

HiSeR: Hybrid Item Encoding for Transformer-based Sequential Recommendation

Nguyen Thi Yen

Banking Academy of Vietnam,
Machine Learning and Applications Lab,
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
yennt@hvn.edu.vn, yennt.NCS@ptit.edu.vn

Tu Minh Phuong*

Machine Learning and Applications Lab,
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
phuongtm@ptit.edu.vn

Abstract—Sequential recommendation systems play a crucial role in predicting the next item a user is likely to interact with. Although recent models have achieved remarkable progress in modeling interaction sequences, their ability to exploit and integrate semantic information from user reviews remains limited. Moreover, effectively incorporating these heterogeneous features into a unified representation for sequential modeling remains an open challenge. In this study, we propose HiSeR: Hybrid Item Encoding for Transformer-based Sequential Recommendation that introduces a novel approach to multimodal information fusion. Unlike prior methods, HiSeR employs a Hybrid Item Encoder capable of deeply synthesizing features from both item IDs and textual reviews, resulting in rich and unified semantic representations for each item. These hybrid representations are then fed into a Transformer-based sequential encoder to model the temporal dynamics of user preferences. Extensive experiments on Amazon Product Reviews dataset demonstrate that HiSeR significantly outperforms several state-of-the-art baselines, confirming the effectiveness of the proposed multimodal fusion strategy in enhancing sequential recommendation quality.

Index Terms—sequential recommendation, hybrid representation learning, transformer-based modeling, review-aware embedding.

I. INTRODUCTION

With the rapid growth of e-commerce and digital platforms, users face an overwhelming number of product and content options. Recommender systems help users quickly find relevant items based on their historical interactions. These systems not only improve user experience, but also improve sales and customer retention for businesses.

Sequential recommendation aims to predict the next item with which a user will interact by modeling the sequence of past behaviors. Each user is represented as a sequence of item identifiers (IDs) corresponding to viewed or purchased items. This approach captures temporal patterns and evolving user preferences more effectively than traditional collaborative filtering, which relies only on user-item matrices. Several models have advanced this area. GRU4Rec [1] employs recurrent neural networks to capture short-term dependencies. Caser [2] uses convolutional layers to identify local sequential patterns. SASRec [3] and BERT4Rec [4] apply self-attention to model both short- and long-term dependencies. Despite

strong performance, these methods depend mainly on item IDs and ignore semantic information available in user reviews.

ID-based models lack interpretability and fail to understand item semantics. In contrast, user reviews contain rich information about item attributes, functionality, and user experiences. For example, a review describing a laptop as “stable performance”, “comfortable keyboard”, and “long battery life” provides explicit signals about user satisfaction and product characteristics. Several studies, such as [5], [6], have explored combining the review text with sequential recommendation. However, most of them treat reviews as auxiliary features rather than an equal source of information. The joint modeling of behavioral and semantic signals within a unified sequential framework remains underexplored.

To address this gap, we propose HiSeR (Hybrid Item Encoding for Transformer-based Sequential Recommendation), a new architecture that integrates sequential and semantic information in a balanced and unified manner. HiSeR introduces a Hybrid Item Encoder that fuses item ID embeddings with semantic representations extracted from pre-trained language models. The resulting hybrid embeddings are fed into a Transformer-based sequential encoder to model user preference dynamics and predict future interactions.

The main contributions of this work are as follows.

- We introduce HiSeR, a unified hybrid sequential recommendation framework that jointly leverages user interaction sequences and textual reviews within a single model.
- We design a Hybrid Item Encoder that fuses item ID with semantic representations from pre-trained language models, and integrate these enriched item embeddings into a Transformer-based sequential encoder to effectively capture both short-term transitions and global contextual dependencies in user preference dynamics.
- We conducted extensive experiments on two Amazon Review datasets to validate the effectiveness and robustness of HiSeR, showing consistent performance improvements in multiple evaluation metrics.

The rest of this paper is organized as follows. Section II reviews related work. Section III describes the proposed HiSeR architecture. Section IV presents the experiments and analysis. Section V concludes the paper and discusses future directions.

*Corresponding author.

II. RELATED WORK

Research on sequential recommendation has evolved substantially over the past decade, progressing from early Markov-based formulations to modern deep learning architectures. Classical models assume that a user's next action depends primarily on recent interactions. Representative methods such as FPMC [7] integrate Markov Chains with Matrix Factorization to capture both short-term transitions and long-term preferences, while Fossil [8] further enhance this direction by incorporating item similarity and hybrid long-short patterns. However, these approaches struggle to model complex nonlinear dependencies. With the introduction of deep learning, GRU4Rec [1] became the first model to apply recurrent neural networks to sequential recommendation, enabling the capture of long-range temporal signals. CNN-based models—including Caser [2], NextItNet [9], and RecConRec [10]—exploit local behavioral patterns via convolutional operations, effectively learning short-term and mid-range dependencies. More recent self-attention-based approaches such as SASRec [3] and BERT4Rec [4] have further advanced the field by learning bidirectional dependencies across the entire sequence, marking a paradigm shift from RNNs toward Transformer-based sequential modeling.

