios [10, 12]. Modeling user interests over such long-term histories has been shown to substantially improve click-through rate (CTR) prediction by capturing persistent, diverse, and evolving user preferences [3, 4, 7, 18, 21, 31]. Nevertheless, directly leveraging lifelong behaviors is computationally challenging, due to their extreme length and the strict latency constraints in online systems [4, 12, 18].

Existing long-sequence recommendation methods can be broadly categorized into two paradigms. The first line of work attempts to model long behavioral sequences end-to-end by designing efficient attention or compression mechanisms [2, 5, 8]. The second line adopts a two-stage paradigm [3, 4, 7, 18], where the first General Search Unit (GSU) stage retrieves a short target-relevant subsequence from the full behavior history, and the second Exact Search Unit (ESU) stage performs user interest modeling over the retrieved behaviors. Due to its favorable efficiency and deployment flexibility, the two-stage paradigm has become a widely adopted solution in large-scale industrial recommendation systems.



(a) Separate Modeling Paradigm **(b) Unified Modeling Paradigm**

Figure 1: Comparison between separate and unified paradigms for multi-modal lifelong user interest modeling. Existing methods usually encode ID-based behavior sequences and multi-modal sequences separately, followed by late fusion after sequence aggregation, which limits fine-grained interaction between collaborative and semantic signals. SIREN instead performs item-level early fusion, allowing multi-modal semantics and collaborative features to interact within a unified sequential modeling framework.

Recent advances in multi-modal foundation models enable high-quality content representation from diverse item modalities [1, 19, 22], motivating their adoption in lifelong user interest modeling [9, 20, 25]. Compared with ID-based signals, multi-modal semantics provide stronger generalization ability in long-tail and cold-start scenarios. However, effectively integrating multi-modal signals into recommender systems remains challenging due to the misalignment between multi-modal and collaborative spaces [14, 23]. Specifically, pre-trained multi-modal embeddings mainly capture content similarity rather than collaborative signals, and their distributions are often incompatible with ID embeddings, making naive integration prone to introducing noise and degrading recommendation performance [13, 15, 23].

To bridge this gap, existing studies typically adopt a separate modeling paradigm, where multi-modal sequences and behavior sequences are modeled individually, and a late fusion process is applied, as illustrated in Fig. 1(a). Despite the effectiveness, the prevailing separate modeling paradigm suffers from two issues:

(1) **Multi-modal signals and collaborative signals are not aligned within a unified representation space, which restricts feature interaction.** Prior methods [25] typically treat multi-modal information as an auxiliary branch and fuse it with ID-based user representations only after sequence aggregation.

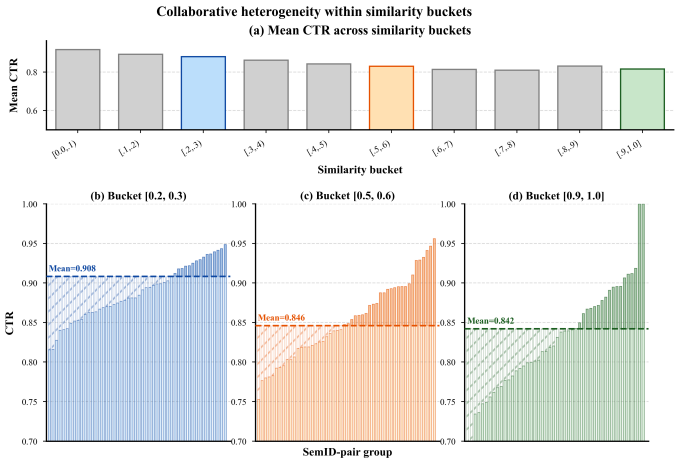


Figure 2: CTR distributions of Semantic ID groups within target–behavior similarity buckets. Although pairs in the same bucket have similar multi-modal proximity to the target item, their CTRs vary substantially across Semantic ID groups, especially in high-similarity regions. This within-bucket heterogeneity shows that similarity buckets provide only coarse target-aware relevance, motivating the use of fine-grained SemIDs to preserve semantic and collaborative distinctions.

Such late-stage fusion prevents multi-modal signals from sufficiently interacting with collaborative signals during sequential modeling. As a result, multi-modal information mainly serves as attention modulation or sequence-level augmentation [20, 25], rather than directly participating in item-level representation learning, thereby limiting the model's ability to capture more discriminative user interest representations.

(2) **Target–behavior similarity provides only a coarse-grained view and overlooks collaborative heterogeneity within multimodal proximity.** Although target–behavior similarity is effective for target-aware multi-modal interest extraction, it compresses the rich behavior–target relationship into a scalar or bucketized proximity measure [20]. This coarse representation implicitly treats behavior–target pairs with similar multi-modal proximity as similarly informative for predicting user responses. However, Fig. 2 shows that pairs within the same similarity bucket can still exhibit markedly different CTR patterns across Semantic ID groups, especially in high-similarity regions. This indicates that *multi-modal proximity captures coarse target relevance, but does not preserve the collaborative structure that governs user feedback.* Therefore, similarity-based representations alone struggle to distinguish behavior–target pairs that are close in the multi-modal space yet behave differently in the collaborative CTR space, limiting their ability to model fine-grained user interests.

