Computer Science

Capstone Report - Fall 2024

DreamRec: Towards Multi-Modality Multi-Behavior Sequential Recommendation

Jiacheng Shen,
Wenluetao Wu

supervised by
Qiaoyu Tan

Declaration

I declare that this senior capstone was composed entirely by myself with the guidance of my advisor, and that it has not been submitted, in whole or in part, to any other application for a degree. Except where it is acknowledged through reference or citation, the work presented in this capstone is entirely my own.

Preface

As technology develops rapidly, recommendation tasks have become more and more common and important in our everyday life. We see them in e-commerce platforms (e.g. Amazon), social platforms (e.g. Instagram), video platforms (e.g. Netflix), and many other real-life scenarios. However, the ever-growing complexity of user interactions has always been a challenge in recommendation tasks. Therefore, understanding and predicting user preferences through diverse data sources has become crucial in performing better recommendation to the users. Motivated by this, we put our focus on addressing the existing gap in the integration of multi-modality and multi-behavior information in sequential recommendation systems. By proposing a novel model, DreamRec, this project aims to demonstrate the potential of combining diverse modalities and behaviors for enhanced recommendation performance.

This project is designed for an audience that includes researchers and students with backgrounds in machine learning, deep learning, and recommender systems, as well as professionals in industries relying on advanced recommendation technologies. We hope this work provides both theoretical advancements and practical insights for future research in this area.

Acknowledgements

We would like to express our heartfelt thanks to our dear friend Weichen Liu for providing warmth and support during our most vulnerable moments. We are also deeply grateful to Doris Yuxuan Zhang, Helen Zihan Xu, Vivian Chen, Jinting Liu, Jiajin Liu, and Lihan Feng for sharing their time, laughter, and joyful moments with us over the past years.

Special thanks to our bros Liyuan Geng, Yuanhe Guo, Junyi Li, Weilun Wang, Yujun Chen, Haoming Liu, Jiahe Qian, Daokai Liu, and Zhehao Zhang for their unwavering friendship and support throughout these four years. Your presence has made this journey truly unforgettable. BE REAL FOREVER!

We are also incredibly thankful to Howard Lai, Peng Liu, Denny Hu, and Ray Qian for creating lifelong, unforgettable memories with us at NYU Abu Dhabi. From the desert and the Gulf, our friendship will continue to span the rest of the world.

We extend our gratitude to our remarkable research collaborators, Kai Shao, Haneen Alsuradi, and Yi Fang. Your guidance and teamwork have been pivotal in enabling us to submit and publish academic papers—an invaluable experience during our undergraduate studies.

Our deepest thanks go to Professor Mohamad Eid, Professor Mathieu Lauriere, and Professor Qiaoyu Tan for mentoring us in research and fostering our critical thinking. Your support has been instrumental in our academic growth.

Finally, we are forever indebted to our parents for their unwavering love and support throughout these four years. You have been our constant source of strength and encouragement, and we cannot thank you enough.

Abstract

Recommender systems play a crucial role in information retrieval by predicting users' future consumption decisions based on user-item interactions. Sequential Recommendation focuses on capturing user preferences by leveraging sequential patterns among items. While significant advancements have been made by incorporating either multi-modality information or multi-behavior modeling, no existing research has addressed the integration of both multi-modality and multi-behavior information.

To address this gap, we propose the Multi-Modality Multi-Behavior Sequential Recommendation problem and introduce a novel model, DreamRec. For our first goal, DreamRec employs a Q-former-style model, leveraging self-attention within each modality and cross-attention between modalities to capture intra-modality and inter-modality patterns. For our second goal, we use MB-HET as the backbone to propagate item-behavior hypergraphs and model relationships between sequential behavior patterns and items. For our third goal, we design a Cloze task by masking the item with the target behavior, enabling simultaneous learning from both modality and behavior information.

Experimental evaluation on ML-100k, demonstrates that DreamRec outperforms several domain-leading models. Furthermore, our approach offers valuable insights into the integration of multi-modality and multi-behavior fusion, showcasing its potential to advance sequential recommendation tasks. Our code can be found here

Keywords

Sequential Recommendation; Multi-Behavior Sequential Recommendation; Multi-Modal Learning

Contents

1 Introduction

Recommender system(RS) plays a vital role in information retrieval. It predicts user's future consumption decision based on user-item interactions. Sequential Recommendation(SR) [1, 2, 3, 4] is one of the core tasks in the recommender system, which targets capturing user preference over time series, leveraging sequential patterns among items. There have been two important main directions for improvement in Sequential Recommendation. One way is to incorporate multiple modalities into SR, making use of information from various modalities[5, 6, 7], as in tasks like e-commerce, short-video apps, and news recommendation, different modality information including vision and text information is highly involved. The other way is to incorporate multiple behaviors, modeling relationships between multiple types of behaviors and items for recommendation tasks in more complex real-life scenarios where users often exhibit diverse behaviors that reflect varying intents or levels of interest[8, 9]. In light of recent advancements in deep learning, subsequent progress in both directions has been made successively with various new models and architectures proposed, indicating the potential of both multi-modality and multi-behavior methods in boosting the model's performance in sequential recommendation tasks.