Parallel to ID-sequence modeling, another research line focuses on leveraging textual reviews to learn rich semantic representations of users and items. Early review-based models such as DeepCoNN [11] utilize dual CNN encoders to extract textual features, while NARRE [12] introduces an attention mechanism to identify informative reviews. Aspect-aware models like ANR [13] and RPR [14] incorporate aspect-level or sentiment-level information to better capture fine-grained preferences. With the rise of pre-trained language models, U-BERT [15] adopts BERT-style pre-training and fine-tuning for review modeling, whereas SBRec [16] employs Sentence-BERT to generate sentence-level embeddings that better reflect semantic similarity. Despite their effectiveness, these approaches generally lack the ability to capture sequential behavioral dynamics.

To integrate both behavioral and semantic signals, several hybrid models have been proposed, combining item IDs with review information. RNS [5] enhances user preference modeling by jointly learning aspect-aware CNN representations and hierarchical attention over reviews. PIRSP [17] fuses interaction sequences and review sequences through a gating mechanism to model short-term preference shifts. RETSR [18] enriches sequential representations by injecting review-derived semantics at each timestep, capturing contextual and temporal variations. More recent work such as CCA [6] employs cascaded cross-attention to model intricate relationships between product sequences, review texts, and candidate items. Although these hybrid frameworks provide significant improvements, many still struggle to maintain a balanced integration between ID-based behavioral signals and review-based semantic cues, often leading to overemphasis on one modality.

Motivated by this gap, the present study introduces HiSeR, a unified architecture designed to jointly and coherently incorporate sequential ID interactions and semantically enriched review representations, ensuring a more balanced and comprehensive modeling of user preference dynamics.

III. PROPOSED METHOD

A. Problem Definition

Let $\mathcal{U}$ denote the set of users and $\mathcal{I}$ denote the set of items. Each user $u \in \mathcal{U}$ is associated with a chronological interaction sequence $S_u = [i_1, i_2, \dots, i_{|S_u|}]$, where $i_t \in \mathcal{I}$ represents the t -th item that user u has interacted with (e.g., viewed, purchased, or reviewed). The sequence S_u reflects the temporal evolution of the user's preferences and serves as the primary signal in sequential recommendation tasks.

Unlike models that rely solely on IDs, this study additionally leverages semantic information extracted from textual reviews to enhance item representations. For each item $i_t \in S_u$, two types of information are considered: (1) Item ID, a unique identifier representing the item in the interaction space; and (2) Review set $\mathcal{R}(i_t) = \{r_1, r_2, \dots, r_{|\mathcal{R}(i_t)|}\}$, consisting of user-generated reviews describing item characteristics, opinions, and user experiences.

Hence, the input sequence for user u is represented as a set of ordered pairs $X_u = \{(i_t, \mathcal{R}(i_t)) \mid t = 1, 2, \dots, |S_u|\}$, where the ordering preserves temporal dependencies among interactions.

The goal of sequential recommendation is to predict the next item $i_{|S_u|+1}$ that user u is most likely to interact with by jointly exploiting behavioral signals from item ID sequences and semantic cues from the corresponding reviews. The system then ranks all candidate items and recommends the top- K items with the highest predicted scores.

B. Method Overview

Figure 1 illustrates the overall architecture of the proposed HiSeR (Hybrid Item Encoding for Transformer-based Sequential Recommendation) model. The main objective of HiSeR is to jointly integrate two complementary types of information: behavioral signals from sequential item IDs interactions and semantic signals extracted from textual reviews. By combining these modalities, the model can more accurately capture the temporal evolution of user preferences.

Each item in a user's interaction sequence is first encoded by the Hybrid Item Encoder (HIE), which serves as the core component of HiSeR. The HIE consists of two parallel branches: (1) an **ID-based branch**, which transforms the item identifier into a dense embedding vector that reflects collaborative behavioral patterns; and (2) a **review-based branch**, which encodes textual reviews associated with the item using a pre-trained language model to extract semantic representations. The two embeddings are then fused to form a unified hybrid item representation that preserves both collaborative and semantic characteristics.

Next, positional embeddings are added to these hybrid item representations to encode the order and temporal relationships

among user interactions. The ordered sequence is then fed into a Transformer-based Sequential Encoder, which captures dependencies between items and learns a dynamic user representation that reflects the user's current intent within the interaction context.

For each candidate item, HiSeR also applies the same Hybrid Item Encoder to obtain its corresponding representation. The model then computes the similarity between the user representation and each candidate item embedding to estimate the relevance score. Items with the highest scores are returned as recommendations.

To train the model, HiSeR employs negative sampling for next-item ranking while jointly optimizing an auxiliary rating regression loss, enabling the system to leverage both implicit behavioral signals and explicit preference information.