To address these limitations, we propose SIREN, a unified multi-granularity framework that bridges multi-modal semantics and collaborative user interests for lifelong user modeling. Following the industrial GSU-ESU paradigm, SIREN incorporates multi-modal signals into both retrieval and exact interest modeling. In GSU, SIREN provides two complementary retrieval mechanisms: similarity-based soft retrieval, which uses dense multi-modal embeddings to retrieve target-relevant behaviors, and SemID-based hard retrieval, which uses discrete Semantic IDs as retrieval keys to enable efficient lookup in large-scale serving systems. In ESU, SIREN introduces a unified *target-conditioned Transformer* [7, 29, 32] for explicit behavior–target interaction. Instead of fusing separately modeled multi-modal and collaborative representations at a late stage, SIREN performs item-level early fusion by combining ID features, coarse target-aware similarity buckets, and fine-grained prefix-encoded SemIDs within the same target-conditioned backbone. Importantly, these signals are not only used to modulate attention weights, but are also integrated into behavior representations before sequence aggregation, *allowing behavior–target interactions to shape both attention and representation learning.* As a result, SIREN jointly captures collaborative dependencies, coarse multi-modal relevance, and fine-grained semantic heterogeneity in a unified interaction representation space.

We conduct extensive experiments on the Taobao-MM dataset and online A/B tests in the Weixin advertising system. Offline results show that SIREN consistently outperforms strong baselines in terms of GAUC, achieving the best GAUC with a relative improvement of +2.48%. Further analyses demonstrate that SIREN learns more discriminative user representations and better captures semantic heterogeneity beyond similarity-based methods. For online evaluation, SIREN achieves consistent GMV improvements across multiple Weixin production scenarios, with gains of +2.28% in Moments, +3.87% in Official Accounts, and +1.61% in Channels. The improvements are more pronounced in cold-start settings, including low-activity users and newly launched ads. These results validate the effectiveness and robustness of SIREN in real-world large-scale deployment.

In summary, our contributions are as follows:

- We propose a lifelong user interest modeling framework that effectively incorporates multi-modal signals into the industrial GSU–ESU two-stage paradigm.
- We investigate both multi-modal similarity-based soft retrieval and SemID-based hard retrieval in the GSU stage, providing a practical trade-off between retrieval quality and online serving efficiency.
- We introduce target-aware similarity buckets and prefix-encoded SemID as complementary multi-modal features for ESU modeling, enabling fine-grained early fusion with ID-based collaborative features upon a target-conditioned Transformer.
- Extensive offline experiments, representation analyses and online A/B tests demonstrate the superior effectiveness of SIREN in large-scale industrial scenarios.

2 Related Work

2.1 Lifelong User Interest Modeling

Modeling long-term user interests from lifelong behavior sequences is a fundamental challenge in industrial recommender systems. To address the scalability issue of ultra-long sequences, industrial recommenders widely adopt a two-stage paradigm (GSU and ESU)

introduced by SIM [18]. Subsequent works further improve retrieval efficiency and scalability through techniques such as locality-sensitive hashing [4], decoupled representation learning [3], hierarchical sequence compression [21], and decoupled embeddings for retrieval and ranking objectives [7]. More recently, several studies have explored end-to-end long-sequence modeling by redesigning attention mechanisms and improving computational efficiency [2, 8, 11, 12]. Despite their strong performance, these methods are still predominantly based on ID-centric representations, which often suffer from semantic cold-start issues and limited cross-domain generalization ability.

2.2 Multi-Modal Sequential Recommendation

Multi-modal signals provide semantic information beyond ID-based collaborative signals, but their integration into sequential recommendation remains challenging due to the misalignment with collaborative spaces. One line of research focuses on semantic tokenization and alignment via semantic ID learning or quantization [17, 24, 26, 27], while paying limited attention to how semantic signals are explicitly incorporated into sequential interest modeling. Another line of work utilizes target-aware multi-modal similarity to enhance sequence modeling. SimTier [20] transforms target–behavior multi-modal similarities into histogram-style representations. The state-of-the-art lifelong interest modeling method MUSE [25] further introduces Semantic-Aware Target Attention (SA-TA), which enhances target-aware interest extraction by integrating multi-modal similarity signals with behavior attention scores. Nevertheless, existing methods typically rely on coarse sequence-level aggregation and late fusion, limiting fine-grained interactions between semantic and collaborative features.

3 Preliminaries

We study lifelong user interest modeling for CTR prediction in a multi-modal setting. Given contextual information c, a user u, and a target item v_t, let $H = (b_1, \ldots, b_N)$ denote the full behavior history of length N. Following the industrial two-stage paradigm, the General Search Unit (GSU) first retrieves a target-relevant subsequence $H_t = (b_1, \ldots, b_L)$ with $L \ll N$, and the Exact Search Unit (ESU) then performs fine-grained interest modeling over H_t.

Each item b_i is represented by $x_i = (\mathbf{z}_i^{id}, e_i^{mm})$, where $\mathbf{z}_i^{id}$ denotes ID-based categorical features and $e_i^{mm} \in \mathbb{R}^d$ is a pre-trained multi-modal embedding aligned with collaborative signals through large-scale user interactions. The objective is to learn a predictive function $\hat{y}_t = P(y_t = 1 \mid H_t, v_t, u, c)$ by minimizing the binary cross-entropy loss with the ground-truth label $y_t \in \{0, 1\}$.

4 Overall Architecture

In this section, we first introduce the multi-modal feature construction in SIREN, and then present the overall two-stage architecture:

- **GSU:** We introduce two multi-modal retrieval strategies, namely *similarity-based soft search* and *SemID-based hard search*—to provide a trade-off between retrieval quality and deployment efficiency.
- **ESU:** We present a unified sequential modeling framework that seamlessly incorporates SemIDs and similarity buckets into sequence modeling, enriched by target-conditioned interaction.

