# Multimodal Recommender Systems: A Survey

QIDONG LIU*, Xi'an Jiaotong University, City University of Hong Kong, China
JIAXI HU*, City University of Hong Kong, Hong Kong SAR
YUTIAN XIAO*, City University of Hong Kong, Hong Kong SAR
JINGTONG GAO, City University of Hong Kong, Hong Kong SAR
XIANGYU ZHAO†, City University of Hong Kong, Hong Kong SAR

The recommender system (RS) has been an integral toolkit of online services. They are equipped with various deep learning techniques to model user preference based on identifier and attribute information. With the emergence of multimedia services, such as short video, news and *etc.*, understanding these contents while recommending becomes critical. Besides, multimodal features are also helpful in alleviating the problem of data sparsity in RS. Thus, **M**ultimodal **R**ecommender **S**ystem (MRS) has attracted much attention from both academia and industry recently. In this paper, we will give a comprehensive survey of the MRS models, mainly from technical views. First, we conclude the general procedures and major challenges for MRS. Then, we introduce the existing MRS models according to three categories, *i.e.,* Feature Interaction, Feature Enhancement and Model Optimization. To make it convenient for those who want to research this field, we also summarize the dataset and code resources. Finally, we discuss some promising future directions of MRS and conclude this paper.

CCS Concepts: • **Information systems** → **Data mining**; **Multimedia information systems**.

Additional Key Words and Phrases: Recommender Systems, Multi-Modal, Multi-Media

**ACM Reference Format:**
Qidong Liu, Jiaxi Hu, Yutian Xiao, Jingtong Gao, and Xiangyu Zhao. 2023. Multimodal Recommender Systems: A Survey. In . ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

With the advancement of the internet, many multimedia online services are emerging, such as fashion recommendation [9], music recommendation [11] and *etc.* Thanks to the development of multimodal research [1], multimodal recommender systems (MRS) have been devised and applied in recent years. On the one hand, MRS can handle different modality information, which is inherent in multimedia services. On the other hand, MRS can also make use of abundant multimodal information of items to alleviate the problems of data sparsity and cold start, which widely exist in recommender systems.

In general, recommender systems utilize collaborative or side information, which refers to the identifier (abbreviated to id) and tabular features of items. By comparison, in an MRS, multimodal features, such as image, audio and text, play a vital role. For simplicity, we define the MRS as: *a recommender system for the items with multimodal features*.

*These authors contributed equally to this research.
†Corresponding Author.

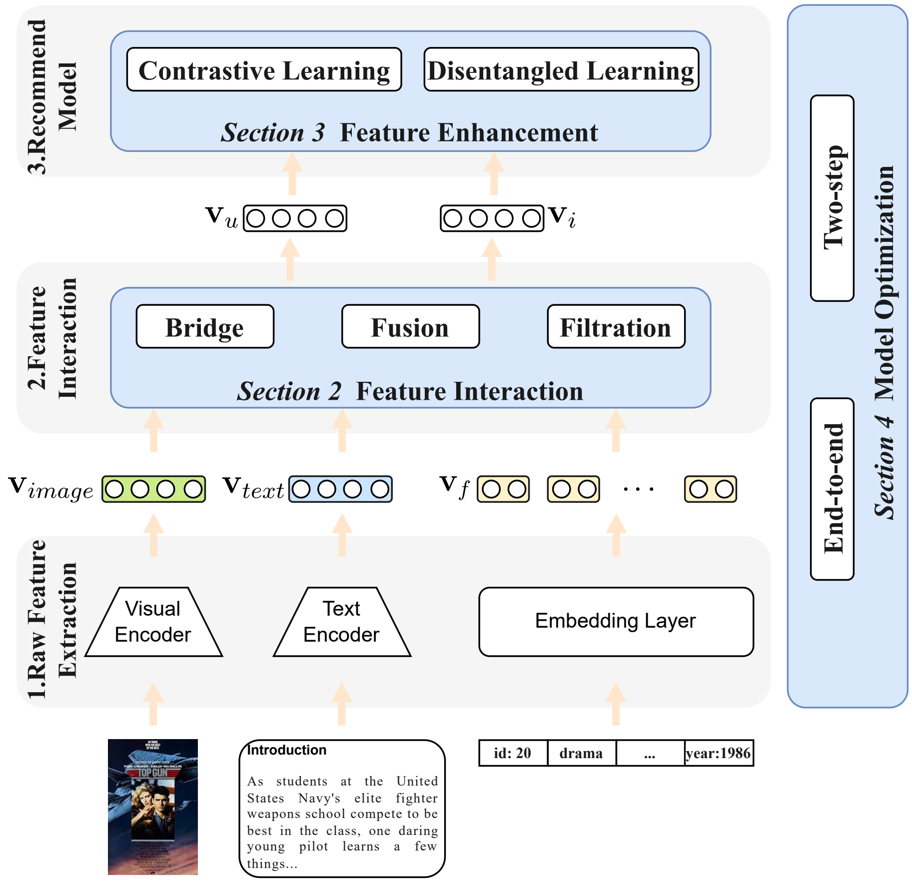



Fig. 1. The general procedures of multimodal recommendation.

