Spatial-Temporal Perceiving: Deciphering User Hierarchical Intent in Session-Based Recommendation

Xiao Wang, **Tingting Dai**, **Qiao Liu*** and **Shuang Liang**
University of Electronic Science and Technology of China
wangxiao16@std.uestc.edu.cn, ttdai_18@outlook.com, qliu@uestc.edu.cn, shuangliang@uestc.edu.cn

Abstract

Session-based recommendation (SBR) aims to predict the next-interacted item based on anonymous users' behavior sequences. The main challenge is how to recognize the user intent with limited interactions to achieve a more accurate inference of user behavior. Existing works usually regard several consecutive items in the interaction session as intent units. However, we argue such intent generation based on temporal transition ignores the fact that each item also has its semantically connected items in the feature space, which can be regarded as spatial intent. The limited consideration of intent fails to capture complex behavioral patterns in real-world scenarios, leading to sub-optimal solutions. To address this issue, we propose the **H**ie**r**archical **Int**ent Perceiving Contrastive Learning Framework (**HearInt**) for SBR, which proposes a hierarchical consideration of intents from both temporal and spatial perspective. Specifically, we first propose that the user's temporal intents are mutually exclusive while the spatial intents are mutually compatible. Following these analyses, we design a Temporal Intent Decoupling module to mitigate the mutual interference of long-term and short-term intents, and a Cross-scale Contrastive Learning task to enhance the consistency of intents across different spatial scales. Experimental results on three real-world datasets exhibit that HearInt achieves state-of-the-art performance.

1 Introduction

Recommender systems are pivotal in managing information overload by effectively understanding user profiles and long-term behavior [Xie *et al.*, 2022; Sun *et al.*, 2023]. Unfortunately, such information may be unavailable in real-world scenarios due to privacy concerns [Hidasi *et al.*, 2016; Gao *et al.*, 2022]. In response, SBR emerges, and its goal is to predict the next item based on the anonymous user's interaction sequence (denoted as a session) [Latifi *et al.*, 2021].

In the past decade, many approaches have emerged in SBR and have demonstrated superiority in capturing the dependency between items within the session. For example, Recurrent neural networks (RNNs) based models [Hidasi *et al.*, 2016; Li *et al.*, 2017; Liu *et al.*, 2018; Qiu *et al.*, 2022] are devoted to learning sequential dependency of consecutive items, while graph neural networks (GNNs) based models [Wu *et al.*, 2019; Qiu *et al.*, 2019; Wang *et al.*, 2020; Pan *et al.*, 2020; Xia *et al.*, 2021b; Xia *et al.*, 2021a; Lai *et al.*, 2022] convert the current session into a graph and capture complex dependency between adjacent items. However, these approaches only take individual items as basic units to extract user preference, neglecting to dive deeply into the user intent arising from the combination of items.

The user intent is the driving force for generating interaction sequences, which can be generally represented by a collection of items that exhibit some shared characteristics. Recently, some studies [Guo *et al.*, 2022; Li *et al.*, 2022; Zhang *et al.*, 2023] extract segments that contain different consecutive items to represent various intents for better inferring user preferences. However, the generation of these intents only considers the temporal dependencies of items, which is insufficient to fully express the user's complex intents in real-world scenarios. For instance, consider a session: *Mouse - Computer - Headphone - Smartphone - Smartphone - Smartphone*. If a model extracts intents solely based on a segment of consecutive items, it may classify the user as a smartphone enthusiast, neglecting the possibility that the user is actually an electronics aficionado. This limitation may constrain the scope of related recommendations. In contrast, if we consider semantically related items as intents in the feature space, we can observe that all items in this session belong to the category of electronic products. Therefore, it is crucial to extract intent by integrating distant items in the session that are semantically related or even related items from other sessions, which is neglected by existing works.

To address this limitation, we propose a hierarchical consideration of user intent. As illustrated in Figure 1, the session consists of different intents from different hierarchical levels:

(1) From the temporal perspective, the user's behavior can be represented by long-term and short-term intents. Generally, since short-term intents can evolve rapidly according to some external factors, long-term and short-term intents tend to be mutually exclusive [Liu *et al.*, 2018; Zheng *et al.*, 2022].

*Corresponding author.

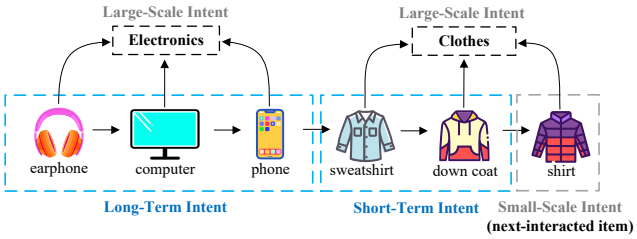


Figure 1: A toy example of the hierarchical intent hidden in the session, where the blue font represents intents from the temporal level and the gray font represents intents from the spatial level.

For example, in Figure1, the long-term intent of this session is *Electronics*, and its short-term intent has shifted to *Clothes* due to recent temperature changes. Therefore, simply conflating these intents during the training process cannot avoid their mutual interference, resulting in sub-optimal session representations learned by the model.

(2) From the spatial perspective, we consider the set of semantically similar items in the feature space as user intents. Based on the number of items included, we refine them into two categories, including the small-scale intent (e.g., the next-interacted item in Figure 1) and the large-scale intent (e.g., the next-interacted category in Figure 1). The large-scale intent is composed of many small-scale intents and can be regarded as a representative of their shared properties. Therefore, large-scale and small-scale intents are compatible with each other.