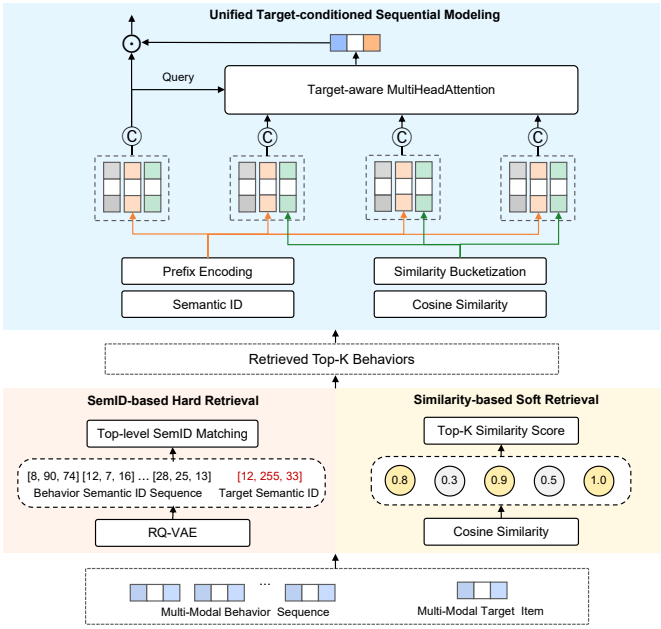


Figure 3: Overview of SIREN. SIREN follows the two-stage GSU–ESU paradigm. The GSU retrieves target-relevant behaviors via similarity-based soft retrieval or SemID-based hard retrieval. The ESU incorporates SemIDs and similarity buckets as side information for unified target-conditioned sequence modeling.

4.1 Multi-modal Feature Construction

To enable unified multi-modal interest modeling, SIREN constructs two complementary feature types from pre-trained multi-modal embeddings: SemID and target-aware similarity bucket.

4.1.1 Semantic ID Construction. To make continuous multi-modal embeddings compatible with ID-centric recommendation systems, we transform each item embedding into a discrete Semantic ID (SemID) using RQ-VAE [28]. Given $e_i^{mm} \in \mathbb{R}^d$, RQ-VAE produces a hierarchical code sequence:

$$\text{SemID}_i = \left(c_i^{(1)}, c_i^{(2)}, \ldots, c_i^{(M)}\right), \tag{1}$$

where different codes capture semantic information at different quantization levels, forming a coarse-to-fine representation of the item.

Rather than embedding each code independently, we adopt prefix encoding [30]. For SemID_i, we construct prefix tokens

$$\mathcal{P}_i = \left\{c_i^{(1)}, (c_i^{(1)}, c_i^{(2)}), \ldots, (c_i^{(1)}, \ldots, c_i^{(K)})\right\}, \tag{2}$$

where $K \le M$ is the maximum prefix depth. Each prefix $p \in \mathcal{P}_i$ is mapped to a learnable embedding via a shared lookup table, and the final semantic representation is

$$e_i^{sem} = \text{Concat}\left(\{\mathbf{E}_{\text{prefix}}[p] \mid p \in \mathcal{P}_i\}\right). \tag{3}$$

This prefix-based representation preserves hierarchical multi-modal semantics while providing a discrete interface for both efficient GSU retrieval and fine-grained ESU representation.

4.1.2 Target-aware Similarity Bucketization. While SemIDs encode item semantics, ESU modeling also requires explicit target-conditioned relevance between each historical behavior and the target item. For each behavior item b_i and target item v_t, we compute their multi-modal cosine similarity:

$$s_{i,t} = \text{sim}(e_i^{mm}, e_t^{mm}) = \frac{(e_i^{mm})^\top e_t^{mm}}{\|e_i^{mm}\| \cdot \|e_t^{mm}\|}. \tag{4}$$

The continuous score is then discretized into a bucket index:

$$q_{i,t} = \mathcal{B}(s_{i,t}) = \left\lfloor \frac{s_{i,t} - s_{\min}}{s_{\max} - s_{\min}} \cdot B \right\rfloor, \tag{5}$$

where B is the number of buckets, and $[s_{\min}, s_{\max}]$ is the effective similarity range estimated from data statistics. Values outside this range are clipped to boundary buckets. Each bucket index is mapped to a learnable embedding:

$$e_{i,t}^{Sim} = \text{Emb}^{sim}(q_{i,t}). \tag{6}$$

4.2 General Search Unit

A key requirement of the GSU stage is to balance retrieval efficiency with relevance quality when operating over ultra-long sequences. Conventional ID-based retrieval strategies suffer from limited semantic expressiveness and fail to capture content-level relevance. To address these limitations, we explore two multi-modal retrieval strategies to enhance the quality of candidate behavior:

Similarity-based Soft Retrieval. Following [25], we retrieve the top-K behaviors based on the cosine similarity between the multi-modal embeddings of historical items and the target item:

$$\mathcal{S}_{\text{sim}} = \text{Top-K}_{b_i \in \mathcal{B}} \ \text{sim}(e_i^{mm}, \ e_t^{mm}), \tag{7}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. While effectively capturing dense content relevance, this strategy incurs heavy online computational overhead due to full-sequence similarity matching. Moreover, maintaining real-time multi-modal embedding indices introduces significant storage and system complexity, limiting its scalability in large-scale production systems.

Semantic ID-based Retrieval. Driven by these limitations, we further explore another retrieval strategy, where SemIDs serve as efficient retrieval keys. Specifically, we use the top-level semantic code $c_t^{(1)}$ of the target item to query an inverted index, retrieving all historical behaviors that share an identical top-level code:

$$\mathcal{S}_{\text{SemID}} = \left\{ b_i \in \mathcal{B} \mid c_i^{(1)} = c_t^{(1)} \right\}. \tag{8}$$

Compared to soft retrieval, this strategy offers two key advantages. First, it replaces dense similarity computations with inverted index lookups, reducing online complexity from $O(|\mathcal{B}| \cdot d)$ to near-constant time. Second, it eliminates storing and transmitting high-dimensional embeddings, significantly lowering memory and bandwidth overhead.

