Attribute Association Driven Multi-Task Learning for Session-based Recommendation

Xinyao Wang[1] , **Zhizhi Yu**[1*] , **Dongxiao He**[1] , **Liang Yang**[2] , **Jianguo Wei**[1] , **Di Jin**[1,3*]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]School of Artifcial Intelligence, Hebei University of Technology, Tianjin, China
[3]Key Laboratory of Artificial Intelligence Application Technology, Qinghai Minzu University, Xining, 810007, China
{wxy2023, yuzhizhi, hedongxiao, jianguo, jindi}@tju.edu.cn, yangliang@vip.qq.com

Abstract

Session-based Recommendation (SBR) aims to predict users' next interaction based on their current session without relying on long-term profiles. Despite its effectiveness in privacy-preserving and real-time scenarios, SBR remains challenging due to limited behavioral signals. Prior methods often overfit co-occurrence patterns, neglecting semantic priors like item attributes. Recent studies have attempted to incorporate item attributes (e.g., category) by assigning fixed embeddings shared across all sessions. However, such approaches suffer from two key limitations: 1) Static attribute encoding fails to reflect semantic shifts under different session contexts. 2) Semantic misalignment between attribute and item ID embeddings. To address these issues, we propose attribute association driven multi-task learning for SBR, dubbed A^{2}D-MTL. It explicitly models item categories using cross-session context to capture user potential interests and designs an adaptive sparse attention mechanism to suppress noise. Experimental results on three public datasets demonstrate the superiority of our method in recommendation accuracy (P@20) and ranking quality (MRR@20), validating the model's effectiveness.

1 Introduction

Session-based Recommendation (SBR) has become a critical task in recommendation systems. A session refers to a sequence of consecutive user interactions [Tan *et al.*, 2016; Tuan and Phuong, 2017], which typically completes within a short time frame and independent of broader context, reflecting the user's preferences or needs in a specific situation. Unlike traditional recommendation systems that rely on long-term user profiles, SBR focuses on leveraging the information within a single session to predict the user's next behavior or preference. This approach is particularly valuable in scenarios where user information cannot be obtained due to privacy constraints or in real-time recommendation contexts, such as overseas e-commerce websites or video-streaming platforms.

*Corresponding authors.

The primary challenges faced by session-based recommendation systems stem from information scarcity caused by short sequences and the interference of noise in the data. First, sessions typically focus only on the ongoing interaction, without considering long-term interaction history. As a result, the sequences are short, and explicit user signals (such as clicks, inputs, or feedback) are limited, leading to data sparsity. Moreover, users may click on items that are irrelevant to their current interests out of curiosity, exploration, or random clicks without clear purchase intent. These behaviors may not reflect the user's true needs and introduce noise, making it difficult to accurately infer user intentions. Consequently, the model must effectively capture potential user intentions even with limited sequence data while maintaining robustness to noise to prevent performance degradation.

Earlier works in session-based recommendation primarily rely on the temporal characteristics of sessions, focusing on single-session modeling. For instance, Recurrent Neural Network (RNN)-based methods [Hidasi *et al.*, 2016; Tan *et al.*, 2016; Tuan and Phuong, 2017; Li *et al.*, 2017], such as GRU4Rec [Hidasi *et al.*, 2016] and its variants using Gated Recurrent Units (GRUs), encode the sequential information of items within sessions. These methods predict the next item the user may interact with based on the temporal dependencies between items in the session data. Self-attention-based methods, such as NARM [Li *et al.*, 2017], overcome the limitations of RNNs by capturing dependencies not just between adjacent items, but across all items in the sequence, improving the model's ability to recognize long-range relationships. However, these methods typically rely solely on session-level context and treat each session independently, neglecting cross-session context. This limits their ability to effectively utilize complex transfer relationships between items, thus reducing recommendation performance. Not only that, self-attention mechanisms often focus excessively on certain parts of the sequence, ignoring the global relevance of user behaviors and making it harder to capture diverse user intentions.

In recent years, some approaches have sought to incorporate both within-session and cross-session context information to improve session-based recommendations. Graph Neural Network (GNN)-based methods [Chen and Wong, 2020; Pan *et al.*, 2020; Xia *et al.*, 2021], such as GCE-GNN [Wang

et al., 2020] and MTD [Huang *et al.*, 2021], model the data of all historical sessions as a session graph, capturing item correlations within the same session and across different sessions. This dual modeling approach creates a complex correlation network between items, facilitating a more comprehensive understanding of user behavior patterns and improving the accuracy of recommendations through message passing on the session graph. Each node (representing an item) aggregates information from its neighboring nodes, enhancing the model's ability to capture complex dependencies. However, these methods often focus solely on item co-occurrence patterns and overlook other valuable information, such as item attributes like categories, resulting in biased recommendations that overemphasize frequently co-purchased items. For instance, as shown in Figure 1, a promotional campaign associates hiking backpacks with children's schoolbags, artificially elevating the latter in the ranking list. Yet this outcome deviates from the user's genuine interests. Based on category-level semantics, the user exhibits preferences for both outdoor and electronic items—implying a stronger interest in items like power banks, which bridge these two scenarios. Such recommendations align more closely with the user's underlying intent than those shaped by incidental promotional associations, highlighting the importance of leveraging attribute information to improve recommendation relevance.

