

Article

MintSR: Multi-Intent Perception with Time-Based Segmentation for Session-Based Recommendation

Yuan Huang, Ruizhi Yin * , Xiaozheng Zhou and Pengwei Shi

School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056009, China

* Correspondence: 13513491314@163.com

Abstract

In an anonymous session recommendation scenario, the key to predicting the user's next interaction item is accurately recognizing the user's intent. Traditional methods focusing on the item Identity Document (ID) modeling have limitations. They often only consider the current preference based on the target session and the last clicked item, neglecting the user's long-term intention tendencies and dynamic changes in intent over time. Moreover, solely relying on ID modeling can result in biased recommendations towards popular items, overlooking subtle shifts in user preferences. To address these issues, this paper introduces the MintSR session recommendation algorithm. MintSR utilizes a hierarchical decoupling model based on time dimensions, categorizing user intent into long-term, short-term, and instantaneous levels. The global context information is mapped to long-term intent, and in the long-term intent, we use the Transformer encoder to eliminate the problem of long-distance dependency and design a unique intent capture module that captures the intent through the category similarity, which can be used to capture the user's intent through category similarity. The same capture module is also used to capture the user's short-term and transient intent through category similarity. Through this hierarchical division, the algorithm can fully explore the user's potential intent and deeply integrate the item category information for modeling. Based on the integration of the three levels of intent, the category comparison learning mechanism is introduced to improve the recommendation performance. Finally, the MintSR algorithm is tested on three real datasets, and the results show that its accuracy is better than other mainstream algorithms.



Academic Editor: Stefan Fischer

Received: 4 November 2025

Revised: 4 December 2025

Accepted: 9 December 2025

Published: 11 December 2025

Citation: Huang, Y.; Yin, R.; Zhou, X.; Shi, P. MintSR: Multi-Intent Perception with Time-Based Segmentation for Session-Based Recommendation. *Appl. Sci.* **2025**, *15*, 13078. <https://doi.org/10.3390/app152413078>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: session-based recommendation; multi-intentional; hierarchical decoupling

1. Introduction

In recent years, the proliferation of big data and smart devices has led to a significant issue of information overload among users. Recommender systems play a vital role in assisting users in discovering items of interest amidst this deluge of data. By analyzing users' historical behavioral data, these systems offer personalized recommendations of products, content, or services, thereby improving user experience and driving economic gains for the platform. However, conventional approaches primarily rely on long-term user behavior and historical data. Particularly in scenarios involving anonymous or new users, these traditional methods [1] often struggle to effectively capture users' short-term interests and intentions, leading to suboptimal recommendation outcomes.

With the increasing awareness of personal data privacy protection, ensuring the security and privacy of user information has become an important consideration in recommender system design [2]. In many cases, user identity information is difficult to obtain,

especially in session-based recommendation (SBR) scenarios, where user interactions are usually anonymous, short-term, and lack long-term historical data. Conversation-based recommender systems (SBR) have become the focus of both academic and industrial fields due to their unique advantage of making recommendations despite the lack of user attributes and long-term behavioral data [3]. The core goal of SBR is to predict the user's interaction behaviors in the future by analyzing the user's short-term session data to recommend the most relevant items or content for users.

In recent years, researchers have begun to explore the application of deep learning in session recommendation to develop more advanced algorithmic models. Some of the important advances and models are listed below. With the widespread adoption of deep learning techniques, researchers have turned to the field to develop algorithmic models that fit the needs of session recommendation. This shift has led to significant advances. Hidasi et al. [4] first applied gated recurrent units (GRUs) in recurrent neural networks (RNNs) to session recommendation by proposing the GRU4Rec model. However, RNN-based approaches mainly focus on modeling sequential transitions of adjacent items and inferring user preferences through the temporal order of sequences, and still have difficulties in modeling complex item transition patterns [5]. With the application of graph neural networks in various deep learning applications, it has proved its effectiveness. One of the main advantages of graph neural networks is their adaptability to non-Euclidean spatial data, which explicitly integrates information about neighbors into the central node representation learning. There has been a considerable amount of research using graph neural networks to solve the task of session-based recommendation. The SR-GNN proposed by Wu et al. [6]. However, with the application of graph neural networks in various deep learning applications, they have proved their effectiveness. One of the main advantages of graph neural networks is their adaptability to non-Euclidean spatial data, which explicitly integrates information about neighbors into the central node representation learning. There has been a considerable amount of research using graph neural networks to solve the task of session-based recommendation. The SR-GNN proposed by Wu et al. [7] suggested the GC-SAN model based on SR-GNN, which learns remote dependencies between item embeddings collected by a GNN via a self-attention method. Yu et al. [8] proposed the TAGNN model, which assumes that different historical behaviors have specific impacts on different items, and it uses an attentional mechanism to capture the given specific target item's correlation with historical behaviors and activate the corresponding historical session based on the interest. Qiu et al. [9] applied graph attention networks to session recommendation and proposed FGNN, which uses graph attention networks with multiple heads of attention and views session recommendation as a graph classification problem. Wang et al. [10] suggested the GCE-GNN model, which aims to obtain richer item representations by capturing the current global session information outside of the current session in order to integrate the global graph into local graph propagation. Although the aforementioned techniques are useful for determining the user's purpose, they are insufficient to completely address how the user's propensity may alter in practical situations. SBR still faces two challenges; the first one is that relying on the information of a single session may lead to insufficient data extraction, while merging the global information increases the risk of long-distance dependency. The second one is that the lack of clarity on the intent segmentation may cause the users' intentions to interfere with each other, and the intentions cannot be extracted separately in an optimal way.

For the first challenge, GCE-GNN was the first to propose combining global context to incorporate global information models for better extraction of user information [11]. However, most of the subsequent models based on global information use GNNs for embedding

extraction, which can create the problem of long-distance dependency generating too much noise, thus affecting the recommendation performance.

The second challenge, mostly for short-term and instantaneous intent disentanglement, is to use only the last click as auxiliary information for recommendation [12], which is not sufficient to fully express the complexity of the user's intent in real-world scenarios. For example, if a user misclicks, considering only the last item may bias the inference of the user's true intent. For example, in Figure 1, consider a session: laptop-mouse-keyboard-hiking shoes-punchbowl. The user may have two intentions: The first is a user misclassification, where the user intends to purchase the electronics category. The second is that the user produces a transient intent change to purchase the clothing category instead. If the model uses the last click for assisted recommendation, it may assume that the user wants to buy clothes, which may be due to the information generated by the user's misclick. However, it is also possible that the user generated a transient intent, and using only the last item click would result in a recommendation error. Modeling the session as if all items assume that the user is more interested in the electronics category, which would result in a loss of transient intent. Ignoring transient changes in user intent. Since transient intent is often mutually exclusive with short-term intent, an untangling operation is necessary for both intents.

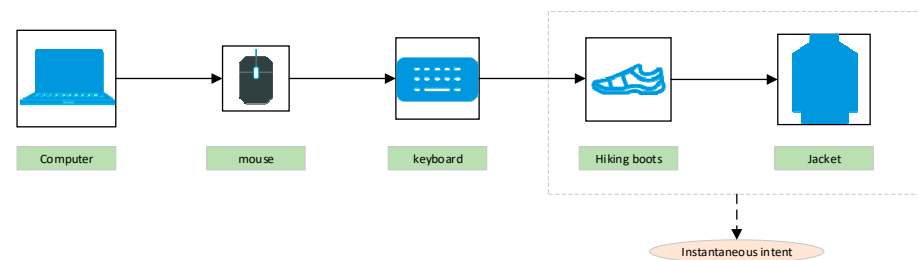


Figure 1. Example of uncertainty in instantaneous intent to change.

The user's intention is the main driver of interaction behavior. The user's habit is fixed for a long time. This model defines it as long-term intention, and the final study divides the user's intention into two major parts: long-term intention and short-term intention. To overcome the limitation, a new multi-level user intent perception method is proposed, which categorizes user intent into long-term, short-term, and instantaneous intent according to the duration of time. For long-term intent, a Transformer-based global context extractor is innovatively designed to generate global context information and global item embedding to accurately capture long-term intent. For transient intent, when short-term intent is generated, traditional methods only focus on the last-clicked item to capture transient intent, but the last-clicked item is likely to be driven by the previously clicked item, and the transient intent may have appeared before. The present model innovatively focuses on the last three item categories to assist in transient intent capture and considers the last clicked item that may have appeared earlier, integrating it with the last three clicked items, and designs attention mechanisms to determine transient intent embedding concerning the different importance of the items. In terms of short-term intent, this model considers that users will only start interacting with an item if they generate targeted intent, and the generation of short-term intent is more likely to be reflected in the first few items of the session, and the category with the highest frequency of occurrence in the target session can also reflect the user's short-term intent. Therefore, the combination of the first three clicked item categories and the most frequently occurring categories in the target session is used for item embedding the most occurring item categories and eventually fused to generate this session representation embedding. The final three unique channels are designed separately for embedding learning, and the main contributions of this work include

(1) This model proposes to divide the model according to the time hierarchy and proposes to design a unique three-channel strategy to process the data separately to avoid mutual interference and fully exploit the user's hidden hierarchical intent. The design of the Transformer-based long-term intent channel enriches and enhances the information extraction and effectively solves the long-distance dependency problem to improve the performance of the recommendation.