4.3 Exact Search Unit

The ESU stage is responsible for fine-grained sequence modeling over the retrieved behaviors. In order to introduce multi-modal signals, a central challenge lies in the misalignment between multi-modal embeddings and ID-based signals.

To address this, we leverage the multi-modal features introduced in Sec . 4.1 as side information. This information is integrated at the item level within the unified target-conditioned modeling framework, enabling fine-grained interaction between historical behaviors and the target item. Fig. 3 illustrates the overall pipeline.

Unified Item Representation. For each behavior item b_i in the sequence H_t, we construct a unified representation by combining collaborative and multi-modal features:

$$h_i = \text{Concat}\big(e_i^{id}, \ e_i^{sem}\big), \tag{9}$$

where e_i^{id} denotes the concatenated embeddings of $\mathbf{z}_i^{id}$.

The target item v_t is represented in the same form, i.e., $h_t = \text{Concat}(e_t^{id}, e_t^{sem})$.

Target-aware Sequential Modeling. Given the unified representations, we employ a target-aware sequence encoder to extract user interest:

$$u_t = f(\{h_1, \ldots, h_L\}, h_t), \tag{10}$$

where $f(\cdot)$ denotes a sequence modeling function. Specifically, f is instantiated as a multi-head target attention mechanism, where the importance of each behavior is conditioned on the target item:

$$\alpha_i = \text{Attn}\big(h_i \oplus e_{i,t}^{Sim}, \ h_t \oplus e_t^{Sim}\big), \qquad u_t = \sum_{i=1}^{L} \alpha_i h_i, \tag{11}$$

where $e_{i,t}^{Sim}$ is the similarity bucket embedding, and e_t^{Sim} is a learnable embedding associated with the target item. α_i denotes the attention weight for behavior b_i, $\oplus$ denotes vector concatenation.

A key property of this design is that all feature types—including ID-based, semantic, and target-aware signals—are integrated prior to sequence aggregation. Consequently, these signals jointly influence both attention weights and behavior representations during interest extraction.

Target-conditioned Interaction. To further align the extracted user interest with the target item, we introduce a lightweight element-wise interaction in the representation space [7, 29, 32]:

$$\tilde{u}_t = u_t \odot h_t, \tag{12}$$

where $\odot$ denotes element-wise multiplication.

The resulting representation is used for final prediction:

$$\hat{y}_t = \sigma\big(g([\tilde{u}_t, h_t, u, c])\big). \tag{13}$$

This interaction captures fine-grained correlations between user interest and target representations, allowing the model to capture fine-grained compatibility patterns beyond attention-based weighting.

5 Experiments

5.1 Experimental Setup

5.1.1 Dataset. We evaluate our method on the Taobao-MM dataset[1], a large-scale public benchmark for multi-modal lifelong user behavior modeling [25]. The dataset is collected from real-world traffic of Taobao's display advertising system and contains long-term user behavior sequences paired with high-quality multi-modal representations. Each item is associated with standard ID-based categorical

[1] https://huggingface.co/datasets/TaoBao-MM/Taobao-MM

features as well as a 128-dimensional pre-trained multi-modal embedding generated by SCL [25].

The dataset contains 99M interaction samples from 8.79M users over 35.4M items. Each user is associated with a historical behavior sequence of up to 1K interactions, making it a realistic benchmark for evaluating long-sequence recommendation methods. Each sample consists of anonymized user features (e.g., user ID, age, gender, location), item features (e.g., item ID, category), and a binary click label.

5.1.2 Baselines. We compare SIREN against a representative set of lifelong user interest modeling methods.

- **DIN** [31]: A target attention model that operates on short-term user behavior sequences without a GSU retrieval stage.
- **SIM-Hard** [18]: A two-stage model where the GSU retrieves behaviors based on exact category match with the target item.
- **SIM-Soft** [18]: A variant of SIM where the GSU retrieves behaviors based on inner-product similarity of item embeddings.
- **TWIN** [3]: A two-stage model that ensures consistency between GSU and ESU by adopting the same target attention mechanism in both stages.
- **MISS** [9]: A multi-modal enhanced retrieval method that introduces a multi-modal GSU alongside the ID-based one.
- **MUSE** [25]: A state-of-the-art multi-modal lifelong interest modeling framework that integrates multi-modal signals into both GSU and ESU stages.
- **SIREN$_{\text{SemID-GSU}}$**: A variant of the SIREN where the GSU stage uses Semantic ID-based hard retrieval, while keeping the ESU unchanged.
- **SIREN$_{\text{Sim-GSU}}$**: A variant of SIREN where the GSU stage adopts similarity-based soft retrieval, while keeping the ESU unchanged.

5.1.3 Implementation Details. For a fair comparison, each model is trained for one epoch following standard practice [3, 25]. For DIN, we use the most recent 50 behaviors, as it is a single-stage model without a GSU retrieval. For all two-stage models (SIM, TWIN, MISS, MUSE, SIREN), the GSU retrieves top-50 behaviors from the lifelong behavior sequence for ESU modeling. For SIM-Hard and SIM-Soft, we adopt the original GSU design while replacing the ESU with the SIREN implementation. By default, SIREN employs similarity-based soft retrieval in the GSU stage. We use AdamW [16] for dense parameters and SparseAdam for sparse embedding parameters, with learning rates set to 2×10^{-4} and 2×10^{-3}, respectively. The batch size is set to 1000. We adopt Group AUC (GAUC) as the primary offline evaluation metric.

