

# Video-Image-Sentence Multi-Modality Sequential Recommendation Model

Guowei Wang<sup>1</sup>[0009-0003-6936-948X], Yicheng Di<sup>2</sup>[0000-0003-3802-2080], and Yuan Liu<sup>3</sup> [0000-0002-2576-1426](✉)

<sup>1</sup> School of Artificial Intelligence and Computer Science, Jiangnan University, No. 1800 Lihu Avenue, Wuxi, 214122, Jiangsu Province, China.

lyuan1800di@jiangnan.edu.cn

**Abstract.** At present, recommendation systems have become an essential tool for users to retrieve information in the Internet of Things (IoT) scenario. Conventional sequential recommendation techniques frequently depend on explicit item identifiers, leading to limitations in data sparsity and domain transfer. Recent research has utilized item modal features as inputs to the model, enabling the transfer of knowledge learned between different modal datasets, thus addressing the issue of data scarcity. In order to achieve this objective, we introduce a pre-training method for modeling multiple modalities, which can effectively integrate information from different modalities. We also propose a new loss calculation to measure the performance of this method. Finally, in order to enhance the model's retrieval performance, we provide a novel sequential recommendation strategy. This strategy utilizes a sequence encoder to record the sequences of user interactions, and uses item encoder to encode information about the items. These two encoders share parameters in order to improve the quality of the encoded information. We evaluate our proposed methods on three public datasets and conduct experiments, the results of which demonstrate an improvement in performance.

**Keywords:** Sequential Recommendation, Multi-Modal, Multimodal Pretraining.

## 1 INTRODUCTION

With the rapid development of the Internet and digital technologies, individuals are increasingly overwhelmed with information and choices. Personalized recommendation systems have become essential tools for helping users quickly find content of interest amid this information overload. Collaborative filtering algorithms are commonly used in traditional recommendation systems. These algorithms propose items to a specific user by analyzing the preferences of other users who have similar interests. However, the lack of user reviews for many items leads to data sparsity, which in turn reduces the accuracy of models based on collaborative filtering algorithms [1].

To address this issue of data sparsity, Berkovsky proposed cross-domain recommendation algorithms that leverage abundant information in a source domain (e.g., observed rating data) to enhance recommendation accuracy in a target domain. Furthermore, with increasing computational power, Multi-Target Collaborative Domain Regression (CDR) [2] has emerged, aiming to improve accuracy further. However, cross-domain recommendations are not without limitations, such as the continuous high computational demands and the added complexity and cost from data collection, transformation, and processing. These limitations are particularly acute when it comes to personalized recommendations that require dynamic user behavior modeling.

Another approach to solving the data sparsity issue is the use of multi-modal data. In the multimedia era, users often have abundant information in a particular domain, presented in various modalities like images and text [3]. Multi-modal recommendation systems [3] can utilize this rich information to enhance algorithmic performance. However, the introduction of multiple sources of information, such as users' social networks, textual descriptions, and images or videos, can also lead to increased noise compared to single-modal data.

The fusion of modal algorithms and sequential recommendation methods [5] can be applied to scenarios requiring both user behavior sequences and multi-modal information. For instance, in video recommendation, combining users' viewing history with the videos' textual descriptions and cover images can offer a more comprehensive understanding of user preferences, thereby providing more accurate recommendations. Although the integration of multi-modal and sequential algorithms shows promise, existing methods for transferring item-related patterns across domains still face significant challenges [6]. One such challenge is the inconsistency at the feature level between model outputs and item embeddings. Directly calculating their similarities would force the alignment of inputs and outputs [7], affecting the effectiveness of model outputs.

In order to tackle these difficulties, we provide a groundbreaking multimodal sequential recommendation model that is founded on the principles of sequential recommendation and the integration of multimodal features. This model enhances the fine-grained discrimination capability of items by conducting contrastive learning between different modalities, aligning features in the latent space across these modalities. The paper's contributions can be succinctly described as follows:

- We present a pre-training technique for modeling multi-modal data that efficiently combines information from many modalities acquired in the Internet of Things (IoT) context.
- We provide a new way for calculating loss to assess the performance of this approach. This method effectively deals with the bias resulting from variations in data modalities between pre-training and real-world application.
- The model we propose is a cross-modal sequential recommendation model that utilizes dual retrieval, which integrates the ideas of sequential recommendation and multi-modal features. This model can encode items while extracting user interaction sequences, addressing the issue of inconsistency between the model's outputs and the features of the items at the feature level.

- • In order to verify the efficiency of our approach, we do experiments on three actual datasets. The results show that FMMS-Rec outperforms the latest benchmark models based on RNN, CNN, and GNN architectures.