To the end, we propose a novel **Hierar**chical **Int**ent Perceiving Contrastive Learning Framework (**HearInt**), which incorporates the hierarchical intent from the temporal and spatial level to comprehensively extract the behavioral patterns. **For temporal intent**, we design a Temporal Intent Decoupling (TID) module to disentangle long-term and short-term intents from the session, aiming to eliminate their mutual interference during the training process. Subsequently, each intent is fed into encoders to obtain its sub-session representation. Following this, we employ a gated mechanism that adaptively combines the above representations to generate session representation for predicting the next item. **For spatial intent**, we first apply a clustering algorithm to all items in the feature space to learn the semantic relevance between items and obtain the item's centroids (i.e., categories). Then, we consider item categories as large-scale intents and items as small-scale intents. Based on it, we introduce a Cross-scale Contrastive Learning task (CCL), whose positive signals are session representation and the next-interacted category. Through this task, semantic relevance between session representation and potentially interacting items is enhanced, promoting the recommendation of more relevant items. The source code and datasets are publicly available on GitHub [1].

In summary, the main contributions of this work include:

- We propose a novel **HearInt** model with spatial and temporal perspectives, which takes full advantage of the hierarchical intent hidden in user behaviors session, promoting recommendation performance.
- We design a TID module to separately learn the representations of long-term and short-term intents, avoiding their mutual interference.
- We propose a cross-scale contrastive learning task, which learns the correlation between the session representation and the large-scale intent to generate more relevant recommendations.
- Experiments on three benchmark datasets show the effectiveness and superiority of HearInt compared with the state-of-the-art approaches.

[1]https://github.com/jarviswww/Code4HearInt

2 Related Work

2.1 Session-based Recommendation

Initial approaches in session-based recommendation make simplistic assumptions and employ Markov chains to extract short-term interest representations of users [Shani *et al.*, 2005; Rendle *et al.*, 2010]. Then, as deep learning demonstrates advantages in modeling complex information, researchers have introduced RNNs [Chung *et al.*, 2014] to learn the temporal relation among items. STAMP [Liu *et al.*, 2018] introduces a short-term memory priority module to emphasize the last click in the session. After that, with the great success of Transformer in other fields, the self-attention mechanism has been introduced to the SBR. For example, SASRec [Kang and McAuley, 2018] firstly leverages the self-attention mechanism to extract the context dependency within sessions. Besides, the significant superiority exhibited by GNNs [Scarselli *et al.*, 2008; Kipf and Welling, 2017] in modeling complex relations, GNNs-based approaches have also gained much attention from researchers in SBR. For example, SRGNN [Wu *et al.*, 2019] innovatively leverages a gated graph neural network to model sessions as graphs. GCE-GNN [Wang *et al.*, 2020] introduces a global graph to capture item transitions across all sessions, acknowledging that a graph based solely on a single session is inadequate for comprehensive modeling of that session. Following this idea, many models are proposed with contrastive learning[Xia *et al.*, 2021b][Xia *et al.*, 2021a]. However, these approaches only take individual items as basic units to extract user preference, neglecting to dive deeply into the user intent arising from the combination of items.

2.2 Intent Perceiving in Session-based Recommendation

In recent years, there has been an emergence of research in session-based recommendation focusing on user intents. The intent is considered to be the engine that drives the generation of the session. Thus, a comprehensive analysis of it is beneficial for modeling the behavioral patterns hidden in sessions. MIHSG [Guo *et al.*, 2022] constructs multi-granular intent graphs by considering different segments of sessions as different user intents, thereby capturing the interactions between intent units of different granularity. HIDE [Li *et al.*, 2022] disentangles the intents under each item click in micro and macro manners to capture the dynamic intents of users and avoid noisy signals. Atten-Mixer [Zhang *et al.*, 2023] leverages different combinations of recent interactions to represent

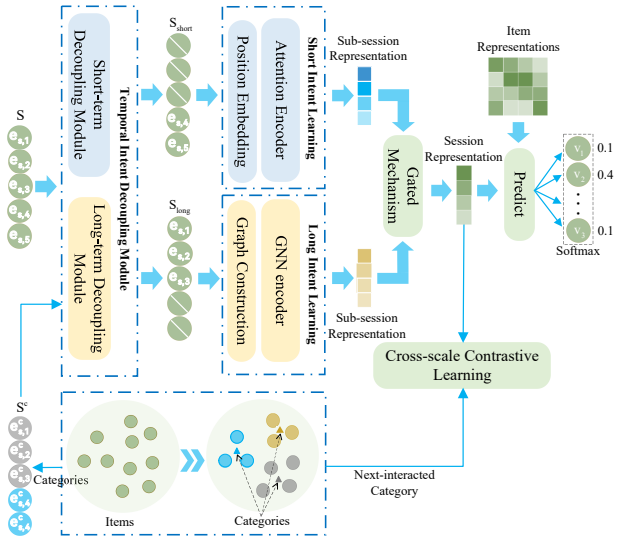


Figure 2: Overall architecture of our proposed HearInt.

multiple user intents, and exploits them to augment the readout module of GNNs. However, although these methods have paved the way for considering user intents in session-based recommendation, they have the limitation of not considering the collaborative effect of intents from different levels. Besides, they do not take into account the mutual interference between different intents, which can affect the accuracy of the recommendations.

