Bridging the Semantic Gap: Leveraging LLMs for Hierarchical Interest Evolution in Sequential Recommendation

Yifan Cao[1,2] , **Rui Wu**[1] , **Xiang Wang**[1] , **Wei Peng**[1] , **Wei Xu**[2] , **Caihong Sun**[2*] and **Jian Zeng**[2]
[1]Meituan, China
[2]School of Information, Renmin University of China, Beijing, China
2024103837@ruc.edu.cn, {wurui13,wangxiang31,pengwei11}@meituan.com,
{weixu,chsun}@ruc.edu.cn, 2024103838@ruc.edu.cn

Abstract

Accurate user behavior modeling is fundamental to the prediction of click-through rates (CTR) in industrial recommendation systems and online advertising. Traditional discriminative models, which rely on isolated ID features, struggle to capture the evolving nature of user intents across multiple channels due to the inherent semantic gap. While Large Language Models (LLMs) offer rich semantic understanding to bridge this gap, their direct application faces two main challenges: modality misalignment between continuous semantic representations and discrete item IDs, and prohibitive computational costs that make them unsuitable for real-time inference. To address these issues, we propose the **H**ierarchical **S**emantic **I**nterest **E**volution **N**etwork (**HSIEN**), a novel generative-discriminative framework. HSIEN leverages an LLM to construct a Markovian state-space model that recursively captures and updates user interests from multi-channel behavior sequences. It then employs a Semantic Mixture-of-Experts (SMoE) strategy to aggregate multi-channel signals into hierarchical intent representations, which are encoded as compact dynamic semantic vectors for integration with downstream discriminative models. Extensive experiments on a public dataset and a large-scale industrial dataset demonstrate that HSIEN significantly alleviates modality misalignment and enhances CTR prediction performance through feature complementarity. This work provides a practical and efficient pathway for leveraging LLMs in large-scale recommendation scenarios.

1 Introduction

In computational advertising and recommendation systems, modeling user behavior intent for accurate prediction of Click-Through Rate (CTR) and Conversion Rate (CVR) is fundamental to maximizing commercial value. However, capturing user intent across behavior data from different channels or domains remains challenging. Figure 1(a) illustrates such a scenario: a user purchases sunglasses and swimwear in an "Instant Retail" channel and subsequently browses beach resorts in a "Hotel & Travel" channel. While the underlying intent—planning an island vacation—is intuitively clear to a human observer, automatically capturing such deep, cross-channel semantic correlations remains a formidable challenge for industrial-scale bidding systems.

For years, mainstream industrial solutions have relied primarily on deep learning-based discriminative models, such as Wide&Deep [Cheng *et al.*, 2016], DeepFM [Guo *et al.*, 2017], DIN [Zhou *et al.*, 2018], and SIM [Pi *et al.*, 2020]. Their efficiency in handling large-scale discrete IDs (Item IDs) and ability to memorize specific co-occurrence patterns (e.g., "users who purchased ID:1001 often click on ID:2002") have made them the cornerstone of real-time systems. However, these models are inherently limited by the Semantic Gap: IDs are non-overlapping across distinct business channels, such as dining, travel, and retail. This often prevents the models from understanding semantic relationships between concepts like "sunscreen" and "resort". Consequently, their performance frequently declines sharply when dealing with cold-start items or long-tail user behaviors, mainly due to insufficient co-occurrence history in the ID space.

To incorporate semantic understanding, recent research has explored using Language Models for recommendation tasks, as exemplified by models like P5 [Geng *et al.*, 2022] and TIGER [Rajput *et al.*, 2023]. These approaches reformulate recommendation as a language generation task. Although they exhibit strong zero-shot reasoning abilities, their practical deployment in high-concurrency advertising bidding environments faces two main challenges. The first is Modality Misalignment. Many current methods require the LLM to directly generate discrete item IDs. This approach underutilizes the LLM's ability to reason in a continuous, open semantic space and forces the model to memorize extensive ID mappings, often leading to hallucination problems and misalignment between semantic representations and the target item space. The second is an Efficiency Bottleneck. Each recommendation request typically needs to process the user's full interaction history through the LLM, resulting in high computational latency. This makes it impractical for real-time systems that demand millisecond-level response times, such as online advertising auctions.

Therefore, we argue that a more effective approach is to

*Corresponding author.