## **2 RELATED WORK**

### **2.1 Multimodal Learning**

The area of multimodal learning has become increasingly important in recent decades. Initially, multimodal techniques were mainly used for voice recognition [8]. Nevertheless, as the internet has become more widespread and advanced gadgets have been developed, a wider range of multimodal data types has been introduced. When applied to recommendation tasks, these diverse data types offer the prospect of enriching the feature sets of recommendation systems, thus alleviating the issue of domain-specific data sparsity in large-scale datasets. Specifically, video datasets are frequently employed in the context of multimodal learning, as they amalgamate multiple data modalities, including images, text, and audio [9]. A critical challenge in this domain lies in the encoding of these various modalities, their projection into a common representational space, and the alignment of features across different modalities.

### **2.2 Sequential Recommendation**

In the early stages of sequential recommendation tasks, Markov Chains (MC) [10] were commonly employed to observe short-term user behaviors. These chains use generated state transition matrices to predict user actions at future time points. Researchers like Rendle and colleagues [11] proposed a hybrid approach combining first-order Markov Chains with matrix factorization, and further explored higher-order Markov Chains [12]. These higher-order variants take into account a greater number of previous items and extract more intricate patterns. However, such methods exhibit considerable limitations in their ability to handle contextual information.

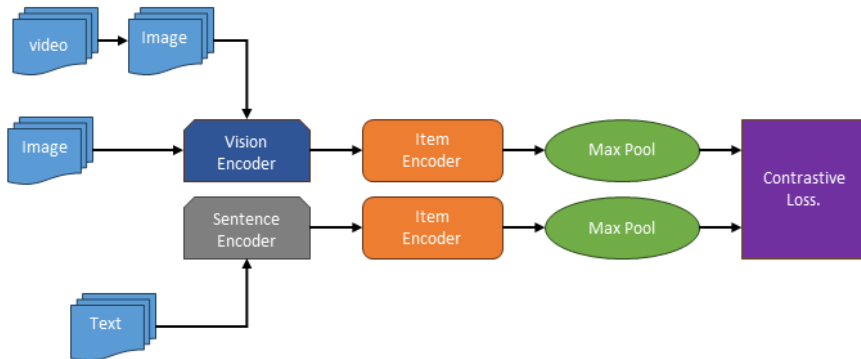
To overcome these shortcomings, Recurrent Neural Networks (RNNs) [13] are an example of deep learning models. were introduced for capturing and processing temporal sequences of user interactions. Variants of RNNs like Long Short-Term Memory (LSTM) [14] networks and Gated Recurrent Units (GRU) [15] are employed to deal with sequences of medium and long lengths. The advent of Transformers, such as BERT (Bidirectional Encoder Representations from Transformers) [16], further elevates the capabilities of recommendation systems by learning richer semantic representations of text, thereby improving the understanding of user interests. Despite the significant improvements these sequential recommendation models have brought about in terms of recommendation efficacy, they predominantly rely on contextual information and largely neglect the relationships between items and users.

### 3 METHOD

#### 3.1 Pretraining Objectives

In this section, we propose a pre-training method for the Video-Image-Sentence Multi-Modality Sequential Recommendation model, which effectively explores the correlations between different modalities of data, thereby learning aligned representations of data from multiple modalities. The input of this method consists of two single-modality encoders, which enhance the model's ability to differentiate items at a fine granularity through contrastive learning between modalities. This alignment of different modal features in the latent space is illustrated in Figure 1. In this experiment, we use visual and textual data as examples.

**3.1.1 Feature Embedding.** In early recommendation algorithms [19], the ability of the models to transfer general knowledge across different domains was limited. To address this issue, we propose the use of multimodal information to connect various domains. We employ pretrained modality models [20,21, 22] to extract modality features



**Fig. 1.** The overview of our Video-Image-Sentence Multi-Modality Sequential Recommendation model. Different modal data undergo feature extraction through different encoders. Specifically, when the item is a video, it is first transformed into a collection of image frames to obtain the representative set of the video.

from items across different domains and map them into a shared semantic space, overcoming the previous models' dependency on explicit item IDs.

Firstly, for data in the visual modality, the input data can be broadly categorized into two types: images and videos. For image processing, pre-trained visual models (PVM) [23] can be utilized for feature extraction. Videos are composed of a sequence of continuous image frames; we can adopt methods used for image processing for video feature extraction as well. We initially sample at a fixed frequency from the image frames

to get a representative set for the video, denoted as  $I_j = i_j^1, i_j^2, \dots, i_j^n$ . Subsequently, we employ PVM to extract features from this set and take their mean to obtain the video's feature representation. Finally, the sequence representation for the visual data,  $E_j^I$ , is computed through a single-layer neural network:

$$E_j^I = f_{NN} \left( \frac{1}{|n|} \sum_{k=1}^{|n|} f_{PVM}(I_j^k) \right) \quad (1)$$

Where  $f_{NN}$  refers to the neural network function, while  $f_{PVM}$  denotes the feature extraction functional.