Subsequent sections provide the technical details of each component and the model training procedure.

C. Hybrid Item Encoder (HIE)

The Hybrid Item Encoder (HIE) is the core component of the proposed HiSeR model. It is designed to seamlessly integrate two complementary sources of information: (1) the behavioral signal derived from item IDs, and (2) the semantic signal extracted from user-generated reviews. The objective of the HIE is to map each item $i_t \in \mathcal{I}$ into a unified representation v_{i_t} that combines both collaborative and semantic knowledge.

a) ID-based Representation: In sequential recommendation, the input consists of a chronologically ordered sequence of item identifiers that a user has interacted with. For each item $i_t \in \mathcal{I}$, its ID-based embedding is defined as

$$e_{i_t}^{\text{ID}} = E_{\text{ID}}[i_t], \quad (1)$$

where $E_{\text{ID}} \in \mathbb{R}^{|\mathcal{I}| \times d}$ is the learnable item embedding matrix containing vector representations for all items. This embedding captures collaborative behavioral relationships learned from user-item interactions during training.

b) Review-based Representation: In addition to the interaction signal, reviews provide rich semantic information that cannot be captured by ID embeddings alone. Each item i_t is associated with a set of textual reviews written by multiple users, denoted as $\mathcal{R}(i_t) = \{r_1, r_2, \dots, r_{|\mathcal{R}(i_t)|}\}$. These reviews describe product characteristics, quality, and user experiences. For example, reviews for a "winter jacket" may include phrases such as "thick and warm material," "ideal for cold weather," or "modern and easy-to-match design," offering semantic cues beyond what the item identifier provides.

To extract semantic features from reviews, prior work has used CNN-based models [11], [12] and pre-trained language model [15]. In HiSeR, we adopt Sentence-BERT (SBERT), a pre-trained language model optimized for semantic similarity and sentence-level embeddings. SBERT [19], [16] inherits BERT's representational power but is fine-tuned with contrastive objectives, enabling efficient mapping of semantically similar sentences into nearby points in the embedding space with reduced computational cost.

For each review r_j , SBERT produces a sentence embedding as

$$m_{i_t,j}^R = \text{SBERT}(r_j), \quad (2)$$

where $m_{i_t,j}^R$ denotes the semantic embedding of review r_j .

Not all reviews contribute equally to product representation. Some reviews contain generic or redundant information (e.g., "fast delivery," "good packaging"), whereas others describe salient product aspects (e.g., "high-quality sound," "elegant design," "battery lasts over 10 hours"). To emphasize more informative reviews, HiSeR employs an attention-based aggregation mechanism that assigns importance weights to each review.

The attention weight for review r_j is defined as

$$\alpha_{i_t,j} = \frac{\exp(m_{i_t,j}^R W_a)}{\sum_{k=1}^{|\mathcal{R}(i_t)|} \exp(m_{i_t,k}^R W_a)}, \quad (3)$$

where W_a is a learnable parameter.

The review-based representation of item i_t is computed as the attention-weighted sum of review embeddings:

$$e_{i_t}^R = \sum_{j=1}^{|\mathcal{R}(i_t)|} \alpha_{i_t,j} m_{i_t,j}^R. \quad (4)$$

This attention pooling mechanism enables HiSeR to emphasize semantically informative reviews while reducing the influence of less relevant ones.

c) Fusion of Dual Representations: To obtain a unified item representation, HiSeR fuses the ID-based and review-based embeddings through vector concatenation:

$$v_{i_t} = [e_{i_t}^{\text{ID}}, e_{i_t}^R], \quad (5)$$

where v_{i_t} denotes the hybrid embedding that jointly preserves both behavioral and semantic information within a shared representation space. This fused representation provides a richer and more comprehensive embedding for each item.

D. Sequential Encoder

The Sequential Encoder aims to model the dependencies among items in a user's interaction history, capturing how previous behaviors influence subsequent user actions.

a) Embedding Layer: Since the length of the interaction sequence $|S_u|$ varies across users, all sequences are normalized to a fixed length L . If $|S_u| < L$, zero-padding is applied to the beginning of the sequence; if $|S_u| > L$, only the most recent L items are retained. This truncation strategy follows prior work such as [3], [4], which demonstrated that recent interactions have stronger predictive power for future behavior.

Each item i_t in the sequence is represented by the Hybrid Item Encoder (HIE), which produces a unified embedding that combines both ID-based and review-based information:

$$v_{i_t} = \text{HIE}(i_t), \quad (6)$$

where $v_{i_t} \in \mathbb{R}^d$ denotes the hybrid item representation.