To alleviate the bias introduced by relying solely on item co-occurrence patterns, some studies have explored incorporating item attribute information to provide additional semantic signals for user preference modeling. Existing methods [Chen *et al.*, 2023; Song *et al.*, 2021] typically assign a fixed embedding to each item attribute before training, which is shared across all sessions. These attribute embeddings are then directly combined with item ID embeddings to construct the overall item representation. However, it suffers from two major limitations: 1) **Static encoding neglects how attribute semantics vary with session contexts:** As shown in Figure 1, the functional and aesthetic attributes emphasized for a hardshell jacket differ significantly between casual wear and outdoor sports scenarios. These context-insensitive embeddings make it difficult to accurately capture user intent across sessions and track evolving interests within a session. 2) **Semantic misalignment between attribute and ID embeddings:** Attribute embeddings encode global, time-invariant semantics, while item ID embeddings emphasize sequential dependencies and local transition patterns. This semantic inconsistency may lead to conflicts when the two are naively fused, introducing noise that ultimately hinders the model's ability to identify fine-grained user preferences and degrades recommendation quality.

To address the aforementioned challenges, we propose **A**ttribute **A**ssociation **D**riven **M**ulti-**T**ask **L**earning for session-based recommendation, termed A^{2}D-MTL. Specifically, we first construct a global category graph based on inter-session information to capture the structural correlations among item categories. On top of this, each item dynamically attends to multiple semantically related categories and adaptively fuses them with the current session context. This design overcomes the limitations of static attribute modeling and enhances the model's ability to generalize user interest transitions and expand recommendation coverage. Then, we introduce an adaptive sparse attention mechanism guided by dynamic category representations. By incorporating item position and contextual state within the session, the attention distribution is adjusted in a data-driven manner to explicitly focus on attribute signals that are most relevant to the user's current intent. This mechanism effectively suppresses irrelevant noise and facilitates accurate modeling of true user preferences. Experimental results on three public datasets demonstrate that our method significantly outperforms existing methods in recommendation accuracy (P@20) and ranking quality (MRR@20), with the rationality and effectiveness of the model design being verified.

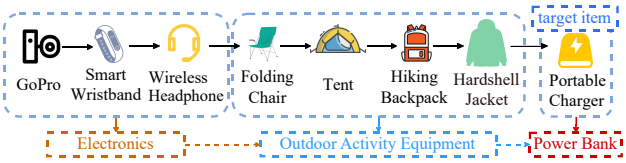

Figure 1: An illustrate example containing multiple categories in a session, where the first half represents electronic items, the second half represents outdoor activity equipment, and the final target item is a power bank.

2 Preliminaries

Notations. Let $V = \{v_1, v_2, \ldots, v_N\}$ denote the set of all items, where N is the total number of items. A session is represented as $S = \{v_1, v_2, \ldots, v_T\}$, where T is the length of the session. Additionally, items are often associated with auxiliary information such as categories. Let $C = \{c_1, c_2, \ldots, c_M\}$ represent the set of categories, where M is the total number of categories. Each item $v \in V$ corresponds to a category $c \in C$. For a session S, the category sequence can be derived as $S_c = \{c_{v_1}, c_{v_2}, \ldots, c_{v_T}\}$, where $c_{v_t} \in C$ denotes the category of item v_t.

Problem Definition. The goal of Session-based Recommendation (SBR) is to predict the next item a user will interact with based on their anonymous historical behavior in a session. Specifically, the task of SBR is to recommend the most likely next item v_{T+1} based on the current session S.

3 Method

We begin with a brief overview of the proposed method, followed by a detailed description of each component.

3.1 Overview

To address the limitations of static category encoding and the semantic misalignment between attributes and item sequences, we propose attribute association driven multi-task learning (A^{2}D-MTL) for SBR, which jointly captures item-level interactions and category-level associations. Dynamic category embeddings guide an adaptive sparse attention mechanism for fine-grained intent modeling. The framework includes two key modules: category relation modeling via a global category graph, and item interaction modeling across