Secondly, for data in the textual modality, pre-trained language models (PLM) can be employed for feature extraction. For item  $j$ , its text token sequence can be represented as  $S_j = s_j^1, s_j^2, \dots, s_j^n$ . A special token [CLS] [24] is prepended to the token sequence, and the resulting concatenated sequence is fed into the PLM. Ultimately, the token located at the [CLS] position in the hidden layer serves as the textual representation, and this representation is then inputted into a single-layer neural network to generate the sequence representation for the text data,  $E_j^S$ :

$$E_j^S = f_{NN} \left( f_{PLM}([CLS], s_j^1, s_j^2, \dots, s_j^n) \right) \quad (2)$$

Where  $f_{PLM}$  denotes the feature extraction function.

After feature extraction, the input item data  $j \in J$  is transformed into  $(E_j^I, E_j^S)$ . Since this input contains two modalities, we enhance its discriminative power by employing both feature extraction and modality embedding methods, and summing both as the input to the model.

**3.1.2 Multimodal Alignment and Fusion.** In each item, the input data may exist in single-modal form or in a combination of two or even three modalities, such as visual and textual forms, denoted as (p, c). Our goal is to design a method that allows items to be input and mapped to a common space regardless of their modality combination, and to use a specific function to ensure that data from different modalities have similar representations, thus achieving cross-modal fusion of heterogeneous data.

Therefore, inspired by the advantages of the transformer in modeling different modalities (e.g., visual, language, and audio) in various multimodal tasks, we designed a multimodal encoder based on transformer modules, which includes multi-head self-attention layers and feedforward layers. Prior to inputting the feature embeddings of items into the model, we incorporate modality embeddings into them. When specific modalities are missing, we substitute the corresponding feature embeddings with a [mask] token. Thanks to the transformer's attention mechanism, it can handle the correlation between different modalities well, thus enabling information sharing and filling in missing modality information. Subsequently, in order to facilitate the model's understanding of the connections between various modules, we project the encoded feature embeddings into a shared hidden space and normalize them. Finally, by the utilization of a max-pooling layer, we are able to preserve crucial feature information from all

modalities. This results in a consolidated representation of the item, thus mitigating the model's tendency to excessively fit the training set:

$$h_i = \max(h_j^i, h_j^s) \quad (3)$$

Here,  $h_j^i, h_j^s$  denote a set containing two values,  $h_j^p$  and  $h_j^c$ . The MaxPooling operation selects the maximum value from this set as the output.

**3.1.3 Contrastive Loss.** VAST establishes multimodal video-caption correspondence during pre-training. However, it is crucial to address the issue of missing modalities in downstream benchmarks and real-world applications, as inconsistencies between modalities used in pre-training and adaptation may have a negative impact. The total loss  $\mathcal{L}$  consists of several components:

$$\mathcal{L} = \mathcal{L}_{MM} + \mathcal{L}_{I-S} + \mathcal{L}_I + \mathcal{L}_S \quad (3)$$

Where  $\mathcal{L}_I, \mathcal{L}_S$  are inspired by the modality grouping strategy proposed by VALOR, designed for modeling visual and textual information, respectively. The calculation of  $\mathcal{L}_{MM}$  is as follows:

$$\mathcal{L}_{MM} = \mathcal{L}_{PWC} + \mathcal{L}_{PWM} \quad (3)$$

$\mathcal{L}_{PWC}$ : Contrastive loss is used to regularize the feature distance between visual and textual modalities. The contrastive loss is defined as follows:

$$\mathcal{L}_{PWC} = -\frac{1}{2} \sum_j^B -\log \frac{\exp(\tau \cdot \text{sim}(i_j, s_j))}{\sum_{j=1}^B \exp(\tau \cdot \text{sim}(i_j, s_j))} - \frac{1}{2} \sum_i^B -\log \frac{\exp(\tau \cdot \text{sim}(i_i, s_i))}{\sum_{i=1}^B \exp(\tau \cdot \text{sim}(i_i, s_i))} \quad (4)$$

where  $\text{sim}(\cdot)$  denotes the dot product, B and  $\tau$  are the batch size and a learnable parameter, respectively.

$\mathcal{L}_{PWM}$ : This loss is used to infer whether the visual and textual modalities match. If OMV and OMC match,  $y = 1$ ; otherwise,  $y = 0$ . The loss function is formulated as:

$$\mathcal{L}_{PWM} = \mathbb{E}_{(I_i, S_i) \sim (I, W)} [y \log p_{ISM} + (1 - y) \log (1 - p_{ISM})] \quad (5)$$

Our model is able to represent objects in a universal way by modeling different types of data in a unified manner. This allows us to overcome the restrictions of the quantity of pre-training datasets in the recommendation sector. Thus, our model demonstrates strong generalization skills by efficiently encoding inputs with various combinations of modalities and aligning items with comparable semantics in a shared undetectable layer space.