Despite this, there has been no existing research focusing on the integration of multi-behavior and multi-modality to the best of our knowledge. Therefore, we propose the multi-modality multi-behavior sequential recommendation problem(see figure 1). To solve this problem, we propose a new model, named **DreamRec**. We aim at

1. capturing intra-modality patterns as well as inter-modality patterns from multi-modality information.

2. capturing relationships between sequential behavior patterns and items.

3. Unifying and integrating the two processes.

For our first goal, we propose a Q-former style model, which applies self-attention within each modality and cross-attention between modalities. For our second goal, we use MB-HET as the backbone to propagate the item-behavior hypergraphs. For our third goal, we design a Cloze task by masking the item with target behavior to learn from both modality and behavior information. To assess the effectiveness of our approach, We set our experiments on ML-100k. Experimental results demonstrate that our approach outperforms 6 baselines across sequential recommendation, multi-modality sequential recommendation, and multi-behavior sequential rec-

Figure 1: demonstrates the motivation of multi-modality multi-behavior sequential recommendation.

ommendation domains. Additionally, our approach provides valuable insights into the integration of multi-modality and multi-behavior fusion.

2 Related Work

2.1 Sequential Recommendation

Sequential recommendation(SR) utilizes temporal information to model user preferences. Previous work [10, 11] models the temporal dependencies with Markov Chain (MC). Markov Chain assumes that the next item a user interacts with only depends on their most recent interactions, typically modeled as a first-order chain. For example, FPMC (Factorizing Personalized Markov Chains) [11] extends the Markov Chain model by incorporating matrix factorization, allowing the model to capture both sequential dynamics and user-item factors in session-based recommendation. Fossil [10] further incorporates Matrix Factorization and the Markov Chain model with a similarity-based method to solve sequential recommendations with a sparse dataset.

With the rise of deep learning, recurrent neural networks (RNNs) became the go-to models for sequential recommendation because of their ability to capture temporal dependencies [1, 12, 13, 14]. GRU4Rec [1] is a pioneering method that applies GRUs to session-based recommendation tasks. LSTM [13] based models have also been widely applied in this context to deal with longer sequences of interactions. Transformer-based models, for example, DMAN [15], SASRec [2], and Bert4Rec [3], are also well developed to model the user-item sequence as a series of interactions through attention mechanisms. Graph Neural Networks (GNNs) have shown great promise in sequential recommendation by modeling user-item interactions as graphs, cap-

turing both local and global dependencies in sequences [16, 17, 18, 19, 20, 21]. Early work like SR-GNN [21] introduced the idea of representing sessions as directed graphs, where GNNs learn session representations by propagating item information. More advanced models, such as MA-GNN [19], incorporated external memory to capture long-term user preferences, while RetagNN [20] introduced relational and temporal attention mechanisms to account for dynamic interactions. Hybrid models like Multi-View GNN-Transformers [18] further enhanced expressiveness by combining GNNs with transformers to capture fine-grained sequential patterns and contextual information.

2.2 Multi-Modality Sequential Recommendation

Traditional recommendation approaches often rely solely on user-item interactions, but these methods can overlook valuable information present in various data modalities, such as text, images, audio, and video. Multi-modality sequential recommendation (MMSR) addresses this limitation by leveraging rich, heterogeneous data sources to capture a more comprehensive understanding of user preferences and item characteristics. Previous works integrate modality information into interaction logs to enable item representation learning beyond id-based through intra and inter-modality learning to enrich item representation. MMMLP[22] is a novel architecture for sequential recommendation that utilizes a Multi-Modal Multi-Layer Perceptron (MLP) with three modules—Feature Mixer Layer, Fusion Mixer Layer, and Prediction Layer—enhancing both efficacy and efficiency in processing multi-modal sequences. MMBR[8] models user history as a graph with intra-modality and inter-modality edges, using dual attention and an update gate to dynamically prioritize early or late fusion. M3SRec [9] is a Multi-Modal Mixture of Experts model designed for sequential recommendation, which leverages rich multi-modal interaction data and employs a novel MoE fusion network along with targeted pre-training tasks to reduce the semantic gap across modalities and enhance the modeling of complex user intents.