5.2 Overall Performance

Table 1 reports the offline performance of all methods. We make the following observations:

Lifelong behavior modeling consistently outperforms short-behavior modeling. DIN, which operates only on recent behaviors, yields the lowest GAUC of 0.6006. In contrast, all two-stage methods leveraging lifelong behavior sequences via GSU retrieval achieve substantial improvements, underscoring the necessity of modeling long-term user interests in industrial recommendation settings.

Table 1: Overall offline performance comparison on GAUC. The best result is highlighted in bold.

Method	GAUC	Relative Lift
DIN	0.6006 (3E-5)	–
TWIN	0.6079 (6E-5)	+1.22%
MISS	0.6087 (1E-5)	+1.35%
SIM-Hard	0.6145 (6E-5)	+2.31%
SIM-Soft	0.6144 (5E-5)	+2.30%
MUSE	0.6148 (7E-5)	+2.36%
SIREN$_{\text{SemID-GSU}}$	0.6148 (7E-5)	+2.36%
SIREN$_{\text{Sim-GSU}}$	**0.6155 (9E-5)**	**+2.48%**

Table 2: Ablation study on GAUC. All variants use similarity-based soft retrieval as the GSU strategy. The base model uses target attention with only ID-based representations over the retrieved lifelong behavior sequence. TI denotes Target-conditioned Interaction.

ESU Configuration	GAUC	Relative Lift
Target Attention (base)	0.6080	–
+ SemID only	0.6095	+0.25%
+ SimBucket only	0.6142	+1.02%
+ SimBucket + SemID	0.6153	+1.20%
SIREN (SimBucket + SemID + TI)	**0.6155**	**+1.23%**

Multi-modal methods outperform ID-only baselines. ID-centric lifelong models (TWIN, MISS, SIM-Hard, and SIM-Soft) underperform compared to multi-modal approaches. Specifically, TWIN's ID-based retrieval generalizes poorly to long-tail items, while MISS and the SIM variants lack sufficient multi-modal integration in the ESU stage, placing them below SIREN.

SIREN achieves the best overall performance. SIREN achieves the highest GAUC of 0.6155, outperforming the strong baseline MUSE by 0.11%. Although the absolute gain over MUSE is modest, AUC/GAUC improvements at the 0.1% level are widely regarded as practically meaningful in large-scale recommendation and CTR prediction systems [4, 6, 7]. Notably, SIREN$_{\text{SemID-GSU}}$achieves comparable performance to MUSE while replacing dense similarity-based retrieval with efficient SemID-based lookup, demonstrating that SemID can serve as an effective and scalable surrogate for multi-modal similarity. The performance gain of the full model stems from SIREN's unified target-conditioned framework, where prefix-encoded SemID and similarity-aware bucket are directly integrated into sequential modeling.

5.3 Ablation Study

To understand the contribution of each component, we conduct ablation experiments by progressively adding components to the base model. Results are reported in Table 2.

Both side-information contribute positively. Adding SemID alone improves GAUC from 0.6080 to 0.6095, while similarity buckets alone yield a larger gain to 0.6142. Their combination further

improves performance to 0.6153, exceeding either component in isolation and demonstrating their complementary effects. And the Target-conditioned Interaction provides additional gain of GAUC. Although the performance gain is modest, the mutual information analysis in Fig. 4 shows that it substantially enhances representation discriminability.

Similarity buckets provide stronger gains, while SemIDs refine semantic distinctions. The larger improvement from similarity buckets is expected, as they directly encode target-aware relevance, which is the most informative signal for interest modeling. In contrast, SemIDs capture target-independent semantic content, helping distinguish behaviors with similar similarity scores but different semantics. Together, they jointly model *how relevant* a behavior is to the target and *what it represents*, enabling more comprehensive interest modeling.

5.4 Representation Discriminability

Beyond performance metrics, we further investigate whether SIREN learns more informative and discriminative user representations. To this end, we evaluate their mutual information (MI) with the click label.

Specifically, we employ the user interest representation $\tilde{u}_t$ defined in Eq. (12). Since $\tilde{u}_t$ is continuous and high-dimensional, we first apply K-means clustering to quantize the representation space, assigning each $\tilde{u}_t$ to a discrete cluster $Q(\tilde{u}_t)$. We then compute the MI between the cluster assignment $Q(\tilde{u}_t)$ and the binary click label Y. This metric serves as a proxy for representation discriminability: a higher MI indicates that the learned representations more effectively separate positive and negative samples.

As shown in Fig. 4(left), we compare representation discriminability across different model configurations and draw the following observations.

SIREN consistently produces more discriminative representations. Across varying numbers of K-means clusters, SIREN achieves consistently higher MI than the MUSE interest representation, as well as its two individual multi-modal components: *SimTier*, which aggregates target-behavior similarity into a global sequence-level histogram, and *SA-TA* (Semantic-Aware Target Attention), which infuses semantic similarities into ID-based attention weights. This improvement can be attributed to SIREN's unified sequential modeling framework with coarse and fine-grained multi-modal signals.

Target-conditioned interaction significantly enhances representation quality. As shown in Fig. 4(b), enabling the target interaction mechanism clearly improves MI. This demonstrates that explicitly modeling the correlation between the extracted user interest and target representation enhances feature interaction and strengthens the discriminability of user representations.

Similarity buckets and SemIDs provide complementary information. Fig. 4(c) shows that combining similarity buckets and SemIDs achieves the highest MI, outperforming either one. This confirms that the two types of side information capture complementary signals of user behavior: similarity buckets encode target-aware relevance, while SemIDs capture item semantic information. Their joint integration enables the model to achieve a more effective fusion of multi-modal features.

Overall, these results provide consistent evidence that SIREN improves not only predictive performance but also the intrinsic quality of learned user representations.