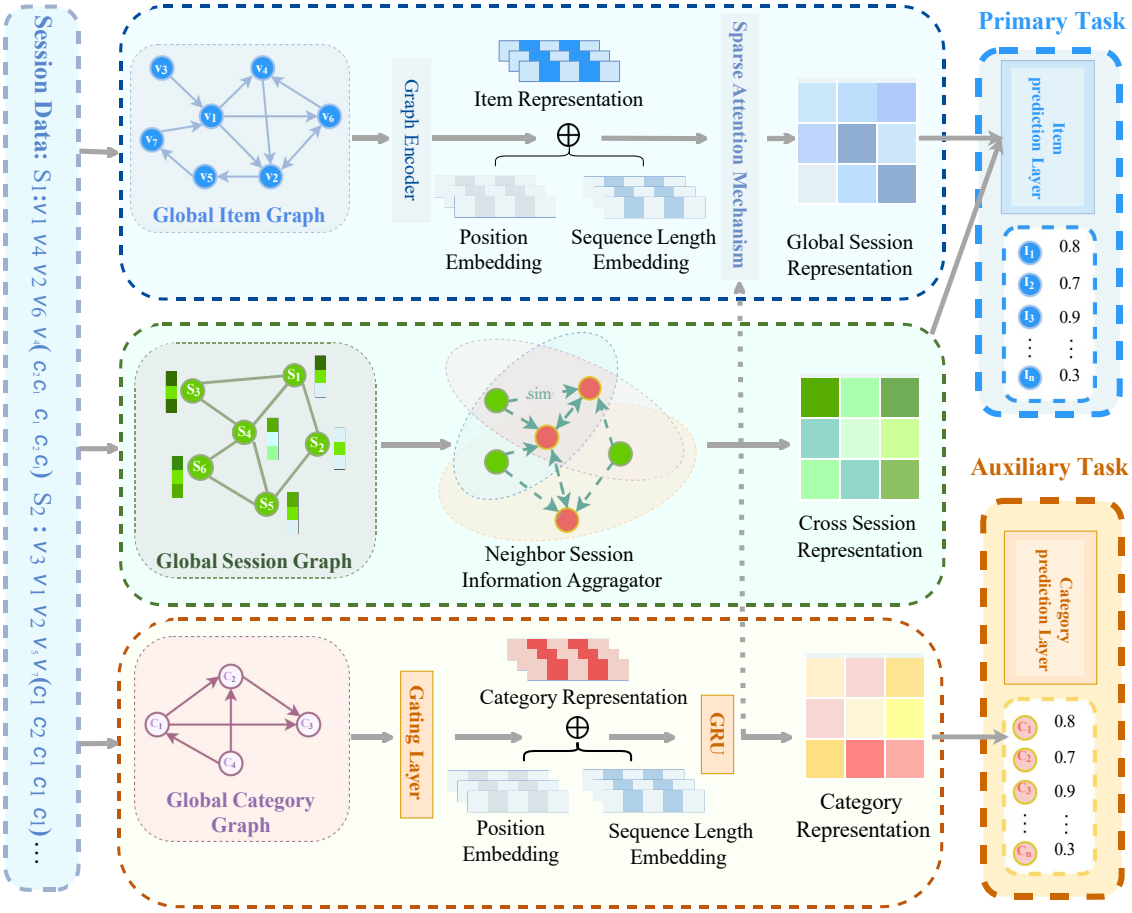

Figure 2: Overall architecture of our proposed A^{2}D-MTL.

and within sessions. A multi-task objective combining category prediction and item recommendation further enhances the semantic guidance and improves model adaptability and generalization.

3.2 Graph Construction

To effectively model the sequential dependencies and category-level relationships in session-based recommendation, we construct two global co-occurrence graphs: Item Graph and Category Graph. These directed graphs are built from the historical session sequences in the dataset. Similar to MSGAT [Qiao *et al.*, 2023], normalized edge weights are used to ensure that the constructed graphs are not dominated by frequent transitions involving popular nodes, maintaining fairness in the graph construction.

Item Graph Construction. The item graph G_i captures the co-occurrence relationships between items across all sessions. For each session sequence $S = \{v_1, v_2, \ldots, v_T\}$, directed edges are created between consecutive items in the order they appear. The edge weight between items v_i and v_j, denoted as w_{ij}, represents the normalized co-occurrence frequency:

$$w_{ij} = \frac{\text{count}(v_i \to v_j)}{\sum_{k \in N_{\text{out}}(v_i)} \text{count}(v_i \to v_k)}, \tag{1}$$

where $\text{count}(v_i \to v_j)$ denotes the number of times v_i is followed by v_j in the session sequences, and $N_{\text{out}}(v_i)$ is the set of items that v_i points to in the graph. It is worth noting that the directed edge $v_i \to v_j$ indicates the order in which items appear, preserving the sequential nature of user interactions.

Category Graph Construction. The category graph G_c models the relationships between item categories. For a given session $S = \{v_1, v_2, \ldots, v_T\}$ and its corresponding category sequence $S_c = \{c_{v_1}, c_{v_2}, \ldots, c_{v_T}\}$, directed edges are created between consecutive categories c_{v_i} and c_{v_j}. Formally, the edge weight $w_{c_{ij}}$ between categories c_i and c_j can be defined as:

$$w_{c_{ij}} = \frac{\text{count}(c_i \to c_j)}{\sum_{k \in N_{\text{out}}(c_i)} \text{count}(c_i \to c_k)}, \tag{2}$$

where $\text{count}(c_i \to c_j)$ is the number of times c_i is followed by c_j in the category sequences, and $N_{\text{out}}(c_i)$ is the set of categories that c_i points to in the graph.

3.3 Category-Level Representation Learning

After constructing the Category Graph, we aim to learn global representations of each category to capture the sequential and dependency relationships between categories. To achieve this, we employ a graph-based approach that first learns the global representations of each category through a Graph Attention Network (GAT) [Veličković *et al.*, 2018].

Given a category sequence $S_c = \{c_1, c_2, \ldots, c_T\}$, the goal is to learn the representation of each category c_t. To model the global dependencies between categories, we introduce neighborhood aggregation to update the representation $h_{c_t}^k$ of each category c_t at the k-th layer, which is formulated as:

$$h_{c_t}^k = \alpha_{c_t,c_t} W_1 h_{c_t}^{k-1} + \textstyle\sum_{c_j \in N(c_t)} \alpha_{c_t,c_j} W_1 h_{c_j}^{k-1}, \quad (3)$$

where α_{c_t,c_j} is the attention score between categories c_t and c_j, representing the importance of category c_t with respect to its neighboring category c_j. The attention score α_{c_t,c_j} is computed as follows:

$$\alpha_{c_t,c_j} = \frac{\exp\left(a^\top \sigma\left(W_1[h_{c_t}^{k-1} \| h_{c_j}^{k-1} \| e_{c_t,c_j}]\right)\right)}{\sum_{c_k \in N(c_t)\cup\{c_t\}} \exp\left(a^\top \sigma\left(W_1[h_{c_t}^{k-1} \| h_{c_k}^{k-1} \| e_{c_t,c_k}]\right)\right)}, \quad (4)$$

where $\|$ denotes the concatenation operation, e_{c_t,c_j} is the edge weight between categories c_t and c_j, a and W_1 are learnable parameters, and σ is the activation function.