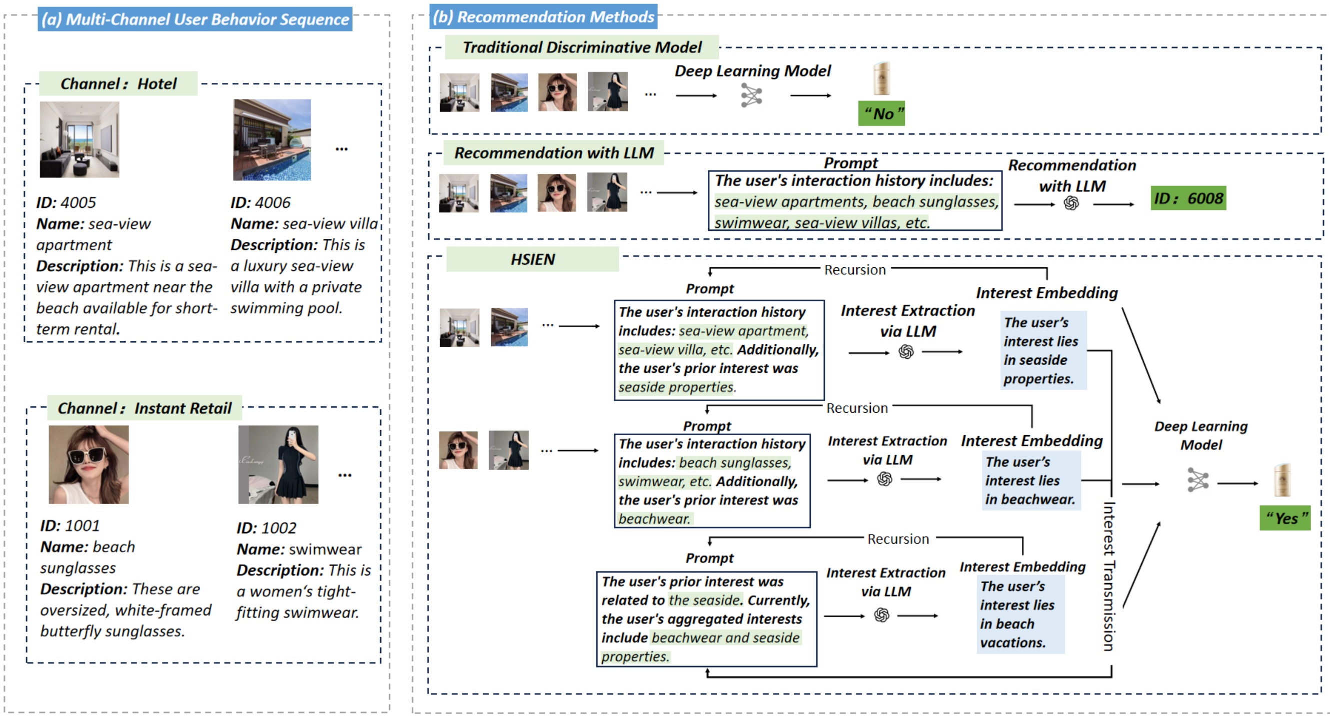

Figure 1: Comparison of three recommendation methods.

leverage LLMs for high-level intent reasoning rather than involving them in low-level ID retrieval and matching tasks. This separation of responsibilities better aligns with the strengths of LLMs and can lead to more scalable and robust systems. To implement this principle and address the aforementioned challenges, we propose the Hierarchical Semantic Interest Evolution Network (HSIEN) as illustrated in Figure 1(b). Our design philosophy is not to replace existing discriminative models with LLMs, but to integrate them synergistically. HSIEN employs an LLM to perform semantic interest extraction from a user's multi-channel behavior sequences. The extracted insights are then fused with a discriminative architecture, achieving a powerful synergy between broad world knowledge and precise collaborative signals. Specifically, HSIEN introduces a collaborative architecture that synergizes Generative Reasoning with Discriminative Decision-Making. It operates in two phases: upstream generative reasoning and downstream discriminative decision-making. In the upstream phase, an LLM builds a Markovian state-space model that recursively converts a user's multi-channel behaviors into explicit semantic interest flows in natural language, summarizing the user's preferences. Downstream, these semantic flows, having been translated into dense vector representations, are injected into discriminative models. This design deftly resolves the mentioned dilemma: it leverages the LLM's world knowledge to fill the semantic gap of ID-based models, while simultaneously avoiding modality misalignment by confining the LLM's output to the semantic space rather than the ID space.

The main contributions of this work are two fold:

- **A Generative-Discriminative Collaborative Architecture:** We introduce a novel hybrid architecture in which an LLM extracts high-level semantic intents from fragmented, multi-channel user behaviors. These intentions are transformed into dense vectors and integrated as features into downstream discriminative CTR/CVR models. This design avoids the modality misalignment of direct ID generation and enhances models with cross-domain reasoning, significantly improving intent prediction accuracy.
- **A Daily Recursive State Update Mechanism:** To ensure efficiency, our model, based on the Markov assumption, updates user interest states incrementally. It uses only the previous day's state and current behaviors, avoiding the computational burden of reprocessing the entire history. This enables low-latency tracking of long-term user interests.

