Capturing User Intent through Integration of Item ID and Modality Information in Session Recommendation

Tiantian Liang, Zhe Yang * [1]
School of Computer Science and Technology, Soochow University, Suzhou, China
ttliang2023@stu.suda.edu.cn, yangzhe@suda.edu.cn

Abstract

Session-based recommendation aims to capture user intent from short-term, anonymous interaction sequences to recommend relevant items. From a cognitive science perspective, understanding user intent is closely tied to how humans process information, allocate attention, and make decisions under limited cognitive resources. While existing session-based methods mainly rely on ID-based modeling, such approaches face severe data sparsity and lack alignment with how users cognitively process information. Incorporating modality information can alleviate this issue. However, simple integration of ID and multimodal information often results in modality underfitting, limiting the effective use of multimodal features. To address these challenges, we propose SRIM(Session-based Recommendation with ID and Modality), a model that integrates ID and multimodal representations through a two-phase strategy: independent training followed by joint optimization. SRIM can better capture session-level intent by simulating users' actual perceptual contexts. Experiments on three real-world datasets demonstrate that SRIM significantly outperforms existing methods in session recommendation. The code for SRIM is available on GitHub https://github.com/liang-tian-tian/SRIM.
Keywords: Artificial Intelligence, Session Intent, Graph Neural Networks.

Introduction

In recent years, predicting the next clicked or purchased item from user-item interaction sequences has become a hot topic in recommendation algorithms (Hou, Hu, et al., 2022; Qiao et al., 2023; X. Zhang et al., 2024).

Drawing an analogy from cognitive science, human decision-making under limited attention and cognitive resources involves selective focus on relevant information, often integrating multiple sensory modalities to form a coherent understanding. Applying this analogy to recommender systems suggests that effective models should not only process sequential interactions but also align and integrate multimodal information to better reflect human-like understanding and prediction of user intent.

In session-based recommendation methods, ID-based approaches currently dominate. Approaches such as RNNs, attention mechanisms, GNNs, and contrastive learning (Hidasi, 2015; Li et al., 2017; S. Wu et al., 2019; Xia et al., 2021) are widely used to model session intent. While these methods are effective and straightforward, they mainly rely on user-item interaction history, overlooking other valuable information. To address these limitations, recent research has explored integrating ID-based methods with multimodal information. This includes incorporating item descriptions (Liu et al., 2020), images (X. Zhang et al., 2023), categories (Lai et al., 2022), and prices (X. Zhang et al., 2022). However, simple information fusion in multimodal integration can lead to underfitting(Hou, Mu, et al., 2022; Wei et al., 2024; T. Zhang et al., 2019).

To address these challenges, we propose SRIM, a model that integrates both ID-based and multimodal information through a two-phase strategy: independent training followed by joint optimization. In the independent training phase, we utilize Graph Convolutional Networks (GCNs) and a direct multi-head self-attention mechanism to generate vector representations for sessions based on ID and multimodal features, respectively. In the joint optimization phase, we employ contrastive learning to align these representations, ensuring semantic consistency across modalities and enhancing the model's ability to capture user intents. SRIM exhibits more accurate recommendations compared to previous approaches. We summarize the contributions as follows:

- We design an ID and multimodal information alignment module that uses contrastive learning to align multimodal features corresponding to specific IDs while maintaining semantic consistency within both the ID and multimodal domains.
- A joint inference loss function is designed for the session's ID and multimodal representations. When the session label co-occurs with items in the session, the ID-based loss is treated as the primary loss, while the multimodal loss serves as the auxiliary loss; conversely, the multimodal loss becomes primary, with the ID loss as auxiliary.
- The direct multi-head self-attention mechanism is introduced when generating ID and multimodal vector representations to further enhance recommendation performance. Specifically, SRIM improves P@20 and MRR@20 by 12.24% and 5.01% on the Cellphones dataset.

Related Work

Our research aims to enhance session-based recommendation through the integration of multimodal information, drawing inspiration from both cognitive science and modality-based recommendation methods.

[1]* Corresponding author.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Cognitive Effects of Multisensory Integration

Recent findings in cognitive science emphasize that multisensory integration significantly enhances human interest formation and decision-making processes. For example, (Mercier & Cappe, 2020) demonstrate that congruent auditory and visual cues accelerate both early sensory encoding and later decision-related neural activity, suggesting that richer multimodal inputs speed up the dynamics underlying choice. Furthermore, attention to multisensory information has been shown to boost neural processing of the attended modality (Seijdel et al., 2024). Notably, individuals tend to favor the sensory modality in which they have higher subjective confidence (Gao et al., 2025). These cognitive insights imply that effective integration of multimodal information can engage users more deeply, enhance their confidence in decision-making, and guide their interests—an observation highly relevant to recommender systems.