3 Method

In this section, we formulate the problem and elaborate on our proposed HearInt, whose basic structure is given in Figure 2. First, we introduce the Temporal Intent Decoupling (TID) that acquires two sub-sessions from the session. Then, we separately introduce the learning process of each sub-session. After introducing the basic components, we will introduce the two learning tasks, which are the next-item prediction task and the Cross-scale Contrastive Learning (CCL) task.

3.1 Problem Statement

Let $V = \{v_1, v_2, .., v_N\}$ denote a set of unique items, where N is the total number of items. For each session, we represent it as $S = \{v_{s,1}, v_{s,2}, ..., v_{s,t}\}$ where $v_{s,i} \in V (1 \leq i \leq t)$ represents the interacted item at time step i. We embed each item into the same feature space and denote $e_j \in \mathbb{R}^d$ as the vector representation of item $v_j \in V$. The embedded session is denoted as $S = \{e_{s,1}, e_{s,2}, ..., e_{s,t}\}$. The session-based recommendation aims to predict the next-interacted item, namely $v_{s,t+1}$, for a given session S.

The categories (i.e., clustering centroids) of items are acquired through clustering. Let each category be denoted as $c_k \in C$ $(1 \leq k \leq K)$, where C is the category set and K is the number of categories, and its vector representations are denoted as e^c_k. For a given embedded session S, we denote its category version as $S^c = \{e^c_{s,1}, e^c_{s,2}, ..., e^c_{s,t}\}$, where item embeddings are replaced by their category embeddings.

3.2 Temporal Intent Decoupling Module

To alleviate the mutual interference between long-term and short-term intents in the current session, we conduct a decoupling operation under the guidance of item categories. Given a session $S = \{e_{s,1}, e_{s,2}, ..., e_{s,t}\}$ and its category version $S^c = \{e^c_{s,1}, e^c_{s,2}, ..., e^c_{s,t}\}$. For the short-term intent, previous studies [Zheng *et al.*, 2022; Li *et al.*, 2022; Liu *et al.*, 2024] typically focus only on the recently clicked items, overlooking that this intent might have appeared earlier in the session. This leads to a loss of potentially valuable information. Therefore, we innovatively consider the category of the last-interacted item as the short-term intent, allowing us to explore more informative intent representation.

Besides, we define the long-term intent as the category that appears most frequently in the session, different from the previous approach [Zheng *et al.*, 2022] of considering the mean of all item representations as the long-term intent, which also includes irrelevant short-term information. We formulate the definition of long-term and short-term intents as follows:

$$e^c_{\text{long}} = \{\text{most-frequent } e^c_{s,i} | e^c_{s,i} \in S^c\}, \tag{1}$$

$$e^c_{\text{short}} = \{\text{last-interacted } e^c_{s,i} | e^c_{s,i} \in S^c\}. \tag{2}$$

For each intent, we compute its cosine similarity with all item categories in S^c, obtaining similarity scores for all items. Subsequently, under the constraints of a threshold $\alpha \in (0, 1)$ and a mask probability $\beta \in (0, 1)$, items with the cosine similarity higher than α are retained, while those with a cosine similarity lower than α are set to 0 with a probability β. This decoupling operation is performed twice, thereby obtaining the decoupled results for both long-term and short-term intents. This process can be formulated as (Using long-term intent as an example):

$$m_i = \begin{cases} 1, & \cos(e^c_{\text{long}}, e^c_{s,i}) \geq \alpha \\ 0, & \cos(e^c_{\text{long}}, e^c_{s,i}) < \alpha \quad \text{and} \quad \hat{\beta}_i < \beta \\ 1, & \cos(e^c_{\text{long}}, e^c_{s,i}) < \alpha \quad \text{and} \quad \hat{\beta}_i > \beta, \end{cases} \tag{3}$$

$$S_{\text{long}} = M \odot S, \tag{4}$$

where $\hat{\beta}_i$ is the randomly generated probability for item $e^c_{s,i}$, and M is the mask sequence. These operations enable the decoupling of the input session into two sub-sessions S_{long} and S_{short} related to long and short-term intents, respectively.

3.3 Representation Learning of Decoupled Session

We next present how to learn sub-session representations through two different networks. Here, we first describe the learning of short-term intent-oriented sub-session representation for S_{short}, whose core is neural attention, and then introduce the learning of long-term intent-oriented sub-session representation for S_{long}, whose core is GNN.

Short-term Intent-oriented Sub-session Representation

Traditional methods merely model the last click and ignore other semantically similar items in the session, resulting in incomplete and inaccurate short-term user preferences. To alleviate the problem, based on the generated sub-session S_{short}, which contains semantically similar items with the last click,

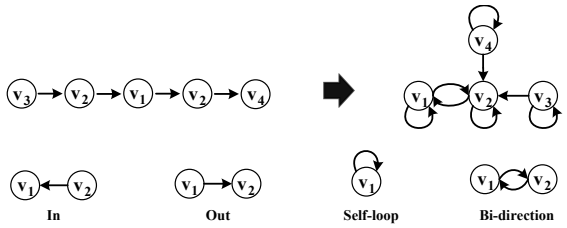


Figure 3: Illustration of the direction-aware graph.

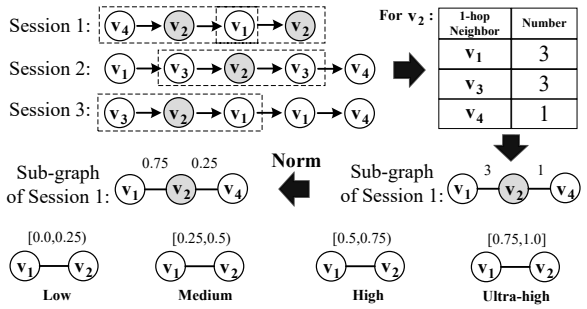


Figure 4: Illustration of the frequency-aware graph. Here, we only give the construction of the subgraph around v_2 in Session 1 as an example.

we generate the transformed information of each item in the sub-session using the self-attention mechanism.