## 3.2 Sequential Representation Learning

Publicly available large-scale pre-training datasets are scarce in the recommendation sector due to the sensitivity of user interactions and tight privacy regulations. Additionally, previous sequence-based recommendation methods relied on single-modal

datasets for pretraining, resulting in limited generalization and performance degradation when transferred across domains. Therefore, to address these issues, we propose a model that can extract effective information from multiple modal datasets to overcome the limitation of limited datasets. Our objective is to prevent the inclusion of domain-specific biases in the model, considering the intricate nature of user behavior patterns across various recommendation domains. Therefore, our objective is to get precise knowledge about sequence interaction patterns pertaining to each area.

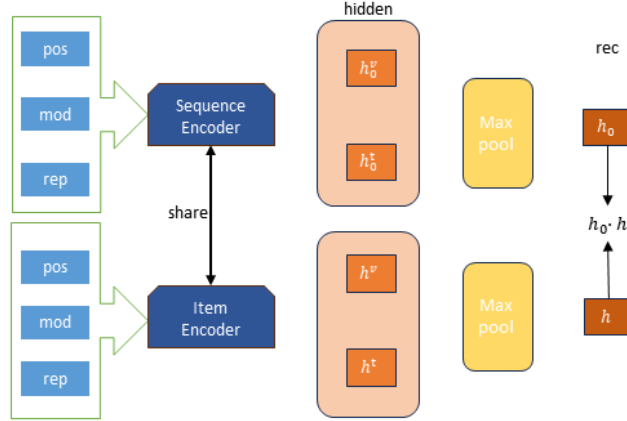


Fig. 2. The sequential recommendation system employs a dual-tower retrieval design.

**3.2.1 Input Representation.** During the pretraining retrieve task, the model does not consider the relative positional relationships between input items. However, in downstream sequential recommendation tasks, the model must utilize the user's prior interactions sequence given input. The components are organized based on their interaction times, resulting in clear positional correlations. In order to represent the temporal order connection, we incorporate positional embeddings as an extra element of the model input. The input embedding of a particular item  $i$  is obtained by adding together the feature embedding, modality embedding, and positional embedding.

In this study, we enhance the model's performance by replacing fixed sinusoidal embeddings with a learnable position-al embeddings matrix. The positional embeddings matrix allows our framework to capture the contextual connections and interaction order between each item in the input sequence, thus enhancing the representation of the user sequence. Furthermore, when the length of the input sequence  $O = \{o_1, o_2, \dots, o_{|L|}\}$  exceeds the maximum length  $N$  that the model can accommodate, we truncate the sequence and keep only the last  $N$  items, denoted as  $O_{retain} = \{o_{|L|-N+1}, \dots, o_{|L|}\}$ .

**3.2.2 Multimodal Sequence Encoder.** The primary goals of the sequence encoder are twofold: firstly, to encode objects by considering their multimodal information, and secondly, to improve the contextual representation of each item by integrating the interaction links between items in the sequence. Given that the item encoder has already

acquired a generalized representation of the items during the pretraining phase, the sequence encoder directly utilizes the model parameters of the item encoder to specifically learn the interaction relationships inside the sequence. The architecture of the sequence encoder closely resembles that of the item encoder, employing transformer blocks that are extensively utilized in diverse applications. The mechanism of attention is crucial for the model's impressive performance as it enables each item in the sequence to concentrate on contextual information. This, in turn, helps the model deduce the representation of each item by considering contextual hints inside the sequence.

**3.2.3 Masked Item Prediction.** The temporal encoder integrates contextual information, predicts future items, infers user preferences based on historical interaction data, and predicts the next item. To achieve this goal, we introduce a task called "masked item prediction," inspired by the concept of masked language models[26].

In this task, given an input sequence  $I = i_1, i_2, \dots, i_{|S|}$ , at the testing stage, as the objective is to predict the next potential item, we append a special token [mask] to the end of the user's historical interaction sequence.

In the testing phase, in order to forecast the next probable interaction item, it is necessary to add a token [mask] at the conclusion of the user's historical interaction sequence. This enables the model to anticipate the subsequent element in the sequence that the user might engage with.