Modality-based Session Recommendation Methods

Building on these cognitive insights, recent session-based recommendation studies have introduced modality information to enrich item representations. Modality-based methods incorporate rich textual and visual information to characterize fine-grained item features and user preferences. Some studies have found that simply merging multimodal data does not always outperform single-modal approaches (Du et al., 2023; Peng et al., 2022; N. Wu et al., 2022; Huang et al., 2022; Wang et al., 2020) and may lead to underfitting in one of the modalities. For example, (Li et al., 2024) propose alternating training of ID and textual information to enhance the complementary learning ability between them, addressing training imbalances. (X. Zhang et al., 2023) integrate item image and price information to build deterministic and probabilistic models, thoroughly mining user intent. (X. Zhang et al., 2024) decouple ID and multimodal data, enhancing both recommendation effectiveness and the interpretability of session recommendation. Currently, there is still insufficient research on effectively integrating ID and multimodal information, lacking in-depth exploration of interactions and information fusion between modalities.

SRIM Model

As shown in Figure 1, the proposed SRIM model consists of several key components: (1) preprocessing and modeling of ID and multimodal information, aiming to learn ID vector representations and multimodal vector representations that fit the session context; (2) an ID and modality alignment strategy that effectively integrates information between different modalities using contrastive learning while maintaining semantic consistency within modalities; (3) a joint inference strategy that dynamically adjusts the weights of primary recommendation reasons, enhancing their impact during the recommendation process, supplemented by secondary reasons to optimize overall recommendation performance; (4) the prediction phase of the model and the backward optimization process.

ID and Modality Preprocessing and Modeling

Modality Data Preprocessing Due to the significant semantic gap between textual and visual information, we follow the processing method in (X. Zhang et al., 2024). For each item, we use the GoogleNet (Szegedy et al., 2015) model to classify its categories, denoted as x_i^{ctxt}. We then fuse these category text descriptions with the original text descriptions (e.g., title, brand) to form x_i^{mo}, which is input into the Bert (Kenton & Toutanova, 2019) model to generate the multimodal vector representation of the item. This process effectively transforms visual information into text form, achieving a unified representation across modalities. The final multimodal information representation of an item is as follows:

$$x_i^{ctxt} = \text{GoogleNet}(x_i^{img}) \tag{1}$$

$$x_i^{mo} = \{w_1, w_2, \ldots, w_t, w'_1, w'_2\} \tag{2}$$

$$e_i^{mo} = \text{Bert}(x_i^{mo}) \tag{3}$$

where x_i^{img} represents the item's image information, $\{w_1, w_2, \ldots, w_t\}$ represents the item's original text descriptions, $x_i^{ctxt} = \{w'_1, w'_2\}$ represents the category text descriptions generated from the image, and e_i^{mo} is the multimodal vector representation generated by the Bert model.

ID and Modality Representation Modeling To effectively process item ID information and multimodal information, we first employ a simple GCN. The initial ID vectors for the items are randomly initialized and denoted as $E_{id} = \{e_1^{id}, e_2^{id}, \ldots, e_N^{id}\}$. The multimodal information comes from the Bert model and is denoted as $E_{mo} = \{e_1^{mo}, e_2^{mo}, \ldots, e_N^{mo}\}$(where N is the total number of items). The update processes for the ID vectors and the multimodal vectors are defined as follows:

$$E_{id}^{(l+1)} = D^{-1} A E_{id}^{(l)} W_{id}^{l} \tag{4}$$

where A is the adjacency matrix, representing the co-occurrence matrix of items, and D is the degree matrix with elements $D_{p,p} = \sum_{q=1}^{N} A_{p,q}$, which represents the degree of item p as the total co-occurrence count of the item. Similarly, the multimodal vectors follow the same update process, where $E_{mo}^{(l+1)} = D^{-1} A E_{mo}^{(l)} W_{mo}^{l}$, with $E_{mo}^{(l)}$ and W_{mo}^{l} denoting the multimodal vector and weight matrix at layer l. The corresponding ID and multimodal session representations are given by:

$$s^{id} = \frac{1}{m} \sum_{t=1}^{m} e_t^{id} \tag{5}$$

where m is the session length. The multimodal session representation s^{mo} is computed similarly, with e_t^{mo} in place of e_t^{id}.