5.5 Necessity of Fine-Grained Semantic IDs

We further analyze why fine-grained Semantic IDs are necessary beyond coarse similarity buckets. Prefix-encoded SemIDs and similarity buckets encode multi-modal information at different granularities: the former captures item-level semantic structure, while the latter provides target-conditioned relevance. We examine their discriminative power and complementarity from three perspectives.

5.5.1 Conditional Entropy Analysis. We first evaluate how different grouping strategies reduce click-label uncertainty. Given a grouping variable G, we compute the information gain

$$I(Y;G) = H(Y) - H(Y|G), \tag{14}$$

where Y denotes the click label. A larger $I(Y;G)$ indicates that the grouping better explains label variation and thus induces more discriminative user-interest partitions.

Semantic-ID grouping achieves $I(Y;\text{SID}) = 0.0195$, which is substantially higher than the information gain of similarity-bucket grouping, $I(Y;\text{Sim}) = 0.0056$. This result shows that SemIDs preserve more click-related information than scalar similarity buckets, suggesting that they provide finer-grained partitions that better align with user feedback.

5.5.2 Within-Bucket CTR Distribution Analysis. We further examine whether similarity buckets can sufficiently distinguish behavior–target pairs with similar multi-modal proximity. As shown in Figure 2, even within the same similarity bucket, CTRs vary substantially across Semantic ID groups. Moreover, this within-bucket dispersion becomes larger in higher-similarity regions: the CTR range increases from $[0.815, 0.949]$ in bucket $[0.2, 0.3)$ to $[0.755, 0.956]$ in bucket $[0.5, 0.6)$, and further to $[0.735, 1.000]$ in bucket $[0.9, 1.0]$. The corresponding standard deviation also increases from about 0.035 to 0.052 and 0.068, respectively.

These results indicate that *high target–behavior similarity does not imply homogeneous user responses*. Behavior–target pairs that are close in the multi-modal space can still differ significantly in the collaborative CTR space. Therefore, *coarse similarity buckets alone are insufficient to capture the fine-grained heterogeneity required for accurate interest modeling, especially in high-similarity regions*.

5.5.3 Limitation of Increasing Bucket Granularity. A natural question is whether using more similarity buckets can recover the missing information. To answer this, we evaluate MUSE with different bucket granularities.

As shown in Fig. 4(right), increasing the number of buckets from 20 to 40 improves GAUC, but further increasing the granularity leads to performance degradation. This suggests that the limitation is not merely caused by insufficient bucket resolution. Instead, histogram-style similarity summarization compresses item-level similarity sequences into global statistics, discarding item identity and temporal structure. In contrast, SIREN preserves both similarity and SemID signals at the item level within the ESU.

Overall, these analyses show that similarity buckets and SemIDs are complementary: similarity buckets encode coarse target-conditioned

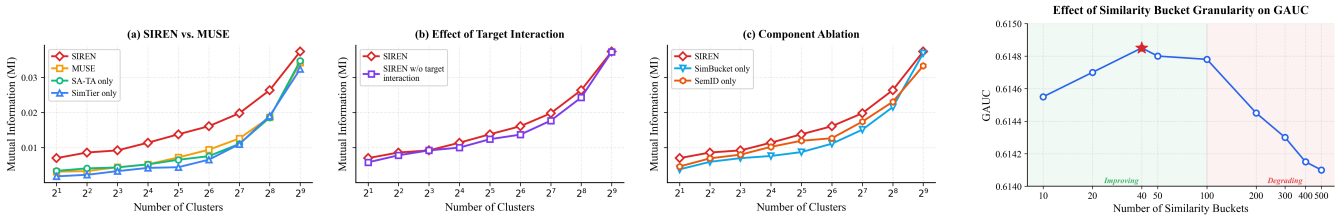


Figure 4: Analysis of representation discriminability and similarity-bucket granularity. Left: representation discriminability measured by mutual information between clustered user interest representations and click labels, including comparisons between SIREN and MUSE, the effect of target-aware interaction, and the ablation of SimBucket and SemID. Right: effect of similarity-bucket granularity on GAUC, where performance peaks at 40 buckets and degrades with further increases in granularity.

Table 3: Online A/B test results of SIREN across different advertising scenarios and pipeline stages.

Scenario	Stage	Feature	GMV Lift
Weixin Moments	pCTR	Similarity	+1.58%
	pCTR	SemID	+0.70%
Weixin Official Accounts	pCTR	Similarity	+1.64%
	pCVR	SemID	+2.23%
Weixin Channels	pCTR	Similarity	+0.87%
	LTR	SemID	+0.74%

relevance, while SemIDs capture fine-grained semantic and collaborative heterogeneity. Such complementarity cannot be recovered by simply refining similarity bucket granularity.

5.6 Online A/B Tests

We deployed SIREN in Tencent's Weixin online advertising system, one of the largest online advertising platforms serving tens of billions of ad requests per day across multiple production scenarios. To capture lifelong user interests, the production model contains cross-domain behavior sequences spanning advertisements, Channels, and content feeds, covering up to two years of user interactions with a maximum sequence length of 4,000 per domain.

We conduct A/B tests across three major advertising scenarios with 20% of traffic over a maximum period of 14 days per experiment. The production baseline was based on SIM-Hard [18]. All reported improvements have been verified through rigorous significance testing under varying traffic ratios, and SIREN has since been fully deployed across all evaluated scenarios. Table 3 summarizes the online results.

SIREN consistently improves GMV across all three scenarios, achieving lifts of +2.28% on Moments, +3.87% on Official Accounts, and +1.61% on Channels. The gains span multiple pipeline stages, demonstrating that the proposed framework generalizes well to diverse real scenarios.