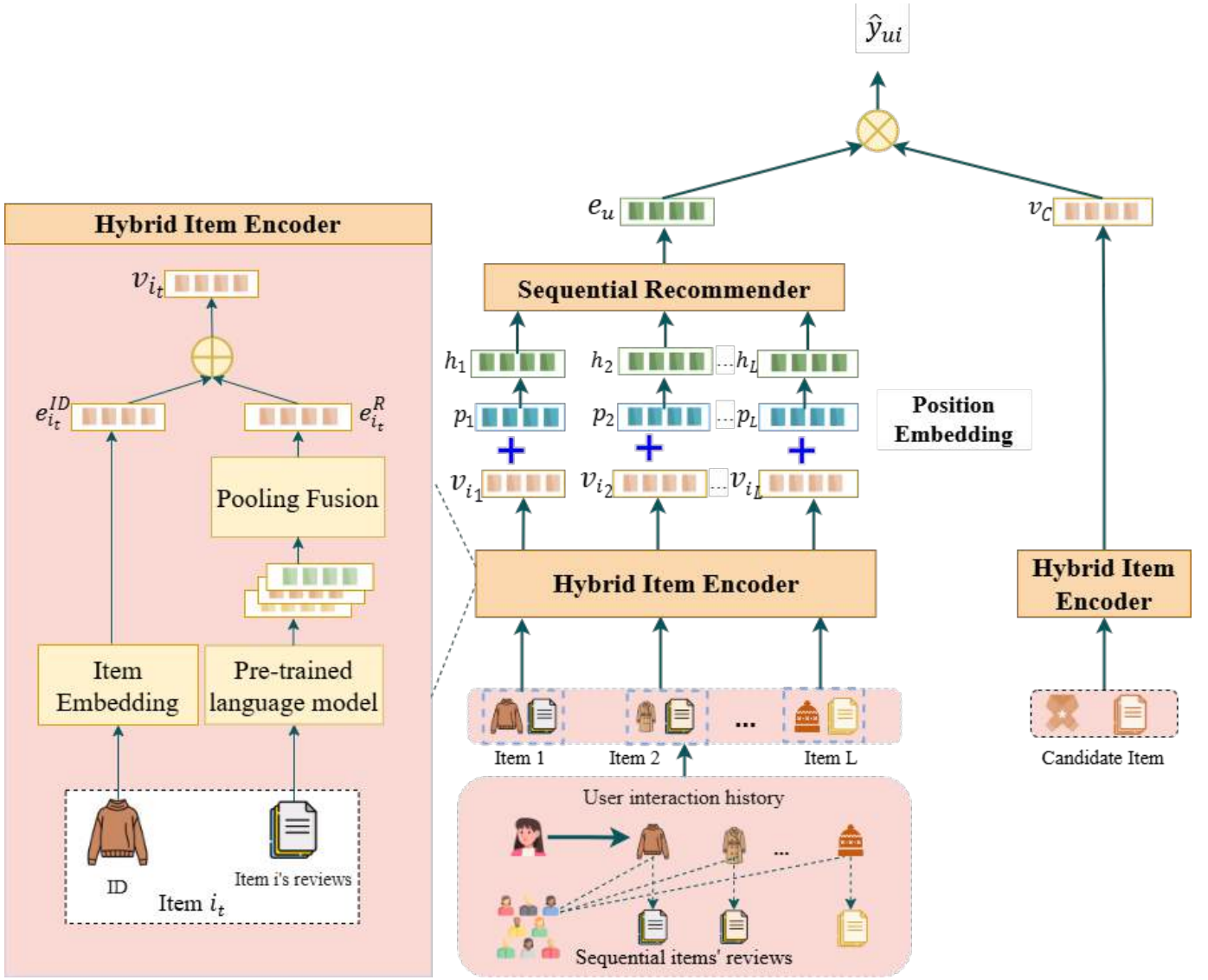


Fig. 1. The network architecture of HiSeR.

Position Embedding. Because the Transformer architecture lacks recurrence or convolutional operations, it cannot inherently model the order of sequence elements. To address this limitation, a positional embedding $p_t \in \mathbb{R}^d$ is added to each item representation:

$$h_t = v_{i_t} + p_t, \quad (7)$$

where h_t denotes the position-aware input vector for the t -th item.

The resulting sequence of vectors forms the input to the sequential encoder:

$$H_u = [h_1, h_2, \dots, h_L]. \quad (8)$$

b) Sequential Modeling: Various approaches have been proposed to model user sequential behaviors in recommendation tasks. RNN-based models [1] simulate temporal dynamics through recurrent state propagation, enabling the capture of

short-term dependencies but suffering from limited ability to model long sequences due to the vanishing gradient problem. CNN-based models [2], [9] are effective in extracting local behavioral patterns but lack flexibility in capturing long-range interactions.

Recently, Transformer-based architectures such as SASRec [3] and BERT4Rec [4] have achieved superior performance by leveraging the self-attention mechanism, which enables learning global dependencies across sequence elements without relying on step-by-step processing. Building upon this advantage, HiSeR adopts a transformer encoder as the backbone of its sequential encoder to jointly capture short-term and long-term item dependencies while benefiting from highly parallelizable training.