2.3 Multi-Behavior Sequential Recommendation

As most of the Sequential Recommendation models focus merely on sequential patterns embedded behind singular behavior type, Multi-behavior sequential recommendation (MBSR) is proposed to overcome SR's drawbacks that it goes against real-life scenarios where people are often engaged with multiple behaviors rather than one, and it may suffer from performance degradation once the sequences get short. MBSR predicts a user's next item of interest based on their interaction

history across different behavior types. It takes into account both sequentiality and homogeneity of user behaviors. Previous works have come up with different paradigms to boost model's performance in mainly two aspects: capturing the relationships between different behavior patterns and different categories of items and item transitions, modeling global dependencies across multiple behaviors. In early works, paradigms without deep learning algorithms were used to tackle the problem, for example, TransRec++ [23] is a matrix factorization-based method, which introduced behavior transition vectors to capture the dynamics of users' heterogeneous behaviors and the items. Later, as deep learning developed, researchers started to incorporate deep learning algorithms to leverage the performance. Models like RLBL [24], IARS [25] and, BINN [26] are all RNN (Recurrent Neural Network)-based architectures for solving MBSR problems. MGNN-SPred [27], GPG4HSR [28], and BA-GNN [29] are examples that utilize Graphic Neural Network (GNN). Transformer-based architectures were also introduced to solve MBSR problems. MBHT [6] incorporates a multi-scale Transformer with low-rank self-attention to better encode behavior-aware sequential patterns, and PBAT [30] redesigns the self-attention layer using a behavior-aware collaboration extractor, which combines behavior and time effects to improve how transitions between items are understood. Some works further develop hybrid-methods-based architectures. Models like MBGNN [31] and MKM-SR [32] combines RNN and GNN, while models like KHGT [33] and TGT [34] combines Transformer and GNN. In some works, contrastive learning algorithm is utilized in order to alleviate the data sparsity problem in MBSR, for instance, MBCCL [35] proposed a multi-behavior collaborative contrastive learning model which effectively integrates user and item attributes with temporal-aware attention to enhance the capture of users' multi-behavior preferences and mitigate data sparsity issues. Other recent works including END4Rec [36] and GHTID [37] also focus on subjects like denoising and efficiency improving, for there often exists noise caused by auxiliary user behaviours and heterogeneous item transitions in the context of MBSR. A recent work [38] which incorporates multi-interest sequential recommendation (MISR) into MBSR , further indicates the possibility and potential of combining MBSR with other sub areas of recommendation system.

3 Methodology

Our proposed DreamRec is demonstrated in figure 2

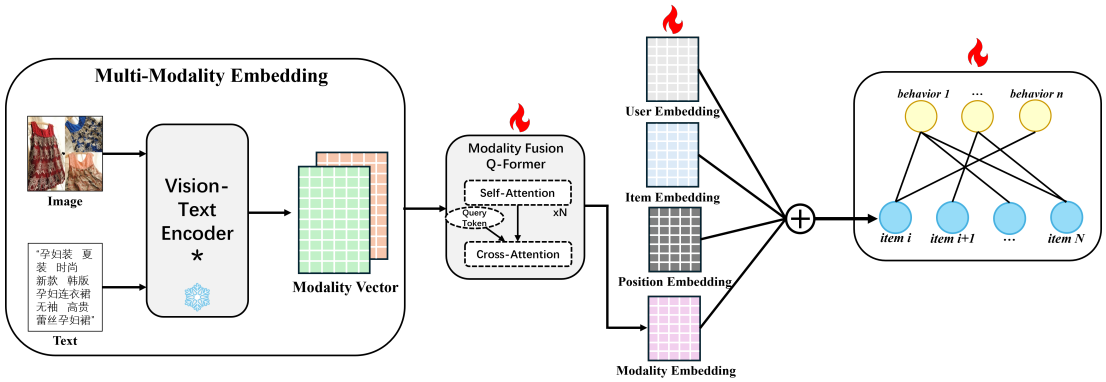

Figure 2: shows the model architecture of DreamRec

3.1 Problem Formulation

Given a user's interaction sequence, the objective is to predict the next item the user is likely to interact with under target behavior by modeling item features, historical behaviors, and modality information through behavior-item hypergraph propagation.

3.2 Input Representation

Let:

- B: Batch size.
- T: Sequence length.
- C: Number of candidate items.
- d: Embedding dimension.
- d_{img}, d_{text}: Dimensions of image and text features, respectively.

We define the following inputs:

$$\text{Item Sequence:} \quad \mathbf{x} = [x_1, x_2, \ldots, x_T] \in \mathbb{R}^{B \times T}, \tag{1}$$

$$\text{Behavior Sequence:} \quad \mathbf{b} = [b_1, b_2, \ldots, b_T] \in \mathbb{R}^{B \times T}, \tag{2}$$

$$\text{User IDs:} \quad \mathbf{u} \in \mathbb{R}^{B}, \tag{3}$$

$$\text{Candidate Set:} \quad \mathbf{c} \in \mathbb{R}^{B \times C}. \tag{4}$$

Here, x_t represents the item ID at position t in the sequence, b_t denotes the behavior type associated with x_t (e.g., click, purchase), and $\mathbf{u}$ contains the user IDs for the batch.