Then, we apply a gating mechanism to adaptively fuse the initial and graph-based category representations as follows:

$$\tilde{h}_{c_t} = \beta_{c_t} \cdot h_{c_t}^{(0)} + (1 - \beta_{c_t}) \cdot h_{c_t}^{\text{global}}, \quad (5)$$

where $\tilde{h}_{c_t}$ denotes the final fused embedding of category c_t, $h_{c_t}^{(0)}$ is the initial category embedding, and $h_{c_t}^{\text{global}}$ is the global representation obtained via the graph attention network. The fusion weight $\beta_{c_t} \in [0, 1]$ is computed through a gating function as:

$$\beta_{c_t} = \sigma\left(W_2\left[h_{c_t}^{(0)} \| h_{c_t}^{\text{global}}\right] + b_2\right), \quad (6)$$

where $\|$ denotes concatenation, W_2 and b_2 are learnable parameters, and $\sigma(\cdot)$ is the sigmoid activation function that ensures the fusion weight lies within $[0, 1]$. This gating mechanism adaptively balances semantic priors and structure-aware representations.

To enhance temporal modeling, the fused category representation is combined with positional and session-length embeddings and encoded by a GRU [Cho *et al.*, 2014]:

$$S_{\text{cate}} = \text{GRU}(\text{concat}(\tilde{h}_{c_t},\ p_t,\ l_t)), \quad (7)$$

where $\tilde{h}_{c_t}$ is the fused category embedding at time step t, p_t and l_t denote the positional and length embeddings, respectively. In this way, the latent dependencies among categories can be effectively modeled by integrating global structure-aware semantics with sequence-aware positional and contextual signals, thus improving both the expressiveness and predictive performance of the category sequence encoder.

3.4 Item Relation Modeling

Intra-Session Representation Learning. We apply a GAT the previously constructed item graph G_i to obtain global item representations $H_{item}^{global} = \{h_{v_1}, h_{v_2}, ..., h_{v_N}\}$. To further encode the current session, we extract multiple types of session-level features, including the embedding of the last interacted item h_{v_T}, which reflects the user's most recent interest; the full item embedding sequence $H_{\text{item}} \in \mathbb{R}^{T\times D}$, where T denotes the session length and D is the embedding dimension; the average-pooled item embedding $H_{\text{item}}^{\text{avg}}$; and the average-pooled category embedding $H_{\text{cate}}^{\text{avg}}$.

To unify feature dimensions, we apply separate linear transformations to each input as follows:

$$q_{v_T} = W_1 h_{v_T} + b_1, \quad Q_{\text{seq}} = H_{item} W_2 + b_2, \quad (8)$$

$$q_{\text{item}} = W_3 H_{item}^{\text{avg}} + b_3, \quad q_{\text{cate}} = W_4 H_{cate}^{\text{avg}} + b_4, \quad (9)$$

where $W_i \in \mathbb{R}^{d\times D}$ and $b_i \in \mathbb{R}^d$ (for $i = 1, 2, 3, 4$) are learnable parameters. Here, q_{v_T}, q_{item}, and $q_{\text{cate}} \in \mathbb{R}^{1\times d}$ denote global session-level features, while $Q_{\text{seq}} \in \mathbb{R}^{T\times d}$ encodes local item-level representations.

Global features are broadcast-added to local representations, followed by a sigmoid activation to obtain the enhanced session representation $S_{\text{seq}} \in \mathbb{R}^{T\times d}$:

$$S_{\text{seq}} = \sigma(Q_{\text{seq}} + q_{v_T} + q_{\text{item}} + q_{\text{cate}}). \quad (10)$$

Then, we apply an Adaptive Sparse Attention Mechanism (ASAM) to model contextual dependencies. The enhanced session representation S_{seq} is first transformed via a multi-layer perceptron (MLP):

$$S_{\text{trans}} = \text{Dropout}(\text{ReLU}(W_{\text{MLP}} S_{\text{seq}} + b_{\text{MLP}})), \quad (11)$$

where $W_{\text{MLP}} \in \mathbb{R}^{d'\times d}$ and $b_{\text{MLP}} \in \mathbb{R}^{d'}$ are the weight and bias parameters, and d' denotes the output dimension.

The transformed representation S_{trans} serves as the query, while the original sequence S_{seq} is used as both key and value in the attention computation:

$$\alpha_{\text{att}} = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \quad (12)$$

where d_k is the key dimensionality, and $\alpha_{\text{att}} \in \mathbb{R}^{T\times d'}$ denotes the context-aware representation weighted by similarity between queries and keys.

