# MISSRec: Pre-training and Transferring Multi-modal Interest-aware Sequence Representation for Recommendation

Jinpeng Wang
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
wjp20@mails.tsinghua.edu.cn

Ziyun Zeng
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
zengzy21@mails.tsinghua.edu.cn

Yunxiao Wang
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
wang-yx20@mails.tsinghua.edu.cn

Yuting Wang
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
wangyt22@mails.tsinghua.edu.cn

Xingyu Lu
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
luxy22@mails.tsinghua.edu.cn

Tianxiang Li
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
litx21@mails.tsinghua.edu.cn

Jun Yuan
Huawei Noah's Ark Lab
Shenzhen, China
yuanjun25@huawei.com

Rui Zhang✉
www.ruizhang.info
rayteam@yeah.net

Hai-Tao Zheng
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
Peng Cheng Laboratory
Shenzhen, China
zheng.haitao@sz.tsinghua.edu.cn

Shu-Tao Xia✉
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
Research Center of Artificial Intelligence, Peng Cheng Laboratory
Shenzhen, China
xiast@sz.tsinghua.edu.cn

## ABSTRACT

The goal of sequential recommendation (SR) is to predict a user's potential interested items based on her/his historical interaction sequences. Most existing sequential recommenders are developed based on ID features, which, despite their widespread use, often underperform with sparse IDs and struggle with the cold-start problem. Besides, inconsistent ID mappings hinder the model's transferability, isolating similar recommendation domains that could have been co-optimized. This paper aims to address these issues by exploring the potential of multi-modal information in learning robust and generalizable sequence representations. We propose **MISSRec**, a multi-modal pre-training and transfer learning framework for SR. On the user side, we design a Transformer-based encoder-decoder model, where the contextual encoder learns to capture the sequence-level multi-modal synergy while a novel interest-aware decoder is developed to grasp item-modality-interest relations for better sequence representation. On the candidate item side, we adopt a dynamic fusion module to produce user-adaptive item representation, providing more precise matching between users and items. We pre-train the model with contrastive learning objectives and fine-tune it in an efficient manner. Extensive experiments demonstrate the effectiveness and flexibility of MISSRec, promising an practical solution for real-world recommendation scenarios.

## CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Multimedia information systems**; **Personalization**.

## KEYWORDS

multi-modal sequential recommendation, pre-training, parameter-efficient fine-tuning, interest-aware sequence representation

**ACM Reference Format:**
Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang✉, Hai-Tao Zheng, and Shu-Tao Xia✉. 2023.

✉ Corresponding authors.



*MM '23, October 29-November 3, 2023, Ottawa, ON, Canada*

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3611967

MISSRec: Pre-training and Transferring Multi-modal Interest-aware Sequence Representation for Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611967

## 1 INTRODUCTION

Recommendation systems [23, 53, 59, 60] are an important component in various application domains, from e-commerce to content platforms, improving user engagement and satisfaction. Sequential recommendation (SR) [50, 72], a popular methodology of recommendation, aims to predict users' potential interested items according to their historical interaction sequences. Typically, this task is formulated as a representation learning problem, where user embeddings are learned to capture underlying preferences behind the interacted item sequence and to match with suitable item embeddings.

Current SR approaches [7, 31, 38, 61, 95] predominantly rely on ID features as input, which, despite their widespread uses in practice, exhibit two primary shortcomings. (**i**) The robustness and scalability largely depend on the distribution of user-item interaction. Due to the sparsity of interaction data, it is challenging to learn accurate item and sequence representations based on ID information, let alone the cold-start problem [54] for new items. Meanwhile, ID-based approaches often exhibit a popularity bias toward popular IDs [1, 90], posing the fairness issue for the majority of infrequent IDs. (**ii**) Transferring knowledge to new scenarios is hard because of inconsistent ID mappings. It isolates similar domains that can be co-optimized, limiting the applicability of ID-based models.

Given the ubiquity of text and images in real-world scenarios, leveraging multi-modal data is promising to remedy the above shortcomings [80]. Particularly, both text and visual information play crucial roles in attracting user attention – compelling titles can boost user engagement, while visual elements like colors and shapes, can influence user decisions. Considering remarkable progress in computer vision [17, 32] and natural language understanding [39, 92], they offer robust and generalizable semantic extraction capability to item content. Therefore, we believe it necessary and also feasible to harness multi-modal content understanding to the SR. However, such investigation remains in a fledgling stage in literature, as we will discuss in Related Works (§2). *How does multi-modal information impact SR? How to make better use of such information in sequence modeling?* These two questions drive our study in this paper.

In general, utilizing multi-modal information in SR poses two challenges. (**i**) The multi-modal synergy within each item is user-dependent and dynamic. Users may interact with the same item for various reasons, and different modalities contribute unequally to user interest in an item. Even for the same user-item pair, such patterns can be time- and context-varying, making it difficult to design effective multi-modal fusion. (**ii**) Information redundancy can overwhelm essential user interests. Interaction sequences typically exhibit imbalanced interest distribution. A user's sequence usually contains lots of homogeneous items, *e.g.*, daily necessities, while other informative items, *e.g.*, sports goods, may sparsely appear. If we treat all interactions equally in user behavior modeling, it will result in overemphasis on specific kinds of items and insufficient focus on others. We give an illustration of these challenges in fig. 1.

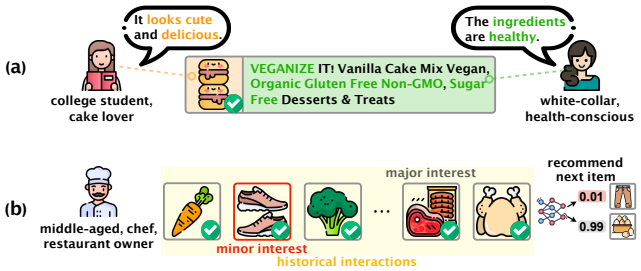


**Figure 1: Challenges of multi-modal sequential recommendation. (a) The multi-modal synergy within each item can be user-dependent and dynamic. Users may interact with the same item focusing on different modalities. (b) Redundant interaction may overwhelm sparse but important interests. Modeling interactions equally will lead to biased predictions.**

To tackle these challenges, we propose **MISSRec** (Multi-modal IntereSt-aware Sequential Recommendation), a multi-modal pre-training and transferring framework for SR. In general, MISSRec takes interaction sequences with multi-modal item information as input, learns ID-agnostic representations with an interest-aware encoder-decoder model via self-supervised pre-training, and can be efficiently adapted for multiple domains to enhance recommendation. Specifically, (**i**) to bridge the semantic gap between general domains and recommendation, we incorporate multi-modal feature adapters [2, 27] to distill personalized semantics of general multi-modal features in a parameter-efficient manner. (**ii**) To explore the multi-modal synergy, we design a Transformer-based contextual encoder for multi-modal sequence processing. Compared to item-level static fusion (*e.g.*, vector addition), our approach can adaptively capture useful modality cues in each item for personalization. On the candidate item side, we also introduce a lightweight dynamic fusion strategy to produce user-specific item representations. (**iii**) To mitigate the adverse effect of information redundancy on sequence modeling, we introduce an interest discovery module to mine global multi-modal interests among users. By associating items with relevant interests via adaptive clustering, we convert multi-modal sequences into a series of interest tokens. Subsequently, we propose an interest-aware Transformer decoder, which consumes the output sequence of the encoder as keys and values, and de-duplicated user interests as queries, to grasp essential item-modality-interest patterns for precise and comprehensive sequence representation.

We conduct detailed experiments and show the effectiveness of MISSRec in leveraging multi-modal data to learn robust and transferable sequence representation. In particular, it exhibits better generalizability than ID-based and text-based approaches and further alleviates the item cold-start issue.

To summarize, we make the following contributions in this paper.

- We highlight the significance and challenges of exploiting multi-modal information for SR and propose an effective pre-training and efficient transfer learning framework for it.
- To capture the contextual and dynamic multi-modal synergy, we design a Transformer-based contextual encoder for multi-modal sequential modeling and adopt a lightweight dynamic fusion module to produce user-adaptive candidate item representation.