(2) Design the intent capture module to capture intent in a unique way that fully considers the dynamic nature of user intent changes and focuses more on effective intent. It also designs category-based comparison learning to improve recommendation accuracy.

(3) Experimental results on three benchmark datasets show that the algorithm has high performance.

2. Related Work

Recommender systems are now an important tool for enhancing user experience and business value. Traditional recommender systems mainly rely on users' historical behaviors and explicit feedback, such as ratings and reviews, to predict users' interests and preferences. However, in many real-world application scenarios, users may visit anonymously or lack long-term historical interaction records. In this case, recommender systems need to rely on the user's interaction behavior in the current session to make predictions. Session recommender systems are proposed to address this problem by analyzing the user's clicking behavior in the current session and predicting the goods or content in which the user may be interested in.

2.1. Early Session-Based Recommendation

Matrix decomposition has been a fundamental technique in recommender systems, particularly in the context of early recommendation methods. The essence of this approach lies in breaking down the user's item assessment matrix into two low-rank matrices that reflect the item's and the user's fundamental characteristics. This makes it possible to unveil implicit relationships between users and items, facilitating the prediction of user interest in unrated items. Mnih et al. innovatively introduced a probabilistic model [13] into matrix decomposition, improving the system's capacity to handle the difficulties presented by sparse and uncertain rating data. Koren et al. further developed this concept by introducing the implicit factor model [14], mapping users and corresponding items to vectors mapped to low latitude, predicting scores from user and item vectors. They have limitations in dealing with item order and temporal dependencies. Markov chain-based recommendation techniques mimic the user's decision-making process by using the user's past behavior to forecast the next potential behavior. Shani et al. first applied Markov chains to recommender systems [15]. The FPMC model, introduced by Rendle et al., offers a novel method for capturing user behavior's sequential dependency. It is a hybrid recommendation approach that merges matrix decomposition with first-order Markov chain techniques [16]. The FPMC model, introduced by Rendle et al., offers a novel method for capturing user behavior's sequential dependency. It is a hybrid recommendation approach that merges matrix decomposition with first-order Markov chain techniques [17]. However, these approaches still face challenges in capturing higher-order information and complex sequence dependencies in session data, which can result in their suboptimal performance in session-based recommendation tasks.

2.2. Deep Learn Session-Based Recommendation

Deep learning methods have been extensively used in recommender systems in recent years, especially excelling in handling complex sequences of user behaviors. Recurrent

Neural Network (RNN)-based models, such as GRU4Rec [18], significantly improve the performance of session recommendation by capturing the temporal dependencies of user behaviors. In addition, Convolutional Neural Networks (CNNs) are also used for session recommendation to capture the short-term and current interests of users by learning different-order features of user behavior sequences. For example, the Caser model [19] uses a convolutional filter to learn different-order features of user interactions and uses them for current interest modeling. The application of Attention Mechanism (AM) in session recommendation has also achieved remarkable results. The NARM model [6] introduces the Attention Mechanism into RNNs to capture users' long-term interests. The STAMP model [10] captures the overall and current interests of users, respectively, using Attention Memory Networks. These models improve the accuracy and robustness of recommendations by increasing the utilization of session context information. Graph Neural Network (GNN), which models the session as a graph structure in order to capture intricate linkages in user behavior, has gained much attention in session recommendation in recent years. For example, the SR-GNN model [20] uses the graph structure and models each session as a directed graph to capture the contextual information in a session. The GC-SAN model [21] introduces the self-attention mechanism on this basis and dynamically constructs the graph structure. The GCE-GNN model [22] efficiently avoids the noise introduction by constructing a global unweighted graph and embedding each node's neighbors by utilizing the session-aware attention mechanism. These models enhance the performance of session recommendation by utilizing contextual information in the graph structure. The majority of GNN models, however, only take into account the items' initial interaction order during a session and disregard the attribute information of the items across sessions, which restricts the models' capacity to mine users' possible interests. With the increasing complexity of user behavior data, multi-intent session recommendation has become an important research direction. Multi-intent session recommendation aims to capture multiple intentions that a user may have in the same session, enhancing the recommendation's precision and personalization. For example, a user may want to buy a certain product and also want to learn relevant information about the product in a shopping session. Therefore, multi-intent session recommendation needs to model multiple user intents at the same time and reflect the weights of these intents in the recommendation results. Multi-intent session recommendation can be achieved by a variety of techniques, including Multi-task Learning, Multi-head Attention Mechanism, and hybrid modeling approaches. Multi-task Learning captures the multi-intentional nature of user behavior by optimizing multiple tasks (e.g., click prediction and intent classification) simultaneously. Multi-head Attention Mechanism improves the expressiveness and robustness of the model by capturing different intents through multiple attention heads separately. Hybrid modeling approaches, on the other hand, integrate the benefits of many models, such as the combination of RNN and GNN, to model user behavior more comprehensively [11]. Recent research has started to concentrate on the attribute information of in-session items in order to better capture users' potential interests. For example, the MGS model [23] presents a bidirectional updating method that uses the intrinsic attribute information of objects during discussions to improve contextual information learning. It constructs directed graphs for individual sessions and their corresponding mirror graphs, which aggregate items sharing similar attributes. By integrating information from the mirror graphs with item data, the model enhances item representation capabilities. Furthermore, the MGS model incorporates contrast learning to oversee the mirror graph construction process and prevent the incorporation of unnecessary noise. This approach of utilizing attribute information has shown significant results in solving the data sparsity problem and improving the model's capacity to identify the interests of potential users. However, the majority of current models continue to concentrate on

exploiting attribute information of items within a session, and global interest modeling of a session has not been explored in depth. To better capture the global interest of a session, global modeling approaches have received increasing attention in recent years. Global modeling constructs global graph structures by exploiting the connections between multiple sessions to enhance the ability to model user interests. For example, the GCE-GNN model effectively avoids the introduction of noise by constructing a global unweighted graph and embedding the neighbors of each node using the session-aware attention mechanism. The COTREC model [24] reduces the issue of data sparsity by mining the relationships inside and outside the sessions and using comparative learning to supervise one another. The global modeling approach improves the performance of session recommendation by constructing global context information by utilizing the item transformation information in multiple sessions. This method better captures users' overall interests and efficiently resolves the issue of data sparsity in a single session. An approach called reverse location awareness seeks to determine how a user's attention is impacted by the order in which they click on objects during a session. By reversing the perspective on item position vectors, reverse position awareness is better able to identify a user's hidden interests throughout a session. Wang et al. highlighted that items clicked later in a session tend to be more indicative and significant in reflecting the user's present interest. Therefore, reverse location awareness has been widely adopted in session representation learning to help enhance the model's comprehension of the user's present interests.

Over the past few years, research on session recommendation systems has made significant progress, especially with deep learning and graph neural network techniques that have excelled in handling complex user behaviors. Multi-intent session recommendation and the utilization of attribute information provide new ideas to improve the model's capacity to identify potential user interests and deal with data sparsity. Techniques such as global modeling and inverse location awareness also contribute significantly to enhancing model performance. Future research can further explore how these techniques can be organically combined to achieve more accurate and personalized session recommendations [25].

Table 1 presents the comparative analysis with state-of-the-art architectures: ① global map + contrastive learning, ② attribute information, ③ multi-semantic decoupling. Symbol legend: ✓ = feature present, ✗ = feature absent, δ = absolute improvement over MintSR (on the same dataset).

Table 1. Comparative analysis with the latest architecture.

Method	Year	Global Graph	CL	Attribute	Multi-Intent	Tmall Recall@20	Vs. MintSR
COTREC	2021	✓	✓	✗	✗	38.4	−3.3
DHCN	2021	✓	✓	✗	✗	39.2	−2.5
MGS	2022	✗	✓	✓	✗	37.9	−3.8
TRHMCI	2024	✓	✗	✗	✗	40.5	−1.2
MintSR	2025	✓	✓	✓	✓	41.7	—

3. MintSR Model

This section introduces MintSR, which is a recommendation model that uses novel intent segmentation items and unique intent capture combined with category information. Figure 2 illustrates the complete architecture of MintSR, including three main channels: the long-term intent channel, the short-term intent channel, and the transient intent channel.

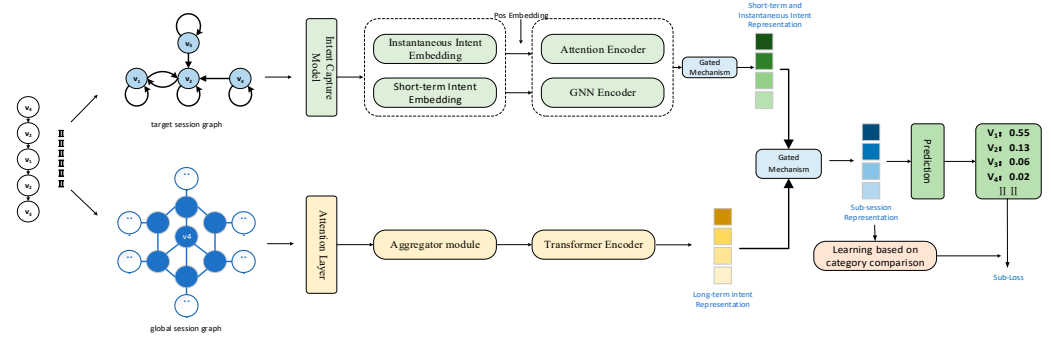


Figure 2. MintSR model diagram.