3.3 Embedding Layers

To capture semantic meanings, we convert discrete inputs into continuous embeddings:

$$\mathbf{E}_x = \text{Embed}_x(\mathbf{x}) \in \mathbb{R}^{B \times T \times d}, \tag{5}$$

$$\mathbf{E}_b = \text{Embed}_b(\mathbf{b}) \in \mathbb{R}^{B \times T \times d}, \tag{6}$$

$$\mathbf{E}_u = \text{Embed}_u(\mathbf{u}) \in \mathbb{R}^{B \times d}, \tag{7}$$

$$\mathbf{E}_{\text{pos}} = \text{Embed}_{\text{pos}}(\mathbf{p}) \in \mathbb{R}^{T \times d}. \tag{8}$$

Here, Embed_x, Embed_b, Embed_u, and $\text{Embed}_{\text{pos}}$ are embedding functions for items, behaviors, users, and positions, respectively. The user embedding $\mathbf{E}_u$ is expanded to match the sequence length:

$$\mathbf{E}'_u = \mathbf{E}_u \otimes \mathbf{1}_{1 \times T} \in \mathbb{R}^{B \times T \times d}, \tag{9}$$

where $\otimes$ denotes the outer product, and $\mathbf{1}_{1 \times T}$ is a vector of ones.

3.4 Modality Fusion using Q-Former

To incorporate multi-modal information, we utilize a Q-Former module that fuses image and text features associated with items.

3.4.1 Modality Inputs

For each item in the sequence, we obtain its associated image and text features:

$$\mathbf{F}_{\text{img}} = [\mathbf{f}_{\text{img},1}, \mathbf{f}_{\text{img},2}, \ldots, \mathbf{f}_{\text{img},T}] \in \mathbb{R}^{B \times T \times d_{\text{img}}}, \tag{10}$$

$$\mathbf{F}_{\text{text}} = [\mathbf{f}_{\text{text},1}, \mathbf{f}_{\text{text},2}, \ldots, \mathbf{f}_{\text{text},T}] \in \mathbb{R}^{B \times T \times d_{\text{text}}}. \tag{11}$$

3.4.2 Feature Projection

We project the image and text features into a common embedding space of dimension d:

$$\mathbf{F}'_{\text{img}} = \mathbf{F}_{\text{img}} \mathbf{W}_{\text{img}} + \mathbf{b}_{\text{img}} \in \mathbb{R}^{B \times T \times d}, \tag{12}$$

$$\mathbf{F}'_{\text{text}} = \mathbf{F}_{\text{text}} \mathbf{W}_{\text{text}} + \mathbf{b}_{\text{text}} \in \mathbb{R}^{B \times T \times d}, \tag{13}$$

where $\mathbf{W}_{\text{img}} \in \mathbb{R}^{d_{\text{img}} \times d}$ and $\mathbf{W}_{\text{text}} \in \mathbb{R}^{d_{\text{text}} \times d}$ are learnable projection matrices, and $\mathbf{b}_{\text{img}}$, $\mathbf{b}_{\text{text}}$ are bias terms.

3.4.3 Self-Attention within Each Modality

To capture intra-modality dependencies, we apply self-attention mechanisms:

$$\mathbf{A}_{\text{img}} = \text{SelfAttn}(\mathbf{F}'_{\text{img}}) \in \mathbb{R}^{B \times T \times d}, \tag{14}$$

$$\mathbf{A}_{\text{text}} = \text{SelfAttn}(\mathbf{F}'_{\text{text}}) \in \mathbb{R}^{B \times T \times d}. \tag{15}$$

The self-attention function computes attention weights to model dependencies among elements within the same modality.

3.4.4 Cross-Attention between Modalities

We fuse the modalities using cross-attention:

$$\mathbf{F}_{\text{cat}} = [\mathbf{A}_{\text{img}}; \mathbf{A}_{\text{text}}] \in \mathbb{R}^{B \times T \times 2d}, \tag{16}$$

$$\mathbf{Q} = \mathbf{W}_q \mathbf{q} + \mathbf{b}_q \in \mathbb{R}^{1 \times d_q}, \tag{17}$$

$$\mathbf{K} = \mathbf{F}_{\text{cat}} \mathbf{W}_k + \mathbf{b}_k \in \mathbb{R}^{B \times T \times d_k}, \tag{18}$$

$$\mathbf{V} = \mathbf{F}_{\text{cat}} \mathbf{W}_v + \mathbf{b}_v \in \mathbb{R}^{B \times T \times d_v}, \tag{19}$$

$$\mathbf{A}_{\text{cross}} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \in \mathbb{R}^{B \times 1 \times d_v}. \tag{20}$$

Here, $\mathbf{q} \in \mathbb{R}^{1 \times d_q}$ is a learnable query vector, $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are projection matrices for queries, keys, and values, respectively, and $d_k = d_v = d$. The cross-attention mechanism allows the model to attend to relevant features across modalities.