- We introduce a multi-modal interest discovery module, based on which we construct an interest-aware decoder to model item-modality-interest relations for better sequence representation.
- Extensive experiments of pre-training and downstream domain adaptation justify the comprehensive merits of our approach.

## 2 RELATED WORKS

### 2.1 Sequential Recommendation

Sequential recommendation (SR) aims to predict the next suitable item for a given user based on her/his interacted item sequence [50, 72]. Early SR approaches [21, 52] exploited the Markov chain mechanism for sequential modeling, while deep SR approaches explored CNN- [63, 83], RNN- [24, 46], MLP- [37, 96] and attention-based [31, 43, 61, 89] architectures to model the transition patterns in the interaction sequences. Besides, interest modeling [4, 34, 38, 88, 93] has been another popular methodology for SR, where user interests were usually implemented by attention or clustering. From other perspectives, extensive efforts have been devoted to designing effective learning strategies for SR, including temporal-aware learning tasks [35, 65, 91] and contrastive learning objectives [7, 48, 68, 78, 78, 95]. Note that most existing SR approaches are designed with ID features, *e.g.*, item IDs or attribute IDs. They usually suffer from the cold-start problem and fall short of transferability. Several SR approaches have leveraged multi-modal item contents to mitigate these issues. For instance, for each item, CSAN [28] aggregated multiple information including multi-modal features and projected the result to obtain an embedding, providing better item representation than the ID embedding alone. MM-Rec [76] adopted pre-trained VL-BERT [58] as the multi-modal item encoder and built a single-tower model for SR. Inspired by the success of MLP-Mixer [66] in CV, MMMLP [37] adapted the architecture to multi-modal SR. On the basis of these works, we further focus on learning universal sequence representations with multi-modal item content. Our MISSRec contributes a pre-training and efficient transfer framework for multi-modal SR, which can benefit real-world practice.

### 2.2 Pre-training and Transfer Learning in Recommendation

The paradigm of "pre-train and transfer" has become increasingly popular in recommendation. Compared to cross-domain recommendation [98], this paradigm is more general as it does not require cross-domain correspondence, *e.g.*, overlapped items. CLUE [8] and PeterRec [82] designed ID-based sequence models and adopted contrastive learning objectives [6, 19] for model learning, where PeterRec further made the subsequent transfer parameter-efficient. UPRec [77] leveraged user profiles and social information to construct auxiliary pre-text tasks. The assumption that both pre-training and target domains share ID vocabulary restricts their application scopes. By contrast, modality-based methods are more flexible. For instance, text-based SR pre-training approaches [25, 26] utilized a pre-trained language model (PLM, *e.g.*, BERT [12]) as the frozen feature extractor in item representation modules, showing efficacy and efficiency. Yuan et al. [84] designed text-based and image-based SR models under different scenarios, demonstrating the improvement from the tunable feature extractors. Wang et al. [70] considered mixed modality sequences where each item is either a text or an image. They also chose to optimize the SR model and feature extractors jointly, improving the item representation but largely downgrading learning efficiency. Furthermore, to capture the multi-modal synergy within each item, two latest works [55, 86] have focused on SR pre-training and transfer learning with multi-modal item representations. Our MISSRec also belongs to this setting, but we have fulfilled two improvements. (**i**) To our best knowledge, we are the first to design interest-aware modeling for multi-modal SR. (**ii**) We design user-adaptive candidate item fusion to model users' dynamic attention to different modalities.

### 2.3 Multi-modal Recommendation

Leveraging multi-modal information (*e.g.*, text and images) has been shown to improve the recommendation efficacy in many applications, *e.g.*, news feed [36, 40, 79, 87], short-video feed [3, 29, 33], fashion e-commerce [22, 56], and can effectively alleviate cold-start issues [16, 47]. Therefore, multi-modal recommendation (MMR) has become an active research topic [10, 11, 94]. Early approaches explored MMR with matrix factorization models [5, 22], while the latter ones mainly developed GNN-based models [62, 74, 75] to demonstrate the efficacy. On this basis, recent works have delved into self-supervised learning strategies [9, 64, 73, 81, 85, 97] to enhance in-domain robustness or pre-training strategies [20, 30, 41, 42, 49, 69] to improve cross-domain generalization. Despite the remarkable progress in MMR, existing solutions were predominantly designed for non-sequential scenarios, *e.g.*, collaborative filtering [44], which are not very suitable for SR. Fortunately, our work uncovers some challenges in multi-modal SR (see §1 and fig. 1) and fulfills effective solutions. We hope it can provide a research basis for future work.

## 3 METHOD

### 3.1 Problem Formulation and Method Overview

Given the historical interaction sequence of the $i$-th user, $S_i = [I_1, I_2, \cdots, I_{T_i}]$, ordered by timestamps, sequential recommendation (SR) aims to predict the next item $I_{T_i+1}$ that the user may interact with by learning representations for $S_i$ and candidate items. Under the multi-modal settings, each item contains a unique item ID and multi-modal content. Without loss of generality, we study leveraging text and image modalities to improve SR in this paper. We propose **MISSRec**, a pre-training and transferring framework, as illustrated in fig. 2. MISSRec consists of 7 main components: (**i**, §3.2.1) Pre-trained, frozen text and image encoders for extracting multi-modal features. (**ii**, §3.2.3) A pair of modality-specific adapters to transform multi-modal features into input tokens, which bridge the semantic gap between general features and personalization. In the fine-tuning stage, they help with efficient model adaptation. (**iii**, §3.2.4) A dynamic fusion module for generating user-adaptive candidate item representations. (**iv**, §3.4.1) A Transformer-based context encoder for the token sequence that dynamically grasps important modality cues for personalization. (**v**, §3.3.1) A multi-modal interest discovery module that generates user interest tokens for the decoder. (**vi**, §3.4.2) An interest-aware Transformer decoder to capture the item-modality-interest relation for better sequence representation. (**vii**, §3.5 and §3.6) The pre-training and downstream fine-tuning tasks to achieve transferable recommendation.

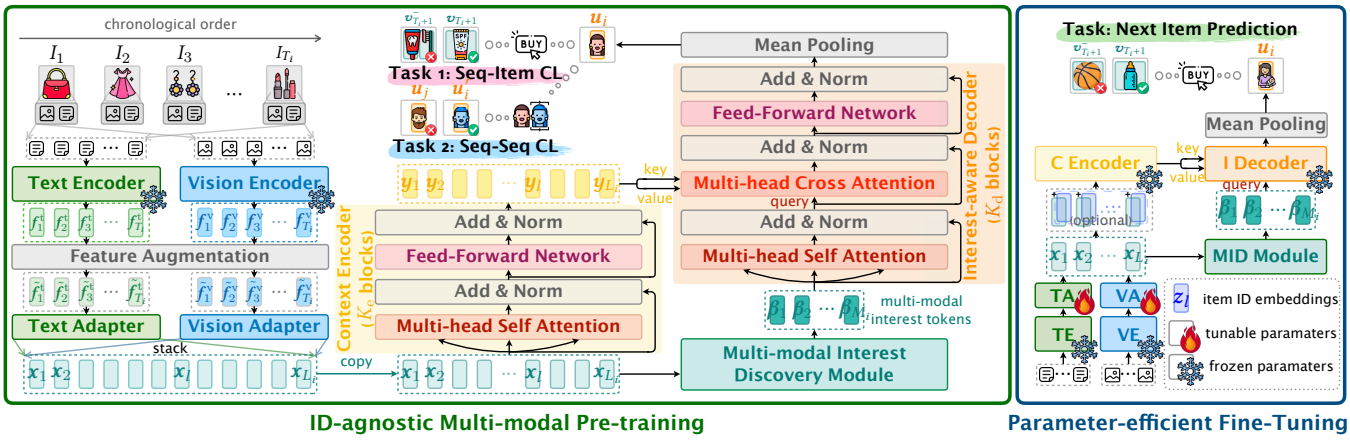


**Figure 2: MISSRec consists of two stages. In the first stage, it first extracts multi-modal features from item content and transforms them into a token sequence. Then, it encodes the sequence to capture contextual multi-modal cues for personalization. Meanwhile, it adaptively converts the token sequence into multi-modal interest tokens. Next, by developing an interest-aware decoding mechanism, it produces a comprehensive sequence embedding. Finally, it adopts sequence-item and sequence-sequence contrastive learning objectives for pre-training. In the second stage, it uses the sequence-item objective for fine-tuning. Both transductive (*i.e.*, w/ item IDs) and inductive (*i.e.*, ID-agnostic) tuning settings are supported. The downstream domain adaptation is parameter-efficient as only the modality-specific adapters need to be tuned. *This figure is best viewed in color.***