For the adjacency matrix A, we define it using the co-occurrence patterns of item ID information. Specifically, we consider that items appearing in the same session have some latent association. Therefore, we construct a co-occurrence matrix for all items in all sessions. If a pair of items co-occur

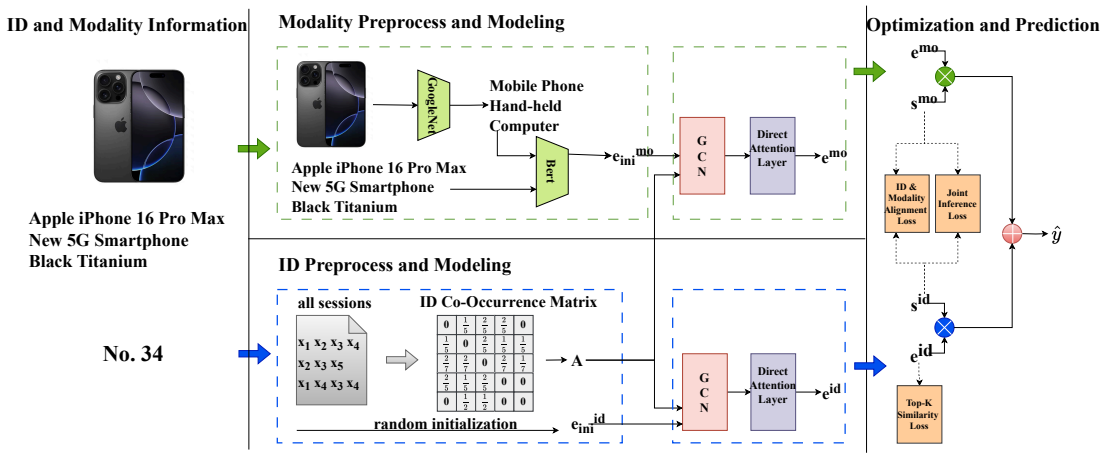

Figure 1: An overview of the proposed SRIM method

in multiple sessions, their co-occurrence value in the matrix is accumulated based on the number of times they co-occurred. If an item has never co-occurred with others in any session, its co-occurrence value is 0. We define the co-occurrence matrix as $A \in \mathbb{R}^{N \times N}$, where N is the total number of items. The co-occurrence count of items i and j across all sessions is represented as:

$$A_{ij} = \sum_{k=1}^{M} \mathbb{I}(i \in S_k \wedge j \in S_k) \tag{6}$$

where A_{ij} denotes the co-occurrence value of items i and j, M is the total number of sessions, S_k is the set of items in the k-th session, and $\mathbb{I}$ is the indicator function that takes the value 1 when i and j co-occur in session S_k, otherwise it is 0.

After several layers of GCN convolution, we employ a direct multi-head self-attention mechanism to adaptively compute the relationship weights among all items in the session, assigning greater weight to items that are more closely aligned with the session's intent. The direct multi-head self-attention mechanism is defined as follows:

$$\hat{S}_{id} = \text{Softmax}\left(\frac{\left(Q_{id} W_{id}^{Q}\right) K_{id}^{T}}{\sqrt{d_{id}}}\right) V_{id} \tag{7}$$

where Q_{id}, K_{id} and V_{id} denote the original session ID sequence $S_{id} = \{e_1^{id}, e_2^{id}, \cdots, e_m^{id}\}$, with each item representation e_i^{id} learned through GCN. Similarly, the multimodal sequence, $S_{mo} = \{e_1^{mo}, e_2^{mo}, \cdots, e_m^{mo}\}$, follows the same attention computation process, with the equation for $\hat{S}_{mo}$ analogous to $\hat{S}_{id}$. Here, m represents the session length. The dot product operation in the self-attention mechanism satisfies the commutative property, meaning if $Q = K$, then $QK^T = KQ^T$, which may lead to indistinguishable features between different roles. Swapping Q and K can shift the focus of the attention computation to different input features, resulting in varied contextual representations. To avoid this issue, we apply linear transformations only to Q.

After processing with the direct multi-head self-attention mechanism, we obtain the session ID sequence representation $\hat{S}_{id} = \{\hat{e}_1^{id}, \hat{e}_2^{id}, \cdots, \hat{e}_m^{id}\}$ and the multimodal sequence representation $\hat{S}_{mo} = \{\hat{e}_1^{mo}, \hat{e}_2^{mo}, \cdots, \hat{e}_m^{mo}\}$. The final session representation for the ID sequence is calculated as follows:

$$\hat{s}^{id} = \frac{1}{m} \sum_{t=1}^{m} \hat{e}_t^{id} \tag{8}$$

where m the length of the session. Similarly, the final multimodal representation $\hat{s}^{mo}$ is obtained using the same formula, with $\hat{e}_t^{mo}$ replacing $\hat{e}_t^{id}$.