To further enhance model expressiveness, we refine the attention output using residual connections and layer normalization. The residual path preserves original information and mitigates gradient vanishing, while the normalization stabilizes the output distribution:

$$a = \text{LN}(\alpha_{\text{att}} + \text{MLP}(\alpha_{\text{att}})). \quad (13)$$

We apply the Entmax activation function [Yuan *et al.*, 2021] to normalize the attention weights, promoting sparsity and encouraging the model to focus on the most relevant items in the session:

$$\alpha = \text{Entmax}(\alpha, \alpha_{\text{ent}}), \quad (14)$$

where α_{ent} is a hyperparameter controlling the degree of sparsity. Finally, the session-level representation is computed as a weighted sum over context features:

$$S_{\text{item}}^{\text{intra}} = \alpha^\top v, \quad (15)$$

where v is a matrix containing external features (e.g., item or category embeddings), and $S_{\text{item}}^{\text{intra}} \in \mathbb{R}^d$ serves as the final session representation for downstream prediction.

Inter-Session Representation Learning. To enhance the model's generalization ability, we incorporate *inter-session representation learning* based on session-level interaction similarity. For a given session n, we first compute the average-pooled representation of all items within the session:

$$s_n^{\text{avg}} = \frac{1}{N_n} \sum_{i=1}^{N_n} h_{v_i^{(n)}}, \tag{16}$$

where N_n denotes the number of items in session n, and $h_{v_i^{(n)}}$ is the embedding of the i-th item in that session.

Then, we compute the cosine similarity between session n and all other sessions:

$$\text{sim}(n, m) = \frac{s_n^{\text{avg}} \cdot s_m^{\text{avg}}}{\|s_n^{\text{avg}}\| \, \|s_m^{\text{avg}}\|}. \tag{17}$$

We select the top-R most similar sessions to form an auxiliary session set $\mathcal{R}(n)$. The inter-session representation is then obtained by aggregating the reference sessions with similarity-based weights:

$$S_{\text{item}}^{\text{inter}} = \sum_{m \in \mathcal{R}(n)} \gamma_{n,m} s_m^{\text{avg}}, \tag{18}$$

where the attention weights $\gamma_{n,m}$ are computed via a softmax function:

$$\gamma_{n,m} = \frac{\exp(\text{sim}(n, m))}{\sum_{m' \in \mathcal{R}(n)} \exp(\text{sim}(n, m'))}. \tag{19}$$

By leveraging contextual information from similar sessions, the model captures broader behavioral patterns, thereby enhancing the robustness of item prediction.

Final Item Representation Learning. After obtaining the intra-session representation $S_{\text{item}}^{\text{intra}}$ and the inter-session representation $S_{\text{item}}^{\text{inter}}$, we compute the final item-level representation via weighted summation:

$$S_{\text{item}} = S_{\text{item}}^{\text{intra}} + S_{\text{item}}^{\text{inter}}. \tag{20}$$

The resulting representation S_{item} is then used for item prediction in the session-based recommendation task.

3.5 Prediction

For item prediction, we compute the dot product of the final item representation S_{item} with the embeddings H_{v_i} of the item i to obtain the predicted scores $\hat{y}_i$:

$$\hat{y}_i = \text{softmax}\left(\langle S_{\text{item}}, H_{v_i} \rangle\right), \tag{21}$$

where $\hat{y}_i$ represents the likelihood of item v_i being the next item in the session. We use the cross-entropy loss function to measure the discrepancy between the predicted and true item labels, where y_i denotes the one-hot encoding vector of the ground truth value:

$$L_{item} = -\sum_{i=1}^{M} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i), \tag{22}$$

where M is the total number of items in the session.

For category prediction, we follow a similar process. The predicted category score $\hat{y}_c$ is computed by taking the dot product of the final session representation S_{item} with all category embeddings H_{c_i} in the category set C:

$$\hat{y}_c = \text{softmax}\left(\langle S_{\text{cate}}, H_{c_i} \rangle\right). \tag{23}$$

We also use the cross-entropy loss for category prediction, where y_c denotes the one-hot encoding vector of the true category:

$$L_{cate} = -\sum_{i=1}^{N} y_{c_i} \log(\hat{y}_{c_i}) + (1 - y_{c_i}) \log(1 - \hat{y}_{c_i}), \tag{24}$$

where N is the number of categories.

The total loss function is then a combination of both item and category loss:

$$L = L_{item} + \lambda \cdot L_{cate}, \tag{25}$$

where λ is a hyperparameter that balances the two loss terms.

4 Experiment

In our study, we conduct an extensive set of experiments on three real-world datasets to investigate three key research questions:

- **RQ1:** Can our proposed A^{2}D-MTL outperform the baselines for session-based recommendation?
- **RQ2:** What is the role of each component in driving the recommendation performance of A^{2}D-MTL?
- **RQ3:** How is the performance of A^{2}D-MTL affected by different parameter settings?

Dataset	Diginetica	Tmall	Yoochoose1_64
#Training Sessions	719,470	351,268	369,859
#Test Sessions	60,858	25,898	55,898
#Items	43,097	40,728	16,766
#Categories	995	711	341
Average Length	5.12	6.69	6.16

Table 1: The statistical results of the datasets.

4.1 Experimental Settings

Datasets. Following recent studies on session-based recommendation systems [Hou *et al.*, 2022; Zhang *et al.*, 2023b; Wang *et al.*, 2024b], three widely-used public benchmark datasets are adopted in our work: **Diginetica**, **Tmall**, and **Yoochoose1_64**.