## 3.2 Universal Multi-modal Item Representation

*3.2.1 Multi-modal Feature Extraction with Pre-trained Transformers.* MISSRec utilizes multi-modal features to achieve universal representations. Concretely, we take advantage of the pre-trained transformers for item content understanding. Specifically, we use BERT [12] and ViT [14] for text and visual feature extraction, respectively. We adopt the cross-modal pre-trained version [51] following Zhang et al. [86]. For an item $I$, we obtain its text feature vector by $\boldsymbol{f}^{\mathrm{t}} = \phi^{\mathrm{t}}(I)$ and the image feature vector by $\boldsymbol{f}^{\mathrm{v}} = \phi^{\mathrm{v}}(I)$, where $\phi^{\mathrm{t}}$ and $\phi^{\mathrm{v}}$ denotes the text encoder the visual encoder, respectively.

*3.2.2 Feature Augmentation.* To enhance the robustness of the sequence representation model, we apply dropout [57] as the feature augmentation. Given the text features $\boldsymbol{F}_i^{\mathrm{t}} = [\boldsymbol{f}_1^{\mathrm{t}}, \boldsymbol{f}_2^{\mathrm{t}}, \cdots, \boldsymbol{f}_{T_i}^{\mathrm{t}}]$, we apply twice the feature augmentation and subsequently obtain $\tilde{\boldsymbol{F}}_i^{\mathrm{t}} = [\tilde{\boldsymbol{f}}_1^{\mathrm{t}}, \tilde{\boldsymbol{f}}_2^{\mathrm{t}}, \cdots, \tilde{\boldsymbol{f}}_{T_i}^{\mathrm{t}}]$ and $\tilde{\boldsymbol{F}}_i^{\mathrm{t}'} = [\tilde{\boldsymbol{f}}_1^{\mathrm{t}'}, \tilde{\boldsymbol{f}}_2^{\mathrm{t}'}, \cdots, \tilde{\boldsymbol{f}}_{T_i}^{\mathrm{t}'}]$ as the augmented text features. They serve as two positive views *w.r.t.* $\boldsymbol{F}_i^{\mathrm{t}}$. The image features $\boldsymbol{F}_i^{\mathrm{v}}$ follow analogous augmentation procedure.

*3.2.3 Bridge Domain Gap with Modality-specific Adapters.* We adopt modality-specific adapters [2, 27] to reduce irrelevant semantics in multi-modal features and enhance the personalization. We transform the augmented features and concatenate the results as item tokens, for example, $\boldsymbol{X}_i = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{L_i}] = [\psi^{\mathrm{t}}(\tilde{\boldsymbol{F}}_i^{\mathrm{t}}); \psi^{\mathrm{v}}(\tilde{\boldsymbol{F}}_i^{\mathrm{v}})]$. $\psi^{\mathrm{t}}$ and $\psi^{\mathrm{v}}$ denote the text and visual adapters, respectively.

In contrast to tunable feature extractors, adapters with frozen extractors exhibits two strengths. **(i)** It allows (pre-) training of the sequence model based on pre-extracted features, reducing time and memory overhead. **(ii)** Adapters enable parameter-efficient transfer in downstream domains, which is preferable to fine-tuning.

*3.2.4 Candidate Item Representation with Dynamic Fusion.* To model users' dynamic attention to different modalities of candidate items, we design a lightweight fusion module to generate user-adaptive item representations. Given a candidate item $I_k$, we first encode it with modality encoders and adapters to obtain modality embeddings, *i.e.*, $\boldsymbol{x}_k^{\mathrm{t}}$ and $\boldsymbol{x}_k^{\mathrm{v}}$. For the $i$-th user, we compute its sequence representation $\boldsymbol{u}_i$ according to the procedures in §3.3 and §3.4. Then, the adaptive representation of $I_k$ is defined by the weighted fusion of modality embeddings, namely

$$\boldsymbol{v}_k = \frac{e^{\alpha \cdot \langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\mathrm{t}} \rangle} \cdot \boldsymbol{x}_k^{\mathrm{t}} + e^{\alpha \cdot \langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\mathrm{v}} \rangle} \cdot \boldsymbol{x}_k^{\mathrm{v}}}{e^{\alpha \cdot \langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\mathrm{t}} \rangle} + e^{\alpha \cdot \langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\mathrm{v}} \rangle}}. \tag{1}$$

$\langle \cdot, \cdot \rangle$ denotes the inner product. $\alpha \geq 0$ is a concentration factor to balance two modalities. Besides the efficacy, the fusion keeps efficiency as well, because it is factorizable when computing matching scores. Specifically, let us define $s_{i,k}^{\mathrm{t}} = \langle \boldsymbol{u}_i, \boldsymbol{v}_j^{\mathrm{t}} \rangle$ and $s_{i,k}^{\mathrm{v}} = \langle \boldsymbol{u}_i, \boldsymbol{v}_j^{\mathrm{v}} \rangle$, the matching score between $\boldsymbol{u}_i$ and $\boldsymbol{v}_k$ can be compute by

$$\langle \boldsymbol{u}_i, \boldsymbol{v}_k \rangle = \frac{s_{i,k}^{\mathrm{t}} \cdot e^{\alpha \cdot s_{i,k}^{\mathrm{t}}} + s_{i,k}^{\mathrm{v}} \cdot e^{\alpha \cdot s_{i,k}^{\mathrm{v}}}}{e^{\alpha \cdot s_{i,k}^{\mathrm{t}}} + e^{\alpha \cdot s_{i,k}^{\mathrm{v}}}}. \tag{2}$$

We can equivalently pre-compute modality-specific scores and fuse them to get the overall score. In particular, dynamic fusion exhibits a transition between mean and max poolings: $\langle \boldsymbol{u}_i, \boldsymbol{v}_k \rangle = (s_{i,k}^{\mathrm{t}} + s_{i,k}^{\mathrm{v}})/2$ when $\alpha = 0$, while $\langle \boldsymbol{u}_i, \boldsymbol{v}_k \rangle = \max(s_{i,k}^{\mathrm{t}}, s_{i,k}^{\mathrm{v}})$ when $\alpha \to +\infty$.

## 3.3 Discover Multi-modal User Interests

Typically, item sequences reflect users' implicit interests that are important for SR. Interest modeling can reduce the negative impact of information redundancy, helping to capture personalized semantics for precise representation. In this sub-section, we introduce a multi-modal interest discovery (MID) module to assist this goal.

*3.3.1 Excavate and Index Interests via Adaptive Clustering.* During the training process, item tokens are expected to convey both transferable multi-modal information and personalized semantics.

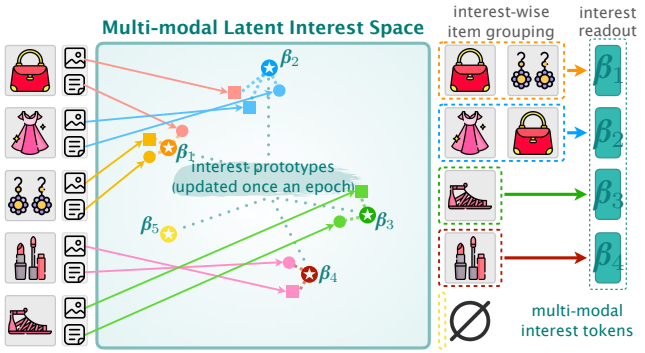


**Figure 3: Convert item tokens to interest tokens via the multi-modal interest discovery (MID) module. An item can be converted to different interests according to multiple modalities.**

Hence, we adopt a variant of $k$-nearest neighbor-based density peaks clustering algorithm (DPC-KNN) [15] on the whole item token set $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N_I}$, to explore the distribution of user interests before every training epoch starts. First, for the $i$-th token, we compute its local density by considering its $k$-nearest neighbors:

$$\rho_i = \exp\left(-\frac{1}{k}\sum_{\boldsymbol{x}_j \in k\mathrm{NN}(\boldsymbol{x}_i)} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2\right). \tag{3}$$

Then, for the $i$-th token, we compute its minimum distance to any other token with higher local density, namely

$$\delta_i = \begin{cases} \min_{j:\rho_j > \rho_i} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i, \\ \max_j \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2, & \text{otherwise.} \end{cases} \tag{4}$$

Next, we select $K_\mathrm{c}$ cluster centroids with the highest $\rho_i \times \delta_i$ scores as the interest prototypes. Finally, we index each token by assigning it to the nearest centroid (*i.e.*, the interest prototype).