2 Related Work

Discriminative User Modeling. The accurate prediction of CTR and CVR serves as the cornerstone of industrial advertising and recommendation systems. Early research, including Wide & Deep [Cheng *et al.*, 2016] and DeepFM [Guo *et al.*, 2017], primarily focused on capturing explicit and implicit feature interactions. To model user behavior sequences, approaches have evolved from foundational methods like YouTube DNN [Covington *et al.*, 2016], which employed mean pooling, to more advanced architectures. Models such as DIN [Zhou *et al.*, 2018], DSIN [Feng *et al.*, 2019], BST [Chen

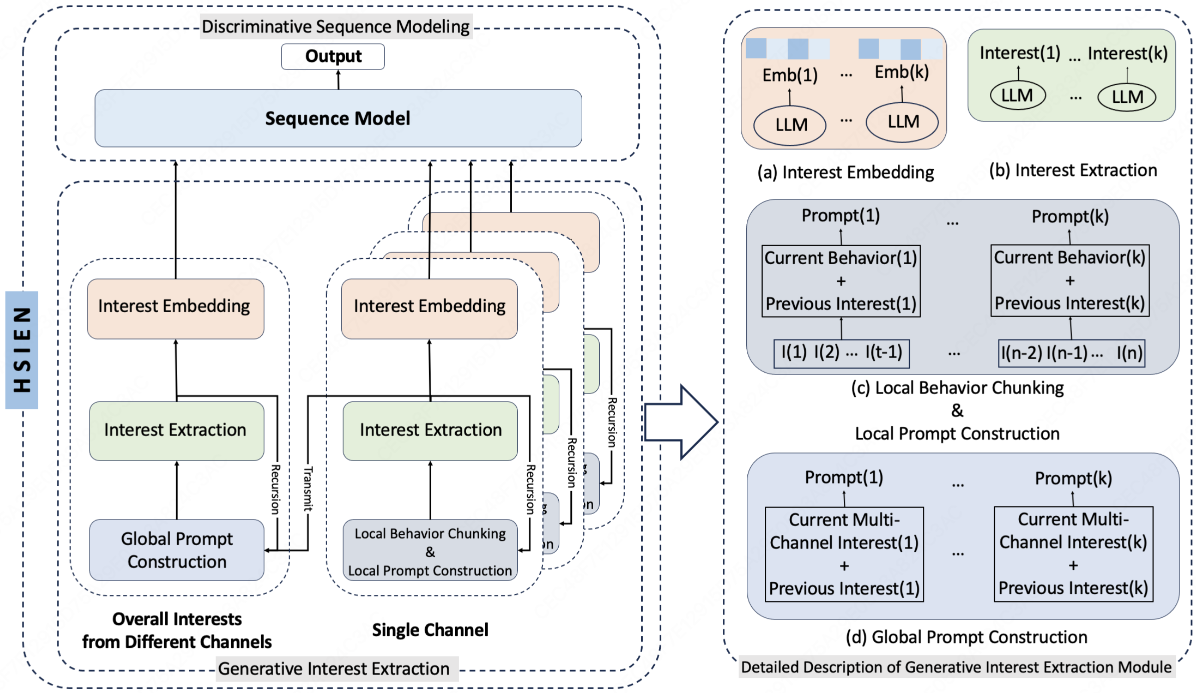

Figure 2: HSIEN combines LLM-based generative interest extraction with discriminative sequence modeling. It recursively updates Global and Local semantic interests from current behaviors and previous summaries, then embeds them into dense vectors for downstream prediction.

et al., 2019], PBAT [Su *et al.*, 2023], MBHT [Yang *et al.*, 2022], SASRec and BERT4Rec [Kang and McAuley, 2018] leverage attention or graph mechanisms to dynamically weight and model user interest flows. More recently, to tackle the challenge of ultra-long sequences, retrieval-based methods such as UBR4CTR [Qin *et al.*, 2020], ETA [Chen *et al.*, 2021], SDIM [Cao *et al.*, 2022], and the TWIN [Chang *et al.*, 2023; Si *et al.*, 2024] family have been proposed for efficient modeling. Although these models excel in ID matching, they are fundamentally constrained by the semantic gap, treating items as isolated ID embeddings. Lacking world knowledge, these models fail to infer deep multi-channel interest correlations.

Recommendation with Language Models. Recently, the emergence of LLMs has provided a new avenue for semantically modeling user interests. Approaches in this field typically fine-tune LLMs to directly generate target item IDs (e.g., HLLM [Chen *et al.*, 2024], LLaRA [Liao *et al.*, 2024]) or textual descriptions (e.g., TALLRec [Bao *et al.*, 2023]). However, these paradigms face significant challenges in lifelong user behavior modeling. They force continuous semantics into a discrete ID space, leading to modality misalignment. Moreover, the computational overhead of real-time LLM inference on long behavior sequences renders these methods impractical for low-latency systems. In contrast, our proposed HSIEN adopts a recursive "Generative Reasoning + Discriminative Decision" architecture to achieve online $O(1)$ inference complexity, effectively eliminating semantic and modality misalignment.

3 Methodology

This section presents the proposed HSIEN framework, as illustrated in Figure 2.

3.1 Problem Definition and Notations

Industrial Context. In large-scale industrial advertising and recommendation systems, developing high-precision CTR and CVR prediction models is a decisive factor in enhancing Return on Investment (ROI). HSIEN significantly improves the accuracy of CTR and CVR estimation.