More and more researches focus on MRS, so a review to survey and categorize them is in urgent need. Though an existing review [10] has taken a good step, they organize the research following the different modalities in real applications. Unlike that review, we organize this survey from the aspect of techniques used in MRS. Also, we try to collect all recent works to help readers to know about the most recent advancement in this field. Then, we will introduce the general procedures and our taxonomy in the rest of this section to make the survey more readable.

### 1.1 Procedures of MRS

Based on the input items of MRS, we conclude the unified procedures for MRS, as Figure 1 shows. There are three procedures: **Raw Feature Extracting**, **Feature interaction** and **Recommendation**. We take the movie recommendation as an example to illustrate it as follows:

**Raw Feature Extracting**. Each movie has two types of features, *i.e.,* tabular features and multimodal features, including the poster image and textual introduction. The embedding layers are used to handle tabular features, which is similar to common content-based RS [19]. The multimodal features are fed into different modality encoders. The modality encoders extract the representations and are general architectures used in other fields, such as ViT [13] for images and

Table 1. Category for Multimodal Recommender Systems.

| **Applications** | **Model** | **Feature Interaction** (Section 2) | **Feature Enhancement** (Section 3) |
|---|---|---|---|
| **General** | [34] | Coarse-grained Attention | CL |
| | **[40]** | Coarse-grained Attention | None |
| | [6], [21] | Fine-gained Attention | None |
| | [30], [27], [57] | Combined Attention | None |
| | [44], **[39]** | User-item Graph + Fine-gained Attention | None |
| | [56] | User-item Graph | CL |
| | [59] | Item-item Graph | CL |
| | [58], [38] | Item-item Graph | None |
| | **[33]** | Item-item Graph + Fine-gained Attention | None |
| | [50], [45] | Knowledge Graph | None |
| | [2], [46] | Knowledge Graph | CL |
| | [8] | Knowledge Graph + Fine-gained Attention | None |
| | [43] | Knowledge Graph + Filtration (graph) | None |
| | [63], [55], [31] | Filtration (graph) | None |
| | [49], [4] | MLP / Concat | DRL |
| | [15], [28] | Fine-gained Attention | DRL |
| | [61], [36], [48] | None | DRL |
| **Video** | [32], [7] | Coarse-grained Attention | None |
| | [21], [26] | Fine-gained Attention | None |
| | **[24]** | Fine-gained Attention | CL |
| | [47], [51] | User-item Graph | None |
| | [54] | User-item Graph | CL |
| | [29] | Knowledge Graph + Fine-gained Attention | None |
| **Fashion** | [3], [37], [59] | Item-item Graph | CL |
| | **[5]**, [18],[25] | Fine-gained Attention | None |
| | [64] | Fine-gained Attention | CL |
| | [35] | Linear Layer | None |
| **News** | [14] | Combined Attention | None |
| | [52] | Fine-gained Attention | None |
| **Restaurant** | [60], [53] | MLP / Concat | None |

[1] Model Optimization (Section 4): all of the two-step training models are bolded in reference, while others are optimized in End-to-End manner.

Bert [12] for texts. Then, we can get the representations of tabular features and multimodal features (*i.e.,* image and text) for each item, denoted as $\mathbf{v}_f, \mathbf{v}_{image}$ and $\mathbf{v}_{text}$.

**Feature Interaction**. We get the representations of different modalities for each item, but they are in different semantic spaces. Besides, different users also have various preferences for modalities. Therefore, in this procedure, MRS seeks to fuse and interact multimodal representations $\mathbf{v}_f, \mathbf{v}_{image}$ and $\mathbf{v}_{text}$ to get the item and user representations, which are important for recommendation models.

**Recommendation**. After the second procedure, we get the representations of the user and item, denoted as $\mathbf{v}_u$ and $\mathbf{v}_i$. The general recommendation models absorb these two representations and give the recommendation probability, such as

MF [22]. However, the problem of data sparsity always degrades the recommendation performance. Therefore, many researches propose to enhance the representations by incorporating multimodal information.