*3.3.2 Translate Multi-modal Item Tokens into Interest Tokens.* As shown in figs. 2 and 3, the MID module receives item tokens, finds the nearest interest prototype for each token, and returns irredundant interest tokens. To accelerate token conversion, we can build the interest index of each item token after clustering for fast lookup.

## 3.4 Multi-modal Interest-aware Sequence Model

We design an encoder-decoder model for sequence representation, which unifies sequence modeling and multi-modal fusion.

*3.4.1 Transformer-based Multi-modal Context Encoder.* Instead of applying a context-free fusion strategy shared by all items, we expect the model to adaptively capture crucial modality cues for personalization, considering the multi-modal interaction sequence as context. To implement this *fusion-in-context* idea, we build a context encoder with $K_\mathrm{e}$ Transformer [67] encoder blocks. We add item-wise positional embeddings $\boldsymbol{P}_i = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{L_i}]$ to the item tokens $\boldsymbol{X}_i = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{L_i}]$ and form the encoder input. The context encoder processes the input and generates an encoded token sequence $\boldsymbol{Y}_i = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{L_i}]$ *w.r.t.* $\boldsymbol{X}_i$ for next-step decoding.

*3.4.2 Transformer-based Interest-aware Decoder.* After obtaining the encoded tokens sequence $\boldsymbol{Y}_i$, we design an interest-aware decoding process to model item-modality-interest relations for precise and comprehensive sequence representation. Specifically, we construct a decoder with $K_\mathrm{d}$ Transformer decoder blocks, each of which follows the standard design [67] except changing the autoregressive attention into the parallel one, *i.e.*, making the decoding process permutation-invariant. For the multi-head cross-attention module in each decoder block, we take $\boldsymbol{Y}_i$ as the key and value. The interest tokens $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_{M_i}$ converted by the MID module (§3.3.2) serve as the decoding queries. We aggregate the decoded embeddings $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_{M_i}$ *w.r.t.* $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_{M_i}$ by mean pooling, resulting in an embedding $\boldsymbol{u}_i$ (*i.e.*, the sequence representation) for the $i$-th user.

## 3.5 Self-supervised Contrastive Pre-training

To pursue universal representation for multi-modal interaction sequences, we design two self-supervised contrastive learning (CL) tasks for pre-training the model.

*3.5.1 Sequence-item Contrastive Learning.* To capture the correspondence between interaction sequences and candidate items, we take sequence-item CL as a pre-training task. The objective *w.r.t.* $\boldsymbol{u}_i$:

$$\ell_i^{S-I} = -\log \frac{\exp(\langle \boldsymbol{u}_i, \boldsymbol{v}_{T_i+1} \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle \boldsymbol{u}_i, \boldsymbol{v}_{T_j+1} \rangle / \tau)}, \tag{5}$$

where $B$ denotes the size of mini-batch. $\tau > 0$ us a temperature factor. The user-item matching scores are computed by eq. (2).

*3.5.2 Sequence-sequence Contrastive Learning.* Apart from sequence-item contrast, we leverage semantic invariance as a self-supervising signal to enhance the robustness of the sequence model. Given $\boldsymbol{u}_i, \boldsymbol{u}_i'$ as the representations of two augmented sequences (see §3.2.2) *w.r.t.* the $i$-th user, we define the sequence-sequence CL by

$$\ell_i^{S-S} = -\log \frac{\exp(\langle \boldsymbol{u}_i, \boldsymbol{u}_i' \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle / \tau) + \exp(\langle \boldsymbol{u}_i, \boldsymbol{u}_j' \rangle / \tau)}. \tag{6}$$

*3.5.3 Overall Objectives.* We sum up pre-train objectives as follows

$$\mathcal{L}_{\text{pre-train}} = \frac{1}{B}\sum_{i=1}^{B}\left[\ell_i^{S-I} + \lambda \cdot \ell_i^{S-S} + \frac{\gamma}{M_i^2}\sum_{m,m'=1}^{M_i}\langle \boldsymbol{\xi}_m, \boldsymbol{\xi}_{m'} \rangle\right]. \tag{7}$$

The last term is an orthogonal regularization to diversify interest-aware decoded results. $\lambda, \gamma > 0$ are the weights for different terms.

## 3.6 Efficient Fine-tuning

In order to transfer the recommendation knowledge to downstream domains and enhance their performance, we further study efficient fine-tuning based on the pre-trained sequence model. For the tuning objectives, we adopt sequence-item CL and orthogonal regularization to directly optimize the next-item prediction, namely,

$$\mathcal{L}_{\text{fine-tune}} = \frac{1}{B}\sum_{i=1}^{B}\left[\ell_i^{S-I} + \frac{\gamma}{M_i^2}\sum_{m,m'=1}^{M_i}\langle \boldsymbol{\xi}_m, \boldsymbol{\xi}_{m'} \rangle\right]. \tag{8}$$

As shown in the right block of fig. 2, we disable the feature augmentation and the sequence-sequence contrast, so the sequence model only encodes once per user, which is computation-efficient.

**Table 1: Statistics of Pre-processed Datasets. "Cover." denotes the image coverage among the item set. "Avg. SL" denotes the average length of interaction sequences.**

| Datasets | #Users | #Items | #Img. (Cover./%) | #Inters. | Avg. SL. |
|---|---|---|---|---|---|
| *Pre-trained* | 1,361,408 | 446,975 | 94,151 (21.06%) | 14,029,229 | 13.51 |
| - Food | 115,349 | 39,670 | 29,990 (75.60%) | 1,027,413 | 8.91 |
| - CDs | 94,010 | 64,439 | 21,166 (32.85%) | 1,118,563 | 12.64 |
| - Kindle | 138,436 | 98,111 | 0 (0%) | 2,204,596 | 15.93 |
| - Movies | 281,700 | 59.203 | 8,675 (14.65%) | 3,226,731 | 11.45 |
| - Home | 731,913 | 185,552 | 34,320 (18.50%) | 6,451,926 | 8.82 |
| Scientific | 8,442 | 4,385 | 1,585 (36.15%) | 59,427 | 7.04 |
| Pantry | 13,101 | 4,898 | 4,587 (93.65%) | 126,962 | 9.69 |
| Instruments | 24,962 | 9,964 | 6,289 (63.12%) | 208,926 | 8.37 |
| Arts | 45,486 | 21,019 | 9,437 (44.90%) | 395,150 | 8.69 |
| Office | 87,436 | 25,986 | 16,628 (63.99%) | 684,837 | 7.84 |

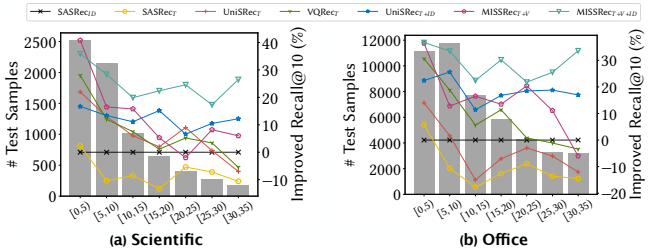


**Figure 4: Performance comparison on long-tailed items. The histogram represents the number of test samples (*i.e.,* interacted items) of different popularities. The line chart shows the relative improvement to $\text{SASRec}_{\text{ID}}$ on Recall@10.**

Besides, only the item ID embedding table (if included) and modality-specific adapters need to be tuned while other modules keep frozen, which is also parameter-efficient [18]. Following Hou et al. [26], we support *inductive* and *transductive* transfer settings for MISSRec.