3.1. Problem Statement

Let $V = \{v_1, v_2, \dots, v_m\}$ be the set consisting of all items that have appeared in the session; $S = \{v_1^s, v_2^s, \dots, v_l^s\}$ be the anonymized session ordered by timestamp, and its length is denoted as $|s|$; each item $v_i \in V$ and session s are embedded into the same vector space, denoted by $s \in \mathbb{R}^d$ and $x_i \in \mathbb{R}^d$, respectively, with d denoting the dimension of the vector. For a given target session $s_0 = [v_1, v_2, \dots, v_{|s_0|}]$, the goal of the SRS task is to predict the next possible interacting item $v_{|s_0|+1}$. The recommendation model generates preference scores for all items in V and selects the top K items with the highest scores for the final recommendation.

The set of all categories is defined as $C = \{c_1, c_2, \dots, c_m\}$ where n is the number of items and m is the number of categories, typically $n \geq m$. Each item $v_i \in V$ corresponds to a particular category $c_k \in C$, usually with multiple items in the same category.

3.2. Session Graph Construction

Long-term intent: Driven by the rational decision-making system, reflecting stable user preferences (e.g., sustained interest in electronics). Characterized by extended decision cycles and high behavioral consistency. **Short-term intent:** Driven by the goal-oriented rational system, representing core session-based needs (e.g., purchasing computer accessories during this session). Features continuous behavioral sequences with clear objectives. **Instantaneous intent:** Triggered by the intuitive decision system, representing immediate reactions to temporary needs or external stimuli (e.g., sudden interest), with short-lived behaviors.

Users tend to have fixed behavioral patterns and intent preferences when interacting with items in different sessions, so considering the construction of global item graphs as well as item transitions is beneficial for capturing users' long-term intent preferences. Targeted session maps provide better access to the changing needs of the current user and better capture short-term intent.

Global graph: In the Global graph model, a methodology that closely resembles the approach used in GCE-GNN is employed to construct the initial global graph. In this model, the neighbor of item v_i^p in session S_p is defined as $N_\epsilon(v_i^p) = \{v_j^q | v_i^p = v_{i_d}^q \in S_p \cap S_q; v_j^q \in S_q; j \in [i_d - \epsilon, i_d + \epsilon]; S_p \neq S_q\}$, with i_d representing the location of item v_i^p , and ϵ serving as the hyperparameter regulating the transfer range between items. The global session graph, denoted as $G_g = (v_g, \epsilon_g)$, is defined as follows: v_g represents the set of nodes encompassing all items in the session sequence, ϵ_g signifies the set of edges denoted as $\epsilon_g = \{e_{ij}^g | (v_i, v_j) | v_i \in V, v_j \in N_\epsilon(v_i)\}$, where each edge comprises all pairs of items in the session. The construction of the global session graph is illustrated in Figure 3.

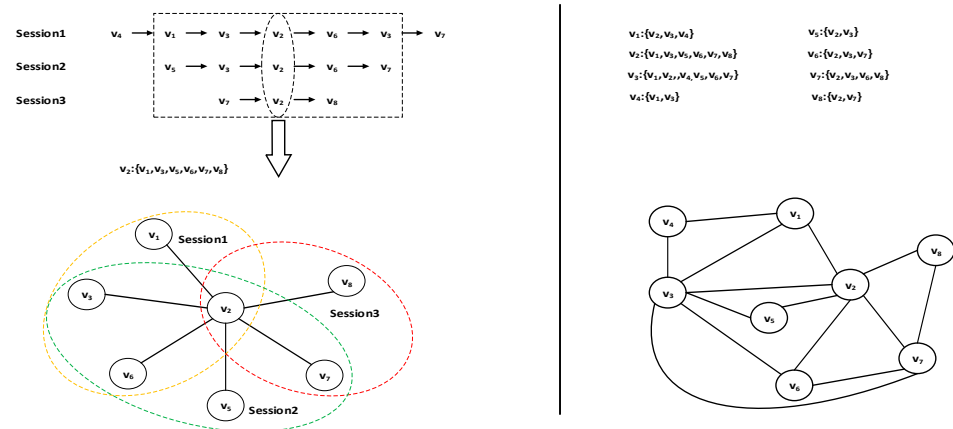


Figure 3. Example of global session graph construction (Different colored dashed lines divide different sessions).

Target Session Graph: This model constructs four kinds of edges according to different transformation patterns during the construction of the target session graph, which depends on the relationship between items, including in, out, in-out, and self. Given a session $S = \{v_1^s, v_2^s, \dots, v_l^s\}$, this model forms a directed edge-aware graph $G_s = (V_s, \varepsilon_s)$, where $V_s \in V$ refers to items in S . For edge (v_i^s, v_j^s) , it means that there are only edges pointing from v_j^s to v_i^s , and out means that there are only edges pointing from v_i^s to v_j^s . And in-out denotes that there are edges from v_j^s to v_i^s as well as edges from v_i^s to v_j^s . Self denotes the self-transformation of an item. Secondly, this model designs the target session co-occurrence graph. For each item in the session, a local undirected graph is constructed according to the connection of the items in the session, and the co-occurrence number is calculated and normalized in the local undirected graph to determine the edge weights. The edges are finally classified into four types and given different weights. The target session graph is constructed as shown in Figure 4.

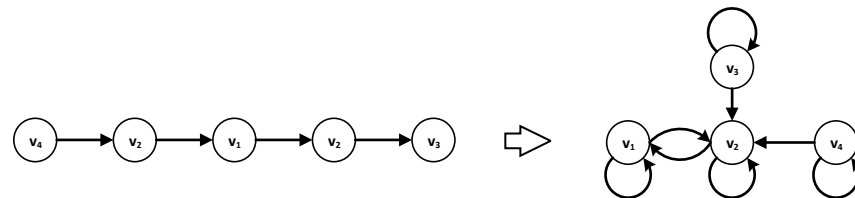


Figure 4. Example of target session graph construction.

3.3. Intent Capture Module

To address the data sparsity and the fixed interaction tendencies of users who have interacted with items for a long period of time, this model combines the global context to deeply mine the users' long-term potential intention tendencies.

In the target session, this model divides the user intent into short-term intent and transient intent. To do this, the model is designed for the target session intent segmentation. Given a session $S = \{e_1^s, e_2^s, \dots, e_l^s\}$, each item corresponds to a one-hot categorical vector representation as $S^c = \{e_{s,1}^c, e_{s,2}^c, \dots, e_{s,t}^c\}$. For transient intent, most of the previous studies focus only on the most recently clicked items, but the last clicked items are probably driven by the previous clicked items, so this model innovatively focuses on the last three item categories to aid the capture of transient intent. Additionally, this model takes into account that the last clicked items may have appeared earlier, so this model takes the combined

category of the last three item categories as the instantaneous intent. Suppose there are j items in total. Then the instantaneous intent can be expressed as follows:

$$e_{\text{instant}}^c = \left\{ \sum_{i=(j-2)}^j e_{s,i}^c / 3 \mid e_{s,i}^c \in S^c \right\} \quad (1)$$

For short-term intent, this model assumes that users will only begin to interact with items if they generate targeted intent, and the generation of short-term intent is more likely to be reflected in the first few items. In addition, the category with the most frequent occurrence of the target session can also reflect the user's short-term intention, so this model takes the first three items and the combined category with the most frequent occurrence as the short-term intention. Assuming that the category with the most frequent occurrence of the target session is $\text{most_frequent_}e_{s,i}^c$, the short-term intention can be expressed as

$$e_{\text{three}}^c = \left\{ \sum_{i=1}^3 e_{s,i}^c \mid e_{s,i}^c \in S^c \right\} \quad (2)$$

$$e_{\text{short}}^c = \left\{ (e_{\text{three}}^c + \text{most_frequent_}e_{s,i}^c) / 2 \mid e_{s,i}^c \in S^c \right\} \quad (3)$$

For optional items v_j , their category vectors are defined as follows: c_j ; with category similarity calculated using cosine similarity:

$$s(v_j, u) = \frac{u^T c_j}{\|u\| \|c_j\|} \in [0, 1], u \in \{e_{\text{instant}}^c, e_{\text{short}}^c\} \quad (4)$$

3.4. Project Embedded Representation Learning

In this subsection, we explain how this model introduces the item embedding representation learning components from the global session and target session, respectively. This model first introduces the global item embedding learning with the Transformer as the core, and then introduces the target item embedding learning with a neural network and attention mechanism as the core.