- **Diginetica**[1]: A dataset of anonymous user transaction records from an e-commerce search engine's logs over five months, provided by the CIKM Cup 2016.
- **Tmall**[2]: A dataset of anonymized shopping logs from the Tmall platform, released for the IJCAI15 competition.

[1] http://cikm2016.cs.iupui.edu/cikm-cup

[2] https://tianchi.aliyun.com/dataset/42

Datasets	Diginetica		Tmall		Yoochoose1_64	
Metrics	P@20	MRR@20	P@20	MRR@20	P@20	MRR@20
GRU4Rec	29.49	8.22	10.93	5.89	60.64	22.89
STAMP	46.62	15.13	26.47	13.36	68.76	29.47
SR-GNN	50.73	17.59	27.57	13.72	70.54	30.97
GCE-GNN	54.22	19.04	33.42	15.42	70.91	30.62
MTD	51.82	17.26	29.12	13.73	71.88	31.32
MSGIFSR	56.08	19.13	36.54	16.09	72.44	<u>32.03</u>
AttenMixer	54.56	19.06	32.49	15.21	65.09	30.92
MiaSRec	54.85	18.93	40.12	<u>16.45</u>	72.93	32.02
MGS	54.85	19.23	40.27	16.62	<u>73.94</u>	31.33
CLHHN	<u>55.67</u>	<u>19.58</u>	<u>41.20</u>	16.39	72.57	30.90
A^{2}D-MTL	**61.94**	**23.25**	**41.38**	**17.43**	**76.06**	**33.94**

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

- **Yoochoose1_64**[3]: A dataset of user click events on an e-commerce platform, created for the RecSys Challenge 2015, using the latest 1/64 portion of training sessions .

Following the approach in [Wu *et al.*, 2019; Wang *et al.*, 2020], we preprocess the three datasets. Specifically, the most recent week's historical sessions are used as the test set, while the remaining sessions are used as the training set. Additionally, items that did not appear in the training set are filtered out from the test set to avoid the influence of cold-start factors for new items. Finally, sessions with a length of 1 and items that appeared fewer than five times are removed. The details of the datasets after preprocessing are summarized in Table 1.

Baselines. We categorize the baselines into two types: those that do not use side information (e.g., GRU4Rec [Hidasi *et al.*, 2016], STAMP [Liu *et al.*, 2018], SR-GNN [Wu *et al.*, 2019], GCE-GNN [Wang *et al.*, 2020], MTD [Huang *et al.*, 2021], MSGIFSR [Guo *et al.*, 2022], Atten-Mixer [Zhang *et al.*, 2023a], and MiaSRec [Choi *et al.*, 2024]), and those that use side information (e.g., MGS [Lai *et al.*, 2022] and CLHHN [Ma *et al.*, 2024]). MGS uses side information through a mirror graph, while CLHHN incorporates it via a heterogeneous hypergraph to improve item representations.

Metrics. We adopt two widely used evaluation metrics in information retrieval: *Precision* (*P@20*) and *Mean Reciprocal Rank* (*MRR@20*) for evaluating the performance.

Hyper-parameter Setup. Following [Wu *et al.*, 2019; Wang *et al.*, 2020], the dimension of the latent vectors is fixed to 256, and the batch size is set to 100. We use the Adam optimizer with the initial learning rate of 0.001, which decays by 0.8 after every 3 epochs. The l_2 penalty is set to 10^{-5}.

4.2 Performance Comparison

To evaluate the effectiveness of A^{2}D-MTL, we report the comparison results with the state-of-the-art baselines. From Table 2, we draw the following observations:

- A^{2}D-MTL outperforms all RNN-based methods on all datasets. Unlike GRU4Rec, which uses GRU to model sequential data without focusing on key time points, A^{2}D-MTL employs an attention mechanism to dynamically highlight important items at each time step. Overall, GNN-based models, including A^{2}D-MTL, outperform traditional sequential models by capturing more complex item transitions. Additionally, models like GCE-GNN and MTD perform better than SR-GNN, suggesting that integrating item transitions from other sessions improves interest prediction.
- Methods leveraging attribute information as side information consistently outperform GNN-based and multi-level models, highlighting its effectiveness in handling sparse interactions in short sequences.
- Our model outperforms all baseline models across all metrics on three datasets, demonstrating its superiority. The performance improvement over methods like MGS, which directly integrates category information into item embeddings, or CLHHN, which links items and categories through hyperedges, further highlights that constructing a global graph with categories as independent nodes better preserves the independence and flexibility of category information, enhances global information modeling, and avoids excessive coupling between items and categories.

4.3 Ablation Study

To investigate the necessity of each component of A^{2}D-MTL, we design four variants and conduct ablation experiments across three datasets. Specifically, we use the A^{2}D-MTL variant without side information (category information) **(A^{2}D-MTL w/o CATE)** for modeling; the variant that directly integrates category into item embeddings **(A^{2}D-MTL w/o EXP)**; the variant without sparse attention **(A^{2}D-MTL w/o SPA)**, which only uses GAT as an encoder; and the variant without adaptive adjustment of sparse attention coefficients **(A^{2}D-MTL w/o AdaSPA)**.