Given the input sequence $H_u = [h_1, h_2, \dots, h_L]$, the model processes it through a stack of transformer encoder layers. Each layer consists of two main components: (1) *Multi-*

Head Self-Attention, which learns attention weights among all items in the sequence; and (2) *Feed-Forward Network*, which performs nonlinear transformations on the extracted contextual features. These components are linked through residual connections and layer normalization to stabilize training and preserve contextual information across layers.

After passing through ℓ transformer layers, the model obtains the contextualized representations

$$H_u^{(\ell)} = [h_1^{(\ell)}, h_2^{(\ell)}, \dots, h_L^{(\ell)}], \quad (9)$$

where each vector $h_t^{(\ell)}$ reflects both the behavioral signal at position t and the aggregated influence of the entire interaction history.

In HiSeR, the overall representation of the user's current state is computed by averaging all contextual vectors in the encoded sequence:

$$e_u = \frac{1}{L} \sum_{t=1}^L h_t^{(\ell)}, \quad (10)$$

where $e_u \in \mathbb{R}^d$ denotes the mean-pooled user embedding. This strategy leverages the transformer's ability to encode global information at every position through self-attention, resulting in a stable and balanced representation that fully reflects the user's interaction history without requiring special tokens or additional pooling mechanisms.

E. Prediction Layer

After obtaining the user representation e_u from the Sequential Encoder, HiSeR predicts the relevance between the user and each candidate item. This prediction is computed by combining two sources of representations: (1) the user representation e_u , which reflects the user's current preference state; and (2) the candidate item representation v_C , extracted by the Hybrid Item Encoder using both the item ID and its associated reviews.

The final prediction score, indicating the likelihood that user u will interact with candidate item C , is obtained via the dot product between e_u and v_C :

$$\hat{y}_{u,i} = e_u^\top v_C, \quad (11)$$

where a larger value of $\hat{y}_{u,i}$ implies a higher degree of alignment between the user's current interests and the characteristics of the candidate item.

F. Model Optimization

HiSeR is trained with a multi-objective learning scheme that combines a ranking-based loss for next-item prediction and a regression loss for rating estimation. This joint optimization allows the model to leverage both implicit sequential behavior and explicit preference signals.

For each user u , the interaction sequence S_u is segmented into a set of sliding windows $\mathcal{C}_u$, where each window $w_j \in \mathcal{C}_u$ has length $L + 1$. The first L items in w_j are used as historical context, and the last item is treated as the positive target. For every window w_j , we randomly sample a set of negative items

N_j that do not appear in w_j . Let $\hat{y}_u^{(j)}$ denote the predicted relevance score for the positive target in window w_j , and $\hat{y}_{u,i}$ the score for a negative item $i \in N_j$. The ranking loss is defined as

$$\mathcal{L}_{\text{rec}} = \sum_u \sum_{w_j \in \mathcal{C}_u} \left[-\log \sigma(\hat{y}_u^{(j)}) - \sum_{i \in N_j} \log (1 - \sigma(\hat{y}_{u,i})) \right], \quad (12)$$

where $\sigma(\cdot)$ is the sigmoid function.

In addition to the ranking objective, HiSeR also exploits explicit ratings as an auxiliary supervision signal. For each window w_j , we concatenate the user representation and the corresponding positive item embedding to predict the rating $\hat{r}_u^{(j)}$. Let $r_u^{(j)}$ denote the ground-truth rating associated with this interaction. The rating regression loss is given by the mean squared error:

$$\mathcal{L}_{\text{rating}} = \frac{1}{|\mathcal{D}_r|} \sum_{(u,j) \in \mathcal{D}_r} (\hat{r}_u^{(j)} - r_u^{(j)})^2, \quad (13)$$

where $\mathcal{D}_r$ is the set of all valid user-target pairs with observed ratings.

The final training objective combines the two loss terms with ℓ_2 regularization on the model parameters:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{rating}} + \lambda \|\Theta\|, \quad (14)$$

where α controls the contribution of the rating prediction task (set to $\beta = 10$ in our experiments) and λ is the regularization coefficient. This joint optimization encourages HiSeR to accurately rank candidate items while also aligning its representations with explicit user ratings.

IV. EXPERIMENTS

This section describes the experimental setup used to evaluate the effectiveness of the proposed method and compare it with existing approaches.

A. Dataset

We conduct experiments on two Amazon Product Reviews datasets: Pet Supplies and Tools and Home Improvement, representing two distinct categories of consumer behavior.

- **Pet Supplies:** includes user reviews related to pet-care products.
- **Tools and Home Improvement:** contains reviews focusing on technical, repair, and home-improvement products.

Each record in the datasets consists of a user ID, item ID, timestamp, rating score, and review text. Following the pre-processing protocol described in prior work [5], we filter out users with fewer than 10 interactions to ensure sufficiently long histories for sequential modeling and to extract meaningful semantic features from review texts.