Long-term intent embedding learning: For global item embedding, for each item in the global graph, information about its neighbor nodes is first aggregated. Since each item may have different neighboring items, each item plays a different role. Therefore, this model aggregates the neighbor information in the global graph for each node and linearly combines the neighbor information based on the session-related attention weights as

$$h_i^a = \sum_{j \in A_i} a_{ij} h_j \quad (5)$$

$$a_{ij} = \frac{\exp(q_1^T \text{LeakyReLU}(W_1[(s \bullet h_j); w_{ij}])))}{\sum_{k \in N_i} \exp(q_1^T \text{LeakyReLU}(W_1[(s \bullet h_k); w_{ij}])))} \quad (6)$$

where N_i is the set of neighbors of node i , h_j is the item embedding vector of node j , w_{ij} is the number of times node i and node j co-occur in all sessions, $\bullet$ is the splicing operation, W_1 and q_1 are the trainable parameters, and s is the session information, computed as the vector average of all item embeddings. Based on the computed attention weights a_{ij} , the embedding vectors of neighboring nodes are linearly combined to obtain the neighbor information embedding vector h_i^a of node i .

Multiple aggregation layers are stacked to capture more neighbor information related to the current node. The k aggregation layer is defined as

$$h_i^{a,k} = \text{agg}(h_i^{a,k-1}, s) \quad (7)$$

where $h_i^{a,k}$ is the item embedding vector obtained in the previous layer and s is the session information.

In addition, to better capture the contextual information in the input sequence, especially when dealing with long-distance dependencies. The model introduces the Transformer to utilize its multi-attention mechanism, allowing the model to consider all other input elements while processing each one. This mechanism can be viewed as aggregating neighbor information, as each element can directly “focus” on the other elements, thus effectively integrating contextual information. It is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\overline{Q}K^T}{\sqrt{d}}\right) \bullet V \quad (8)$$

The final weighted aggregation of the value matrix using the normalized attention weights is used to obtain the session global context representation.

Finally, to mitigate bias and overfitting problems caused by multi-layer stacking, information is dynamically selected to be retained or ignored based on the context of the input, handling noise and irrelevant information. This model employs a gated network to obtain the final long-term intent embedding. This gate dynamically suppresses cross-session edges inconsistent with the current session’s intent based on the current session representation:

$$z = \sigma(W_2 h_i + W_3 h_i^{a,k}) \quad (9)$$

$$h_i^g = (1 - z) \bullet h_i + z \bullet h_i^{a,k} \quad (10)$$

where W_2 and W_3 are trainable parameters. Finally, by aggregating all item embeddings and weighing them according to their importance:

$$\gamma_i = g_1^T \text{sigmoid}(W_4 \hat{s}_i + W_5 \hat{s}_i' + b_1) \quad (11)$$

$$h_{long} = \sum_{i=1}^n \gamma_i h_i \quad (12)$$

γ_i denotes the attention weight of the i th item, which is used to measure the importance of the item in the long-term intentional sub-session. g_1 is a learnable vector for weighing and summing the outputs of the sigmoid function. W_4 and W_5 are learnable weight matrices, which are used to linearly transform the item representations and the average representations. b_1 denotes a learnable bias vector used to adjust the output of the sigmoid function. h_{long} denotes the final long-term intentional sub-session representation, obtained by weighting and summing the representations of all the items by the attentional weights.

Short-term intent embedding learning: In the short-term intent learning, to fully learn the target session embedding information and better extract the information to avoid the omission of information contextual relationship, this model firstly learns the short-term intent information through two perceptual graphs, which are dubbed the directed and undirected graphs. The directed graph can introduce directionality and transition patterns through various types of edges to enhance the model’s expression ability. This model shows the semantic correlation between items by referencing the undirected graph and incorporating co-occurrence frequencies.

Firstly, after this model constructs the target session graph, it uses GNN to aggregate item information to learn the embedding representation of the items, and considering the

different importance of different neighbors, this model adopts the attention mechanism to learn the weights between neighbors:

$$\beta_{ij} = \frac{\exp(\text{LeakyReLU}(a_{r_{ij}}^T(h_i \bullet h_j)))}{\sum_{k \in A_i} \exp(\text{LeakyReLU}(a_{r_{ik}}^T(h_i \bullet h_k)))} \quad (13)$$

where A_i is the set of neighbor nodes directly connected to node i in the session graph, h_j is the embedding of node j , a_{ij} is the weight of node j , LeakyReLU is the activation function, and r_{ij} refers to the four kinds of edges referred to in the above session graph building blocks. Afterwards, this model updates the embedding representation of node i by a linear combination of neighbors associated with the attention score.

$$h_i^l = \sum_{j \in A_i} a_{ij} h_j \quad (14)$$

Equations (10) and (11) are used for both the directed and undirected graphs, and the sub-session representations h_1 and h_2 are obtained, after which gating techniques are used to aggregate the two representations:

$$\delta = \text{sigmoid}(W_6[h_1 || h_2]) \quad (15)$$

$$h_{\text{short}} = (1 - \delta) \bullet h_1 + \delta \bullet h_2 \quad (16)$$

where W_6 denotes trainable parameters. The weights are dynamically assigned to each item in the short-term intention sub-session through the soft attention mechanism, emphasizing the important items and reducing the influence of irrelevant information. The final generated h_{short} can reflect the user's short-term intention more accurately and improve the accuracy and robustness of the recommendation. Equations (10) and (11) end up with the short-term intent sub-session representation h_{short} .

Instantaneous intent embedding learning: In instantaneous intent, this model first obtains the cis-schematic category embedding, using Equation (1). However, the transient intent may have occurred before, so this model looks for items that are similar to the transient intent and uses the attention mechanism to generate item transformation information. First, this model uses cosine similarity to measure the similarity between items and transient intentions:

$$\text{similarity} = \frac{e_{\text{instant}}^c \bullet e_i^c}{||e_{\text{instant}}^c|| ||e_i^c||} \quad (17)$$

where e_{instant}^c is the transient intent category and e_i^c is the category of the remaining items in this session. Ultimately, this model filters out items related to transient intentions based on similarity scores and queuing values α .

$$m_i = \begin{cases} 1, & \text{if } \text{cosine_similarity}(e_{\text{short}}, e_{c,s,i}) \geq \partial \\ 0, & \text{if } \text{cosine_similarity}(e_{\text{short}}, e_{c,s,i}) < \partial \end{cases} \quad (18)$$

where m_i is a mask sequence. To capture the location information of the items in the session, a learnable location embedding P is first initialized and element-by-element summed with the instantaneous intent sub-session h_{instant} . After masking and location information embedding, by using the self-attention mechanism model, this model can capture the sequential and semantic interdependence between the objects in the sub-session. The specific steps are as follows:

Query Q: obtained by a linear transformation of a learnable weight matrix W_7 and a bias vector b_0 on h_{instant} :

$$Q = \text{SELU}(W_7 h_{\text{instant}} + b_0) \quad (19)$$

Key K and value V: direct use of h_{instant} as key and value subsequently computed as attention score:

$$M = \text{softmax}\left(\frac{\overline{Q}K^T}{\sqrt{d}}\right) \bullet V \quad (20)$$

where d_k is the dimension of the embedding. To increase the nonlinear expressiveness of the model, the attention scores are transformed using a location-aware feedforward network:

$$h_{\text{instant}} = \text{SELU}(MW_8 + b_1)W_9 + b_2 \quad (21)$$

where W_8, W_9 are the learnable weight matrix and b_1, b_2 are the learnable bias vector.

3.5. Integration and Prediction

This model first fuses the short-time intent and the instantaneous intent extracted from the target session, and first computes the gating coefficients:

$$\delta = \text{sigmoid}(W_{10}[h_{\text{instant}} \parallel h_{\text{short}}]) \quad (22)$$

where W_8 is the learnable weight matrix, and δ is the gating coefficient to control the proportion of contributions of short-term intentions and instantaneous intentions. The final target session is as follows:

$$h = (1 - \delta) \odot h_{\text{instant}} + \delta \odot h_{\text{short}} \quad (23)$$

where h is the final goal session representation obtained by weighting and summing the instantaneous intentional sub-session representation and the short-term intentional sub-session representation by a gating factor. For the finally obtained global session representation h_{long} similarly, using gating aggregation for the final session representation h , the model finally computes the probability that each item v_j , which is predicted to be the next interacting item by the following formula:

$$\hat{y}_i = \text{softmax}(h^T e_j) \quad (24)$$

where e_j is the embedding representation of item v_j and $\hat{y}_i$ is the probability that item is predicted to be the next interacting item. Eventually, the two predictions are dynamically fused using a gating technique to obtain the final probability $\hat{y}_i$. The model is trained by computing the loss for the local learning task and the global learning task independently using the cross-entropy loss function. The difference between the true label and the model's predicted likelihood is measured by the cross-entropy loss. The final loss function is obtained by adding the losses of the local and global jobs, and the time back propagation procedure is then used to optimize the model's parameters. The following is the formulation of the cross-entropy loss function:

$$L = -\sum_{i=1}^m y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \quad (25)$$

where y_i is the true label indicating whether item i is actually interacted with by the user; $\hat{y}_i$ is the interaction probability of item i predicted by the model.