3.5 Combined Sequence Representation and Projection Layer

We combine all embeddings to form the final sequence representation:

$$\mathbf{S} = \mathbf{E}_x + \mathbf{E}'_u + \mathbf{E}_{\text{pos}} + \mathbf{A}_{\text{cross}} \in \mathbb{R}^{B \times T \times d}. \tag{21}$$

To capture higher-level features, we apply a projection layer:

$$\mathbf{H} = \text{ReLU}(\mathbf{S}\mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}}) \in \mathbb{R}^{B\times T\times d}, \tag{22}$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d\times d}$ is a weight matrix, and $\mathbf{b}_{\text{proj}}$ is a bias term.

3.6 Item-Behavior Hypergraph Learning

3.6.1 Hypergraph Construction

We construct a hypergraph G to model the complex relationships between items and behaviors within a user's interaction sequence. The hypergraph consists of:

- **Nodes**: $\mathcal{V} = \{\mathbf{v}_i^{\text{item}}, \mathbf{v}_j^{\text{beh}}, \}$, where $\mathbf{v}_i^{\text{item}}$ represents item nodes and $\mathbf{v}_j^{\text{beh}}$ represents behavior nodes.
- **Edges**: $\mathcal{E} = \{(\mathbf{v}_i^{\text{item}}, \mathbf{v}_j^{\text{beh}}), (\mathbf{v}_j^{\text{beh}}, \mathbf{v}_i^{\text{item}})\}$.

The hypergraph captures pairwise interactions between items and behaviors to facilitate information flow.

3.6.2 Item-behavior Hypergraph Propagation

We utilize MB-HET[6] as a backbone to propagate information through the hypergraph. The initial node features are:

$$\mathbf{H}^{(0)} = [\mathbf{h}_1^{\text{item}}, \mathbf{h}_2^{\text{item}}, \ldots, \mathbf{h}_M^{\text{item}}, \mathbf{h}_1^{\text{beh}}, \mathbf{h}_2^{\text{beh}}, \ldots, \mathbf{h}_N^{\text{beh}},] \in \mathbb{R}^{(M+N+1)\times d}, \tag{23}$$

where M and N are the numbers of item and behavior nodes, respectively.

The propagator updates node representations through L layers:

$$\mathbf{H}^{(l+1)} = f(\mathbf{H}^{(l)}, \mathbf{A}), \quad l = 0, 1, \ldots, L-1, \tag{24}$$

where $\mathbf{A}$ is the adjacency matrix of the hypergraph.

3.7 Prediction and Loss Computation

3.7.1 Training Phase

During training, we perform the following steps:

1. **Edge Masking**: We apply the Cloze task to train our model, which masks an item where the item has edges with the target behavior.

2. **Item-behavior Hypergraph Propagation**: Update node embeddings using the propagator to obtain $\mathbf{H} = \mathbf{H}^{(L)}$.

3. **Logits Computation**: Compute logits for predicting the target behavior:

$$\mathbf{z} = \mathbf{E}_{\text{beh}}\mathbf{h}, \quad \mathbf{z} \in \mathbb{R}^K. \tag{25}$$

   h is the embedding of the masking item by averaging the embedding of its neighbors.

4. **Loss Function**: Apply the cross-entropy loss:

$$\mathcal{L} = -\sum_{k=1}^{K} y_k \log \sigma(z_k), \tag{26}$$

   where $y_k = 1$ if $k = b_{\text{target}}$ and 0 otherwise, and $\sigma(\cdot)$ denotes the softmax function over the logits.

3.7.2 Inference Phase

During inference, we predict the next item as follows:

1. **Candidate Embeddings**: Obtain embeddings for candidate items:

$$\mathbf{E}_{\text{cand}} = \text{Embed}_x(\mathbf{c}) \in \mathbb{R}^{B \times C \times d}. \tag{27}$$

2. **Score Computation**: Append a dream token at the end of sequence together with the previous sequence to f. Compute scores between the dream token and each candidate item:

$$s_{in} = \mathbf{h}_i^{\text{dream token}^\top} \mathbf{e}_{\text{cand},in}, \quad n = 1, 2, \ldots, C, \tag{28}$$

   where $\mathbf{e}_{\text{cand},in}$ is the embedding of the k-th candidate item for user i.

3. **Recommendation**: Rank the candidate items based on the scores s_{in} and recommend the top items.