ID and Modality Alignment

We obtain the initial ID representation s_{id} and multimodal representation s_{mo} of the session through GCN. We then use a direct multi-head self-attention mechanism to derive the final session ID representation $\hat{s}_{id}$ and multimodal representation $\hat{s}_{mo}$. To effectively fuse cross-modal information while maintaining internal consistency within their respective modality, we design a contrastive loss function to handle feature alignment both within and between modalities.

Within the same modality, we enhance semantic coherence and consistency by maximizing the cosine similarity between the session representations before and after applying the direct multi-head self-attention mechanism. The loss functions are computed as follows:

$$L_{align\_id} = \log\left(1 + \frac{\exp\left(\text{sim}(W_3 s^{id}, W_4 s^{mo})\right)}{1 + \exp\left(\text{sim}(W_1 s^{id}, W_2 \hat{s}^{id})\right)}\right) \tag{9}$$

Here, W_1 to W_6 are learnable linear transformation matrices, and $sim(\cdot)$ denotes similarity computation function. The alignment loss for the multimodal representation, L_{align_mo} follows the same structure, with $W_5 s^{mo}$ and $W_6 \hat{s}^{mo}$ used in place of $W_1 s^{id}$ and $W_2 \hat{s}^{mo}$ in the denominator.

To ensure the effective fusion of ID and multimodal information, we employ contrastive learning to align relevant features between the session ID representation and the multimodal representation while maintaining internal consistency within each modality. The inter-modal alignment loss L_{id_mo} is defined as:

$$L_{id\_mo} = \log\left(1 + \frac{\exp\left(\text{sim}(W_1 s^{id}, W_2 \hat{s}^{id})\right) + \exp(\text{sim}(W_5 s^{mo}, W_6 \hat{s}^{mo}))}{1 + \exp\left(\text{sim}(W_3 s^{id}, W_4 s^{mo})\right)}\right) \tag{10}$$

The total alignment loss L_{align} is defined as the sum of the above three alignment losses:

$$L_{align} = L_{align\_id} + L_{align\_mo} + L_{id\_mo} \tag{11}$$

Through this loss design, we ensure feature consistency within each modality and enhance feature fusion across different modalities through inter-modal contrastive loss.

Joint Inference

In the joint inference phase, we utilize a co-occurrence matrix to determine whether the session label co-occurs with items in the session. This co-occurrence relationship is used to select the primary loss function. If the label and items co-occur, we set the $flag$ to 1 during data processing. In this case, the ID loss is treated as the primary loss and the multimodal loss as the auxiliary loss, jointly optimizing both through backpropagation. Conversely, if the label and items do not co-occur, we consider the multimodal features to play a primary role, treating the multimodal loss as the primary loss and the ID loss as the auxiliary loss. The joint loss for the ID L_{joint_id} is calculated as follows:

$$L_{joint\_id} = \log\left(\exp(\text{sim}(s^{id}, e^{id}_{m+1})) + \exp(\text{sim}(s^{mo}, e^{mo}_{m+1}))\right) - \text{sim}(s^{id}, e^{id}_{m+1}) \tag{12}$$

where $sim(\cdot)$ represents the similarity computation function. The joint multimodal loss L_{joint_mo} with $\text{sim}(s^{mo}, e^{mo}_{m+1})$ replacing $\text{sim}(s^{id}, e^{id}_{m+1})$ in the subtraction term.

The final joint loss L_{joint} is determined based on the value of the flag, combining the primary and auxiliary losses as follows:

$$L_{joint} = \begin{cases} L_{joint\_id} + 0.5 \cdot L_{joint\_mo}, & \text{if } flag = 1 \\ L_{joint\_mo} + 0.5 \cdot L_{joint\_id}, & \text{otherwise} \end{cases} \tag{13}$$

This design dynamically adjusts the combination of primary and auxiliary losses, increasing the weight on the main causes of the recommendation results while treating other causes as auxiliary, thereby leveraging ID and multimodal information more comprehensively to enhance the model's recommendation performance.