Cold-start analysis. To better understand where the gains originate, we perform a fine-grained analysis over user activity levels and ad cold-start scenarios on Weixin Moments. As shown in Table 4, the GMV lift increases monotonically as user activity decreases, with SIREN delivering pronounced gains for both *low-activity users* (<50 interactions) and *cold-start users* (<10 interactions). In these scenarios where sparse interaction data renders traditional ID-based signals unreliable, our multi-modal side information substantially improves interest modeling. Similarly, on the item side, SIREN effectively leverages visual and textual content to enhance interest matching, thereby alleviating the cold-start problem for new ads launched within the past 24 hours.

Table 4: Relative GMV lift compared to the overall average on Weixin Moments.

Side	Segment	Relative Lift vs. Overall
User-side	Low-activity users	~1.7×
	Cold-start users	~3.6×
Item-side	New ads (first day)	~1.4×

Deployment efficiency of SemID-based hard retrieval. We further analyze SemID-based as an alternative retrieval strategy in the GSU stage. Compared to similarity-based retrieval, which requires maintaining dense embedding indices and computing pairwise similarities over high-dimensional vectors, SemID-based retrieval significantly reduces online serving cost by over 90% in both latency and storage. Importantly, this efficiency gain comes with minimal performance degradation. These results suggest that SemID-based strategy provides a practical efficiency–performance trade-off, offering a cost-effective alternative for large-scale deployment.

6 Conclusion

In this paper, we propose SIREN, a unified multi-granularity semantic interaction framework for multi-modal lifelong user interest modeling. In the GSU stage, similarity-based soft retrieval and SemID-based hard retrieval are explored to provide a flexible trade-off between retrieval effectiveness and serving efficiency. In the ESU stage, prefix-encoded SemIDs and similarity buckets are incorporated as complementary multi-granular side information, enabling unified target-conditioned sequential modeling alongside ID-based features. Experimental results on both offline benchmarks and large-scale real-world online systems demonstrate that SIREN effectively integrates multi-modal signals into lifelong interest modeling, and significantly enhance recommendation performance in industrial scenarios.

References

[1] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. In *ACL*. 13590–13618.

[2] Zheng Chai, Qin Ren, Xijun Xiao, Huizhi Yang, Bo Han, Sijun Zhang, Di Chen, Hui Lu, Wenlin Zhao, Lele Yu, Xionghang Xie, Shiru Ren, Xiang Sun, Yaocheng Tan, Peng Xu, Yuchao Zheng, and Di Wu. 2025. LONGER: Scaling Up Long Sequence Modeling in Industrial Recommenders. In *RecSys*. 247–256.

[3] Jianxin Chang, Chenbin Zhang, Zhiyi Fu, Xiaoxue Zang, Lin Guan, Jing Lu, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, and Kun Gai. 2023. TWIN: TWo-stage Interest Network for Lifelong User Behavior Modeling in CTR Prediction at Kuaishou. In *KDD*. 3785–3794.

[4] Qiwei Chen, Changhua Pei, Shanshan Lv, Chao Li, Junfeng Ge, and Wenwu Ou. 2021. End-to-End User Behavior Retrieval in Click-Through RatePrediction Model. *arXiv Preprint* https://arxiv.org/abs/2108.04468 (2021).

[5] Qin Ding, Kevin Course, Linjian Ma, Jianhui Sun, Ruochen Liu, Zhao Zhu, Chunxing Yin, Wei Li, Dai Li, Yu Shi, Xuan Cao, Ze Yang, Han Li, Xing Liu, Bi Xue, Hongwei Li, Rui Jian, Daisy Shi He, Jing Qian, Matt Ma, Qunshu Zhang, and Rui Li. 2026. Bending the Scaling Law Curve in Large-Scale Recommendation Systems. *arXiv Preprint* https://arxiv.org/abs/2602.16986 (2026).

[6] Zhifang Fan, Dan Ou, Yulong Gu, Bairan Fu, Xiang Li, Wentian Bao, Xin-Yu Dai, Xiaoyi Zeng, Tao Zhuang, and Qingwen Liu. 2022. Modeling Users' Contextualized Page-wise Feedback for Click-Through Rate Prediction in E-commerce Search. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 262–270.

[7] Ningya Feng, Junwei Pan, Jialong Wu, Baixu Chen, Ximei Wang, Qian Li, Xian Hu, Jie Jiang, and Mingsheng Long. 2025. Long-Sequence Recommendation Models Need Decoupled Embeddings. In *ICLR*.

[8] Lin Guan, Jia-Qi Yang, Zhishan Zhao, Beichuan Zhang, Bo Sun, Xuanyuan Luo, Jinan Ni, Xiaowen Li, Yuhang Qi, Zhifang Fan, Hangyu Wang, Qiwei Chen, Yi Cheng, Feng Zhang, and Xiao Yang. 2025. Make It Long, Keep It Fast: End-to-End 10k-Sequence Modeling at Billion Scale on Douyin. *arXiv Preprint* https://arxiv.org/abs/2511.06077 (2025).

[9] Chengcheng Guo, Junda She, Kuo Cai, Shiyao Wang, Qigen Hu, Qiang Luo, Guorui Zhou, and Kun Gai. 2025. MISS: Multi-Modal Tree Indexing and Searching with Lifelong Sequential Behavior for Retrieval Recommendation. In *CIKM*. 5683–5690.

[10] Zhicheng He, Weiwen Liu, Wei Guo, Jiarui Qin, Yingxue Zhang, Yaochen Hu, and Ruiming Tang. 2023. A Survey on User Behavior Modeling in Recommender Systems. In *IJCAI*. ijcai.org, 6656–6664.