Given the sub-session S_{short}, we initialize a learnable position embedding P and perform an item-wise addition between P and S_{short}, which is formulated as $S_{\text{short}} = [S_{\text{short}} + P]$. After that, we employ a scaled dot-product attention network to extract the semantic and sequential dependency between items within the session, and incorporate a position-wise feed-forward network after it to introduce more non-linearity. This process can be formulated as:

$$M' = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{5}$$

$$\hat{S}_{\text{short}} = \text{SELU}(M'W_1 + b_1)W_2 + b_2, \tag{6}$$

where Q is the query, K is the key matrix, V is the value matrix, $W_1, W_2 \in \mathbb{R}^{d\times d}$ are weight matrices, $b_1, b_2 \in \mathbb{R}^d$ are bias vectors, $K = V = S_{\text{short}}$, and $Q = \text{SELU}(W_0 S_{\text{short}} + b_0)$ where $W_0 \in \mathbb{R}^{d\times d}$ and $b_0 \in \mathbb{R}^d$ refer to weight matrix and bias vector, respectively. After aggregating the information of semantically related neighbors of the last-interacted item, we consider its representation $e_{s,t} \in \mathbb{R}^d$ as the sub-session representation, and denote it as R_{short}.

Long-term Intent-oriented Sub-session Representation

Graph Construction. To better aggregate information of the long-term intent-oriented sub-session, we first reconstruct S_{long} into two edge-aware graphs to explore more intricate edge transition relationships. The proposed two edge-aware graphs are defined as:

(i) *Direction-aware Graph*, which constructs four types of edges according to different item transition patterns, including *in*, *out*, *bidirection*, and *self-loop*, as shown in Figure 3. Given a session $S = \{v_{s,1}, v_{s,2}, ..., v_{s,t}\}$, we form a directed edge-aware graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, where $\mathcal{V}_s \subseteq V$ refers to the set of items in S, $\mathcal{E}_s = \{\boldsymbol{e}_{ij}\}$ is the set of edge which indicates the item-transition pattern between $v_{s,i}$ and $v_{s,j}$.

(ii) *Frequency-aware Graph*, whose construction process is shown in Figure 4. For a specific item in a session, we initially construct a local undirected graph based on the connections between items in the sub-session. To determine the edge weights in this graph, we calculate the co-occurrence times of each item with its neighbors across all sessions in datasets and then normalize them within the local undirected graph to ascertain the weights. Four types of edges are assigned to this graph according to the normalized weights. This approach enables the utilization of global information to guide the aggregation of information in the local undirected graph, thereby complementing the directed graph that relies solely on item transitions in the local session. Given a session $S = \{v_{s,1}, v_{s,2}, ..., v_{s,t}\}$, we form a frequency-aware graph $\mathcal{G}_f = (\mathcal{V}_f, \mathcal{E}_f)$, where $\mathcal{V}_f \subseteq V$ refers to the set of items in S, $\mathcal{E}_f = \{\boldsymbol{e}^f_{ij}\}$ is the set of frequency edges from $v_{s,i}$ to $v_{s,j}$.

GNN Encoder. After constructing edge-aware graphs, we employ the GNN to refine the information aggregation process by assigning different learnable parameters to different types of edges. Considering the different importance of neighboring items, we employ attention mechanisms to learn the weights between adjacent item pairs. The attention score can be calculated as:

$$a_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T_{r_{ij}}(e_{s,i} \odot e_{s,j})\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T_{r_{ik}}(e_{s,i} \odot e_{s,k})\right)\right)}, \tag{7}$$

where a_{ij} estimates the importance weight from item j to item i, r_{ij} is the type of edge from item i to item j, $\mathbf{a}^T_{r_{ij}}$ is the learnable parameter assigned to edge r_{ij}, $\mathcal{N}_i$ is the 1-order neighbors around item i in sub-session S_{long}, and $e_{s,i}$,$e_{s,j} \in S_{\text{long}}$. Then, we update the representation of item i by performing a linear combination of neighbors associated with the attention scores:

$$\hat{s}_i = \sum_{j\in\mathcal{N}_i} a_{ij} e_{s,j}. \tag{8}$$

For both edge-aware graphs, we employ the same transformation process as illustrated in Equ. (7) and (8), and denote the obtained sub-sessions as $\hat{S}^1_{\text{long}}$ and $\hat{S}^2_{\text{long}}$. A gate mechanism is then employed to aggregate information from both graphs as follows, where W_3 are trainable parameters.

$$\theta = \text{sigmoid}\left(W_3[\hat{S}^1_{\text{long}} || \hat{S}^2_{\text{long}}]\right), \tag{9}$$

$$\hat{S}_{\text{long}} = (1-\theta) \odot \hat{S}^1_{\text{long}} + \theta \odot \hat{S}^2_{\text{long}}, \tag{10}$$

To distinguish and aggregate the contribution made by each item to the current sub-session, we adopt the soft attention mechanism to learn the sub-session representation.

$$\gamma_i = g_1^T \text{sigmoid}\left(W_5 \hat{s}''_{s,i} + W_6 \hat{s}'_{s,i} + b_4\right), \tag{11}$$

$$R_{\text{long}} = \sum_{i=1}^{n} \gamma_i \hat{s}_{s,i}, \tag{12}$$

where $\hat{s}_{s,i} \in \hat{S}_{\text{long}}$, $\hat{s}'_{s,i}$ represents average representation in $\hat{S}_{\text{long}}$, and $\hat{s}''_{s,i} = \tanh\left(W_4\left[\hat{s}_{s,i}||p_i\right] + b_3\right)$ where $p_i \in P$. $W_4 \in \mathbb{R}^{d\times 2d}, W_5, W_6 \in \mathbb{R}^{d\times d}, b_3, b_4 \in \mathbb{R}^d$ are trainable parameters, and γ_i represents the attention weight.