3.6. Learning Based on Category Comparison

The main role of the category-based comparative learning task is to enhance the consistency between session representations and item categories through comparative learning, thereby improving the relevance and accuracy of recommendations. Specifically, it enables the model to better understand users' behavioral patterns and intentions by maximizing the similarity between the session representation and the positive sample categories while minimizing the similarity between the session representation and the negative sample categories. Positive samples are defined as the next interaction category, and negative samples are defined as other categories that are not the next interaction. First, this model calculates the similarity between the session representation h and the positive and negative samples separately:

$$\text{similarity}_{\text{pos}} = h \cdot e_{c,\text{pos}} \quad (26)$$

$$\text{similarity}_{\text{neg}} = h \cdot e_{c,\text{neg}} \quad (27)$$

where $e_{c,\text{pos}}$ is the positive sample embedding representation, and $e_{c,\text{neg}}$ is the negative sample embedding representation. By increasing the similarity of positive samples and decreasing the similarity of negative samples, you may define the contrast learning loss function.

$$L_c = -\log \frac{\exp(\text{similarity}_{\text{pos}}/\tau)}{\exp(\text{similarity}_{\text{pos}}/\tau) + \sum_{e_{c,\text{neg}} \in C_{\text{neg}}} \exp(\text{similarity}_{\text{neg}}/\tau)} \quad (28)$$

τ is a temperature parameter that controls the differentiation of negative samples. Combining the contrast learning task with the next item prediction task controls the weight of the contrast learning loss by introducing the hyperparameter λ :

$$L = L_p + \lambda L_c \quad (29)$$

where L_p is the next item prediction task's cross-entropy loss and L_c is the loss for the contrast learning task. The model can increase the accuracy of recommendations by using the comparison learning task to better grasp the user's intents and behavioral patterns.

4. Experimentation

In this part, we explain the experimental procedures and how the model establishes the baseline. To validate MintSR, we carried out a number of studies to address the following research questions:

Question 1: How does MintSR perform compared to the benchmark model?

Question 2: Did each of the major modules have a positive impact?

Question 3: What was the effect of key parameters on MintSR?

4.1. Datasets and Pre-Processing

The experiments use three public datasets: Diginetica, Tmall, and Nowplaying, which are representative of the session recommendation domain and can better validate the performance of the model. Diginetica: the data comes from e-commerce websites and contains the clicking behaviors of the users on the products. Tmall: the data comes from the e-commerce platform of Tmall and records the users' shopping behavior. Nowplaying: the data comes from Twitter, extracting the music playing behavior shared by users. Session lengths exhibit a long-tail distribution, with average durations of 6.7, 5.1, and 7.4 clicks, respectively; over 95% of sessions involve ≤ 20 clicks. In the temporal dimension, during

preprocessing, we removed duplicate items, retained products appearing ≥ 5 times, and discarded sessions < 2 in length. Subsequently, training sub-sequences with a maximum length of 50 were generated in increments of 1, padding with zeros where necessary. The training phase employed 1:1 random negative sampling, while the testing phase calculated recall and MRR by ranking all products. To generate subsequences, for each session, $S = [s_1, s_2, \dots, s_n]$ generate multiple subsequences and their corresponding labels, e.g., $[s_1, s_2, \dots, s_{n-1}], s_n$ to increase the diversity of the training data. Table 2 shows the statistics of the three datasets:

Table 2. Statistics for the three datasets.

Statistics	Tmall	Diginetica	Nowplaying
clicks	818,479	982,961	1,367,963
training sessions	351,268	719,470	825,304
test sessions	25,898	60,858	89,824
items	40,728	43,097	60,417
Average length	6.69	5.12	7.42

4.2. Baseline Algorithm

To assess the efficacy of the MintSR model, several representative models were chosen for comparative analysis:

POP [26]: It chooses the elements from the training set that occur most frequently.

Item-KNN [17]: It determines how similar the current object is to other objects and suggests the top K items.

FPMC [16]: A hybrid model that mixes Markov chain and matrix decomposition techniques to address data sparsity difficulties, with a low-rank estimation approach to mitigate overfitting.

GRU4REC [18]: An RNN-based model that encodes sequence items into one-hot codes and utilizes GRU to predict the likelihood of each item being selected next.

NARM [20]: Uses attention mechanisms and RNNs to extract users' general interests.

STAMP [22]: Uses the attention process to highlight short-term memory.

SR-GNN [6]: The pioneering application of GNN in session-based recommendation to learn item representations, coupled with an attention network to capture users' immediate interests.

FGNN [9]: Creates a graph classification problem for the session's next item recommendation.

GCE-GNN [10]: The novel incorporation of a global graph for recommendation purposes, considering the transformation relationships between paired items at both global and session levels.

DHCN [27]: A dual-channel hypergraph convolutional network for session-based recommendation is introduced for the first time. In order to improve feature extraction and recommendation performance, it blends self-supervised learning with hypergraph modeling to capture high-order item associations.

CM-GNN [28]: It uses an advanced GNN to handle transitions between projects and uses contrastive learning to improve performance.

TRHMC-GNN [29]: It constructs global and self-cyclic star graphs to illustrate the relationships between topics and proposes a multi-level cross-item information graph neural network to improve suggestion efficacy.

EAT-SGNN [30]: It makes use of an improved attention module that uses graph data at the global and session levels to provide recommendations.

4.3. Realization Details

In terms of hyperparameter settings, the embedding size is set to 128, the batch size is set to 128, the initial learning rate is set to 0.001, and the attenuation is set to 0.6 per 1 epoch. The clustering algorithm adopts min k-means, and the number of clusters is set to 100. The cosine similarity threshold α is set to 0 across all three datasets. For instantaneous and short-term intent categories, items should align with the target item ($\cos \geq 0$) to be retained; those in the opposite direction ($\cos < 0$) are considered distractors. The masking probability β is set to 0.4 on the Tmall dataset and 0.1 on the Diginetica and Nowplaying datasets. These parameters, which aim to increase the model's capacity for generalization and recommendation accuracy, are determined by carefully weighing the model's performance, computational effectiveness, and dataset properties. The model evaluates the outcomes using the commonly used metrics P@K (precision) and MRR@K (mean reversed rank) with $K \in \{10, 20\}$.

The percentage of accurately predicted target items in the first K results is indicated by P@K. This metric was selected for this model in order to favorably assess the SBR model's recommendation performance.

Since $MRR(i) = \frac{1}{rank_i}$ indicates the rank of the item in the first K listings, MRR@K computes the average of the inverse rank of the correct item in the first K recommendation results. This statistic was selected for this model to assess how well the SBR model predicts ranks.

4.4. Experimental Structure and Analysis

RQ1 performance comparison: Three real-world datasets for the MintSR model proposed in this study have been used to thoroughly assess the model, and they have been contrasted with a number of representative baseline techniques, as shown in Table 3.

Table 3. Performance of the comparison method on three datasets. Bold is the best result, and underlining is the next best result.

Method	Tmall				Diginetica				NowPlaying			
	P@10	MRR@10	P@20	MRR@20	P@10	MRR@10	P@20	MRR@20	P@10	MRR@10	P@20	MRR@20
POP	1.67	0.88	2.00	0.90	0.76	0.26	1.18	0.28	1.86	0.83	2.28	0.86
Item-KNN	6.65	3.11	9.15	3.31	25.07	10.77	35.75	11.57	10.96	4.55	15.94	4.91
FPMC	13.10	7.12	16.06	7.32	15.43	6.20	22.14	6.66	5.28	2.68	7.36	2.82
GRU4REC	9.47	5.78	10.93	5.89	17.93	7.73	30.79	8.22	6.74	4.40	7.92	4.48
NARM	19.17	13.12	26.47	13.36	35.44	15.13	48.32	16.00	13.60	6.62	18.59	6.93
STAMP	22.63	13.12	26.47	13.36	33.98	14.26	46.62	15.13	13.22	6.57	17.66	6.88
SR-GNN	23.41	13.45	27.57	13.72	36.86	15.52	50.73	17.59	14.17	7.15	18.87	7.47
FGNN	20.67	10.07	25.24	10.39	37.72	15.95	50.58	16.84	13.89	6.8	18.78	7.15
GCE-GNN	28.01	15.08	33.42	15.42	41.16	18.15	54.22	19.04	16.94	8.03	22.37	8.40
S ² -DHCN	26.22	14.60	31.42	15.05	40.21	17.59	53.66	18.51	17.35	7.87	<u>23.50</u>	8.18
CM-GNN	29.72	16.06	34.69	16.41	<u>41.56</u>	<u>18.28</u>	<u>54.67</u>	<u>19.19</u>	<u>17.50</u>	8.05	23.30	8.44
TRHMCI-GNN	30.56	<u>16.73</u>	<u>37.98</u>	<u>17.45</u>	40.51	17.52	53.84	18.43	15.87	7.12	21.93	7.53
EAT-SGNN	<u>31.47</u>	16.69	37.66	17.08	38.85	16.62	52.15	<u>18.67</u>	17.32	8.09	23.33	8.64
MintSR	35.51	19.16	41.74	19.60	41.79	18.69	55.34	19.57	18.06	<u>8.06</u>	24.02	<u>8.47</u>

Based on the results of five replicate experiments using different random seeds, the statistical analysis findings are presented below to validate the stability and reliability of the model's performance according to Tables 4 and 5.

Table 4. Random seed result.