Optimization and Prediction

Based on the learned session ID representation and multimodal representation, the score for candidate items in each session can be computed through the corresponding dot product summed up as follows:

$$\hat{y} = \hat{s}^{id} \cdot e^{id} + \hat{s}_{mo} \cdot e^{mo} \tag{14}$$

We utilize the cross-entropy loss function to optimize the model, defined as:

$$L_{main} = -\sum_{i=1}^{N} (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{15}$$

The final total loss function is:

$$L = L_{main} + \beta \left(L_{align} + L_{joint} + L_{co}\right) \tag{16}$$

where β is a hyperparameter controlling the scale of the contrastive loss functions. The loss function L_{co} selects the top-l items with the highest co-occurrence values for each item e^{id}_i as positive samples, denoted by $\{e^{id+}_1, e^{id+}_2, \cdots, e^{id+}_l\}$. Similarly, it selects l items with the lowest co-occurrence values as negative samples, denoted by $\{e^{id-}_1, e^{id-}_2, \cdots, e^{id-}_l\}$. This loss function aims to minimize the distance between frequently co-occurring positive samples while maximizing the distance between items that co-occur less frequently, defined as follows:

$$L_{co} = -\frac{sim(e^{id}_i,\ \bar{e}^{+}_i)}{\sum_{k=1}^{l} sim(e^{id}_i,\ e^{id-}_k)} \tag{17}$$

where $sim(\cdot)$ denotes the cosine similarity, and $\bar{e}^{+}_i = \frac{1}{l}\sum_{k=1}^{l} e^{id+}_k$ represents the average representation of the positive samples with higher co-occurrence values.

Experiment

Dataset

We evaluate our SRIM model and all baseline models on three datasets sourced from Amazon[2], covering different domains including Cellphones, Grocery, and Sports. Following the experiment environment in (X. Zhang et al., 2024), we treat the last item in each session as the prediction target, while the remaining items are used to model session intent. Additionally, we filter out sessions of length 1 and items that apeared fewer than 5 times. Items with missing or invalid images/text are also removed. Finally, we split each dataset into training, validation, and test sets in a 7:2:1 ratio, maintaining temporal order. Statistics for all datasets are shown in Table 2.

Baseline Methods

We select two groups of competitive methods for performance comparison. ID-based models primarily provide recommendations by mining co-occurrence patterns of items. NARM (Li et al., 2017) is an RNN model based on attention mechanisms designed to capture the user's main intent. Bert4Rec (Sun et al., 2019) employs a bidirectional attention mechanism to model user behavior sequences. SR-GNN (S. Wu et al., 2019)

[2] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

Table 1: Performance comparison of SRIM and all baseline methods on three datasets. The best results are highlighted in bold, and the second-best results are underlined.

Methods	Cellphones				Grocery				Sports			
	P@10	M@10	P@20	M@20	P@10	M@10	P@20	M@20	P@10	M@10	P@20	M@20
NARM(CIKM'17) (Li et al., 2017)	15.42	12.43	16.80	12.53	45.67	40.39	47.14	40.59	35.55	33.40	36.67	33.57
BERT4Rec(CIKM'19) (Sun et al., 2019)	18.31	11.96	22.44	12.25	45.82	38.33	49.07	38.56	38.13	33.89	40.34	34.01
SR-GNN(AAAI'19) (S. Wu et al., 2019)	16.36	12.96	18.11	13.09	44.33	39.44	46.24	39.64	36.31	33.36	37.69	33.66
MSGIFSR(WSDM'22) (Guo et al., 2022)	17.80	12.40	21.16	12.64	45.45	38.16	48.15	38.35	36.27	30.36	39.65	30.59
Atten-Mixer(WSDM'23) (P. Zhang et al., 2023)	19.51	14.54	22.28	14.71	47.65	40.71	49.56	40.84	37.30	33.63	39.19	33.86
MSGAT(CIKM'23) (Qiao et al., 2023)	17.22	13.41	20.01	13.67	45.20	39.98	47.01	40.12	37.19	33.69	38.63	33.91
MGS(SIGIR'22) (Lai et al., 2022)	21.54	14.24	25.02	14.48	46.59	38.83	48.37	38.98	36.79	32.39	48.45	32.50
UniSRec(SIGKDD'22) (Hou, Mu, et al., 2022)	20.30	14.32	23.78	14.56	47.95	40.90	50.21	41.05	38.31	33.50	40.62	33.76
CoHHN(SIGIR'22) (X. Zhang et al., 2022)	23.60	15.77	27.71	15.96	41.58	35.33	43.59	35.58	32.12	27.13	35.02	27.31
MMSBR(TKDE'23) (X. Zhang et al., 2023)	20.59	13.94	22.82	14.13	46.05	39.01	47.89	39.23	36.69	32.52	38.29	32.73
DIMO(SIGIR'24) (Li et al., 2024)	31.66	16.98	38.81	17.36	53.03	41.81	57.01	41.98	45.07	34.86	49.86	35.15
SRIM	**34.84**	**17.64**	**43.56**	**18.23**	**56.99**	**42.76**	**62.25**	**43.12**	**47.70**	**35.49**	**54.07**	**35.91**
Impro.	10.04%	3.88%	12.24%	5.01%	7.47%	2.27%	9.19%	2.72%	5.84%	1.81%	8.44%	2.16%