*3.6.1 Inductive Transfer.* Given a target domain with lots of cold items, making robust predictions is tricky for ID-based recommenders. We support an ID-agnostic solution to deal with extreme sparsity, where the model predicts next item for the $i$-th user by

$$\Pr_{\text{ind}}(I_{T_i+1} \mid I_1, I_2, \cdots, I_{T_i}) = \text{softmax}(\langle \boldsymbol{u}_i, \boldsymbol{v}_{T_i+1} \rangle). \quad (9)$$

*3.6.2 Transductive Transfer.* Given a target domain of mostly warm items, we pursue more precise recommendations by utilizing item IDs. Specifically, the model predicts next item for the $i$-th user by

$$\Pr_{\text{trd}}(I_{T_i+1} \mid I_1, I_2, \cdots, I_{T_i}) = \text{softmax}(\langle \bar{\boldsymbol{u}}_i, \boldsymbol{v}_{T_i+1} + \boldsymbol{z}_{T_i+1} \rangle), \quad (10)$$

where $\boldsymbol{z}_{T_i+1}$ is the ID embeddings of $I_{T_i+1}$. $\bar{\boldsymbol{u}}_i$ denotes the encoded sequence representation with item IDs (see the right block in fig. 2).

# 4 EXPERIMENTS

## 4.1 Research Questions

We evaluate the proposed method by conducting experiments on three datasets. We aim to answer the following research questions:

**RQ1:** Compared with state-of-the-art sequential recommendation models using various types of information, can MISSRec achieve competitive performance in downstream domains?

**RQ2:** Is MISSRec better than other modality-based baselines in utilizing different modalities? How much impact do multi-modal information and pre-training have on its efficacy?

**RQ3:** How do different designs contribute to MISSRec's efficacy?

## 4.2 Experimental Setup

*4.2.1 Datasets.* We adopt 10 domains including "*Grocery and Gourmet Food*", "*Home and Kitchen*", "*CDs and Vinyl*", "*Kindle Store*", "*Movies and TV*", "*Prime Pantry*", "*Industrial and Scientific*", "*Musical Instruments*", "*Arts, Crafts and Sewing*", and "*Office Products*", from the standard benchmark dataset, **Amazon Review** [45]. To provide extensive evaluations of the transferability, We divide the former 5 and the latter 5 into pre-training and downstream target domains, respectively. We follow Hou et al. [25, 26] to pre-process for interaction data and retrieve text information including *title*, *categories* and *brand* from the metadata. To support multi-modal inputs, we further crawl item images according to the URLs in the metadata. The dataset statistics are reported in table 1. While texts are fully available in the metadata, there are many missing images in the dataset due to expired products or URLs. We still retain the items without images in our experiments to keep a fair comparison with previous works [25, 26], As we will show below, our MISSRec can exhibit robust improvement on incomplete multi-modal datasets.

*4.2.2 Metrics.* Following previous works [25, 26, 95], we adopt two standard metrics, *i.e.,* **Recall** (**R@**$K$) and **NDCG** (**N@**$K$), to evaluate the retrieval performance. We set $K$ to 10 and 50 for showcases.

*4.2.3 Baselines.* We compare our method with 7 state-of-the-art sequential recommenders, including (**i**) 2 pure ID-based recommenders: SASRec [31] and BERT4Rec [61]; (**ii**) 1 attribute-aware recommender, $S^3$-Rec [95]; (**iii**) 4 text-based recommenders, FDSA [89], ZESRec [13], UniSRec [26], and VQRec [25]. We derive an extra text-feature-based variant from SASRec regarding its simplicity and effectiveness. For $S^3$-Rec, we inherit previous works [25, 26] to embed the tags with textual features so as to support universal representation. We adopt both transductive and inductive variants of UniSRec for comparison. Therefore, we have 10 baselines in total.

*4.2.4 Implementation Details.* The training batch size is 2048. We take Adam as the common optimizer in our comparison, and we carefully choose suitable learning rates from $\{3e^{-4}, 1e^{-3}, 3e^{-3}, 1e^{-2}\}$ for different baselines. We take 100 epochs for pre-training. And for fine-tuning, we implement early stopping with a patience of 10 to prevent over-fitting. We adopt CLIP-B/32 [51, 71] as the basic feature encoder to extract `[CLS]` features for each text and image. The modality adapters then transform the features into the 300-d latent space for sequence representation. To keep a consistent parameter scale with baselines, we build our model with 1 Transformer encoder block and 1 decoder block, *i.e.,* $K_{\text{e}} = K_{\text{d}} = 1$, each with 2 attention heads. The loss weights in eqs. (7) and (8) are set $\lambda = 1$ and $\gamma = 0.1$, respectively. The temperature factor for contrastive learning objectives, *e.g.*, eqs. (5) and (6), is set $\tau = 0.075$. The concentration factor $\alpha$ in §3.2.4 is set to a learnable variable.

**Table 2: Comparisons on different target datasets. "T" and "V" stands for text and visual features. "*Improv.*" denotes the relative improvement of MISSRec to the best baselines. The best and second-best results are in bold and underlined, respectively.**

| Input Type & Model → | | ID | | T+ID | | | T+V+ID | *Improv.* w/ ID | T | | | | T+V | *Improv.* w/o ID |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Metric | SASRec | BERT4Rec | FDSA | $S^3$-Rec | UniSRec | MISSRec | | SASRec | ZESRec | UniSRec | VQRec | MISSRec | |
| Scientific | R@10 | 0.1080 | 0.0488 | 0.0899 | 0.0525 | 0.1235 | **0.1304** | 5.58% | 0.0994 | 0.0851 | 0.1188 | 0.1211 | **0.1277** | 5.42% |
| | N@10 | 0.0553 | 0.0243 | 0.0580 | 0.0275 | 0.0634 | **0.0675** | 6.42% | 0.0561 | 0.0475 | 0.0641 | 0.0643 | **0.0652** | 1.47% |
| | R@50 | 0.2042 | 0.1185 | 0.1732 | 0.1418 | 0.2473 | **0.2556** | 3.37% | 0.2162 | 0.1746 | 0.2394 | 0.2369 | **0.2496** | 4.25% |
| | N@50 | 0.0760 | 0.0393 | 0.0759 | 0.0468 | 0.0904 | **0.0947** | 4.80% | 0.0815 | 0.0670 | 0.0903 | 0.0897 | **0.0919** | 1.78% |
| Pantry | R@10 | 0.0501 | 0.0308 | 0.0395 | 0.0444 | 0.0693 | **0.0743** | 7.27% | 0.0585 | 0.0454 | 0.0636 | 0.0660 | **0.0743** | 12.63% |
| | N@10 | 0.0218 | 0.0152 | 0.0209 | 0.0214 | 0.0311 | **0.0342** | 10.09% | 0.0285 | 0.0230 | 0.0306 | 0.0293 | **0.0359** | 17.17% |
| | R@50 | 0.1322 | 0.1030 | 0.1151 | 0.1315 | 0.1827 | **0.1948** | 6.64% | 0.1647 | 0.1141 | 0.1658 | 0.1753 | **0.1904** | 8.61% |
| | N@50 | 0.0394 | 0.0305 | 0.0370 | 0.0400 | 0.0556 | **0.0602** | 8.27% | 0.0523 | 0.0378 | 0.0527 | 0.0527 | **0.0588** | 11.54% |
| Instruments | R@10 | 0.1118 | 0.0813 | 0.1070 | 0.1056 | 0.1267 | **0.1314** | 3.71% | 0.1127 | 0.0783 | 0.1189 | 0.1222 | **0.1264** | 3.40% |
| | N@10 | 0.0612 | 0.0620 | **0.0796** | 0.0713 | 0.0748 | 0.0754 | - | 0.0661 | 0.0497 | 0.0680 | 0.0758 | **0.0773** | 2.02% |
| | R@50 | 0.2106 | 0.1454 | 0.1890 | 0.1927 | 0.2387 | **0.2463** | 3.20% | 0.2104 | 0.1387 | 0.2255 | 0.2343 | **0.2412** | 2.94% |
| | N@50 | 0.0826 | 0.0756 | 0.0972 | 0.0901 | 0.0991 | **0.1004** | 1.31% | 0.0873 | 0.0627 | 0.0912 | 0.1002 | **0.1021** | 1.87% |
| Arts | R@10 | 0.1108 | 0.0722 | 0.1002 | 0.1003 | 0.1239 | **0.1301** | 4.99% | 0.0977 | 0.0664 | 0.1066 | 0.1189 | **0.1230** | 3.46% |
| | N@10 | 0.0587 | 0.0479 | 0.0714 | 0.0601 | 0.0712 | **0.0718** | 0.58% | 0.0562 | 0.0375 | 0.0586 | 0.0703 | **0.0706** | 0.41% |
| | R@50 | 0.2030 | 0.1367 | 0.1779 | 0.1888 | 0.2347 | **0.2459** | 4.79% | 0.1916 | 0.1323 | 0.2049 | 0.2249 | **0.2384** | 5.98% |
| | N@50 | 0.0788 | 0.0619 | 0.0883 | 0.0793 | 0.0955 | **0.0971** | 1.63% | 0.0766 | 0.0518 | 0.0799 | 0.0935 | **0.0947** | 1.32% |
| Office | R@10 | 0.1056 | 0.0825 | 0.1118 | 0.1030 | 0.1280 | **0.1301** | 1.63% | 0.0929 | 0.0641 | 0.1013 | 0.1236 | **0.1258** | 1.77% |
| | N@10 | 0.0710 | 0.0634 | **0.0868** | 0.0653 | 0.0831 | 0.0842 | - | 0.0582 | 0.0391 | 0.0619 | **0.0814** | 0.0795 | - |
| | R@50 | 0.1627 | 0.1227 | 0.1665 | 0.1613 | 0.2016 | **0.2091** | 3.71% | 0.1580 | 0.1113 | 0.1702 | 0.1957 | **0.2010** | 2.73% |
| | N@50 | 0.0835 | 0.0721 | 0.0987 | 0.0780 | 0.0991 | **0.1006** | 1.50% | 0.0723 | 0.0493 | 0.0769 | 0.0972 | **0.0977** | 0.51% |