Random Seed	Recall@5	Recall@10	Recall@20	MRR@5	MRR@10	MRR@20
1	0.2762	0.3475	0.4126	0.1779	0.1875	0.1920
2	0.2814	0.3516	0.4143	0.1803	0.1897	0.1940
3	0.2821	0.3525	0.4163	0.1797	0.1891	0.1936
4	0.2737	0.3477	0.4134	0.1757	0.1857	0.1903
5	0.2831	0.3545	0.4172	0.1819	0.1915	0.1958

Table 5. Confidence interval.

Evaluation Indicator	Average	Standard Deviation	95% Confidence Interval
Recall@5	0.2793	0.0037	[0.2745, 0.2841]
Recall@10	0.3507	0.0033	[0.3463, 0.3551]
Recall@20	0.4148	0.0018	[0.4124, 0.4172]
MRR@5	0.1791	0.0023	[0.1761, 0.1821]
MRR@10	0.1887	0.0022	[0.1858, 0.1916]
MRR@20	0.1931	0.0020	[0.1905, 0.1957]

The standard deviation for all metrics is less than 0.004, with Recall@20 exhibiting the smallest standard deviation (0.0018) and MRR@5 the largest (0.0023). This indicates minimal performance fluctuation across different random seeds, demonstrating excellent stability.

Confidence interval analysis: The 95% confidence interval spans for all metrics are less than 0.009. For example, the confidence interval for Recall@5 is only [0.2745, 0.2841], indicating low estimation error in the experimental results and high statistical reliability.

Reproducibility verification: The core metric (Recall@20) across five experiments fluctuated only between 0.4126 and 0.4172, with a difference of just 0.0046. This proves that the model's performance is unaffected by random seeds and exhibits strong reproducibility.

Two decimal places are used to round all measurements. The model underlines the second-best outcome and highlights the best and second-best in bold text. On almost every criterion, the present model produced the best results. Based on the results, the model can draw the following conclusions:

(1) Deep learning methods significantly outperform traditional methods (e.g., POP, Item-KNN, and FPMC) in session recommender systems because traditional methods have difficulty in mining complex hidden features in item sequences, while user clicking intentions are dynamically changing. Graph neural network (GNN)-based models (e.g., SR-GNN, GCE-GNN) usually outperform RNN- or LSTM-based models (e.g., GRU4Rec, NARM, STAMP) because RNNs and LSTMs have limitations in capturing the complex relationships of non-adjacent items in a session. At the same time, GNNs can capture the user's intention more accurately by modeling the higher-order relationships. In addition, models combining session graphs and global graphs perform better than baseline approaches, indicating the importance of global contextual information for enhanced item representation and user intent understanding. This pattern indicates that session graphs have been widely used in session recommender systems since the proposal of SR-GNN, but it also raises concerns that depending solely on session graphs may restrict the investigation of item similarity and user intent fluctuations. Introducing global contextual information can leverage potential associations between different sessions to generate more accurate representations of user intentions and items, providing deeper insights into session recommendation systems.

(2) The poor performance of most models on the NowPlaying dataset is closely related to the characteristics of its data. This dataset is based on users' listening histories, and implicit characteristics like artist, genre, and release date frequently affect users' musical

preferences. However, due to the limitation in available data attributes, this key information cannot be accessed, resulting in the model's inability to utilize these features, which negatively affects recommendation performance. In contrast, users' behavior of purchasing goods is easier to model through actions because the purchase behavior is more direct and the intention is easier to identify through the user's interaction records (e.g., clicks, purchases, etc.). By comparing the NowPlaying dataset with other purchase behavior datasets, the importance of user intent in a session recommendation system can be seen. The MintSR model successfully overcomes these challenges by comprehensively modeling dynamic intent and can capture the changes in user intent across sessions and generate more accurate recommendation results by combining the contextual information, which validates the importance of in-depth modeling of user intent in session recommendation systems.

(3) Three datasets show that the suggested MintSR model performs better than the current models, highlighting the effectiveness of its hierarchical intent structure. This layered intent approach offers two key advantages: (i) implementing multi-level intent segmentation and ensuring effective disentanglement to facilitate independent learning of different user intents by each module. Additionally, incorporating pertinent information elements in short-term and immediate intents improves the model's capacity to record representations of sessions. (ii) Introducing a category comparison learning task that reinforces the alignment between session representations and subsequent user clicks, thereby strengthening the semantic relationship during model training. These components collectively enable MintSR to capture more informative session representations, consequently improving the relevance of recommendation outcomes.

RQ2 ablation study: The performance improvement achieved by MintSR can be attributed to the combined effect of combining the three intentions, as well as comparative learning. To examine how various modules affect MintSR, four variants of MintSR were developed for this model:

(1) MintSR-L: This variant focuses just on the target session and eliminates the long-term intent channel of the model.

(2) MintSR-S: In this variant, the short-term intent channel of this model is removed, concentrating solely on the target session's recently clicked item and long-term goal.

(3) MintSR-I: In this variant, the instantaneous intent channel of the model is removed and is no longer extracted for recently clicked items.

(4) MintSR-C: Removes the comparative learning module of this model.

(5) MintSR-FM: Removes the intent fusion module, replacing gated fusion with simple weighted summation.

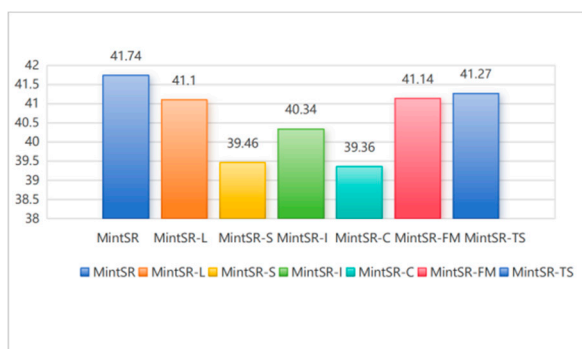
(6) MintSR-TS: Removes the temporal segmentation module, modeling all intents using full-session modeling.

Table 6 and Figure 5 compare the outcomes of MintSR with its six variations on the three datasets. The primary contribution of MintSR is the innovative multi-level intent segmentation scheme, which enables each module to learn the user's long-term, short-term, and transient intents in a targeted manner. It is evident from the experimental findings that when sequentially removing the sections corresponding to long-term intent, short-term intent, momentary intent, intent fusion, and time segmentation, the performance of the model decreases to different degrees. This phenomenon strongly proves that all five kinds of intent components have indispensable contributions to the overall performance of MintSR. In the validation experiments with the contrast learning module, the performance of the variant model with the contrast learning module removed shows a significant decline compared to the full MintSR model. This outcome strongly implies that the category comparison learning module can be successfully implemented to enhance the model's performance and recommendation accuracy. In the experiments on the NowPlaying dataset, the MintSR-L

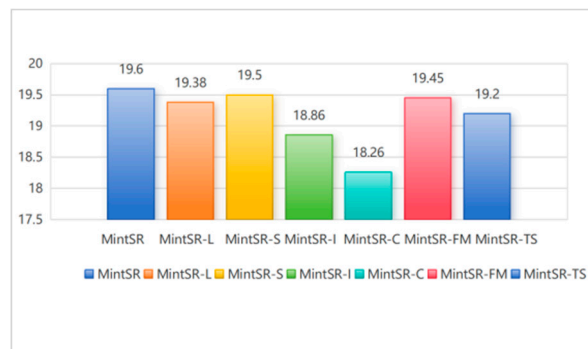
model obtained by removing the long-term intent channel has a more significant decrease in P@20 and MRR@20 metrics compared to the full MintSR model when the session length is long, and the number of items is high. This result further highlights the important role of the long-term intention channel in enriching information extraction as well as effectively solving the long-distance dependency problem, suggesting that the long-term intention channel is better able to capture the user's stable preferences over a long time span.

Table 6. Ablation experimental studies.

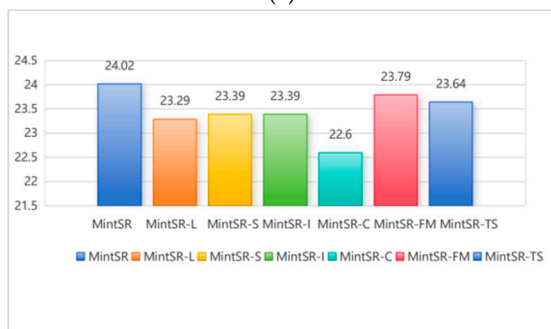
Datasets	Tmall		NowPlaying		Diginetica	
	P@20	MRR@20	P@20	MRR@20	P@20	MRR@20
MintSR	41.74	19.60	24.00	8.45	55.34	19.68
MintSR-L	41.10	19.38	23.29	7.89	55.06	19.60
MintSR-S	39.46	19.50	23.36	8.44	54.66	19.13
MintSR-I	40.34	18.86	23.39	8.40	54.18	18.93
MintSR-C	39.36	18.26	22.60	8.01	54.89	19.30
MintSR-FM	41.14	19.45	23.79	7.95	54.89	19.52
MintSR-TS	41.27	19.20	23.64	8.02	54.90	19.36