### 1.2 Taxonomy

According to the procedures mentioned above, we conclude three challenges as follows:

- *Challenge 1.* How to fuse the modality features in different semantic spaces and get the preferences for each modality.
- *Challenge 2.* How to get comprehensive representations for recommendation models under the data-sparse condition.
- *Challenge 3.* How to optimize the lightweight recommendation models and parameterized modality encoder.

The most recent works aim to propose efficient techniques to solve these three challenges, so we categorize them into three corresponding taxonomies: **Feature Interaction**, **Feature Enhancement** and **Model Optimization**. The first category focuses on *Challenge 1*, which fuses and captures various modality features mainly by GNN and attention techniques. As shown in Figure 1, these works always aim at the second procedure in MRS. The researches on feature enhancement propose using contrastive learning and disentangled learning to enhance the representations of item and user with limited data, which refers to *Challenge 2*. It is worth noting that these works are always closely related to the recommendation procedure in MRS. To handle *Challenge 3*, the model optimization works design efficient methods to train the model with lightweight recommendation models and parameterized modality encoder. These works are involved in the whole process of the MRS. As we know, we are the first to organize MRS works in such a technical way. We conclude existing works based on the category in Table 1.

## 2 FEATURE INTERACTION

Multimodal data refers to various modalities of description information. Since they are sparse and in different semantic spaces, connecting them to the recommendation task is essential. The feature interaction can realize the nonlinear transformation of feature space to common space, finally elevating performance and generalization of recommendation model. As shown in Figure 2, we categorize interactions into three types: **Bridge**, **Fusion**, and **Filtration**. These three types of techniques implement interaction from various views, so they can be applied to one MRS model simultaneously.

### 2.1 Bridge

Here the Bridge refers to the construction of a multimodal information transfer channel. It focus on capturing the inter-relationship between users and items considering multimodal information. The difference between multimedia recommendation and traditional recommendation is that the items contain rich multimedia information. Most early works simply use multimodal content to enhance the item expression, but they often ignore the interactions between users and items. The message-passing mechanism of graph neural network can enhance user representation through information exchange between user and item and further capture the user's preference for different modal information. Figure 2(a) gives an example: many works get user 1 preference by aggregating interacted items for each modality. Besides, the modality representation of movie 1 can be derived from the latent item-item graph. In this subsection, we will introduce the methods for how to build bridges in MRS.

*2.1.1 User-item Graph.* Leveraging the information exchange between users and items, users' preferences for different modalities can be captured. Therefore, some works utilize the user-item graph. MMGCN [51] establishes a user-item bipartite graph for each modality. For each node, the topology of adjacent nodes and the modality information of the item can be used to update the feature expression of the node. Based on MMGCN, GRCN [55] improves the performance

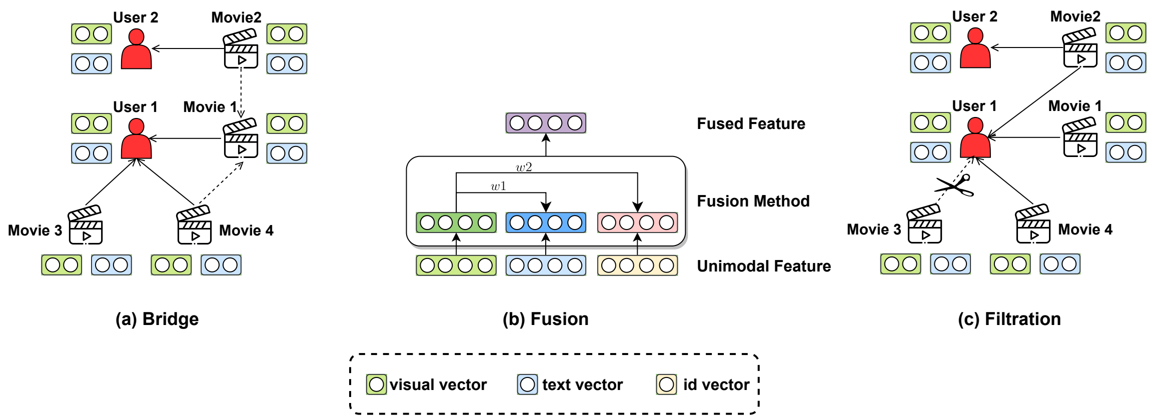



Fig. 2. The illustration to three types of feature interaction.