## 4.3 Comparasion with State-of-the-arts (RQ1)

*4.3.1 Overall Performance.* The overall comparison results on five different downstream datasets are shown in table 2, from which we obtain three findings. (**i**) Text features can be exploited as an effective alternative or supplement to ID features, and pre-training can further enhance their efficacy. Comparing the SASRec variants with different types of input, the text-based variant achieves competitive performance and outperforms the ID-based variant on "Pantry" and "Instruments" datasets. Besides, text-enhanced approaches FDSA and $S^3$-Rec outperform ID-based BERT4Rec by large margins. By adopting pre-training, UniSRec and VQRec achieve further improvement. (**ii**) ID information is still important for personalization. Transductive methods generally perform better than inductive ones. Although $\text{VQRec}_\text{T}$ outperforms $\text{UniSRec}_\text{T}$ with text-only input, it is still inferior to $\text{UniSRec}_{\text{T+ID}}$ that can access ID information. (**iii**) Under both transductive and inductive settings, our MISSRec outperforms state-of-the-art baselines in most cases, and the margins are considerable. We attribute the performance gain to three factors. First, multi-modal information reflects more precise and comprehensive user preferences, benefiting personalization. Second, the synergy of different designs in MISSRec contributes to the efficacy of multi-modal sequence representation. Third, pre-training further strengthened the above advantages.

*4.3.2 Performance* w.r.t. *Different Item Frequencies.* In this sub-section, we select the best performers from table 2 and compare their efficacy *w.r.t.* long-tailed items. As shown in fig. 4, all modality-based methods demonstrate effectiveness to ID-based SASRec on the least frequent items, while the performance gains begin to vanish and even turn to degradation as the items warm up. Compared with other baselines, MISSRec with multi-modal learning shows the best few-short performance and also keeps robust improvement with more frequent items, suggesting its feasibility in addressing long-tailed and cold-start recommendation issues.

**Table 3: Model comparison *w.r.t.* various input types on the full-modality data subset. We can see the superior capability of MISSRec in leveraging modality features.**

| | | Scientific | | | | Office | | | |
|---|---|---|---|---|---|---|---|---|---|
| Input Type | Variant | R@10 | R@50 | N@10 | N@50 | R@10 | R@50 | N@10 | N@50 |
| ID+T | UniSRec | 0.1612 | 0.3223 | 0.0777 | 0.1132 | 0.1340 | 0.2150 | 0.0846 | 0.1021 |
| | MISSRec | **0.1633** | **0.3272** | **0.0793** | **0.1150** | **0.1362** | **0.2193** | **0.0860** | **0.1037** |
| ID+V | UniSRec | 0.1569 | 0.3134 | 0.0764 | 0.1113 | 0.1290 | 0.2069 | 0.0831 | 0.1004 |
| | MISSRec | **0.1594** | **0.3187** | **0.0774** | **0.1123** | **0.1339** | **0.2150** | **0.0854** | **0.1029** |
| ID+T+V | UniSRec | 0.1616 | 0.3232 | 0.0771 | 0.1119 | 0.1370 | 0.2201 | 0.0864 | 0.1047 |
| | MISSRec | **0.1646** | **0.3288** | **0.0793** | **0.1150** | **0.1384** | **0.2221** | **0.0872** | **0.1052** |
| T | UniSRec | 0.1602 | 0.3201 | 0.0772 | 0.1119 | 0.1342 | 0.2154 | 0.0840 | 0.1017 |
| | VQRec | 0.1607 | 0.3219 | 0.0778 | 0.1133 | 0.1329 | 0.2146 | 0.0833 | 0.1019 |
| | MISSRec | **0.1619** | **0.3238** | **0.0780** | **0.1136** | **0.1345** | **0.2161** | **0.0855** | **0.1036** |
| V | UniSRec | 0.1568 | 0.3134 | 0.0749 | 0.1090 | 0.1293 | 0.2074 | 0.0829 | 0.1000 |
| | VQRec | 0.1571 | **0.3149** | 0.0752 | **0.1100** | 0.1294 | 0.2084 | 0.0824 | 0.0997 |
| | MISSRec | **0.1575** | 0.3143 | **0.0756** | 0.1090 | **0.1306** | **0.2103** | **0.0838** | **0.1014** |
| T+V | UniSRec | 0.1609 | 0.3215 | 0.0774 | 0.1112 | 0.1367 | 0.2195 | 0.0860 | 0.1038 |
| | VQRec | 0.1617 | 0.3240 | 0.0769 | 0.1116 | 0.1354 | 0.2174 | 0.0862 | 0.1045 |
| | MISSRec | **0.1633** | **0.3264** | **0.0778** | **0.1131** | **0.1371** | **0.2203** | **0.0867** | **0.1047** |

## 4.4 Multi-modal & Pre-training Analyses (RQ2)

*4.4.1 Comparison with Different Input Modalities.* In this sub-section, we investigate the capacities of different modality-based sequence models in leveraging multi-modal information. Specifically, we filter out the items without images in "Scientific" and "Office" datasets.

**Table 4: Analysis of the effect of pre-training.**

| Model: MISSRec | | Scientific | | | | Office | | | |
|---|---|---|---|---|---|---|---|---|---|
| w/ ID? | Pre-train? | R@10 | R@50 | N@10 | N@50 | R@10 | R@50 | N@10 | N@50 |
| ✔ | ✔ | 0.1304 | 0.2556 | 0.0675 | 0.0947 | 0.1301 | 0.2091 | 0.0842 | 0.1006 |
| | ✗ | 0.1228 | 0.2426 | 0.0637 | 0.0905 | 0.1263 | 0.2045 | 0.0799 | 0.0975 |
| *Improv.* w/ ID | | 6.16% | 5.38% | 5.99% | 4.66% | 3.04% | 2.25% | 5.34% | 3.17% |
| ✗ | ✔ | 0.1277 | 0.2496 | 0.0652 | 0.0919 | 0.1258 | 0.2010 | 0.0795 | 0.0977 |
| | ✗ | 0.1216 | 0.2405 | 0.0626 | 0.0889 | 0.1243 | 0.1989 | 0.0789 | 0.0974 |
| *Improv.* w/o ID | | 5.00% | 3.77% | 4.23% | 3.39% | 1.19% | 1.08% | 0.72% | 0.26% |

Then we conduct experiments on the filtered subsets to enable horizontal comparison among different settings, *e.g.*, visual-only *vs* text-only. To ensure objective evaluation, we adopt the same CLIP-B/32 feature extractors for all comparison methods. We report the results in table 3. Interestingly, although utilizing visual modality alone does not lead to satisfactory results, the multi-modal synergy creates an effect of "1+1>2", which highlights the significance of multi-modal sequential modeling. Besides, MISSRec demonstrates robust and competitive performance under various settings.