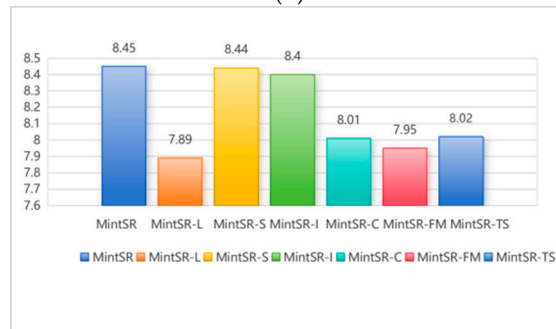
(a)



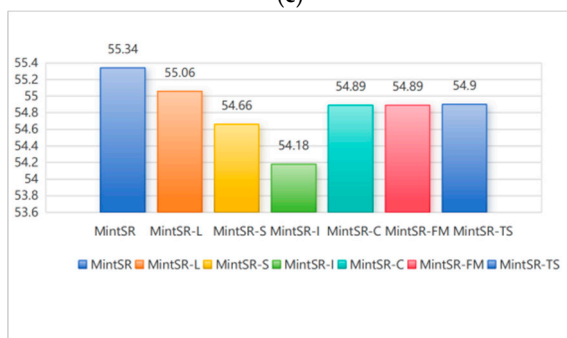
(b)



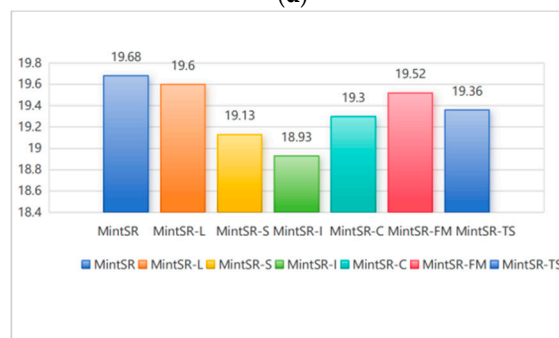
(c)



(d)



(e)



(f)

Figure 5. Ablation test results. (a) P@20 on Tmall (b) MRR@20 on Tmall. (c) P@20 on NowPlaying (d) MRR@20 on NowPlaying. (e) P@20 on Diginetica (f) MRR@20 on Diginetica.

RQ3 key parameter study: In this subsection, the impact on MintSR performance by adjusting the values of two key parameters is explored. This includes the effect of global intent and the impact of the number of iterative layers.

Effect of long-term intent: In the MintSR model, the impact of long-term intentions is explored in depth. The model introduces a Transformer-based long-term intent channel, which is meant to be utilized as supplemental data to increase the precision of user preference predictions. During the experiments, a representative set of long-term intent ratios was selected as a way to assess their actual impact on recommendation effectiveness. From the experimental results in Table 7 and Figure 6, on the shopping platform dataset Tmall, the model can achieve the best recommendation effect when the proportion of long-term intention is set to 0.5; whereas on the music platform dataset NowPlaying, this optimal effect occurs when the proportion of long-term intention is 0.1. This phenomenon fully demonstrates that different types of datasets require different proportions of long-term intentions to achieve optimal recommendation performance and also proves that a moderate amount of long-term intentions helps the model to mine deeper user preference information, thus improving recommendation accuracy. However, the experimental results also reveal a key point that over-reliance on long-term intent may introduce noisy information. When the proportion of long-term intent is too high, the model may capture information that is inconsistent with the intent of the target session, leading to a decrease in recommendation performance instead. This suggests that in practical applications, the proportion of long-term intent needs to be reasonably adjusted according to specific dataset characteristics and user behavior patterns to achieve the best balance of recommendation performance.

Table 7. Global intent impact study.

Datasets	Tmall	NowPlaying	Diginetica
	P@20	P@20	P@20
0.1	41.22	23.88	55.22
0.2	41.32	23.81	55.34
0.3	41.57	23.70	55.20
0.4	41.60	23.77	55.17
0.5	41.74	23.69	55.02

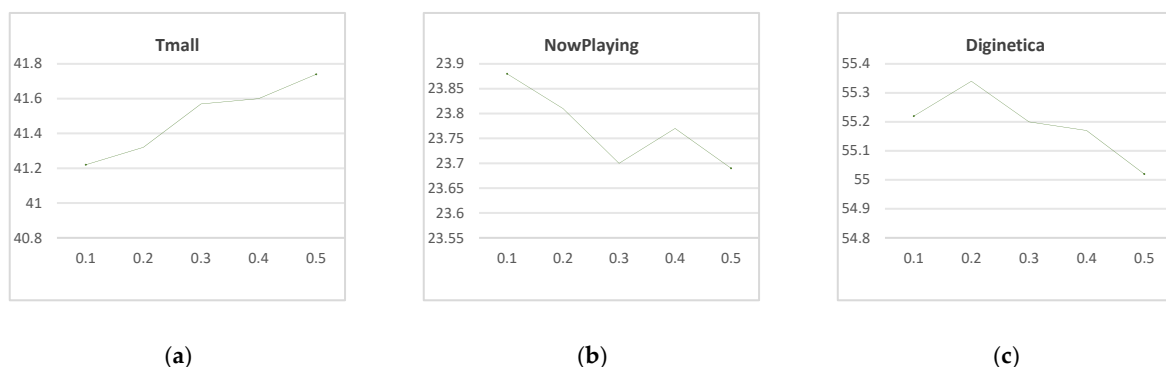


Figure 6. Global intent affects the results. (a) Tmall (b) NowPlaying (c) Diginetica.

The number of iteration layers affects the hyperparameter nl specifies the number of iterations in the multi-layer iterative session graph that is built in the local channel to mine local information. To determine their effect on recommendation effectiveness, this model employs a set of representative nl values {1, 2, 3, 4, 5}. The performance of SBR can be enhanced by selecting the right number of iterations. Traditional methods usually ignore the deep relationships in localized sessions. From the experimental results

in Table 8 and Figure 7, on the datasets NowPlaying and RetailRocket, the performance of the model improves significantly when the number of iterations is more than one (i.e., $nl > 1$) as compared to only a single extraction ($nl = 1$). This is strong evidence that proper multiple iterations can help the model to mine deeper information and thus improve the effectiveness of recommendations. However, the experimental results also reveal an issue that requires attention: when the number of iterations is too high, since the session graph will inevitably contain noisy information, the model may extract this unnecessary noise through excessive iterations, leading to gradient conflicts with the global channel and ultimately impairing the model's overall performance.

Table 8. Iteration layer impact study.

Datasets	Tmall	NowPlaying	Diginetica
	P@20	P@20	P@20
1	41.74	23.86	55.02
2	41.61	23.88	55.34
3	41.49	23.87	55.02
4	40.91	24.02	54.94
5	41.21	23.95	54.85

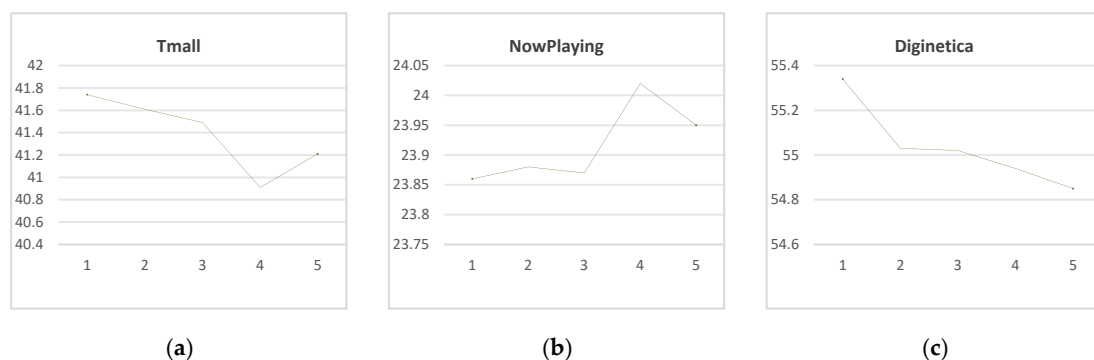
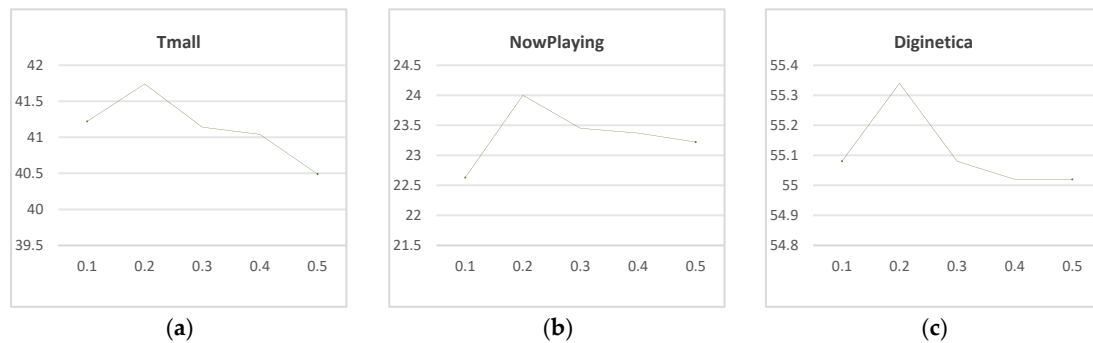


Figure 7. The number of iteration layers affects the results. (a) Tmall (b) NowPlaying (c) Diginetica.