4 Results

4.1 Experimentation protocol

4.1.1 Datasets

We evaluate our proposed DreamRec with the following public dataset:

ML-100k. The movie lens dataset is a user rating dataset about movie list.We differentiate the rating data into three behaviors which are like, neutral, and dislike according to the rating following [30, 7]. We treat like as the target behavior in this dataset.

We filter the dataset by 5-core filtering strategy to exclude users and items appearing less than 5 times.

Table 1: Dataset statistics after filtering

Datasets	**MovieLens 100k**
#Users	943
#Items	1326
#Interactions	97844
#Average Length	103.758
#Behavior Types	{Like neutral dislike}
#Modality	Text Image

4.1.2 Modality Extraction

MovieLens 100k dataset: The visual modality in ML-100k dataset is the poster of each film and the textual modality which we scrape from the IMDB website. We utilize Blip-2 [39] to extract the images and text modality in ML-100k dataset which guarantees the alignment of latent representations.

4.1.3 Evaluation Protocols

Similarly to the setting in [3, 2, 30], we utilize the leave-one-out evaluation strategy that considers the last two interactions as validation and test data and the data before as the training data. We pair each positive sample with 100 negative instances based on item popularity

[3]. We utilize three evaluation metrics: *Hit Ratio (HR@N)*, *Normalized Discounted Cumulative Gain(NDCG@N)* and *Mean Reciprocal Rank(MRR@N)*

4.1.4 Baselines

We consider the folowing baselins as comparing with our multi-behavior **sequential recommendation** problem.

(a) **SASRec**[2] SASRec employs a Transformer architecture with self-attention layers to model the sequence of user-item interactions. It focuses on capturing relevant user behaviors without the need for recurrent or convolutional layers, making it effective for long sequences.

(b) **Bert4Rec**[3] BERT4Rec is a sequential recommendation model that uses bidirectional Transformers, inspired by BERT in NLP. It learns user behavior by modeling sequences in both directions and predicting masked items within a sequence, capturing complex dependencies in user interactions.

(c) **Caser**[4] Caser leverages convolutional neural networks to capture both short-term and long-term user preferences by modeling the sequence of users' past interactions, enabling it to predict the next item in a sequence. Caser stands out for its ability to learn personalized sequential patterns while being flexible enough to handle diverse recommendation scenarios.

(d) **GRU4Rec**[1] GRU4Rec uses Gated Recurrent Units to model sequential user interactions. As an RNN-based model, it processes user behavior over time step-by-step, handling sequential dependencies, and is one of the first neural network approaches for session-based recommendation.

(e) **MB-STR**[6] MB-STR integrates hypergraph structures into a transformer framework to capture complex relationships between different user behaviors in sequential recommendation tasks. It aims to enhance recommendation accuracy by leveraging multi-behavior interactions in a hypergraph setting.

(f) **MMSR**[8] MMSR combines and adapts multiple modalities of data to improve recommendations over time. The model dynamically integrates these diverse data sources to better capture user preferences and provide more accurate sequential recommendations.

Table 2: ml-100k results of baselines and DreamRec

Model	NDCG@5	NDCG@10	NDCG@30	Hit@5	Hit@10	Hit@30	MRR@5	MRR@10	MRR@ 30
Caser	0.2212	0.2782	0.3476	0.3401	0.5160	0.8049	0.1822	0.2057	0.2231
Bert4Rec	0.2713	0.3231	0.3866	0.4179	0.5768	0.8422	0.2234	0.2450	0.2608
GRU4Rec	0.2594	0.3188	0.3813	0.4009	0.5842	0.8433	0.2131	0.2377	0.2535
SASRec	0.2790	0.3351	0.3954	0.4243	0.5959	0.8454	0.2314	0.2548	0.2701
MB-STR	0.2988	0.3453	0.4072	0.4291	0.5740	0.8315	0.2558	0.2748	0.2904
MMSR	0.2546	0.3028	0.3642	0.3739	0.5257	0.8035	0.2161	0.2363	0.2513
DreamRec*	**0.3025**	**0.3532**	**0.4123**	**0.4918**	**0.6210**	**0.8568**	**0.2623**	**0.2835**	**0.3002**

4.2 Experiment result

The experimental results for DreamRec are presented in Table 2. As observed, DreamRec outperforms traditional sequential recommendation models and multi-modality sequential recommendation systems by a substantial margin. Additionally, DreamRec surpasses the Multi-Behavior Sequential Recommendation model, MB-STR.

Notably, DreamRec demonstrates a significant improvement over other sequential recommendation models in terms of Hit Rate, which suggests that it effectively enhances the exposure of items aligned with the target behavior. However, when evaluated using NDCG and MRR, DreamRec shows a smaller gap in performance, indicating that while the model excels in item exposure, there is still room for improvement in terms of ranking quality. This suggests that future work may focus on enhancing DreamRec's ability to refine the ranking of recommended items, particularly by addressing the temporal relationships between behaviors and modalities.