*4.4.2 Pre-training Analysis.* Here we examine the effectiveness of pre-training to MISSRec. As shown in table 4, we can see that pre-training enhances the downstream recommendation in target domains in both transductive and inductive settings.

## 4.5 Model Analyses (RQ3)

To investigate the efficacy of different components, we construct 5 variants of $\text{MISSRec}_{\text{T+V+ID}}$ for comparison. Specifically, as presented by the leftmost column in table 5, variant (1) removes the sequence-sequence contrastive task (§3.5.2) from pre-training. Variant (2) removes the modality-specific adapters (§3.2.3) so that the sequence model directly consumes 512-d CLIP features. Variant (3) replaces the interest-aware decoder (§3.4.2) with another same context encoder block and produced sequence embedding by aggregating the encoder output. Variant (4) builds a 2-layer interest encoder that consumes interest tokens rather than item tokens for sequence representation. Variant (5) removes the orthogonal regularization in eqs. (7) and (8).

*4.5.1 Effects of Modality Adapters.* Comparing variants (2) with (0), we can see significant performance decays, owing to the semantic gap between general multi-modal features and user interests. As a result, multi-modal features themselves as item representations could not provide sufficient personalization for recommendation, and modality adapters are required. On the other hand, although end-to-end optimizing feature extractors with the recommender also helps to mitigate the gap, as revealed by [70, 84], it will consume much more computation resources. Anyway, parameter-efficient tuning with modality adapters can be preferable.

*4.5.2 Effects of Learning Objectives.* Comparing (1) and (0), we learn that $\ell_i^{\text{S-S}}$ in the pre-training stage can enhance the representation robustness, as it helps to improve downstream recommendation. Comparing (5) and (0), we find that the orthogonal regularization leads to slightly better results on the Scientific dataset but shows a negative impact on the Office dataset. We can infer that the orthogonal regularization is sensitive to the domain or the weighting factor, which requires careful choosing according to specific scenarios.

**Table 5: Ablation study with MISSRec.**

| | Scientific | | | | Office | | | |
|---|---|---|---|---|---|---|---|---|
| Variant | R@10 | R@50 | N@10 | N@50 | R@10 | R@50 | N@10 | N@50 |
| (0) MISSRec | **0.1304** | **0.2556** | **0.0675** | **0.0947** | 0.1301 | 0.2091 | 0.0842 | 0.1006 |
| (1) w/o $\ell_i^{S-S}$ in Eq.(7) | 0.1282 | 0.2513 | 0.0636 | 0.0893 | 0.1299 | 0.2085 | 0.0833 | 0.0996 |
| (2) w/o Modality Adapters | 0.1015 | 0.1994 | 0.0521 | 0.0732 | 0.1050 | 0.1686 | 0.0696 | 0.0832 |
| (3) 2-Layer Context Encoder | 0.1170 | 0.2299 | 0.0609 | 0.0858 | 0.1199 | 0.1928 | 0.0810 | 0.0968 |
| (4) 2-Layer Interest Encoder | 0.1123 | 0.2200 | 0.0579 | 0.0813 | 0.1023 | 0.1643 | 0.0684 | 0.0816 |
| (5) w/o Oth. Reg. in Eqs.(7-8) | 0.1299 | 0.2527 | 0.0668 | 0.0939 | **0.1312** | **0.2106** | **0.0859** | **0.1028** |

*4.5.3 Effects of Interest-aware Decoding and Contextual Encoding.* By comparing variants (3) and (0), we can learn that interest-aware decoding helps to learn better sequence representations, justifying the efficacy of the interest-aware decoder. Besides, variant (4) without accessing encoded context exhibits poor performance among the variants, suggesting the necessity of the collaboration between context encoding and interest-aware decoding.

## 5 CONCLUSIONS

This paper addresses the limitations of existing sequential recommendation models that rely heavily on ID features by exploring the potential of multi-modal information. We propose MISSRec, a novel multi-modal pre-training and transfer learning framework for sequential recommendation, which effectively tackles the cold-start problem and allows for efficient domain adaptation. By utilizing a transformer-based contextual encoder, an interest-aware decoder, a lightweight dynamic fusion module, and a modality-guided negative mining strategy, MISSRec demonstrates improved performance and generalizability compared to existing methods. Extensive experiments also showcase the compatibility and robustness of MISSRec in handling incomplete or missing modalities, reinforcing its pragmatic value for real-world scenarios. More importantly, our paper suggests a promising direction for future research in leveraging multi-modal information for sequential recommendation. We hope MISSRec can provide some inspiration to further explorations.

## ACKNOWLEDGMENTS

We want to thank the anonymous reviewers and the meta-reviewer for their valuable comments and suggestions. This research is supported by National Natural Science Foundation of China (Grant No.62171248 and Grant No.62276154), the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914), Shenzhen Science and Technology Program (JCYJ20220818101012025), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), the Major Key Project of PCL for Experiments and Applications (PCL2021A07, and PCL2021A06), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (HW2021008). We gratefully acknowledge the support of MindSpore[1], which is a new deep learning framework used for this research.

[1]https://www.mindspore.cn

# REFERENCES


[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *RecSys*.

[2] Ankur Bapna and Orhan Firat. 2019. Simple, Scalable Adaptation for Neural Machine Translation. In *EMNLP*.

[3] Desheng Cai, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2021. Heterogeneous hierarchical feature aggregation network for personalized micro-video recommendation. *IEEE TMM* (2021).

[4] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable multi-interest framework for recommendation. In *KDD*.

[5] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

[7] Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. 2022. Intent contrastive learning for sequential recommendation. In *WWW*.

[8] Mingyue Cheng, Fajie Yuan, Qi Liu, Xin Xin, and Enhong Chen. 2021. Learning transferable user representations with sequential behaviors via contrastive pre-training. In *ICDM*.

[9] Quanyu Dai, Yalei Lv, Jieming Zhu, Junjie Ye, Zhenhua Dong, Rui Zhang, Shu-Tao Xia, and Ruiming Tang. 2022. LCD: Adaptive Label Correction for Denoising Music Recommendation. In *CIKM*.

[10] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. *ACM CSUR* (2020).

[11] Yashar Deldjoo, Markus Schedl, Balázs Hidasi, Yinwei Wei, and Xiangnan He. 2021. Multimedia recommender systems: Algorithms and challenges. In *Recommender systems handbook*.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

[13] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2022. Zero-Shot Recommender Systems. In *ICLR Workshops*.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

[15] Mingjing Du, Shifei Ding, and Hongjie Jia. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Elsevier KBS* (2016).

[16] Xiaoyu Du, Xiang Wang, Xiangnan He, Zechao Li, Jinhui Tang, and Tat-Seng Chua. 2020. How to learn item representation for cold-start multimedia recommendation?. In *MM*.

[17] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE TPAMI* (2022).

[18] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a Unified View of Parameter-Efficient Transfer Learning. In *ICLR*.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

[20] Li He, Hongxu Chen, Dingxian Wang, Shoaib Jameel, Philip Yu, and Guandong Xu. 2021. Click-through rate prediction with multi-modal hypergraphs. In *CIKM*.

[21] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *ICDM*.

[22] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*.

[23] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*.

[24] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *ICLR*.

[25] Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning vector-quantized item representation for transferable sequential recommenders. In *WWW*.

[26] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *KDD*.

[27] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *ICML*.

[28] Xiaowen Huang, Shengsheng Qian, Quan Fang, Jitao Sang, and Changsheng Xu. 2018. Csan: Contextual self-attention network for user sequential recommendation. In *MM*.