Table 2: Statistics of datasets

Datasets	Cellphones	Grocery	Sports
items	9,091	7,286	14,650
interaction	123,186	151,251	282,102
session	40,344	43,648	90,492
avg.length	3.05	3.47	3.12

is the first model to use graph neural networks for item vector representations. MSGIFSR (Guo et al., 2022) extracts multi-granularity user intent by segmenting sessions. Atten-Mixer (P. Zhang et al., 2023) captures multi-level user intent by generating multiple attention maps. MSGAT (Qiao et al., 2023) constructs both intra- and inter-session graphs to generate session intent.

Multimodal-based methods leverage additional information to capture users' fine-grained preferences. MGS (Lai et al., 2022) incorporates item category information to better estimate user preferences. UniSRec (Hou, Mu, et al., 2022) includes descriptive information about items to obtain more generalized sequential representations. CoHHN (X. Zhang et al., 2022) integrates item price information to better capture session intent. MMSBR (X. Zhang et al., 2023) is the first model to fuse item text and image information to enhance session recommendation performance. DIMO (Li et al., 2024) optimizes the model by decoupling the roles of ID information and multimodal information.

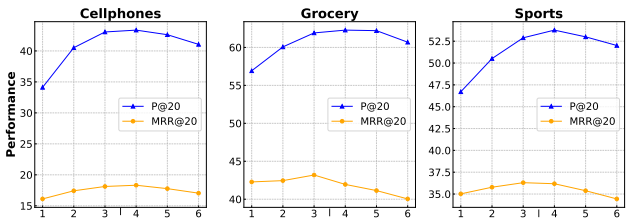

Figure 2: Impact of the GCN layers

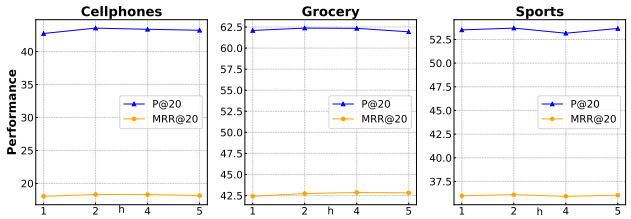

Figure 3: Impact of the attention heads

Hyper-parameter Settings

Following previous work, we set the embedding dimension to 100, the batch size to 50, and the L2 regularization coefficient to 10^{-5}. The model is optimized using the Adam optimizer with a learning rate of 0.001. In the SRIM model, the number of layers in the graph convolutional neural network is set to 4 for the Cellphones and Sports datasets and to 3 for the Grocery dataset, as shown in Fig.2. The number of heads in the direct multi-head self-attention mechanism is set to 2 for the Cellphones and Sports datasets and to 4 for the Grocery dataset, as illustrated in Fig.3. The weight parameter β for modality alignment and co-occurrence loss is set to 0.01 for the Grocery and Sports datasets and to 0.005 for the Cellphones dataset.

Results and Analysis

Overall Performance

Table 1 presents the performance comparison between the SRIM model and all baseline models. Among ID-based methods, different models excel in various metrics: RNN-based methods (NARM (Li et al., 2017), BERT4Rec (Sun et al., 2019)), GNN-based methods (SR-GNN (S. Wu et al., 2019), MSGAT (Qiao et al., 2023)), and multi-intent capturing methods (MSGIFSR (Guo et al., 2022), Atten-Mixer (P. Zhang et al., 2023)) each show strengths in specific aspects. In contrast, our SRIM model significantly outperforms these ID-based baselines in overall performance.