Table 3 presents the comparison results, from which we make the following observations: 1) Compared to **A^{2}D-MTL w/o CATE**, **A^{2}D-MTL w/o EXP** performs better, highlighting the importance of side information. However, **A^{2}D-MTL w/o EXP** underperforms **A^{2}D-MTL**, indicating that constructing a graph with categories as independent nodes yields

[3]http://2015.recsyschallenge.com/challege

Datasets	Diginetica		Tmall		Yoochoose1_64	
Metrics	P@20	MRR@20	P@20	MRR@20	P@20	MRR@20
A^{2}D-MTL w/o CATE	52.92	16.65	35.25	13.34	70.18	30.24
A^{2}D-MTL w/o EXP	54.56	21.13	36.54	16.29	73.03	32.44
A^{2}D-MTL w/o SPA	56.85	19.22	39.57	15.19	73.94	31.02
A^{2}D-MTL w/o AdaSPA	59.21	20.67	40.09	15.78	74.98	31.90
A^{2}D-MTL	61.94	23.25	41.38	17.43	76.06	33.94

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

more accurate recommendations than directly integrating category information into item embeddings. 2) When compared to the variant without the sparse attention mechanism (**A^{2}D-MTL w/o SPA**), the variant without adaptive adjustment of sparse attention coefficients (**A^{2}D-MTL w/o AdaSPA**) performs better, demonstrating the effectiveness of the sparse attention mechanism. However, its MRR value is lower than that of **A^{2}D-MTL**, suggesting that the adaptive sparse attention mechanism helps more accurately identify noise and rank target items higher. 3) **A^{2}D-MTL** achieves the best performance, demonstrating that the integration of commodity category association guidance and the adaptive sparse attention mechanism enhances recommendation relevance.

4.4 Parameter Sensitivity

The regularization parameter λ in Eq.(25) balances the item prediction loss and category prediction loss in A^{2}D-MTL. We evaluate the performance of A^{2}D-MTL under different λ values $\{0.01, 0.1, 0.2, 0.23, 0.25, 0.27, 0.29, 0.3, 0.4, 0.5\}$. As shown in Figure 3, A^{2}D-MTL achieve good results on both datasets. Specifically, increasing the value of λ within the range of 0.01 to 0.25 gradually improved performance metrics. When λ is between 0.25 and 0.29, performance metrics stabilize. However, when λ exceeds 0.3, the performance metrics decline significantly. These observations indicate that selecting an appropriate value of λ can enhance recommendation performance.

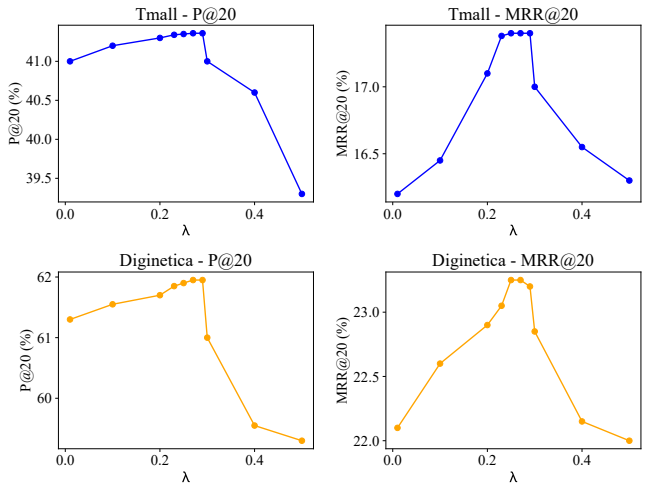

Figure 3: The impact of λ across Tmall and Diginetica datasets in terms of P@20 and MRR@20.

5 Related Work

Early RNN-based session-based recommendation (SBR) methods, such as GRU4Rec [Hidasi *et al.*, 2016], predict user interest by modeling the temporal dependencies in the item sequence. Later, NARM [Li *et al.*, 2017] introduces the attention mechanism to enhance the modeling of item importance, but it still lacked sufficient exploration of category information. As research progressed, attention mechanisms derived from Transformer models were widely applied to SBR. For example, CoSAN [Luo *et al.*, 2020] generates richer item representations by combining neighborhood information with attention mechanisms. Graph Neural Networks (GNNs) have also been extensively used in SBR. SR-GNN [Wu *et al.*, 2019] is the first to apply Gated Graph Neural Networks (GGNN) [Li *et al.*, 2016] to SBR, learning transfer relationships between items by propagating information on the session graph. GCE-GNN [Wang *et al.*, 2020] enhances understanding of complex session structures by learning session representations across multiple graph layers. MGIR [Han *et al.*, 2022] strengthens session representations by learning global item relationships.

Recent studies highlight the importance of item category information in improving recommendation performance, as it is both effective and easy to obtain in practice. Integrating such information into SBR models has become an effective approach to enhancing recommendation performance. For example, CLHHN [Ma *et al.*, 2023] introduces category hyperedges and constructs hypergraphs to capture high-order interactions among categories, thereby revealing more complex item similarities. Additionally, HearInt [Wang *et al.*, 2024a] integrates category information into its intent recognition module to achieve finer-grained semantic partitioning when parsing dynamic user interests, thus accurately capturing users' ever-changing demands.