Multi-Channel Behavioral Space. In recommendation systems, the Meituan platform's multi-channel data constitutes the core context for understanding user dynamic intents. In contrast to single-channel data, the recommendation platform typically has access to interaction records across K channels, reflecting the multifaceted and interrelated nature of user interests. We define a single interaction as a tuple $b = (v, \mathcal{M}_v, \tau)$, containing an item ID v, its metadata $\mathcal{M}_v$ (e.g., category, price), and a timestamp τ. To capture the evolution of user interests, we partition the user's interaction history into a sequence of ordered behavior Blocks, and for user u on block t, the incremental behavior sequence within channel k is $\mathbf{X}_{u,t}^{(k)} = [b_1, \dots, b_m]$. The user's multi-channel incremental history $\mathbf{X}_{u,t}$ is then defined as a set of sequences across all active channels:

$$\mathbf{X}_{u,t} = \left\{ \mathbf{X}_{u,t}^{(k)} \right\}_{k=1}^{K}. \tag{1}$$

Task Definition. The CTR and CVR prediction task is formulated as a binary classification problem. Given a user's cumulative historical data $\mathbf{X}_{u,t}$ (i.e., all behaviors up to the current block t), the target ad features $\mathbf{a}$, and the context $\mathbf{c}$, our objective is to learn a function f_Θ to estimate the conversion probability $y \in \{0, 1\}$:

$$\hat{y}_{u,a} = P(y = 1 \mid \mathbf{X}_{u,t}, \mathbf{a}, \mathbf{c}; \Theta). \tag{2}$$

Dual Challenges. Existing approaches face a dilemma. Traditional discriminative models, which rely on ID-based encoding, struggle to capture semantic evolution in multi-channel scenarios. Conversely, recommendation with Language Models typically attempts to generate discrete item IDs directly from semantic descriptions, resulting in a severe Modality Misalignment between the semantic space and the ID space. Therefore, our proposed HSIEN seeks to balance deep semantic reasoning with precise ID representation.

3.2 Overview of the HSIEN Framework

To address the semantic limitations of discriminative models and the modality misalignment inherent in Recommendation with Language Models, we propose the HSIEN. As illustrated in Figure 2, HSIEN is a bottom-up hierarchical generative architecture that deeply synergizes LLMs and discriminative models. It transforms unstructured data into dense semantic representations through three distinct stages, ultimately performing CTR or CVR estimation for a given target item:

- **Microscopic Level: Local Interest Recursive Evolution.**For a single channel, we construct a state-space model that satisfies the Markov property. This design avoids the linear computational complexity in processing the entire raw behavioral sequence. HSIEN employs the LLM as its local state transition function. Its input is strictly limited to two components: the explicit interest summary of the preceding behavior block and a prompt characterizing the incremental behaviors within the current block. This mechanism ensures that "micro-interest states" are updated recursively on a block-by-block basis, thereby facilitating highly efficient offline inference.
- **Macroscopic Level: Multi-Channel Semantic Aggregation.**A single-channel perspective is inherently limited. To obtain a global view, HSIEN introduces a Semantic Mixture-of-Experts (SMoE) strategy. Here, the LLM acts as a Meta-Reasoner. It first aggregates the current block's micro-states from all active channels, and then recursively incorporates the global summary from the preceding block. Through this joint reasoning process, HSIEN identifies conflicts, performs denoising, and ultimately synthesizes a coherent, explicit semantic flow that captures high-order user intents across channels.
- **Representation Level: Joint Recommendation with Downstream Discriminative Models.**To bridge the gap between semantic reasoning and numerical computation, we employ a semantic alignment strategy. By utilizing a pre-trained text encoder, the global explicit semantic flow is mapped into a fixed-dimensional dense vector. This vector acts as a dynamic global semantic context, which is seamlessly injected into downstream discriminative models for training to complement ID collaborative signals.

Local Prompt
System: Infer current interests from previous interests and current behavior.
Input: Previous: Science Fiction, Space Opera, Dystopia. Current: *Frankenstein*; *Dune*.
Task: Summarize current interests with three comma-separated keywords.

Global Prompt
System: Infer global interests from previous global interests and current multi-channel interests.
Input: Previous: High-Tech Preference, Smart Living, Innovation Seeker. Channel 1: Smart Wearables, Tech Reviews, AR Devices. Channel 2: Cryptocurrency, Blockchain, Tech Sector Stocks.
Task: Summarize global interests with three comma-separated keywords.

Figure 3: Example of prompt.