[29] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. In *MM*.

[30] Xunqiang Jiang, Yuanfu Lu, Yuan Fang, and Chuan Shi. 2021. Contrastive pre-training of gnns on heterogeneous graphs. In *CIKM*.

[31] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*.

[32] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM CSUR* (2022).

[33] Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. 2021. Semi: A sequential multi-modal information transfer network for e-commerce micro-video recommendations. In *KDD*. 3161–3171.

[34] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *CIKM*.

[35] Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In *WSDM*.

[36] Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: multi-interest matching network for news recommendation. In *ACL Findings*.

[37] Jiahao Liang, Xiangyu Zhao, Muyang Li, Zijian Zhang, Wanyu Wang, Haochen Liu, and Zitao Liu. 2023. MMMLP: Multi-modal Multilayer Perceptron for Sequential Recommendations. In *WWW*.

[38] Guanyu Lin, Chen Gao, Yu Zheng, Jianxin Chang, Yanan Niu, Yang Song, Zhiheng Li, Depeng Jin, and Yong Li. 2023. Dual-interest Factorization-heads Attention for Sequential Recommendation. In *WWW*.

[39] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM CSUR* (2023).

[40] Qijiong Liu, Jieming Zhu, Quanyu Dai, and Xiaoming Wu. 2022. Boosting deep ctr prediction with a plug-and-play pre-trainer for news recommendation. In *COLING*.

[41] Yong Liu, Susen Yang, Chenyi Lei, Guoxin Wang, Haihong Tang, Juyong Zhang, Aixin Sun, and Chunyan Miao. 2021. Pre-training graph transformer with multi-modal side information for recommendation. In *MM*.

[42] Zhuang Liu, Yunpu Ma, Matthias Schubert, Yuanxin Ouyang, and Zhang Xiong. 2022. Multi-Modal Contrastive Pre-training for Recommendation. In *ICMR*.

[43] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *KDD*.

[44] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. 2021. SimpleX: A simple and strong baseline for collaborative filtering. In *CIKM*.

[45] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP*.

[46] Yabo Ni, Dan Ou, Shichen Liu, Xiang Li, Wenwu Ou, Anxiang Zeng, and Luo Si. 2018. Perceive your users in depth: Learning universal user representations from multiple e-commerce tasks. In *KDD*.

[47] Xingyu Pan, Yushuo Chen, Changxin Tian, Zihan Lin, Jinpeng Wang, He Hu, and Wayne Xin Zhao. 2022. Multimodal Meta-Learning for Cold-Start Sequential Recommendation. In *CIKM*.

[48] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *CIKM*.

[49] Zhaopeng Qiu, Xian Wu, Jingyue Gao, and Wei Fan. 2021. U-BERT: Pre-training user representations for improved recommendation. In *AAAI*.

[50] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-aware recommender systems. *ACM CSUR* (2018).

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[52] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *WWW*.

[53] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* (1997).

[54] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *SIGIR*.

[55] Kunzhe Song, Qingfeng Sun, Can Xu, Kai Zheng, and Yaming Yang. 2023. Self-Supervised Multi-Modal Sequential Recommendation. *arXiv preprint arXiv:2304.13277* (2023).

[56] Xuemeng Song, Chun Wang, Changchang Sun, Shanshan Feng, Min Zhou, and Liqiang Nie. 2023. MM-FRec: Multi-Modal Enhanced Fashion Item Recommendation. *IEEE TKDE* (2023).

[57] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR* (2014).

[58] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

[59] Yixin Su, Rui Zhang, Sarah Erfani, and Zhenghua Xu. 2021. Detecting beneficial feature interactions for recommender systems. In *AAAI*.

[60] Yixin Su, Yunxiang Zhao, Sarah Erfani, Junhao Gan, and Rui Zhang. 2022. Detecting arbitrary order beneficial feature interactions for recommender systems. In *KDD*.

[61] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM*.

[62] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *CIKM*.

[63] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WSDM*.

[64] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE TMM* (2022).

[65] Changxin Tian, Zihan Lin, Shuqing Bian, Jinpeng Wang, and Wayne Xin Zhao. 2022. Temporal Contrastive Pre-Training for Sequential Recommendation. In *CIKM*.

[66] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *NIPS* (2021).

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

[68] Chenyang Wang, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. Sequential recommendation with multiple contrast signals. *ACM TOIS* (2023).

[69] Hui Wang, Kun Zhou, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Curriculum Pre-Training Heterogeneous Subgraph Transformer for Top-N Recommendation. *ACM TOIS* (2023).

[70] Jie Wang, Fajie Yuan, Mingyue Cheng, Joemon M Jose, Chenyun Yu, Beibei Kong, Zhijin Wang, Bo Hu, and Zang Li. 2022. TransRec: Learning Transferable Recommendation from Mixture-of-Modality Feedback. *arXiv preprint arXiv:2206.06190* (2022).

[71] Jinpeng Wang, Ziyun Zeng, Bin Chen, Yuting Wang, Dongliang Liao, Gongfu Li, Yiru Wang, Shu-Tao Xia, and Peng Cheng Intelligence. 2022. Hugs Are Better Than Handshakes: Unsupervised Cross-Modal Transformer Hashing with Multi-granularity Alignment. In *BMVC*.

[72] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z. Sheng, and Mehmet Orgun. 2019. Sequential Recommender Systems: Challenges, Progress and Prospects. In *IJCAI*.

[73] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. *arXiv preprint arXiv:2302.10632* (2023).

[74] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *MM*.

[75] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *MM*.

[76] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation. In *SIGIR*.

[77] Chaojun Xiao, Ruobing Xie, Yuan Yao, Zhiyuan Liu, Maosong Sun, Xu Zhang, and Leyu Lin. 2021. UPRec: User-aware Pre-training for Recommender Systems. *arXiv preprint arXiv:2102.10989* (2021).

[78] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *ICDE*.

[79] Jiahao Xun, Shengyu Zhang, Zhou Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiuqiang He, Xiaofei He, Tat-Seng Chua, and Fei Wu. 2021. Why do we click: visual impression-aware news recommendation. In *MM*.

[80] Shiquan Yang, Rui Zhang, Sarah M Erfani, and Jey Han Lau. 2021. UniMF: A Unified Framework to Incorporate Multimodal Knowledge Bases into End-to-End Task-Oriented Dialogue Systems.. In *IJCAI*.

[81] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal graph contrastive learning for micro-video recommendation. In *SIGIR*.

[82] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. 2020. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In *SIGIR*.

[83] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *WSDM*.

[84] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to Go Next for Recommender Systems? ID-vs. Modality-based recommender models revisited. In *SIGIR*.

[85] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *MM*.

[86] Lingzi Zhang, Xin Zhou, and Zhiqi Shen. 2023. Multimodal Pre-training Framework for Sequential Recommendation via Contrastive Learning. *arXiv preprint arXiv:2303.11879* (2023).

[87] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. UNBERT: User-News Matching BERT for News Recommendation.. In *IJCAI*.

[88] Shengyu Zhang, Lingxiao Yang, Dong Yao, Yujie Lu, Fuli Feng, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2022. Re4: Learning to Re-contrast, Re-attend, Re-construct for Multi-interest Recommendation. In *WWW*.

[89] Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, Xiaofang Zhou, et al. 2019. Feature-level Deeper Self-Attention Network for Sequential Recommendation.. In *IJCAI*.

[90] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *SIGIR*.

[91] Qihang Zhao. 2022. RESETBERT4Rec: A pre-training model integrating time and user historical behavior for sequential recommendation. In *SIGIR*.

[92] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[93] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *KDD*.

[94] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions. *arXiv preprint arXiv:2302.04473* (2023).

[95] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-Rec: Self-supervised Learning for Sequential Recommendation with Mutual Information Maximization. In *CIKM*.

[96] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Filter-enhanced MLP is all you need for sequential recommendation. In *WWW*.

[97] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multi-Modal Recommendation. In *WWW*.

[98] Feng Zhu, Yan Wang, Chaochao Chen, Jun Zhou, Longfei Li, and Guanfeng Liu. 2021. Cross-domain recommendation: challenges, progress, and prospects. In *IJCAI*.