After acquiring sub-session representations R_{short} and R_{long}, we employ a gate mechanism to adaptively aggregate them to generate the final session representation:

$$\delta = \text{sigmoid}\left(W_7[R_{\text{long}}||R_{\text{short}}]\right), \tag{13}$$

$$R = (1-\delta)\odot R_{\text{long}} + \delta \odot R_{\text{short}}, \tag{14}$$

where $W_7 \in \mathbb{R}^{d\times 2d}$ is the learnable parameter and R denotes the final session representation.

3.4 Next-item Prediction Task

Given session representation R, we multiply it with candidate item embeddings and apply a softmax to calculate the probabilities of each item being the next one:

$$\hat{y_i} = \text{softmax}(R^T e_j), \tag{15}$$

where e_j is the vector representation of $v_j \in V$. Then, we adopt a cross-entropy loss function as the learning objective, which is defined as:

$$\mathcal{L}_p = -\sum_{j=1}^{m} y_i \log(\hat{y_i}), \tag{16}$$

where y_i is the one-hot encoding vector of the ground truth.

3.5 Cross-scale Contrastive Learning Task

To enhance the compatibility within the spatial intents, we employ a cross-scale contrastive learning task. Different from the typical same-scale contrastive learning, here we employ two samples with different scales, which are the next-interacted category $e^c_{s,t+1}$ and the session representation R, as contrastive signals. It is worth noting that since the session representation gradually converges with the representation of the next-interacted item (i.e., small-scale intent) in the next-item prediction task, we also consider it a kind of small-scale signal. Then, we adopt InfoNCE [Oord *et al.*, 2018] to maximize the similarity between R and $e^c_{s,t+1}$ (positive), while minimizing the similarity between R and other categories (negative). This process can be formulated as:

$$\mathcal{L}_c = -\log \frac{\exp\left(R \cdot e^c_{s,t+1}/\tau\right)}{\sum_{e_k \in C} \exp\left(R \cdot e_k/\tau\right)}, \tag{17}$$

where τ is the temperature parameter that controls the discrimination towards negative samples.

To incorporate the cross-scale contrastive learning task into the next-item prediction task, we introduce a multi-task learning strategy. Specifically, a hyper-parameter λ is introduced to control the magnitude of cross-scale contrastive loss. The total learning loss for session S can be expressed as:

$$\mathcal{L} = \mathcal{L}_p + \lambda \mathcal{L}_c. \tag{18}$$

4 Experiments

In this section, we first illustrate experiment setups, including datasets, baselines, evaluation metrics, and hyper-parameters. Then, we analyze comparison experimental results.

Dataset	Tmall	RetailRocket	Diginetica
training sessions	351,268	433,643	719,470
test sessions	25,898	15,132	60,858
# of items	40,728	36,968	43,097
average lengths	6.69	5.43	5.12

Table 1: Statistics of three datasets

4.1 Experiment Setups

Datasets

We evaluate the proposed model on three benchmark datasets, namely $Tmall$[2], $RetailRocket$[3],$Diginetica$[4]. *Tmall* is from the IJCAI-15 competition and consists of shopping logs of unnamed users on Tmall online shopping platform. *RetailRocket* is released by an e-commerce corporation for the Kaggle competition and contains users' browsing activity. *Diginetica* comes from CIKM Cup 2016 and describes the music listening behavior of users. Following [Xia *et al.*, 2021a], we conduct preprocessing over each dataset. Specifically, sessions with a length of 1 and items that appeared fewer than 5 times are excluded. Then, the latest data (such as the data from last week) is set to be a test set, and previous data is used as the training set. Additionally, we use a sequence splitting preprocess method to augment session $S = s_1, s_2, ..., s_n$ in these datasets, and generate multiple sessions with corresponding labels $([s_1]; s_2), ([s_1, s_2]; s_3), ..., ([s_1, s_2, ..., s_{n-1}]; s_n)$. The statistics of the datasets are presented in Table 1.

Metrics

Following [Li *et al.*, 2017], we adopt P@K (Precision) and MRR@K (Mean Reciprocal Rank) to evaluate the recommendation results, where K is 10 or 20.

Hyper-parameters Setup

For the general setting, the embedding size is 100, the batch size is 100. For HearInt, the initial learning rate is 0.001, which will decay by 0.6 after every 1 epoch. We employ the k-means as the clustering algorithm and set the number of clusters to 100. The threshold α of cosine similarity is 0 in three datasets. Mask probability β is 0.4 for *Tmall* , while 0.1 is for *Diginetica* and *Retailrocket*. The θ The best number of layers in both attention and GNN encoders is 1,2,2 for *Tmall*, *Retailrocket*, and *Diginetica*, respectively. For the baseline models, we refer to their best parameter setups reported in the original papers and directly report their results if in line with general settings, evaluation metrics, and datasets. Otherwise, we record the reproduced results under the public code.