In MintSR, the category-level contrastive learning loss is employed to strengthen the “intent-item” semantic association, where its weight λ directly determines the relative importance of the contrastive signal versus the primary recommendation task. We selected the representative range $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ to systematically evaluate λ 's impact on P@20. Table 9 and Figure 8, experimental results show that $\lambda = 0.2$ achieves peak performance across all three datasets: Tmall 41.74%, NowPlaying 24.00%, and Diginetica 55.34%, significantly outperforming the baseline without contrastive learning. This indicates that moderate contrastive constraints widen the distance between different intent categories, enhance the discriminative power of intent representations, and consequently improve recommendation accuracy. However, as λ increases to 0.3 and beyond, performance consistently declines: Tmall drops by 1.25 percentage points at $\lambda = 0.5$, and NowPlaying decreases by 0.78 percentage points. This indicates that overly strong contrast signals can obscure the primary task gradient, weakening the recommendation objective and potentially introducing additional noise. Therefore, in practical deployment, early stopping based on the validation set or automatic parameter tuning should be employed to maintain λ around 0.2, achieving the optimal balance between contrast enhancement and recommendation performance.

Table 9. The contrastive learning impact study.

Datasets	Tmall	NowPlaying	Diginetica
	P@20	P@20	P@20
0.1	41.22	22.63	55.08
0.2	41.74	24.00	55.34
0.3	41.14	23.45	55.08
0.4	41.04	23.37	55.02
0.5	40.49	23.22	55.02

**Figure 8.** The contrastive learning affects the results. (a) Tmall (b) NowPlaying (c) Diginetica.

5. Conclusions

In this paper, a session recommendation algorithm called MintSR is proposed, which centers on deeply mining and exploiting the hierarchical intent of users. After analyzing the relationship between long-term, short-term, and transient intent, it is found that short-term intent and transient intent need to be clearly delineated to avoid mutual interference between the two, while long-term intent can be used as auxiliary information to provide strong support for the recommendation process. Considering the findings above, this paper proposes an intent capture module that can accurately distinguish different levels of intents, thus effectively reducing the mutual interference between them. Meanwhile, to accurately capture long-term intentions, a Transformer-based global context extractor is designed in this study. Furthermore, to further improve the recommendation accuracy and performance, a category comparison learning task is implemented. The effectiveness of the MintSR framework in session recommendation scenarios is validated by experimental results on three datasets, which demonstrate notable increases in real-time and detection accuracy when compared to conventional approaches. However, for the user's intent on item attributes such as price and style, which further affect the user's desire to interact, the next step will be to try to design a module that captures the user's intent on the item's own attributes as auxiliary information for more efficient recommendations.

Author Contributions: Conceptualization, R.Y.; methodology, R.Y.; software, R.Y.; validation, R.Y.; formal analysis, R.Y.; investigation, R.Y.; resources, R.Y.; data curation, R.Y.; writing—original draft preparation, R.Y.; writing—review and editing, R.Y.; visualization, R.Y.; supervision, Y.H.; project administration, R.Y.; funding acquisition, X.Z. and P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in TMALL: <https://tianchi.aliyun.com/dataset/dataDetail?dataId=42> (accessed on 4 November 2025), Diginetica: <https://competitions.codalab.org/competitions/11161> (accessed on 4 November 2025), Nowplaying: <https://www.kaggle.com/datasets/chelseapower/nowplayingrs> (accessed on 4 November 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hsieh, C.K.; Yang, L.; Cui, Y.; Belongie, S.; Estrin, D. Collaborative metric learning. In Proceedings of the 26th International Conference on World Wide Web, Geneva, Switzerland, 3–7 April 2017; pp. 193–201.
2. Gao, C.; Wang, X.; He, X.; Li, Y. Graph neural networks for recommender system. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event, 21–25 February 2022; pp. 1623–1625.
3. Li, A.; Zhu, J.; Li, Z.; Cheng, H. Transition information enhanced disentangled graph neural networks for session-based recommendation. *Expert Syst. Appl.* **2022**, *210*, 118336. [\[CrossRef\]](#)
4. Hidasi, B.; Karatzoglou, A. Recurrent neural networks with top-k gains for session-based recommendations. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, 22–26 October 2018; pp. 843–852.
5. Luo, A.; Zhao, P.; Liu, Y.; Zhuang, F.; Wang, D.; Xu, J.; Fang, J.; Sheng, V.S. Collaborative self-attention network for session-based recommendation. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Main Track, Yokohama, Japan, 11–17 July 2020; pp. 2591–2597.
6. Wu, S.; Tang, Y.; Zhu, Y.; Wang, L.; Xie, X.; Tan, T. Session-based recommendation with graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2019; Volume 33, pp. 346–353.
7. Xu, C.; Zhao, P.; Liu, Y.; Sheng, V.S.; Xu, J.; Zhuang, F.; Fang, J.; Zhou, X. Graph contextualized self-attention network for session-based recommendation. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence Main Track, Macao, China, 10–16 August 2019; Volume 19, pp. 3940–3946.
8. Yu, F.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, L.; Tan, T. TAGNN: Target attentive graph neural networks for session-based recommendation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 25–30 July 2020; pp. 1921–1924.
9. Qiu, R.; Li, J.; Huang, Z.; Yin, H. Rethinking the item order in session-based recommendation with graph neural networks. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing China, 3–7 November 2019; pp. 579–588.
10. Wang, Z.; Wei, W.; Cong, G.; Li, X.; Mao, X.; Qiu, M. Global context enhanced graph neural networks for session-based recommendation. In Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, Virtual Event, 25–30 July 2020; pp. 169–178.
11. Zhang, P.; Guo, J.; Li, C.; Xie, Y.; Kim, J.B.; Zhang, Y.; Xie, X.; Wang, H.; Kim, S. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, Singapore, 27 February–3 March 2023; pp. 168–176.
12. Wang, X.; Dai, T.; Liu, Q.; Liang, S. Spatial-temporal perceiving: Deciphering user hierarchical intent in session-based recommendation. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, Jeju, Korea, 3–9 August 2024; pp. 2415–2423.
13. Mnih, A.; Salakhutdinov, R.R. Probabilistic matrix factorization. *Adv. Neural Inf. Process. Syst.* **2007**, *20*, 1257–1264.
14. Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *42*, 30–37. [\[CrossRef\]](#)
15. Shani, G.; Heckerman, D.; Brafman, R.I. An MDP-based recommender system. *J. Mach. Learn. Res.* **2005**, *6*, 453–460.
16. Rendle, S.; Freudenthaler, C.; Schmidt-Thieme, L. Factorizing personalized markov chains for next-basket recommendation. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 811–820.
17. Sarwar, B.; Karypis, G.; Konstan, J.; Riedl, J. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th International Conference on World Wide Web, Hong Kong, China, 1–5 May 2001; pp. 285–295.
18. Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; Tikk, D. Session-based recommendations with recurrent neural networks. *arXiv* **2015**, arXiv:1511.06939.
19. Kang, W.C.; McAuley, J. Self-attentive sequential recommendation. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 197–206.
20. Li, J.; Ren, P.; Chen, Z.; Ren, Z.; Lian, T.; Ma, J. Neural attentive session-based recommendation. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 1419–1428.
21. Zheng, Y.; Gao, C.; Chang, J.; Niu, Y.; Song, Y.; Jin, D.; Li, Y. Disentangling long and short-term interests for recommendation. In Proceedings of the ACM Web Conference, Sydney, Australia, 28 April–2 May 2022; Volume 2022, pp. 2256–2267.

22. Liu, Q.; Zeng, Y.; Mokhosi, R.; Zhang, H. STAMP: Short-term attention/memory priority model for session-based recommendation. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1831–1839.
23. Pan, Z.; Cai, F.; Chen, W.; Chen, H.; de Rijke, M. Star graph neural networks for session-based recommendation. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, 19–23 October 2020; pp. 1195–1204.
24. Xia, X.; Yin, H.; Yu, J.; Shao, Y.; Cui, L. Self-supervised graph co-training for session-based recommendation. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual Event, 1–5 November 2021; pp. 2180–2190.
25. Liu, H.; Jing, L.; Yu, J.; Ng, M.K. Learning hierarchical preferences for recommendation with mixture intention neural stochastic processes. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3237–3251. [\[CrossRef\]](#)
26. Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749.
27. Xia, X.; Yin, H.; Yu, J.; Wang, Q.; Cui, L.; Zhang, X. Self-supervised hypergraph convolutional networks for session-based recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2021; Volume 35, pp. 4503–4511.
28. Wang, F.; Gao, X.; Chen, Z.; Lyu, L. Contrastive multi-level graph neural networks for session-based recommendation. *IEEE Trans. Multimed.* **2023**, *25*, 9278–9289. [\[CrossRef\]](#)
29. Yang, F.; Peng, D. A graph neural network with topic relation heterogeneous multi-level cross-item information for session-based recommendation. *Inf. Syst.* **2024**, *123*, 102380. [\[CrossRef\]](#)
30. Wang, Q.; Cui, H.; Zhang, J.; Du, Y.; Lu, X. Self-supervised global graph neural networks with enhance-attention for session-based recommendation. *Appl. Soft Comput.* **2024**, *150*, 111026.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.