5 Discussion

5.1 Research Question 1: Does MMMB-SR improve user satisfaction?

In this project, we introduce DreamRec, a novel recommender system designed to address the multi-modality multi-behavior sequential recommendation problem. Compared to existing multi-behavior sequential recommendation systems, our model incorporates a Q-former-style modality fusion module that enhances item representation, making the recommender system modality-aware. As demonstrated in the experimental results, DreamRec outperforms the most recent models in multi-modality sequential recommendation.

Furthermore, unlike traditional models that treat user behavior uniformly, our model adopts a more sophisticated approach by considering multi-behavior scenarios. We achieve this through

the design of an item-behavior hypergraph propagation component, which leads to significant performance improvements over traditional sequential and multi-modal sequential recommendation systems. These traditional systems tend to treat user behavior homogeneously, which does not fully capture the diversity of user interactions.

5.2 Research Question 2: How to effectively model the correlation between modality and behavior in temporal order?

Currently, our model utilizes a Behavior-Item Hypergraph Propagation backbone, which isolates the temporal learning processes associated with modality sequences and behavior sequences. While this architecture allows for the independent learning of each sequence, it might not fully capture the interdependencies between modalities and behaviors in the temporal domain, which could explain the absence of dramatic improvements in ranking quality, as discussed in the results section.

To overcome this limitation, one potential direction for future work is to introduce an additional learning stage that jointly considers raw modality and behavior sequences through an attention mechanism. This would allow the model to better capture the correlation between modality and behavior sequences in temporal order. Furthermore, we could explore alternative fusion strategies—such as delaying fusion until later stages of the model—which may help preserve the temporal patterns within the modality sequence when learning user behavior patterns in the item-behavior hypergraph.

6 Personal Contributions

Jiacheng Shen Enriched the MovieLens 100K dataset with modality information, implemented six baseline models using the enriched dataset, proposed the DreamRec model architecture, and conducted experiments to evaluate its performance.

Wenluetao Wu Processed the Tmall dataset for modality information, contributing to the enrichment of the datasets for modality-aware recommendation tasks.

7 Conclusion

This project introduces a new problem in sequential recommendation: multi-modality multi-behavior sequential recommendation. We designed the Q-former-style modality fusion module

to perform effective modality fusion and constructed an item-behavior propagation graph to model diverse user behaviors using a graph-based model backbone. Our DreamRec model shows significant improvement over six baseline models.

While our current approach demonstrates strong performance, there are still avenues for enhancing this work. Specifically, adding a second-stage learning process that captures the sequential dependencies between behavior and modality sequences could further improve the ranking quality of DreamRec. This could contribute to a more nuanced understanding of the temporal dynamics between different user behaviors and modalities, ultimately increasing the model's effectiveness in real-world applications.

References

[1] D. Jannach and M. Ludewig, "When recurrent neural networks meet the neighborhood for session-based recommendation," in *Proceedings of the eleventh ACM conference on recommender systems*, 2017, pp. 306–310.

[2] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 197–206.

[3] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.

[4] J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 565–573.

[5] L. Xia, C. Huang, Y. Xu, and J. Pei, "Multi-behavior sequential recommendation with temporal graph transformer," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 6099–6112, 2022.

[6] Y. Yang, C. Huang, L. Xia, Y. Liang, Y. Yu, and C. Li, "Multi-behavior hypergraph-enhanced transformer for sequential recommendation," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 2263–2274.

[7] E. Yuan, W. Guo, Z. He, H. Guo, C. Liu, and R. Tang, "Multi-behavior sequential transformer recommender," in *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 1642–1652.

[8] H. Hu, W. Guo, Y. Liu, and M.-Y. Kan, "Adaptive multi-modalities fusion in sequential recommendation systems," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 843–853.

[9] S. Bian, X. Pan, W. X. Zhao, J. Wang, C. Wang, and J.-R. Wen, "Multi-modal mixture of experts represetation learning for sequential recommendation," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 110–119.

[10] R. He and J. McAuley, "Fusing similarity models with markov chains for sparse sequential recommendation," in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 191–200.

[11] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 811–820.

[12] Y. Sun, P. Zhao, and H. Zhang, "Ta4rec: Recurrent neural networks with time attention factors for session-based recommendations," in *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[13] C. Zhao, J. You, X. Wen, and X. Li, "Deep bi-lstm networks for sequential recommendation," *Entropy*, vol. 22, no. 8, p. 870, 2020.

[14] B. Hidasi and A. Karatzoglou, "Recurrent neural networks with top-k gains for session-based recommendations," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 843–852.

[15] Q. Tan, J. Zhang, N. Liu, X. Huang, H. Yang, J. Zhou, and X. Hu, "Dynamic memory based attention network for sequential recommendation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 5, 2021, pp. 4384–4392.