For each remaining user, all interactions are chronologically sorted to construct an interaction sequence that reflects the temporal evolution of user preferences.

The processed datasets are then split in chronological order: the earliest 70% of each user's sequence is used for training,

and the most recent 30% is reserved for testing. This approach preserves temporal causality and aligns well with real-world next-item prediction scenarios.

Table I summarizes the dataset statistics after preprocessing. Both datasets exhibit extremely high sparsity (above 99.7%), which is typical for large-scale online review corpora. On average, each user has approximately 16 interactions, and the review texts contain rich descriptive content, providing valuable semantic context for learning text-enhanced item representations.

TABLE I
DATASET STATISTICS

Statistic	Pet Supplies	Tools and Home Improvement
# Users	7,417	10,076
# Items	33,798	66,710
# Interactions	117,385	169,245
Sparsity	99.95%	99.97%
Avg. actions/user	15.8	16.8

B. Experimental Setup

a) *Evaluation Metrics*: Recommender systems typically present users with a short ranked list of top- K items. In our experiments, we set $K = 5$ and evaluate model performance using four widely adopted metrics: Precision@5, Recall@5, NDCG@5, and Hit Ratio (HR@5).

- **Precision@5** measures the proportion of relevant items appearing in the top-5 recommendations.
- **Recall@5** evaluates the extent to which the ground-truth relevant items are covered by the top-5 list.
- **NDCG@5** accounts for the ranking positions of correctly recommended items, placing higher emphasis on items ranked earlier.
- **HR@5** reflects the probability that at least one ground-truth item appears in the top-5 recommendations.

To reduce computational cost and ensure fair comparison, we follow the negative-sampling strategy used in prior studies [3], [5]. Specifically, for each user, 100 unobserved items are randomly sampled as negatives and ranked together with the ground-truth item, from which the evaluation metrics are computed.

b) *Comparison Methods*: To evaluate the effectiveness of HiSeR, we compare it against a set of representative baselines spanning three major categories: (1) ID-based sequential recommendation, (2) deep sequential models, and (3) hybrid models incorporating review information.

- **ID-based Sequential Models**. These methods model user behavior purely through item interaction sequences. FPMC [7] integrates Matrix Factorization with Markov Chains to jointly capture long-term preferences and sequential transitions, and is trained using Bayesian Personalized Ranking (BPR). Caser [2] applies Convolutional Neural Networks to extract local sequential patterns, while SASRec [3] leverages self-attention to identify and

weight informative interactions, effectively modeling both short-term and long-term dependencies.

- **Deep Sequential Models**. These approaches utilize deep neural architectures to encode interaction sequences. GRU4Rec [1] is a pioneering work employing Gated Recurrent Units (GRU) for session-based recommendation, modeling the full interaction history rather than relying solely on the most recent item. RUM [20] extends this idea by integrating a Memory-Augmented Neural Network (MANN), enabling flexible storage and retrieval of user histories instead of compressing them into a single vector. It includes two variants: RUM-I at the item level and RUM-F at the feature level, enabling more expressive user representations. These models established the foundation for applying deep architectures to sequential behavior learning.
- **Hybrid Models Incorporating Review Information**. These methods combine sequential interaction data with semantic information extracted from user reviews. RNS [5] learns user and item representations from reviews using aspect-based embeddings and hierarchical attention, integrating semantic signals with sequential behaviors. PIRSP [17] adopts two parallel branches—one for IDs and one for reviews—using GRU, CNN, and gating-based fusion to capture both short-term and long-term preferences. RETSR [18] focuses on selecting informative reviews and modeling temporal delays between interactions through dual time-gating mechanisms, while TASR [21] incorporates temporal information and multi-level attention to capture evolving behavioral patterns.

c) *Implementation Details*: For all baseline models, we adopt the parameter configurations recommended in their original papers.

For HiSeR, we employ the Sentence-BERT (all-MiniLM-L6-v2) model to generate review embeddings. The model is optimized using AdamW with a learning rate of 0.001, an embedding size of 384, a training batch size of 32, a testing batch size of 1, and a dropout rate of 10%. To ensure fair and consistent comparison with the baselines, we set the Sequential Encoder to use 2 Transformer layers, 4 attention heads, and a sequence length of $L = 5$.

C. Results and Discussion

The results in Table II show that HiSeR consistently outperforms all baseline models on both datasets, achieving the highest performance across all evaluation metrics. This improvement demonstrates the effectiveness and generalizability of the proposed model across different product domains—from emotionally driven and repetitive-purchase items (Pet Supplies) to more technical, rational-evaluation products (Tools and Home Improvement).

Compared with ID-only sequential models such as FPMC, GRU4Rec, and SASRec, HiSeR exhibits clear advantages by incorporating additional semantic signals extracted from review texts. Although review-based models such as RNS, PIRSP, and RETSR also achieve competitive results, they

TABLE II
PERFORMANCE COMPARISON OF HiSeR WITH SEQUENTIAL AND REVIEW-ENHANCED RECOMMENDATION MODELS. THE HIGHEST VALUES ARE HIGHLIGHTED IN BOLD.