4.2 Baselines

We compare HearInt with the following approaches. :

- **NARM** [Li *et al.*, 2017] adopts RNNs and the attention mechanism to extract the general interest of users.
- **STAMP** [Liu *et al.*, 2018] emphasizes the short-term memory through leveraging the attention mechanism.

[2]https://tianchi.aliyun.com/dataset/dataDetail?dataId=42

[3]https://www.kaggle.com/retailrocket/ecommerce-dataset

[4]https://competitions.codalab.org/competitions/11161

Datasets	RetailRocket				Tmall				Diginetica			
Metrics	P@10	MRR@10	P@20	MRR@20	P@10	MRR@10	P@20	MRR@20	P@10	MRR@10	P@20	MRR@20
NARM	42.07	24.88	50.22	24.59	19.17	10.42	23.30	10.70	35.44	15.13	49.70	16.17
STAMP	42.95	24.61	50.96	25.17	22.63	13.12	26.47	13.36	33.98	14.26	45.64	14.32
SR-GNN	43.21	26.07	50.32	26.57	23.41	13.45	27.57	13.72	36.86	15.52	50.73	17.59
GCE-GNN	46.05	27.48	53.63	28.01	29.19	15.55	33.42	15.42	41.54	<u>18.29</u>	<u>54.22</u>	19.04
S^2-DHCN	46.15	26.85	53.66	27.30	26.22	14.60	31.42	15.05	39.87	17.53	53.18	18.44
COTREC	48.61	<u>29.46</u>	<u>56.17</u>	<u>29.97</u>	30.62	17.65	36.35	18.04	**41.88**	18.16	54.18	<u>19.07</u>
HIDE	43.95	25.70	51.25	26.20	31.10	16.77	37.12	17.19	40.39	17.45	53.72	18.37
Atten-Mixer	<u>48.63</u>	28.05	56.01	28.57	<u>31.62</u>	<u>18.23</u>	<u>37.24</u>	<u>18.62</u>	40.24	17.32	53.86	18.27
HearInt	**49.62**	**30.02**	**57.39**	**30.54**	**33.87**	**18.85**	**39.92**	**19.27**	<u>41.47</u>	**18.60**	**55.02**	**19.52**

Table 2: Performances of comparison approaches on three datasets. The boldface is the best result, and the underline is the second best result.

- **SR-GNN** [Wu *et al.*, 2019] employs GNNs to learn item embeddings and learn the session representation through a soft-attention layer.
- **GCE-GNN** [Wang *et al.*, 2020] consider the extraction of spatial information from global and local views.
- **S^2-DHCN** [Xia *et al.*, 2021b] generates two different views from the original session using the hypergraph, regarding them as self-supervised signals to enhance the session representation.
- **COTREC** [Xia *et al.*, 2021a] forms two independent session representations and adopts co-training to generate positive and negative samples pertaining to the last click from the candidate items.
- **HIDE** [Li *et al.*, 2022] disentangles the intents under each item click in micro and macro manners to capture the dynamic intents of users and avoid noisy signals.
- **Atten-Mixer** [Zhang *et al.*, 2023] considers compositions of the last few items in a session as different intents to employ a multi-level reasoning component for GNNs.

4.3 Overall Performance

To evaluate the effectiveness of HearInt, we report the comparison results with the state-of-the-art baselines. From Table 2, we draw the following observations:

- Among the sequential baseline models, STAMP outperforms NARM on *Tmall* and *RetailRocket* but lags on *Diginetica*. As shown in Table 1, the average session length is longer in *Tmall* and *RetailRocket* compared to *Diginetica*. This phenomenon shows that STAMP explicitly emphasizes short-term intents to alleviate the interference of long-range items, leading to better performance in modeling user interests within longer sessions. It further validates our previous view that long-term and short-term intents can mutually influence each other.
- In general, GNN-based models outperform the sequential models, which reveals that leveraging the captured complex instead of single sequential transitions between items is more beneficial in learning the session representation. Compared with SR-GNN, models that like GCE-GNN, S^2-DHCN, and COTREC exhibit more effectiveness. It indicates that integrating item-transitions information from other sessions into the current session improves the infer of user behavior patterns.
- Our proposed HearInt almost surpasses the overall baselines on these three datasets, demonstrating the usefulness of the proposed hierarchical intent. Compared to existing models, the hierarchical intent benefits the model in two main aspects: (i) devised the TID module allows HearInt to avoid introducing irrelevant information in long-term intent and increase related information(such as semantically similar items) in short-term intent, making learned session representation superior. (ii) incorporating CCL task, whose positive signals are session representation and category of the next click, makes semantic relevance between session representation and the next click closer during the training process. The above modules enable HearInt to acquire session representations containing more relevant information, which enhances the relevance of recommendations results.

4.4 Model Analysis and Discussion

In this subsection, we take an in-depth model analysis study, aiming to further understand the framework of HearInt. Due to the space limit, we only show the analysis results with K=20 in the partial datasets. We have obtained similar experimental results in terms of other metrics and datasets.