3.3 Local Interest Recursive Evolution

To capture microscopic interest evolution while mitigating high time complexity $O(T)$, we construct an efficient state-space generative model. It represents user interests as explicit semantic states that evolve over time. For the k-th channel, we define a state $S_{u,t}^{(k)}$ to represent the user's state at the t-th block, which is generated following the Markov property:

$$S_{u,t}^{(k)} = \mathrm{LLM}_{\mathrm{local}}\left(\mathcal{I}_{\mathrm{prompt}}, \tilde{S}_{u,t-1}^{(k)}, \mathcal{T}\left(\mathbf{X}_{u,t}^{(k)}\right)\right), \tag{3}$$

where $\mathrm{LLM}_{\mathrm{local}}$ denotes the LLM serving as the state-transition function, $\mathcal{I}_{\mathrm{prompt}}$ represents the predefined instruction prompt template, and $\mathcal{T}(\cdot)$ corresponds to the textualized current incremental behavior sequence. This process incorporates three key design components:

Sequence Textualization $\mathcal{T}(\cdot)$. To bridge the semantic gap between IDs and natural language, we convert raw behavior logs into textual narratives. The textualization function $\mathcal{T}(\cdot)$ constructs prompts by leveraging rich metadata, including item descriptions such as identifiers (e.g., `item_name`), business attributes, price, and hierarchical categories. Figure 3 shows an example of a local prompt. This formulation activates the world knowledge inherent in the LLM, allowing it to infer user consumption levels and preferences.

Temporal Sparsity and Backtracking Mechanism $\delta(\cdot)$. Temporal sparsity is a ubiquitous characteristic of user behavior sequences. To effectively organize these behaviors, we partition the sequence into variable-sized blocks, which ensures that the behaviors within each block are semantically correlated, thereby exhibiting relatively unified and consistent interest features. For users with sparse interactions, outdated data may not accurately reflect their current interest preferences. To address this, we define a backtracked state $\tilde{S}_{u,t-1}^{(k)}$ that searches within a window W (e.g., $W = 10$) for the most recent valid state:

$$\begin{gathered} \tilde{S}_{u,t-1}^{(k)} = \delta(S_{u,t-1}^{(k)}, W) = S_{u,\tau}^{(k)}, \\ \tau = \max\{t' < t \mid S_{u,t'}^{(k)} \neq \emptyset \text{ and } t - t' \leq W\}. \end{gathered} \tag{4}$$

This mechanism ensures the continuity of interest flows for long-tail users. By adopting an "offline-inference, online-lookup" architecture, our model achieves an online computational complexity of $O(1)$. This design enables production-ready deployment, delivering deep semantic insights with negligible operational overhead.

Semantic Information Limitation. To mitigate information overload, we impose strict constraints within the prompt template $\mathcal{I}_{\text{prompt}}$. Specifically, the LLM is constrained to compress the state $S_{u,t}^{(k)}$ into a fixed number of keyword summaries (e.g., "three core keywords"). This limitation compels the model to actively perform denoising and distill the core user intent.

3.4 Multi-Channel Semantic Aggregation

A single-channel perspective is inherently limited and may introduce bias. To construct a holistic user profile, it is essential to integrate signals across channels. Meituan's advertising scenario spans heterogeneous channels. Simple concatenation fails to resolve the underlying semantic conflicts across these diverse channels. Therefore, we propose a SMoE aggregation mechanism.

Here, the LLM functions as a meta-reasoner, taking the global state of the preceding behavior block $G_{u,t-1}$ and the current local micro-state:

$$G_{u,t} = \text{LLM}_{\text{global}}\left(\mathcal{I}_{\text{agg}}, G_{u,t-1}, \sum_{k=1}^{K} S_{u,t}^{(k)}\right). \quad (5)$$

This formulation enables two key operations:

Multi-Channel Denoising. HSIEN leverages the global context to filter out local noise. For instance, if a user clicks on a high-priced item in instant retail but consistently exhibits price-sensitive behavior in restaurant and food delivery channels, HSIEN would downweight the high-price preference signal to maintain consistency in the user profile.

High-Order Latent Intent Reasoning. This capability is a key differentiator between HSIEN and traditional ID-based models. The LLM generally infers implicit intents from multi-channel cues. For example, from signals like "resort hotel" (travel), "sunscreen" (health), and "manicure" (service), it can generate a tag such as "planning an island vacation" without relying on ID co-occurrence. Figure 3 shows an example of a global prompt. The output $G_{u,t}$ is a distilled, explicit semantic flow that characterizes the user's multi-channel state. This flow not only summarizes past actions but, more importantly, provides a prospective explanation of future user intentions.

3.5 Joint Recommendation with Downstream Discriminative Models

To leverage the semantic signals $G_{u,t}$ and $S_{u,t}^{(k)}$, we developed a synergistic framework designed for the deep fusion of high-level semantic insights with fine-grained collaborative signals derived from item IDs.

Sequential Model-based Semantic Extraction. These sequences are fed into sequence modules to capture their relevance to the target advertisement $\mathbf{a}$:

Dataset	Users	Items[a]	Categories	Instances
Amazon(Book).	75053	358367	1583	150016
Industrial.	0.12 billion	0.4 billion	157000	9.3 billion

[a] For industrial dataset, items refer to advertisements.

Table 1: Statistics of datasets used in this paper.

- Multi-Channel Interest Extraction: An attention-weighted sum over the multi-channel sequence yields the vector $\mathbf{v}^{\text{global}}$.
- Per-Channel Interest Extraction: Applying attention to each per-channel sequence produces a set of channel-specific vectors $\{\mathbf{v}_{\text{channel}}^{(k)}\}_{k=1}^{K}$.