[16] B. Wang and W. Cai, "Knowledge-enhanced graph neural networks for sequential recommendation," *Information*, vol. 11, no. 8, p. 388, 2020.

[17] Y. Liu, L. Xia, and C. Huang, "Selfgnn: Self-supervised graph neural networks for sequential recommendation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 1609–1618.

[18] T. Luo, Y. Liu, and S. J. Pan, "Collaborative sequential recommendations via multi-view gnn-transformers," *ACM Transactions on Information Systems*, 2024.

[19] C. Ma, L. Ma, Y. Zhang, J. Sun, X. Liu, and M. Coates, "Memory augmented graph neural networks for sequential recommendation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5045–5052.

[20] C. Hsu and C.-T. Li, "Retagnn: Relational temporal attentive graph neural networks for holistic sequential recommendation," in *Proceedings of the web conference 2021*, 2021, pp. 2968–2979.

[21] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 346–353.

[22] J. Liang, X. Zhao, M. Li, Z. Zhang, W. Wang, H. Liu, and Z. Liu, "Mmmlp: Multi-modal multilayer perceptron for sequential recommendations," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 1109–1117.

[23] Z. Zhan, M. He, W. Pan, and Z. Ming, "Transrec++: Translation-based sequential recommendation with heterogeneous feedback." *Frontiers Comput. Sci.*, vol. 16, no. 2, p. 162615, 2022.

[24] Q. Liu, S. Wu, and L. Wang, "Multi-behavioral sequential prediction with recurrent log-bilinear model," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 6, pp. 1254–1267, 2017.

[25] Y. Xu, Y. Zhu, and J. Yu, "Modeling multiple coexisting category-level intentions for next item recommendation," *ACM Transactions on Information Systems (TOIS)*, vol. 39, no. 3, pp. 1–24, 2021.

[26] Z. Li, H. Zhao, Q. Liu, Z. Huang, T. Mei, and E. Chen, "Learning from history and present: Next-item recommendation via discriminatively exploiting user behaviors," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1734–1743.

[27] W. Wang, W. Zhang, S. Liu, Q. Liu, B. Zhang, L. Lin, and H. Zha, "Beyond clicks: Modeling multi-relational item graph for session-based target behavior prediction," in *Proceedings of the web conference 2020*, 2020, pp. 3056–3062.

[28] W. Chen, M. He, Y. Ni, W. Pan, L. Chen, and Z. Ming, "Global and personalized graphs for heterogeneous sequential recommendation by learning behavior transitions and user intentions," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 268–277.

[29] Y. Liang, Q. Song, Z. Zhao, H. Zhou, and M. Gong, "Ba-gnn: Behavior-aware graph neural network for session-based recommendation," *Frontiers of Computer Science*, vol. 17, no. 6, p. 176613, 2023.

[30] J. Su, C. Chen, Z. Lin, X. Li, W. Liu, and X. Zheng, "Personalized behavior-aware transformer for multi-behavior sequential recommendation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6321–6331.

[31] W. Pan and K. Yang, "Multi-behavior graph neural networks for session-based recommendation," in *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. IEEE, 2021, pp. 756–761.

[32] W. Meng, D. Yang, and Y. Xiao, "Incorporating user micro-behaviors and item knowledge into multi-task learning for session-based recommendation," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 1091–1100.

[33] L. Xia, C. Huang, Y. Xu, P. Dai, X. Zhang, H. Yang, J. Pei, and L. Bo, "Knowledge-enhanced hierarchical graph transformer network for multi-behavior recommendation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 5, 2021, pp. 4486–4493.

[34] L. Xia, C. Huang, Y. Xu, and J. Pei, "Multi-behavior sequential recommendation with temporal graph transformer," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 6099–6112, 2022.

[35] Y. Chen, Q. Cao, X. Huang, and S. Zou, "Multi-behavior collaborative contrastive learning for sequential recommendation," *Complex & Intelligent Systems*, pp. 1–16, 2024.

[36] Y. Han, H. Wang, K. Wang, L. Wu, Z. Li, W. Guo, Y. Liu, D. Lian, and E. Chen, "End4rec: Efficient noise-decoupling for multi-behavior sequential recommendation," *arXiv preprint arXiv:2403.17603*, 2024.

[37] X. Li, H. Chen, J. Yu, M. Zhao, T. Xu, W. Zhang, and M. Yu, "Global heterogeneous graph and target interest denoising for multi-behavior sequential recommendation," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 387–395.

[38] B. Wu, Y. Cheng, H. Yuan, and Q. Ma, "When multi-behavior meets multi-interest: Multi-behavior sequential recommendation with multi-interest self-supervised learning," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024, pp. 845–858.

[39] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.