Ablation Study

To profoundly comprehend the contribution of the hierarchical intent component in HearInt, we designed five variants: **HearInt-base**, **HearInt-S**, **HearInt-L**, **HearInt-NT**, and **HearInt-NC**. HearInt-base removes the TID module and CCL. HearInt-NT removes the TID module but contains the CCL module. HearInt-NC includes the TID module but drops the CCL module. HearInt-S discards the entire long-term intent channel (i.e., GNN encoder), while HearInt-L discards the entire short-term intent channel (i.e., attention encoder). Table 3 illustrates the comparison results. From Table 3, we have the following observation: (i) HearInt with the best results while HearInt-base with the worst performance, demonstrating that introducing semantic-related information

Datasets	Tmall		Diginetica	
Metrics	P@20	MRR@20	P@20	MRR@20
HearInt	**39.92**	**19.37**	**55.02**	**19.52**
HearInt-base	37.59	17.82	54.58	18.97
HearInt-S	38.60	19.27	53.80	18.86
HearInt-L	38.04	18.58	54.51	19.24
HearInt-NT	39.68	18.98	54.70	19.33
HearInt-NC	37.83	18.07	54.60	19.20

Table 3: Performance of variant models on P@20 and MRR@20.

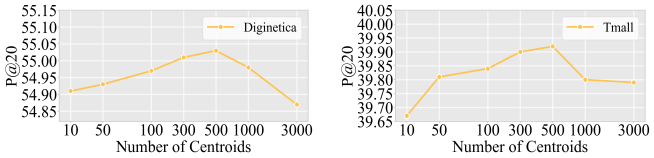


Figure 5: Impact of the number of clustering centroids on P@20.

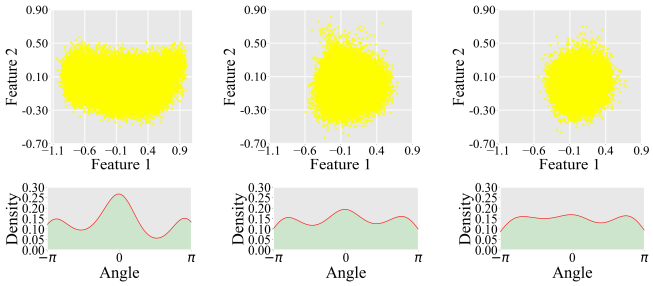
in both spatial and temporal perspectives is effective in enhancing recommendation relevance through TID and CCL. (ii) Compared to HearInt, HearInt-S with short-term intent and HearInt-L with long-term intent make more contributions in *Tmall* and *Dignetica*, respectively. However, the average length of sessions in *Tmall* is longer than in *Dignetica*, as shown in Table 1. This further demonstrates that simultaneously incorporating decoupled intent is positive in learning better session representation. (iii) HearInt outperforms HearInt-NT. It reveals that utilizing contrastive learning with maximum consistency between session representation and the category of the next click improves the recommendation's relevance.

Influence of Clustering Centriods

To explore the impact of clustering centroids, we range the numbers of clustering centroids within $\{10, 50, 100, 300, 500, 1000, 3000\}$. Figure 5 illustrates that as the number of clustering centroids increases, the performance of P@20 on two datasets exhibits a hump-shaped trend. This is because, as the number of clustering centroids increases, the negative samples in contrastive learning gradually increase, causing the constraint in the denominator to shift from 'reducing irrelevant samples' to 'reducing relevant samples'.

Influence of the Cross-scale Contrastive Learning

To evaluate how dose cross-scale contrastive learning enhances the recommendation performance, we visualize the ablation model's item embeddings and their distribution in the feature space. Specifically, we employ the singular value decomposition (SVD) to project the high-dimensional embeddings into a 2-D space and conduct the Gaussian kernel density estimation (KDE) on angles to depict the density of embedding distribution across different orientations. Results are reported in Figure 6. For models like STAMP without contrastive learning, the item embeddings are notably uneven, concentrating around three specific angles. When it comes to HearInt-NC, although its distribution of items is more uniform than STAMP, it still shows a steeper distribution curve



(a) STAMP (b) HearInt-NC (c) HearInt

Figure 6: Item embeddings and distribution on *Diginetica*.

SessionID	No.14		No.15	
Range	Top-50	Top-100	Top-50	Top-100
HearInt	**15**	**23**	**12**	**21**
HearInt-base	11	16	8	15

Table 4: Statistics of items contained the next-interacted category within Top-50 and Top-100 recommendation list on *Diginetica*

than HearInt. It indicates that learning the correlation between the session representation and the large-scale intent enhances the uniformity of items, resulting in a more uniform distribution of item embeddings. The uniform distribution maintains the differences between different items since non-uniformity implies that some items are gathered in feature space. Besides, it can prevent the model from focusing excessively on items gathered in specific regions during training and ignoring items scattered elsewhere in feature space.

Case Study

To straightforwardly perceive that HearInt introduces more semantically relevant items in recommendation lists, we randomly pick the No.14 and No.15 sessions from the test set as cases to explain. The results are shown in Table 4. Compared with HearInt-base, it can be observed that the Top-K recommendation results of HearInt include more items whose categories are the same as the next-interacted category. This proves that our proposed model indeed benefits in recommending more relevant items.

5 Conclusion

In this paper, we propose a novel session-based recommendation framework based on the leveraging of the hierarchical intent. First, we explore the relation between intentions at temporal and spatial hierarchical levels, proposing that the long-term and short-term intents exhibit mutual exclusivity while the large-scale and small-scale intents are mutually compatible. Then, we present a temporal intent decoupling module to mitigate mutual interference among temporal intents and implement a cross-scale contrastive learning task to enhance consistency among spatial intents. Extensive experiments conducted on three datasets confirm that our proposed framework, HearInt, outperforms state-of-the-art methods.

Ethical Statement

We affirm that our manuscript is original, unpublished, and not under consideration elsewhere. No part of our study, including data, has been fabricated or manipulated.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable discussions and constructive feedback. This work was supported by the National Key R&D Program of China(2022YFB4300603), the National Natural Science Foundation of China (U22B2061), the Sichuan Science and Technology Program (2023YFG0151), and the Development of a Big Data-based Platform for Analyzing the Coupling Relationship of Strip Production Processes Project (211129).