In order to maintain uniformity between the inputs used for training and testing, we implement a particular method during training where the [mask] token is only placed at the final position during testing. During the training process, the final position of the input sequence sample is consistently substituted with the [mask] token. For positions in the sequence where the [mask] token appears, we employ three different algorithms for replacement: (1) in 80% of cases, we replace the item with a [mask] token, (2) in 10% of cases, we replace the item with a randomly selected item, and (3) in the remaining 10% of cases, we keep the item unchanged. This strategy resolves the discrepancy between the inputs used for training and testing. It guarantees that the model, once trained, can accurately predict the masked items in the test sequence.

## 4 EXPERIMENT

### 4.1 Datasets

We would like to introduce the datasets employed for our experimental evaluation.

1. **Meituan [27]:** This dataset includes transaction records from the Beijing area on the Meituan platform spanning six years (from January 2014 to January 2020). We use attributes like categories, locations, and keywords extracted from customer reviews.
2. **Amazon Beauty, Sports [28]:** These three datasets are derived from Amazon's review dataset. For this study, we have chosen two specific subcategories, namely "Beauty" and "Sports." We will be focusing on detailed categories and specific product brands as features.



3. **Tmall5 [29]:** A commonly used e-commerce dataset, often utilized in research on recommendation systems and deep learning models. This dataset contains user interaction data on the Tmall platform, including user clicks, purchases, favorites, and other behavioral data, as well as item attribute information. The Tmall5 dataset is characterized by its large scale and rich multimodal information, including text and images. Researchers can use this dataset to study user behavior patterns, recommendation algorithms, and multimodal data processing, among other issues.

We have preprocessed the datasets utilized in our experiments using the following approach: We employ a prior strategy to filter the data, retaining only the records of users and items that have at least five interactions, while filtering out those with fewer interactions. This ensures that each user and item has a history of at least five interactions. Additionally, we restructure the cleaned dataset by grouping it according to user interaction behavior and subsequently sorting the interaction records in ascending order based on timestamps. Through these steps, we aim to highlight the more representative users and items during the data analysis and modeling process, while also maintaining the temporal sequence of interactions.

## 4.2 Evaluation Metrics

We employ Hit Rate (HR)、Normalized Discounted Cumulative Gain (NDCG) [30] as metrics to assess the model's performance.

1. **HR@K:** It measures how many items among the top K recommended items are actually interacted with by the user in the real interaction data.

$$HR = \frac{U_c}{U_t} \quad (6)$$

Where  $U_c$  represents the number of users with clicks, while  $U_t$  signifies the total number of users.

2. **NDCG@K:** It assesses the quality and ranking accuracy of the top K recommended items. The calculation formula for NDCG (Normalized Discounted Cumulative Gain) is as follows:

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (7)$$

Where, IDCG refers to Ideal DCG, which is the Discounted Cumulative Gain in its optimal ordering. To compute IDCG, one first obtains the search results and manually sorts them into the best possible arrangement. The DCG computed from this ideal ranking constitutes the IDCG.

$$CG_p = \sum_{i=1}^p rel_i \quad (8)$$

The term  $rel_i$  represents the relevance at the i-th position.

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2(1+i)} \quad (9)$$

We report the results of HR@5, HR@10, NDCG@5, and NDCG@10 as performance metrics for the model.

### 4.3 Baseline Models

We will evaluate our proposed method by comparing it to the subsequent baseline methods:

1. **S3-Rec [31]:** S<sup>3</sup> employs self-supervised learning tailored for sequence recommendation based on self-attention mechanisms. The main strategy involves using the natural connections in available data to create self-supervised signals. This is followed by employing pre-training approaches to augment data representations, ultimately leading to improved sequence recommendations.
2. **FM [32]:** Factorization Machines (FM) are machine learning models designed to address recommendation and prediction problems under sparse data conditions. By decomposing interactions among features, FM effectively captures relationships between different characteristics.
3. **BERT4Rec [33]:** This involves the application of pre-trained BERT models to recommendation systems. By learning from item text descriptions and user interaction sequences, it aims to boost recommendation accuracy and personalization.
4. **DuoRec [34]:** This method improves recommendation accuracy and personalization by combining user behavior and item information. Using neural networks, it learns user and item representation vectors to predict user preferences. This approach addresses data sparsity and cold-start issues, enhancing recommendation system performance.
5. **UniRec [35]:** UniRec is a single-modality recommendation method that maximizes the use of a single data source to enhance recommendation accuracy and efficiency. It does not involve multi-modal data fusion but focuses solely on optimizing recommendation models based on a single data modality.
6. **MMRec [36]:** This approach aims to improve recommendation system performance by leveraging correlations between different data modalities. It integrates multimodal data to comprehensively capture user and item features, enhancing recommendation accuracy and personalization.