The extracted vectors are then concatenated with sparse ID-based features $\mathbf{v}_{\text{base}}$ (e.g., user ID, ad ID, context) and fed into the downstream discriminative model:

$$\mathbf{v}_{\text{input}} = \text{Concat}\left(\mathbf{v}_{\text{base}}, \mathbf{v}^{\text{global}}, \{\mathbf{v}_{\text{channel}}^{(k)}\}_{k=1}^{K}\right), \quad (6)$$

$$\hat{y}_{u,a} = \mathcal{F}_{\text{discriminative}}(\mathbf{v}_{\text{input}}). \quad (7)$$

Complementarity Enhancement. This component translates the upstream semantic flows into dynamic interest vectors, semantically enriched and aligned with the target advertisement. This design establishes a deep complementarity: the LLM-based generator leverages world knowledge to bridge data sparsity (e.g., inferring a broad "camping" interest from a specific "tent" purchase), while the discriminative model retains precise, instance-level memory of user-item interactions. Consequently, this synergy enhances the system's robustness and expressive power in complex advertising scenarios.

4 Experiments

In this section, we present a detailed account of the experimental process. We evaluate the proposed HSIEN model against state-of-the-art benchmarks on one public dataset and one industrial dataset. Given that HSIEN has been fully deployed in Meituan's online advertising system, we also conducted rigorous online A/B tests.

4.1 Datasets

Model comparisons are conducted on two datasets: the publicly available Amazon dataset and an Industrial Dataset collected from Meituan's online advertising system.Table 1 shows the statistics of all datasets.

Amazon (Book) Dataset. [1] This dataset consists of product reviews and metadata from Amazon [McAuley *et al.*, 2015]. We utilize the Books subset, which contains 75,053 users, 358,367 items, and 1,583 categories. In our experiments, we treat reviews as interaction behaviors and sort them chronologically for each user. Given a user with a total of T behaviors, we use the preceding $T-1$ behaviors to predict whether the user will write the T-th review. To focus on long-sequence user behavior prediction, we filter out samples with behavior sequence lengths less than 20. This preprocessing method is widely adopted in related work.

[1] http://jmcauley.ucsd.edu/data/amazon/

Industrial Dataset. Collected from Meituan's platform, this dataset represents a typical multi-channel scenario encompassing multiple heterogeneous channels such as catering group buying, instant retail, food delivery, and healthcare services. Its samples are extracted from impression logs, with "click" or "non-click" as labels. Following a chronological split, the training set covers 49 consecutive days, while the test set comprises the subsequent day. Notably, user behavior sequences in this dataset span a three-month temporal window, with over 30% of samples exceeding a length of 10,000. This provides an ideal experimental setting for validating the model's capability to process long-term and ultra-long sequences, and to capture multi-channel interest evolution.

4.2 Experimental Settings

We compare HSIEN with several representative and baseline CTR prediction models:

- **Embedding&MLP:** A foundational deep learning model for CTR prediction. It employs a sum pooling operation to aggregate behavior embeddings.
- **GRU4Rec** [Hidasi *et al.*, 2015]: An RNN-based model utilizing recurrent cells for user behavior sequence modeling.
- **DIN** [Zhou *et al.*, 2018]: Introduces an attention mechanism to capture diverse user interests relative to the candidate item.
- **RUM** [Chen *et al.*, 2018]: Achieves sequential modeling via attention-based "soft-writing" and "attentive reading" mechanisms.
- **DIEN** [Zhou *et al.*, 2019]: Extends DIN by incorporating an interest evolution network to capture the dynamic shifts in user interests over time.
- **MIMN** [Pi *et al.*, 2019]: Designed with an incremental update mechanism, it efficiently captures user interests in ultra-long sequences while maintaining constant system latency.
- **ARNN** [Guo *et al.*, 2020]: Extends GRU4Rec by applying an attention mechanism to weight temporal hidden states.

Experiment Setup. For a fair comparison, we adopt experimental settings consistent with related work. We use the Adam optimizer for all models. The learning rate is initialized at 0.001, following an exponential decay strategy with a decay rate of 0.9. The Fully Connected Network (FCN) layers are configured with sizes $200 \times 80 \times 2$. The embedding dimension is set to 16. We use the widely adopted Area Under the ROC Curve (AUC) as the primary evaluation metric.

4.3 Overall Performance

Results on Public Datasets

Table 2 presents the performance of all competing models on the Amazon (Book) dataset. HSIEN consistently outperforms all baseline models, including Embedding&MLP, GRU4Rec, DIN, RUM, DIEN, MIMN and ARNN, across the AUC metric.

We argue that traditional discriminative models, despite their ability to handle ultra-long sequences, fundamentally rely on ID-based co-occurrence matching. In semantic-rich scenarios like Amazon Books, HSIEN transcends this limitation by incorporating LLM-generated explicit semantic flows, which enables the model to capture deep semantic associations that elude ID-based approaches—for instance, inferring an interest in "children's picture books" from a history of "parenting books", even when these items have never co-occurred in the training data. HSIEN delivers substantial AUC improvements over the strongest baseline, MIMN. This performance improvement provides compelling evidence that augmenting discriminative models with high-quality semantic information is pivotal for breaking the performance ceiling of current ID-based paradigms.