References

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS Workshop on Deep Learning*, Quebec, Canada, 2014.

[Gao *et al.*, 2022] Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. Graph neural networks for recommender system. In *WSDM*, pages 1623–1625, AZ, USA, 2022. ACM.

[Guo *et al.*, 2022] Jiayan Guo, Yaming Yang, Xiangchen Song, Yuan Zhang, Yujing Wang, Jing Bai, and Yan Zhang. Learning multi-granularity consecutive user intent unit for session-based recommendation. In *WSDM*, pages 343–352, AZ, USA, 2022. ACM.

[Hidasi *et al.*, 2016] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *ICLR*, San Juan, Puerto Rico, 2016.

[Kang and McAuley, 2018] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *ICDM*, pages 197–206, Singapore, 2018. IEEE.

[Kipf and Welling, 2017] TN Kipf and M Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, Toulon, France, 2017. ICLR.

[Lai *et al.*, 2022] Siqi Lai, Erli Meng, Fan Zhang, Chenliang Li, Bin Wang, and Aixin Sun. An attribute-driven mirror graph network for session-based recommendation. In *SIGIR*, pages 1674–1683, Madrid, Spain, 2022. ACM.

[Latifi *et al.*, 2021] Sara Latifi, Noemi Mauro, and Jannach. Session-aware recommendation: A surprising quest for the state-of-the-art. *Information Sciences*, 573:291–315, 2021.

[Li *et al.*, 2017] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In *CIKM*, page 1419–1428, Singapore, 2017. ACM.

[Li *et al.*, 2022] Yinfeng Li, Chen Gao, Hengliang Luo, Depeng Jin, and Yong Li. Enhancing hypergraph neural networks with intent disentanglement for session-based recommendation. In *SIGIR*, pages 1997–2002, Madrid, Spain, 2022. ACM.

[Liu *et al.*, 2018] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. Stamp: short-term attention/memory priority model for session-based recommendation. In *SIGKDD*, pages 1831–1839, London,United Kingdom, 2018. ACM.

[Liu *et al.*, 2024] Huafeng Liu, Mingjie Zhou, Mingyang Song, Deqiang Ouyang, Yawen Li, Liping Jing, Jian Yu, and Michael K Ng. Learning hierarchical preferences for recommendation with mixture intention neural stochastic processes. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2024.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Pan *et al.*, 2020] Zhiqiang Pan, Fei Cai, Wanyu Chen, Honghui Chen, and Maarten De Rijke. Star graph neural networks for session-based recommendation. In *CIKM*, pages 1195–1204, Virtual Event, 2020. ACM.

[Qiu *et al.*, 2019] Ruihong Qiu, Jingjing Li, Zi Huang, and Hongzhi Yin. Rethinking the item order in session-based recommendation with graph neural networks. In *CIKM*, pages 579–588, Beijing, China, 2019. ACM.

[Qiu *et al.*, 2022] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. Contrastive learning for representation degeneration problem in sequential recommendation. In *WSDM*, pages 813–823, Virtual Event, 2022. ACM.

[Rendle *et al.*, 2010] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *WWW*, pages 811–820, Raleigh North Carolina, USA, 2010. ACM.

[Scarselli *et al.*, 2008] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

[Shani *et al.*, 2005] Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. An mdp-based recommender system. *Journal of Machine Learning Research*, 6(9):1265–1295, 2005.

[Sun *et al.*, 2023] Zhu Sun, Hui Fang, Jie Yang, Xinghua Qu, Hongyang Liu, Di Yu, Yew-Soon Ong, and Jie Zhang. Daisyrec 2.0: Benchmarking recommendation for rigorous evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8206–8226, 2023.

[Wang *et al.*, 2020] Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. Global context enhanced graph neural networks for session-based recommendation. In *SIGIR*, pages 169–178, Virtual Event, 2020. ACM.

[Wu *et al.*, 2019] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *AAAI*, pages 346–353, Hawaii, USA, 2019.

[Xia *et al.*, 2021a] Xin Xia, Hongzhi Yin, Junliang Yu, Yingxia Shao, and Lizhen Cui. Self-supervised graph co-training for session-based recommendation. In *CIKM*, pages 2180–2190, GA, USA, 2021. ACM.

[Xia *et al.*, 2021b] Xin Xia, Hongzhi Yin, Junliang Yu, Qinyong Wang, Lizhen Cui, and Xiangliang Zhang. Self-supervised hypergraph convolutional networks for session-based recommendation. In *AAAI*, pages 4503–4511, Virtual Event, 2021.

[Xie *et al.*, 2022] Yueqi Xie, Peilin Zhou, and Sunghun Kim. Decoupled side information fusion for sequential recommendation. In *SIGIR*, pages 1611–1621, Madrid, Spain, 2022. ACM.

[Zhang *et al.*, 2023] Peiyan Zhang, Jiayan Guo, Chaozhuo Li, Yueqi Xie, Jaeboum Kim, Yan Zhang, Xing Xie, Haohan Wang, and Sunghun Kim. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *WSDM*, pages 168–176, Singapore, 2023. ACM.

[Zheng *et al.*, 2022] Yu Zheng, Chen Gao, Jianxin Chang, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. Disentangling long and short-term interests for recommendation. In *WWW*, pages 2256–2267, Lyon, France, 2022. ACM.