Model	Pet Supplies				Tools and Home Improvement			
	<i>Precision@5</i>	<i>Recall@5</i>	<i>NDCG@5</i>	<i>HR@5</i>	<i>Precision@5</i>	<i>Recall@5</i>	<i>NDCG@5</i>	<i>HR@5</i>
FPMC	0.0985	0.1009	0.1075	0.3761	0.0785	0.0828	0.0911	0.3047
GRU4Rec	0.0943	0.0940	0.1015	0.3646	0.0712	0.0747	0.0821	0.2915
RUM(I)	0.1013	0.1005	0.1086	0.3790	0.0790	0.0853	0.0936	0.3166
RUM(F)	0.1041	0.1047	0.1146	0.3973	0.0803	0.0855	0.0925	0.3199
SASRec	0.1026	0.1105	0.1241	0.3922	0.0820	0.0875	0.1036	0.3384
Caser	0.1038	0.1136	0.1252	0.3975	0.0819	0.0873	0.1029	0.3378
RNS	0.1146	0.1252	0.1388	0.4362	0.0894	0.0943	0.1120	0.3614
PIRSP	0.1264	0.1363	0.1462	0.4575	0.0989	0.1043	0.1213	0.3836
RETSR	0.1254	0.1367	0.1521	0.4650	0.1016	0.1076	0.1274	0.3977
TASR	0.1147	0.1348	0.1479	0.4480	0.0897	0.1022	0.1193	0.3740
HiSeR	0.1528	0.1662	0.1864	0.5367	0.1176	0.1249	0.1449	0.4441

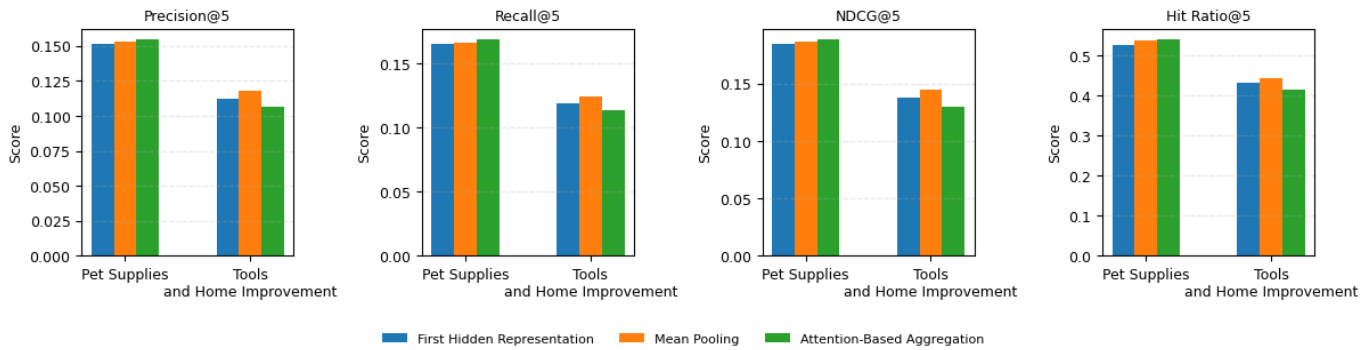


Fig. 2. Comparison of User Representation Aggregation Techniques.

remain limited by the absence of a unified mechanism that jointly learns sequential behavioral patterns and textual semantics within a single architecture.

HiSeR addresses these limitations through two key innovations:

- 1) a Hybrid Item Encoder that integrates ID-based and review-based features into a shared representation space; and
- 2) a Transformer-based Sequential Encoder that effectively captures both short-term and long-term contextual dependencies.

Overall, the experimental results indicate that HiSeR serves as a unified and robust recommendation framework capable of simultaneously leveraging ID signals, semantic information, and sequential dependencies, thereby delivering superior and consistent performance across diverse datasets.

D. Analysis of User Representation Aggregation Techniques

We conduct an ablation study to compare three strategies for aggregating user states after the Transformer encoder: (1) First Hidden Representation, (2) Mean Pooling, and (3) Attention-Based Aggregation. The comparison results are illustrated in Figure 2, which visualizes the performance of all three approaches across the four evaluation metrics (Precision@5,

Recall@5, NDCG@5, and HR@5). The results show that Mean Pooling yields the most stable performance across both datasets.

On the Pet Supplies dataset, the attention-based method occasionally produces slightly higher scores but does not maintain consistency across all metrics. On the Tools and Home Improvement dataset, Mean Pooling achieves the highest values for all four metrics, clearly outperforming the other two methods. In contrast, First Hidden Representation and Attention-Based Aggregation show larger variability between datasets and exhibit less stable performance.