Model	Amazon (mean ± std)
Embedding&MLP	0.7367 ± 0.00043
DIN	0.7419 ± 0.00049
GRU4Rec	0.7411 ± 0.00182
ARNN	0.7420 ± 0.00029
RUM	0.7428 ± 0.00041
DIEN	0.7481 ± 0.00102
MIMN	0.7593 ± 0.00150
HSIEN	$\mathbf{0.7795 \pm 0.00106}$

Table 2: Model performance (AUC) on public datasets.

Results on Industrial Dataset

We further evaluated HSIEN on Meituan's large-scale production dataset to verify its effectiveness in a real-world industrial environment.

Offline Evaluation. As illustrated in Table 3, HSIEN achieves an absolute AUC gain of 0.59 percentage points (PP) compared to the current production baseline. In an industrial setting serving hundreds of millions of users and billions of samples, an improvement of this magnitude is highly significant. It directly demonstrates that in data-sparse or multi-channel scenarios (e.g., transitions from "Food Delivery" to "Hotel & Travel"), the integration of semantic information substantially bolsters the model's generalization capability. This semantic integration effectively mitigates the prediction bias inherent in ID-based models when dealing with distinct ID systems.

Online A/B Testing. We deployed HSIEN in Meituan's online display advertising system for a rigorous A/B test. From December 3, 2025, to January 12, 2026, HSIEN achieved a significant positive gain in ROI and Gross Merchandise Volume (GMV) compared to the baseline model. Specifically, HSIEN increased **ROI** by **3.0 PP** and **GMV** by **2.7 PP** relative

Model	Industrial (AUC)
DIN	0.8442
HSIEN	**0.8501**

Table 3: Ablation study on the Industrial Dataset.

Model	AUC (mean ± std)
Global-only	0.7670 ± 0.00141
Local-only	0.7581 ± 0.00130
Global + Local (Ours)	$\mathbf{0.7795 \pm 0.00106}$

Table 4: Ablation study on semantic aggregation strategies.

Recursive Update	AUC (mean ± std)
HSIEN (No)	0.7587 ± 0.00135
HSIEN (Yes)	$\mathbf{0.7795 \pm 0.00106}$

Table 5: Ablation study on the recursive update mechanism.

to the baseline model. These results confirm that HSIEN is not only theoretically sound but also translates effectively into tangible business value. We attribute this to the significant role played by combining LLM-derived semantic interests with the discriminative model.

4.4 Ablation Study

To verify the effectiveness of the core components of HSIEN, we conducted two types of ablation experiments on the Amazon dataset:

Effectiveness of Aggregation Strategies. We compared different semantic aggregation strategies:

- **Local-only:** Uses only independent interest sequences from each channel (e.g., catering, delivery).
- **Global-only:** Uses only the aggregated global interest sequence.
- **Full HSIEN:** Incorporates both channel-specific (local) and global semantic features.

The results in Table 4 indicate that the full HSIEN achieves peak performance. Notably, the fact that Global-only outperforms Local-only underscores the LLM's capacity as a "meta-reasoner". It demonstrates that multi-channel information fusion—such as synthesizing signals from "dining" and "travel" behaviors to infer a latent "vacation" intent—is pivotal for enhancing predictive accuracy.

Effectiveness of Recursive Evolution. We conducted an ablation study by comparing HSIEN with a variant that lacks the recursive mechanism, where intent generation depends solely on isolated daily increments. The significant drop in AUC observed in this variant, as shown in Table 5, highlights the critical importance of daily recursive updates—grounded in the Markovian assumption—for capturing multi-month interest evolution. Consequently, HSIEN effectively retains historical context while circumventing the computational overhead of re-processing entire sequences.

4.5 Practical Experience For Deployment

In this section, we share our practical experience in deploying HSIEN within a production online serving system. To meet the ultra-low online inference latency requirements of our high-traffic computational advertising scenarios, we adopted an "offline pre-computation and online lookup" strategy to decouple LLM inference from the real-time serving pipeline:

- **Offline Recursive Updates:** We utilize an LLM in daily offline pipelines. Based on the previous block's state and the current block's incremental behavior, the system computes the user's latest explicit semantic states, $S_{u,t}$ and $G_{u,t}$.
- **Vectorized Caching:** The generated semantic text is transformed into dense embeddings via a Text Encoder and stored in high-performance KV storage.
- **Inference-Free:** During real-time requests, the Real-Time Prediction service bypasses LLM inference entirely. It simply retrieves the pre-computed semantic vectors using the User ID and feeds them as standard features into downstream discriminative models.

Through this design, HSIEN integrates the powerful semantic reasoning of LLMs while incurring near-zero additional latency during the online phase (limited only to high-speed KV retrieval), successfully satisfying the stringent requirements of industrial-grade advertising Real-Time Bidding systems for high concurrency and low latency. HSIEN has been fully applied in Meituan's main advertising bidding scenarios.