### 4.4 Experimental Details

We preprocessed the datasets used in the experiments as follows: we adopted a previous strategy to filter the data, keeping only user and item records with at least five interactions, while filtering out those with fewer interactions. This ensured that each user and item had at least five interaction histories. Additionally, we reorganized the cleaned dataset by grouping it based on user interaction behavior and sorting the interaction records in ascending order according to the timestamps. Through these steps, we aimed to highlight users and items that are more representative in the data analysis and modeling process, while maintaining the chronological order of interactions, as shown in Table 1.

We utilize the source code provided by the authors for BERT4Rec, GRU4Rec,  $S^3$ -Rec, FM, and DuoRec. The hyperparameters are configured according to the recommendations given in the original publications. Furthermore, the process of adjusting and optimizing the baseline models is carried out on three different datasets that are used for further analysis. In regards to our proposed model, we use the Adam optimizer with a learning rate of 0.001 and configure the batch size to be 8192. The masking rate for the model is set at 0.2. For the four Amazon datasets, we set the maximum sequence length to be 20, while for the Meituan dataset, we configure the maximum sequence length to be 100. To ensure the fairness of the experiments, we conducted ten repetitions of each experiment. The results of each experiment were averaged, and the standard deviation was included.

**Table 1.** Statistics of the datasets after preprocessing.

<b>Datasets</b>	<b>Meituan</b>	<b>Beauty</b>	<b>Sports</b>	<b>Tmall</b>
Users	13,622	22,363	25,598	66,909
Items	20,062	12,101	18,357	37,367
Actions	747,827	198,502	296,337	427,797
Sparsity	99.73%	99.93%	99.95%	99.95%

#### 4.5 Comparison to the State of the Arts

In Table 2, we compare the proposed method with several baseline models across three public datasets. From the results, several observations can be made:

1. Compared to BERT4Rec, our method across all datasets, we find that our method outperforms BERT4Rec. This is because BERT4Rec primarily uses learned text features as item representations, rather than ID embeddings.
2. Compared to FM, our method shows significant superiority. This is because FM only models interactions between features within a single modality and lacks the ability to integrate multiple modalities.
3. Compared to  $S^3$  – Rec and DuoRec, we observe that models combining modality features with ID embeddings as input perform better than models relying solely on ID embeddings. This can be attributed to the fact that modality features provide additional information for training.
4. By comparing with other multi-modal recommendation methods, such as UniSRec and MMRec, we find that MMRec outperforms UniSRec. This is because UniSRec relies on pre-training methods to learn generic, ID-agnostic representations from text features, while MMRec can model more multi-modal features in feature-level sequences and cross-modal fusion. Additionally, our method introduces a new scoring criterion and models items well, leading to superior results.
5. Compared to all other baseline models, our method shows significant improvements on dense datasets, such as Meituan. Our model diverges from conventional techniques that depend on item embedding-based methods for sequence recommendation. Instead, it employs a self-supervised multimodal pre-training approach to gain

comprehensive item representations, hence improving the model’s ability to generalize. These data demonstrate the efficacy of our model.

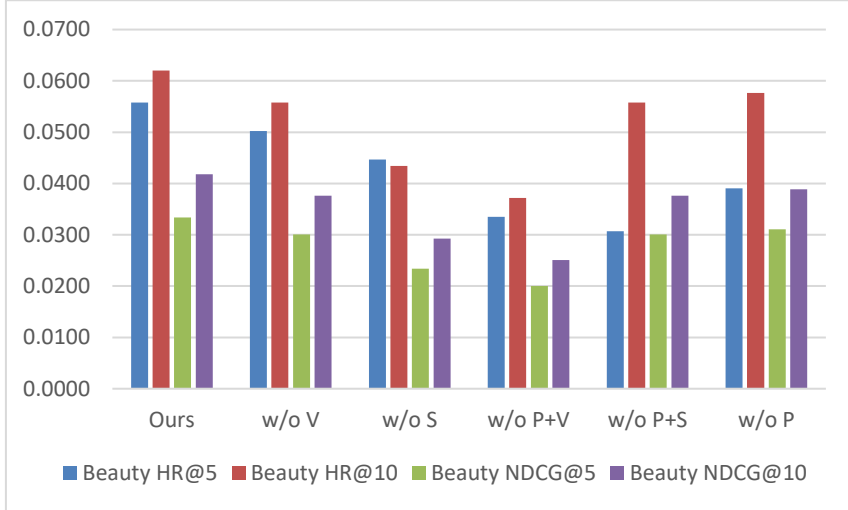
Therefore, these results substantiate the efficacy of our proposed model.