6 Conclusion

This paper proposes A^{2}D-MTL, a novel session-based recommendation method that leverages item and category graphs to capture sequential and categorical relationships. By utilizing a Graph Attention Network (GAT) for global item and category representation learning, and employing a multi-head sparse attention mechanism to capture session context, A^{2}D-MTL significantly improves performance, surpassing existing baselines. Our approach enhances global information modeling while preserving the independence of category information, demonstrating its effectiveness in addressing sparse interaction issues and improving recommendation accuracy.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62402337, No. 92370111, No. 62422210, No. 62272340, and No. 62276187), the Postdoctoral Fellowship Program of CPSF under Grant No. GZC20241207, the China Postdoctoral Science Foundation under Grant No. 2024M752367, and the Hebei Natural Science Foundation under Grant No. F2024202047.

References

[Chen and Wong, 2020] Tianwen Chen and Raymond Chi-Wing Wong. Handling information loss of graph neural networks for session-based recommendation. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1172–1180, 2020.

[Chen *et al.*, 2023] Qian Chen, Jianjun Li, Zhiqiang Guo, Guohui Li, and Zhiying Deng. Attribute-enhanced dual channel representation learning for session-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3793–3797, 2023.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[Choi *et al.*, 2024] Minjin Choi, Hye-young Kim, Hyunsouk Cho, and Jongwuk Lee. Multi-intent-aware session-based recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2532–2536, 2024.

[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 *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*, pages 343–352, 2022.

[Han *et al.*, 2022] Qilong Han, Chi Zhang, Rui Chen, Riwei Lai, Hongtao Song, and Li Li. Multi-faceted global item relation learning for session-based recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1705–1715, 2022.

[Hidasi *et al.*, 2016] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

[Hou *et al.*, 2022] Yupeng Hou, Binbin Hu, Zhiqiang Zhang, and Wayne Xin Zhao. CORE: simple and effective session-based recommendation within consistent representation space. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1796–1801, 2022.

[Huang *et al.*, 2021] Chao Huang, Jiahui Chen, Lianghao Xia, Yong Xu, Peng Dai, Yanqing Chen, Liefeng Bo, Jiashu Zhao, and Jimmy Xiangji Huang. Graph-enhanced multi-task learning of multi-level transition dynamics for session-based recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4123–4130, 2021.

[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 *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1674–1683, 2022.

[Li *et al.*, 2016] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

[Li *et al.*, 2017] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1419–1428, 2017.

[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 *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1831–1839, 2018.

[Luo *et al.*, 2020] Anjing Luo, Pengpeng Zhao, Yanchi Liu, Fuzhen Zhuang, Deqing Wang, Jiajie Xu, Junhua Fang, and Victor S. Sheng. Collaborative self-attention network for session-based recommendation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2591–2597, 2020.

[Ma *et al.*, 2023] Yutao Ma, Zesheng Wang, Liwei Huang, and Jian Wang. Clhhn: Category-aware lossless heterogeneous hypergraph neural network for session-based recommendation. *ACM Transactions on the Web*, 18(1):1–37, 2023.

[Ma *et al.*, 2024] Yutao Ma, Zesheng Wang, Liwei Huang, and Jian Wang. CLHHN: category-aware lossless heterogeneous hypergraph neural network for session-based recommendation. *ACM Trans. Web*, 18(1):12:1–12:37, 2024.

[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 *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 1195–1204, 2020.

[Qiao *et al.*, 2023] Shutong Qiao, Wei Zhou, Junhao Wen, Hongyu Zhang, and Min Gao. Bi-channel multiple sparse graph attention networks for session-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2075–2084, 2023.

[Song *et al.*, 2021] Jiayu Song, Jiajie Xu, Rui Zhou, Lu Chen, Jianxin Li, and Chengfei Liu. Cbml: A cluster-based meta-learning model for session-based recommendation. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 1713–1722, 2021.

[Tan *et al.*, 2016] Yong Kiam Tan, Xinxing Xu, and Yong Liu. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pages 17–22, 2016.

[Tuan and Phuong, 2017] Trinh Xuan Tuan and Tu Minh Phuong. 3d convolutional networks for session-based recommendation with content features. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 138–146, 2017.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[Wang *et al.*, 2020] Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xianling Mao, and Minghui Qiu. 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*, pages 169–178, 2020.

[Wang *et al.*, 2024a] Xiao Wang, Tingting Dai, Qiao Liu, and Shuang Liang. Spatial-temporal perceiving: deciphering user hierarchical intent in session-based recommendation. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 2415–2423, 2024.

[Wang *et al.*, 2024b] Yu Wang, Amin Javari, Janani Balaji, Walid Shalaby, Tyler Derr, and Xiquan Cui. Knowledge graph-based session recommendation with session-adaptive propagation. In *Proceedings of the ACM on Web Conference*, pages 264–273, 2024.

[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 *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 346–353, 2019.

[Xia *et al.*, 2021] Xin Xia, Hongzhi Yin, Junliang Yu, Yingxia Shao, and Lizhen Cui. Self-supervised graph co-training for session-based recommendation. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 2180–2190, 2021.

[Yuan *et al.*, 2021] Jiahao Yuan, Zihan Song, Mingyou Sun, Xiaoling Wang, and Wayne Xin Zhao. Dual sparse attention network for session-based recommendation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 4635–4643, 2021.

[Zhang *et al.*, 2023a] 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 *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pages 168–176, 2023.

[Zhang *et al.*, 2023b] Zhihui Zhang, Jianxiang Yu, and Xiang Li. Context-aware session-based recommendation with graph neural networks. In *Proceedings of the IEEE International Conference on Knowledge Graph*, pages 35–44, 2023.