Table 3: Ablation study

Methods	Cellphones				Grocery				Sports			
	P@10	M@10	P@20	M@20	P@10	M@10	P@20	M@20	P@10	M@10	P@20	M@20
SRIM-IdMo	34.64	17.55	43.01	18.15	56.67	42.26	62.16	42.64	47.62	35.49	53.28	35.89
SRIM-JoIn	34.17	17.34	42.81	17.93	56.97	42.44	61.95	42.76	47.57	**35.69**	53.33	**36.06**
SRIM-DirAtt	30.15	16.31	37.93	16.85	52.37	41.16	56.09	41.40	44.06	34.36	49.26	34.68
SRIM	**34.84**	**17.64**	**43.56**	**18.23**	**56.99**	**42.76**	**62.25**	**43.12**	**47.70**	35.49	**54.07**	35.91

For multimodal methods, DIMO (Li et al., 2024) performs well by separating ID and multimodal features. SRIM further improves performance by learning ID and multimodal representations independently and integrating them via contrastive learning. It achieves notable gains in P@20 and MRR@20, with respective improvements of 12.24% and 5.01% on Cellphones, 9.19% and 2.72% on Grocery, and 8.44% and 2.16% on Sports.

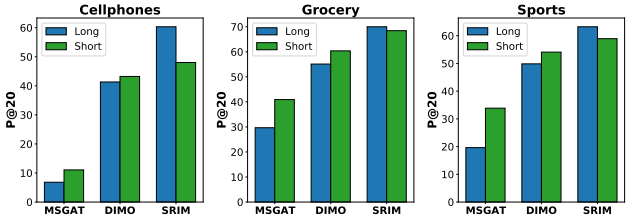

Figure 4: P@20 results on long and short sessions

Ablation Study

In this section, we design three SRIM variants to evaluate the contributions of individual components: SRIM-IdMo, SRIM-JoIn, and SRIM-DirAtt. SRIM-IdMo removes the alignment loss for ID and modality, SRIM-JoIn removes the joint inference loss for ID and modality, and SRIM-DirAtt replaces the direct multi-head self-attention mechanism with a configuration where Q equals K.

As shown in Table 3, SRIM-DirAtt suffers the most significant performance drop, indicating that when Q and K are identical, the attention mechanism can confuse attended and attending elements, leading to biased results. SRIM-JoIn shows a slight improvement in MRR metrics on the Sports dataset, possibly due to the presence of noise in the multimodal information of this dataset, where joint training introduces additional noise, thus affecting recommendation performance. After removing the alignment loss in SRIM-IdMo, all metrics decline, demonstrating the effectiveness of the alignment component.

Performance Analysis across Session Lengths

In real-world scenarios, it is common to encounter sessions of varying lengths. To evaluate model performance, we categorize sessions with lengths of 5 or fewer as short sessions, and those with lengths greater than 5 as long sessions. We compared our model against the latest ID-based model, MSGAT, and the modality-integrated model, DIMO. Figure 4 shows that SRIM outperforms the baseline models across different session lengths and datasets on P@20 metric, demonstrating its adaptability in real-world session-based scenarios.

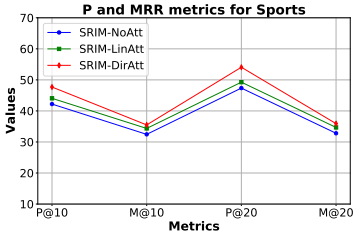

Figure 5: Evaluation of self-attention settings

Evaluation of Self-Attention Settings

To investigate the impact of Query-Key design in self-attention, we designed and compared three methods: SRIM-NoAtt completely removes the self-attention mechanism. SRIM-LinAtt uses identical Q and K, directly employing the session sequence as both the Query and Key. SRIM-DirAtt applies a linear transformation to obtain the Query while using the original session sequence as the Key, ensuring a clear distinction between Q and K.

As shown in Figure 5, SRIM-DirAtt outperforms SRIM-LinAtt across all metrics and datasets, and SRIM-LinAtt performs better than SRIM-NoAtt. This result indicates that separating Q and K through linear transformations within the self-attention mechanism can significantly enhance model performance.

Conclusion

This paper proposes a model named SRIM, which effectively integrates multi-modal information through components such as modality alignment and joint inference to deliver more accurate session-based recommendations. Inspired by cognitive science findings on multisensory integration, where congruent inputs enhance attention, confidence, and decision-making, SRIM is designed to leverage the complementary strengths of different modalities to better capture user intent. Extensive experiments on three public benchmark datasets demonstrate that SRIM consistently outperforms current state-of-the-art methods, validating both its technical effectiveness and its cognitive motivation.

Acknowledgements

This work is supported by the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Du, C., Teng, J., Li, T., Liu, Y., Yuan, T., Wang, Y., ... Zhao, H. (2023). On uni-modal feature learning in supervised multi-modal learning. In *International conference on machine learning* (pp. 8632–8656).