**Table 2.** Performance comparison of different methods on three datasets.

Datasets	Metric	BERT4Rec	FM	$S^3 - \text{Rec}$	DuoRec	UniRec	MMRec	Ours
Meituan	$HR@5$	0.1800	0.0493	0.2219	0.2194	0.1817	0.2206	<b>0.2465</b>
		$\pm 0.0041$	$\pm 0.0013$	$\pm 0.0044$	$\pm 0.0045$	$\pm 0.0037$	$\pm 0.0045$	$\pm 0.0030$
	$HR@10$	0.2007	0.0542	0.2441	0.2414	0.1999	0.2427	<b>0.2712</b>
		$\pm 0.0043$	$\pm 0.0016$	$\pm 0.0042$	$\pm 0.0048$	$\pm 0.0038$	$\pm 0.0048$	$\pm 0.0032$
	$NDCG@5$	0.1079	0.0308	0.1387	0.1371	0.1136	0.1379	<b>0.1541</b>
		$\pm 0.0023$	$\pm 0.0013$	$\pm 0.0026$	$\pm 0.0025$	$\pm 0.0031$	$\pm 0.0036$	$\pm 0.0022$
	$NDCG@10$	0.1367	0.0385	0.1733	0.1714	0.1419	0.1724	<b>0.1926</b>
		$\pm 0.0026$	$\pm 0.0013$	$\pm 0.0027$	$\pm 0.0028$	$\pm 0.0039$	$\pm 0.0039$	$\pm 0.0024$
Beauty	$HR@5$	0.0396	0.0112	0.0503	0.0497	0.0412	0.0500	<b>0.0558</b>
		$\pm 0.0012$	$\pm 0.0003$	$\pm 0.0014$	$\pm 0.0016$	$\pm 0.0014$	$\pm 0.0017$	$\pm 0.0008$
	$HR@10$	0.0440	0.0124	0.0558	0.0552	0.0457	0.0555	<b>0.0620</b>
		$\pm 0.0014$	$\pm 0.0003$	$\pm 0.0014$	$\pm 0.0018$	$\pm 0.0015$	$\pm 0.0018$	$\pm 0.0009$
	$NDCG@5$	0.0244	0.0067	0.0301	0.0298	0.0246	0.0299	<b>0.0334</b>
		$\pm 0.0010$	$\pm 0.0001$	$\pm 0.0012$	$\pm 0.0011$	$\pm 0.0010$	$\pm 0.0012$	$\pm 0.0004$
	$NDCG@10$	0.0305	0.0084	0.0376	0.0372	0.0308	0.0374	<b>0.0418</b>
		$\pm 0.0012$	$\pm 0.0002$	$\pm 0.011$	$\pm 0.0015$	$\pm 0.0012$	$\pm 0.0013$	$\pm 0.0004$
Sports	$HR@5$	0.0325	0.0094	0.0424	0.0419	0.0347	0.0421	<b>0.0471</b>
		$\pm 0.0012$	$\pm 0.0002$	$\pm 0.0017$	$\pm 0.0018$	$\pm 0.0014$	$\pm 0.0014$	$\pm 0.0005$
	$HR@10$	0.0361	0.0105	0.0470	0.0465	0.0385	0.0468	<b>0.0523</b>
		$\pm 0.0013$	$\pm 0.0006$	$\pm 0.0015$	$\pm 0.0015$	$\pm 0.0014$	$\pm 0.0017$	$\pm 0.0007$
	$NDCG@5$	0.0213	0.0057	0.0255	0.0252	0.0209	0.0254	<b>0.0284</b>
		$\pm 0.0009$	$\pm 0.0002$	$\pm 0.0010$	$\pm 0.0011$	$\pm 0.0010$	$\pm 0.0011$	$\pm 0.0002$
	$NDCG@10$	0.0249	0.0065	0.0291	0.0287	0.0238	0.0289	<b>0.0323</b>
		$\pm 0.0011$	$\pm 0.0001$	$\pm 0.0014$	$\pm 0.0014$	$\pm 0.0010$	$\pm 0.0013$	$\pm 0.0002$

#### 4.6 Ablation study3

In this section, we conduct an ablation study on datasets such as Meituan and Sports to analyze the impact of each proposed technique and component on system performance. Specifically, we compared with the following variants: (1) w/o V: without visual modality. (2) w/o S: without textual modality. (3) w/o P: without pretraining. (4) w/o P+V: without pretraining and visual modality. (5) w/o P+S: without pretraining and textual modality.



**Fig. 1.** Performance comparison of different methods on three datasets.

The results are shown in Figure 2. Based on the experimental results of our model, we can infer that:

1. Our pretraining has been shown to improve the performance of the model in downstream recommendation tasks, providing further evidence that our pretraining task effectively learns how to align representations between different modalities and boost their reciprocal reinforcement.
2. Our model design efficiently combines features from multiple modalities. Additionally, when using both visual and textual modalities simultaneously, our model achieves outstanding performance.