5 Conclusion

In this paper, we propose HSIEN, a novel sequential recommendation framework that achieves a deep coupling between LLMs and discriminative models. The core innovation of HSIEN lies in its ability to abstract raw user behavior sequences into explicit semantic interest sequences. This abstraction allows for the integration of multi-channel interaction signals, represented as semantic embeddings, into downstream discriminative models for CTR and CVR prediction. Specifically, by constructing a Markov-based state space for each channel, HSIEN enables the efficient recursive updating of local interests. The aggregation of these cross-channel signals further facilitates the distillation of global user interest preferences, thereby building a holistic user profile from both local and global perspectives. In the downstream phase, traditional sequential modeling techniques are employed to seamlessly fuse ID-based collaborative signals with the generated semantic signals.

Compared to existing methods, HSIEN significantly enhances user profiling by leveraging the extensive open-world knowledge inherent in LLMs. Extensive offline experiments demonstrate that our semantic extraction approach leads to a marked improvement in the accuracy of CTR and CVR predictions. Ablation studies further confirm the critical role of the Markov state space-based recursive mechanism in effectively summarizing long-term user interests. Moreover, online A/B testing conducted in a real-world production environment confirms that HSIEN achieves a significant lift in ROI, which has led to its successful full deployment within our advertising bidding system. In future work, we will explore the impact of different prompting strategies on the extraction of users' interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 72271233), Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies (SZS2023007), Smart Social Governance Technology and Innovative Application Platform (YZCXPT2023101), and Innovation System of the Integration between Industry and Education for Smart Governance (CJRH2024101).

References

[Bao *et al.*, 2023] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM conference on recommender systems*, pages 1007–1014, 2023.

[Cao *et al.*, 2022] Yue Cao, Xiaojiang Zhou, Jiaqi Feng, Peihao Huang, Yao Xiao, Dayao Chen, and Sheng Chen. 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*, pages 2974–2983, 2022.

[Chang *et al.*, 2023] Jianxin Chang, Chenbin Zhang, Zhiyi Fu, Xiaoxue Zang, Lin Guan, Jing Lu, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, et al. 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*, pages 3785–3794, 2023.

[Chen *et al.*, 2018] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Sequential recommendation with user memory networks. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 108–116, 2018.

[Chen *et al.*, 2019] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 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*, pages 1–4, 2019.

[Chen *et al.*, 2021] Qiwei Chen, Changhua Pei, Shanshan Lv, Chao Li, Junfeng Ge, and Wenwu Ou. End-to-end user behavior retrieval in click-through rateprediction model. *arXiv preprint arXiv:2108.04468*, 2021.

[Chen *et al.*, 2024] Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling. *arXiv preprint arXiv:2409.12740*, 2024.

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

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

[Feng *et al.*, 2019] Yufei Feng, Fuyu Lv, Weichen Shen, Menghan Wang, Fei Sun, Yu Zhu, and Keping Yang. Deep session interest network for click-through rate prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 2301–2307. International Joint Conferences on Artificial Intelligence Organization, 2019.

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

[Guo *et al.*, 2017] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: A factorization-machine based neural network for ctr prediction. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 1725–1731. International Joint Conferences on Artificial Intelligence, 2017.

[Guo *et al.*, 2020] Qing Guo, Zhu Sun, Jie Zhang, and Yin-Leng Theng. An attentional recurrent neural network for personalized next location recommendation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 34, pages 83–90, 2020.

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

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

[Liao *et al.*, 2024] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1785–1795, 2024.

[McAuley *et al.*, 2015] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[Pi *et al.*, 2019] Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2671–2679, 2019.

[Pi *et al.*, 2020] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 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, pages 2685–2692, 2020.

[Qin *et al.*, 2020] Jiarui Qin, Weinan Zhang, Xin Wu, Jiarui Jin, Yuchen Fang, and Yong Yu. User behavior retrieval for click-through rate prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2347–2356, 2020.

[Rajput *et al.*, 2023] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

[Si *et al.*, 2024] Zihua Si, Lin Guan, ZhongXiang Sun, Xiaoxue Zang, Jing Lu, Yiqun Hui, Xingchao Cao, Zeyu Yang, Yichen Zheng, Dewei Leng, et al. 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*, pages 4890–4897, 2024.

[Su *et al.*, 2023] Jiajie Su, Chaochao Chen, Zibin Lin, Xi Li, Weiming Liu, and Xiaolin Zheng. Personalized behavior-aware transformer for multi-behavior sequential recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6321–6331, 2023.

[Yang *et al.*, 2022] Yuhao Yang, Chao Huang, Lianghao Xia, Yuxuan Liang, Yanwei Yu, and Chenliang Li. Multi-behavior hypergraph-enhanced transformer for sequential recommendation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2263–2274, 2022.

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

[Zhou *et al.*, 2019] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5941–5948, 2019.