Since each contextual vector in the sequence already incorporates global information through the Transformer's self-attention mechanism, Mean Pooling offers a simple yet robust way to combine these representations, reducing the impact of irregular interaction patterns and yielding more consistent user embeddings across datasets.

V. CONCLUSION

In this paper, we introduced HiSeR, a hybrid sequential recommendation framework that jointly integrates collaborative signals derived from item identifiers with semantic information extracted from textual reviews. By unifying these two complementary modalities through a Hybrid Item Encoder

and modeling user dynamics with a Transformer-based sequential architecture, HiSeR effectively captures both short-term and long-term preference patterns while enriching item representations with meaningful semantic cues.

Extensive experiments on two Amazon Review datasets demonstrate that HiSeR consistently outperforms existing ID-based, review-based, and hybrid baselines across multiple evaluation metrics. The ablation studies further highlight the effectiveness of the hybrid item representation and the robustness of the mean-pooling strategy for user-state aggregation.

Beyond next-item prediction, HiSeR incorporates an auxiliary rating regression objective, enabling the model to leverage fine-grained feedback signals and further enhance representation quality. This multi-task training paradigm contributes to improved ranking performance and provides a more comprehensive understanding of user preferences.

Future work may explore extending HiSeR to multi-modal settings (e.g., images or product attributes), incorporating temporal dynamics more explicitly, and applying large language models to improve semantic encoding of review content. Overall, the proposed framework provides a flexible and effective foundation for building next-generation recommender systems that jointly exploit behavioral and textual information.

ACKNOWLEDGMENT

We sincerely thank the Machine Learning and Applications Lab at the Posts and Telecommunications Institute of Technology and the Banking Academy of Vietnam for their invaluable support.

REFERENCES

- [1] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2015.
- [2] J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 565–573.
- [3] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 197–206.
- [4] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.
- [5] C. Li, X. Niu, X. Luo, Z. Chen, and C. Quan, "A review-driven neural model for sequential recommendation," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, p. 2866–2872.
- [6] B. Huang, J. Luo, W. Du, W. Pan, and Z. Ming, "Cascaded cross attention for review-based sequential recommendation," in *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2023, pp. 170–179.
- [7] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 811–820.
- [8] R. He and J. McAuley, "Fusing similarity models with markov chains for sparse sequential recommendation," IEEE, pp. 191–200, 2016.
- [9] L. Chen, F. Yuan, J. Yang, X. Ao, C. Li, and M. Yang, "A user-adaptive layer selection framework for very deep sequential recommender models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 3984–3991.
- [10] N. X. Bach, D. H. Long, and T. M. Phuong, "Recurrent convolutional networks for session-based recommendations," *Neurocomputing*, vol. 411, pp. 247–258, 2020.
- [11] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in *Proceedings of the tenth ACM international conference on web search and data mining*, 2017, pp. 425–434.
- [12] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural attentional rating regression with review-level explanations," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1583–1592.
- [13] J. Y. Chin, K. Zhao, S. Joty, and G. Cong, "Anr: Aspect-based neural recommender," in *Proceedings of the 27th ACM International conference on information and knowledge management*, 2018, pp. 147–156.
- [14] H. Liu, Y. Guo, J. Yin, Z. Gao, and L. Nie, "Review polarity-wise recommender," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10 039–10 050, 2022.
- [15] Z. Qiu, X. Wu, J. Gao, and W. Fan, "U-bert: Pre-training user representations for improved recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4320–4327.
- [16] N. T. Yen, N. Do Hai, and T. M. Phuong, "Review based recommendation with pre-trained transformer and hierarchical attention networks," in *2024 18th International Conference on Advanced Computing and Analytics (ACOMPA)*. IEEE, 2024, pp. 9–15.
- [17] J. Zhang, X. Mu, P. Zhao, K. Kang, and C. Ma, "Improving current interest with item and review sequential patterns for sequential recommendation," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104348, 2021.
- [18] W. Wang, J. Li, Y. Wang, and Z. Zhang, "Retrs: An effective review-enhanced and time-aware sequential recommendation framework," in *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2022, pp. 826–831.
- [19] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.
- [20] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, and H. Zha, "Sequential recommendation with user memory networks," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 108–116.
- [21] T. Liu, T. Li, B. Song, Y. Du, H. Yang, and L. Ding, "A time-aware sequential recommendation based on attention mechanism," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 2024, pp. 123–134.



Nguyen-Thi Yen is a lecturer at the Banking Academy of Vietnam and a PhD candidate at the Posts and Telecommunications Institute of Technology. Her research interests include machine learning, natural language processing and recommender systems. She is also involved in developing applications for digital transformation in the banking sector.



Tu Minh Phuong is professor of computer science and Chairman of University Council at Posts and Telecommunications Institute of Technology. His research interests include machine learning, recommender systems, natural language processing, and computer vision. He received Ph.D. degree in control in technical systems from the National Academy of Sciences, Uzbekistan, in 1995.