## 5 Conclusion, Broader Impact and Limitation

This research presents a new and innovative sequential recommendation model designed for the Internet of Things (IoT) scenario. In order to allow the model to make predictions about the next item in a sequence based on the contextual information, we combine the feature embeddings, modality embeddings, and position embeddings to create a representation of the item. We use the predicted embeddings from the user encoder for retrieving the embeddings generated by the item encoder. Furthermore, to enhance the retrieval performance of the model, we propose a multi-modal pretraining method. The model's generalization capacity is enhanced by creating contrastive learning challenges that involve various feature combinations. Our model's performance is demonstrated through experiments conducted on three public datasets.

In our future work, we will explore ways to streamline our model structure in order to decrease the computational resources required. We also aim to optimize modal selection to better align with our daily lives. Additionally, we aim to expand the current framework to encompass a wider range of user modeling activities.

## References

1. Zhu F, Wang Y, Chen C, et al. Cross-domain recommendation: challenges, progress, and prospects[J]. arXiv preprint arXiv:2103.01696, 2021..
2. Cui Q, Wei T, Zhang Y, et al. HeroGRAPH: A Heterogeneous Graph Framework for Multi-Target Cross-Domain Recommendation[C]//ORSUM@ RecSys. 2020.
3. Liu M, Nie L, Wang M, et al. Towards micro-video understanding by joint sequential-sparse modeling[C]//Proceedings of the 25th ACM international conference on Multimedia. 2017: 970-978.
4. Min W, Bao B K, Xu C, et al. Cross-platform multi-modal topic modeling for personalized inter-platform recommendation[J]. IEEE Transactions on Multimedia, 2015, 17(10): 1787-1801.
5. Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks[J]. arXiv preprint arXiv:1511.06939, 2015.
6. Ding H, Ma Y, Deoras A, et al. Zero-shot recommender systems[J]. arXiv preprint arXiv:2105.08318, 2021.
7. Song K, Sun Q, Xu C, et al. Self-Supervised Multi-Modal Sequential Recommendation[J]. arXiv preprint arXiv:2304.13277, 2023.
8. B. P. Yuhua, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," IEEE Communications Magazine, 1989.
9. Guzhov A, Raue F, Hees J, et al. Audioclip: Extending clip to image, text and audio[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 976-980.
10. Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In Proceedings of the 19th international conference on World wide web. 811–820.
11. Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback[J]. arXiv preprint arXiv:1205.2618, 2012.
12. Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized markov chains for next-basket recommendation[C]//Proceedings of the 19th international conference on World wide web. 2010: 811-820.
13. Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
14. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
15. Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks[J]. arXiv preprint arXiv:1511.06939, 2015.
16. Sun F, Liu J, Wu J, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 1441-1450.
17. Lever J, Krzyzawinski M, Altman N. Points of significance: model selection and overfitting[J]. Nature methods, 2016, 13(9): 703-705.
18. Samir S Soliman and Mandyam D Srinath. 1990. Continuous and discrete signals and systems. Englewood Cliffs (1990).
19. Sun F, Liu J, Wu J, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 1441-1450.

20. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
21. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
22. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.
23. Sarzynska-Wawer J, Wawer A, Pawlak A, et al. Detecting formal thought disorder by deep contextualized word representations[J]. Psychiatry Research, 2021, 304: 114135.
24. Choi H, Kim J, Joe S, et al. Evaluation of bert and albert sentence embedding performance on downstream nlp tasks[C]//2020 25th International conference on pattern recognition (ICPR). IEEE, 2021: 5482-5487.
25. Schmittfull M, Vlah Z, McDonald P. Fast large scale structure perturbation theory using one-dimensional fast Fourier transforms[J]. Physical Review D, 2016, 93(10): 103528.
26. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
27. <https://www.meituan.com>
28. McAuley J, Targett C, Shi Q, et al. Image-based recommendations on styles and substitutes[C]//Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. 2015: 43-52.
29. <https://tianchi.aliyun.com/dataset/dataDetail?dataId=42>
30. Xie X, Sun F, Liu Z, et al. Contrastive learning for sequential recommendation[C]//2022 IEEE 38th international conference on data engineering (ICDE). IEEE, 2022: 1259-1273.
31. Zhou K, Wang H, Zhao W X, et al. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization[C]//Proceedings of the 29th ACM international conference on information & knowledge management. 2020: 1893-1902.
32. Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks[J]. arXiv preprint arXiv:1511.06939, 2015.
33. Rendle S. Factorization machines[C]//2010 IEEE International conference on data mining. IEEE, 2010: 995-1000.
34. Sun F, Liu J, Wu J, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 1441-1450.
35. Hou Y, Mu S, Zhao W X, et al. Towards universal sequence representation learning for recommender systems[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 585-593.
36. Wu C, Wu F, Qi T, et al. Mm-rec: Visiolinguistic model empowered multimodal news recommendation[C]//Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval. 2022: 2560-2564.