[11] Ruijie Hou, Zhaoyang Yang, Ming Yu, Hongyu Lu, Zhuobin Zheng, Yu Chen, Qinsong Zeng, and Ming Chen. 2024. Cross-Domain LifeLong Sequential Modeling for Online Click-Through Rate Prediction. In *KDD*. 5116–5125.

[12] Xian Hu, Ming Yue, Zhixiang Feng, Junwei Pan, Junjie Zhai, Ximei Wang, Xinrui Miao, Qian Li, Xun Liu, Shangyu Zhang, Letian Wang, Hua Lu, Zijian Zeng, Chen Cai, Wei Wang, Fei Xiong, Pengfei Xiong, Jintao Zhang, Zhiyuan Wu, Chunhui Zhang, Anan Liu, Jiulong You, Chao Deng, Yuekui Yang, Shudong Huang, Dapeng Liu, and Haijie Gu. 2025. Practice on Long Behavior Sequence Modeling in Tencent Advertising. *arXiv Preprint* https://arxiv.org/abs/2510.21714 (2025).

[13] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2025. Multimodal Recommender Systems: A Survey. *ACM Comput. Surv.* 57, 2 (2025), 26:1–26:17.

[14] Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. 2024. AlignRec: Aligning and Training in Multimodal Recommendations. In *CIKM*. 1503–1512.

[15] Alejo Lopez-Avila and Jinhua Du. 2025. A Survey on Large Language Models in Multimodal Recommender Systems. *arXiv Preprint* https://arxiv.org/abs/2505.09777 (2025).

[16] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

[17] Xinchen Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, Changqing Qiu, Jiaqi Zhang, Xu Zhang, Zhiheng Yan, Jingming Zhang, Simin Zhang, Mingxing Wen, Zhaojie Liu, and Guorui Zhou. 2025. QARM: Quantitative Alignment Multi-Modal Recommendation at Kuaishou. In *CIKM*. 5915–5922.

[18] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 2020. Search-based User Interest Modeling with Lifelong Sequential Behavior Data for Click-Through Rate Prediction. *arXiv Preprint* (2020). https://arxiv.org/abs/2006.05639

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*. 8748–8763.

[20] Xiang-Rong Sheng, Feifan Yang, Litong Gong, Biao Wang, Zhangming Chan, Yujing Zhang, Yueyao Cheng, Yong-Nan Zhu, Tiezheng Ge, Han Zhu, Yuning Jiang, Jian Xu, and Bo Zheng. 2024. Enhancing Taobao Display Advertising with Multimodal Representations: Challenges, Approaches and Insights. In *CIKM*. 4858–4865.

[21] Zihua Si, Lin Guan, Zhongxiang Sun, Xiaoxue Zang, Jing Lu, Yiqun Hui, Xingchao Cao, Zeyu Yang, Yichen Zheng, Dewei Leng, Kai Zheng, Chenbin Zhang, Yanan Niu, Yang Song, and Kun Gai. 2024. TWIN V2: Scaling Ultra-Long User Behavior Sequence Modeling for Enhanced CTR Prediction at Kuaishou. In *CIKM*. 4890–4897.

[22] Qwen Team. 2025. Qwen3-VL Technical Report. *arXiv Preprint* https://arxiv.org/abs/2511.21631 (2025).

[23] Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. 2023. MISSRec: Pre-training and Transferring Multi-modal Interest-aware Sequence Representation for Recommendation. In *ACM Multimedia*.

[24] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Learnable Item Tokenization for Generative Recommendation. In *CIKM*. 2400–2409.

[25] Bin Wu, Feifan Yang, Zhangming Chan, Yu-Ran Gu, Jiawei Feng, Chao Yi, Xiang-Rong Sheng, Han Zhu, Jian Xu, Mang Ye, and Bo Zheng. 2025. MUSE: A Simple Yet Effective Multimodal Search-Based Framework for Lifelong User Interest Modeling. *arXiv Preprint* https://arxiv.org/abs/2512.07216 (2025).

[26] Yi Xu, Moyu Zhang, Chenxuan Li, Zhihao Liao, Haibo Xing, Hao Deng, Jinxin Hu, Yu Zhang, Xiaoyi Zeng, and Jing Zhang. 2026. MMQ: Multimodal Mixture-of-Quantization Tokenization for Semantic ID Generation and User Behavioral Adaptation. In *WSDM*. 788–797.

[27] Wencai Ye, Mingjie Sun, Shaoyun Shi, Peng Wang, Wenjin Wu, and Peng Jiang. 2025. DAS: Dual-Aligned Semantic IDs Empowered Industrial Recommender System. In *CIKM*. 6217–6224.

[28] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. SoundStream: An End-to-End Neural Audio Codec. *IEEE ACM Trans. Audio Speech Lang. Process.* 30 (2022), 495–507.

[29] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. 2024. Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations. In *ICML*. 58484–58509.

[30] Carolina Zheng, Minhui Huang, Dmitrii Pedchenko, Kaushik Rangadurai, Siyu Wang, Fan Xia, Gaby Nahum, Jie Lei, Yang Yang, Tao Liu, Zutian Luo, Xiaohan Wei, Dinesh Ramasamy, Jiyan Yang, Yiping Han, Lin Yang, Hangjun Xu, Rong Jin, and Shuang Yang. 2025. Enhancing Embedding Representation Stability in Recommendation Systems with Semantic ID. In *RecSys*. 954–957.

[31] 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 *KDD*.

[32] Haolin Zhou, Junwei Pan, Xinyi Zhou, Xihua Chen, Jie Jiang, Xiaofeng Gao, and Guihai Chen. 2024. Temporal Interest Network for User Response Prediction. In *WWW*. 413–422.