Gao, Y., Xue, K., Odegaard, B., & Rahnev, D. (2025). Automatic multisensory integration follows subjective confidence rather than objective performance. *Communications Psychology*, *3*(1), 38.

Guo, J., Yang, Y., Song, X., Zhang, Y., Wang, Y., Bai, J., & Zhang, Y. (2022). Learning multi-granularity consecutive user intent unit for session-based recommendation. In *Proceedings of the fifteenth acm international conference on web search and data mining* (pp. 343–352).

Hidasi, B. (2015). Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Hou, Y., Hu, B., Zhang, Z., & Zhao, W. X. (2022). 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* (pp. 1796–1801).

Hou, Y., Mu, S., Zhao, W. X., Li, Y., Ding, B., & Wen, J.-R. (2022). Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th acm sigkdd conference on knowledge discovery and data mining* (pp. 585–593).

Huang, Y., Lin, J., Zhou, C., Yang, H., & Huang, L. (2022). Modality competition: What makes joint training of multimodal network fail in deep learning?(provably). In *International conference on machine learning* (pp. 9226–9259).

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-hlt* (Vol. 1, p. 2).

Lai, S., Meng, E., Zhang, F., Li, C., Wang, B., & Sun, A. (2022). 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* (pp. 1674–1683).

Li, J., Han, H., Chen, Z., Shomer, H., Jin, W., Javari, A., & Tang, J. (2024). Enhancing id and text fusion via alternative training in session-based recommendation. *arXiv preprint arXiv:2402.08921*.

Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., & Ma, J. (2017). Neural attentive session-based recommendation. In *Proceedings of the 2017 acm on conference on information and knowledge management* (pp. 1419–1428).

Liu, Y., Ren, Z., Zhang, W.-N., Che, W., Liu, T., & Yin, D. (2020). Keywords generation improves e-commerce session-based recommendation. In *Proceedings of the web conference 2020* (pp. 1604–1614).

Mercier, M. R., & Cappe, C. (2020). The interplay between multisensory integration and perceptual decision making. *NeuroImage*, *222*, 116970.

Peng, X., Wei, Y., Deng, A., Wang, D., & Hu, D. (2022). Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8238–8247).

Qiao, S., Zhou, W., Wen, J., Zhang, H., & Gao, M. (2023). Bi-channel multiple sparse graph attention networks for session-based recommendation. In *Proceedings of the 32nd acm international conference on information and knowledge management* (pp. 2075–2084).

Seijdel, N., Schoffelen, J.-M., Hagoort, P., & Drijvers, L. (2024). Attention drives visual processing and audiovisual integration during multimodal communication. *Journal of Neuroscience*, *44*(10).

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th acm international conference on information and knowledge management* (pp. 1441–1450).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Wang, W., Tran, D., & Feiszli, M. (2020). What makes training multi-modal classification networks hard? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12695–12705).

Wei, W., Ren, X., Tang, J., Wang, Q., Su, L., Cheng, S., ... Huang, C. (2024). Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th acm international conference on web search and data mining* (pp. 806–815).

Wu, N., Jastrzebski, S., Cho, K., & Geras, K. J. (2022). Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International conference on machine learning* (pp. 24043–24055).

Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., & Tan, T. (2019). Session-based recommendation with graph neural networks. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 346–353).

Xia, X., Yin, H., Yu, J., Shao, Y., & Cui, L. (2021). Self-supervised graph co-training for session-based recommendation. In *Proceedings of the 30th acm international conference on information & knowledge management* (pp. 2180–2190).

Zhang, P., Guo, J., Li, C., Xie, Y., Kim, J. B., Zhang, Y., ... Kim, S. (2023). 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* (pp. 168–176).

Zhang, T., Zhao, P., Liu, Y., Sheng, V. S., Xu, J., Wang, D., . . . others (2019). Feature-level deeper self-attention network for sequential recommendation. In *Ijcai* (pp. 4320–4326).

Zhang, X., Xu, B., Ma, F., Li, C., Yang, L., & Lin, H. (2023). Beyond co-occurrence: Multi-modal session-based recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, X., Xu, B., Ren, Z., Wang, X., Lin, H., & Ma, F. (2024). Disentangling id and modality effects for session-based recommendation. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval* (pp. 1883–1892).

Zhang, X., Xu, B., Yang, L., Li, C., Ma, F., Liu, H., & Lin, H. (2022). Price does matter! modeling price and interest preferences in session-based recommendation. In *Proceedings of the 45th international acm sigir conference on research and development in information retrieval* (pp. 1684–1693).