of recommendations by adaptively modifying the graph's structure during model training to delete incorrect interaction data (users clicked uninteresting videos). Although these methods have achieved great success in performance, these methods are still limited by using a unified way to fuse user preferences of different modalities, ignoring the difference in the degree of user preference for different modalities. In other words, giving equal weight to each modality may result in the sub-optimal performance of the model. To solve this problem, DualGNN [47] utilizes the correlation between users to learn user preferences based on the bipartite and user co-occurrence graph. Also, MMGCL [54] designs a new multimodal graph contrastive learning method to solve this problem. MMGCL uses modal edge loss and modal masking to generate user-item graphs and introduces a novel negative sampling technique to learn the correlation between modalities. MGAT [44] introduces an attention mechanism based on MMGCN, which is conducive to adaptively capturing user preferences for different modalities. Moreover, MGAT uses the gated attention mechanism to judge the user's preference for different modalities, which can capture relatively complex interaction patterns hidden in user behavior.

*2.1.2 Item-item Graph.* The above works focus on using multimodal features to model user-item interactions while ignoring latent semantic item-item structures. Reasonable use of item-item structures is conducive to better learning item representation and improving model performance. For instance, LATTICE [58] constructs an item-item graph for each modality based on the user-item bipartite graph. It aggregates them to obtain latent item graphs. MICRO [59] also constructs an item-item graph for each modality. Unlike LATTICE, MICRO adopts a new comparison method to fuse features after performing graph convolution. However, these works do not take into account the differences in preferences between various specific user groups. Furthermore, HCGCN [38] proposes a clustering graph convolutional network, which first groups item-item and user-item graphs and then learns user preferences through dynamic graph clustering. Besides, inspired by the recent success of pre-training models, PMGT [33] proposes a pre-trained graph transformer referring to Bert's structure and provides a unified view of project relationships and their associated side information in a multimodal form. BGCN [3], as a model in bundle recommendation, unifies user-item interaction, user-bundle interaction, and bundle-item affiliation into a heterogeneous graph, using graph convolution to extract fine-gained future. Cross-CBR [37] builds the user-bundle graph, the user-item diagram, and the Item-bundle graph, using contrastive learning to align them from the bundle and item views.

*2.1.3 Knowledge Graph.* Knowledge graphs (KG) are widely used because they can provide auxiliary information for recommender systems. To combine the KG and MRS, many researchers introduce each modality of items to KG as an entity. MKGAT [43] is the first model to introduce a knowledge graph into the multimodal recommendation. MKGAT proposes a multimodal graph attention technique to model multimodal knowledge graph from two aspects of entity information aggregation and entity relationship reasoning, respectively. Furthermore, a novel graph attention network is adopted to aggregate neighboring entities while considering the relations in the knowledge graph. SI-MKR [50] proposes an enhanced multimodal recommendation method based on alternate training and the knowledge graph representation based on MKR [45]. Besides, most multimodal recommender systems ignore the problem of data type diversity. SI-MKR solves it by adding user and item attribute information from the knowledge graph. By comparison, MMKGV [29] adopts a graph attention network for information dissemination and information aggregation on a knowledge graph, which combines multimodal information and uses the triplet reasoning relationship of the knowledge graph. CMCKG [2] treats information from descriptive attributes and structural connections as two modals and learns node representation by maximizing consistency between these two views.

## 2.2 Fusion

In the multimodal recommendation scenario, the types and quantities of multimodal information of users and items are very large. So, it is necessary to fuse the different multimodal information to generate the feature vector for the recommendation task. Compared with bridge, fusion concerns more about the multimodal intra-relationships of items. To be specific, it aims at combining various preferences to modalities. Since the inter- and intra-relationships are vital to learn comprehensive representations, many MRS models [39, 44] even adopt both fusion and bridge. The attention mechanism is the most widely used feature fusion method, which can flexibly fuse multimodal information with different weights and focus. In this subsection, as shown in Figure 2(b), we first divide attention mechanisms by fusion granularity and then introduce some of the other fusion approaches that exist in the MRS.

*2.2.1 Coarse-grained Attention.* Some models apply attention mechanisms to fuse information from multiple modalities at a coarse-grained level. For example, UVCAN [32] divides multimodal information into user-side and item-side, including their respective id information and side information. It uses multimodal data on the user side to generate fusion weights for the item side through self-attention. Based on UVCAN, MCPTR [34] proposes to merge item and user information in parallel. In addition to the user and item sides, some models merge information from different modal aspects. CMBF [7] introduces the cross-attention mechanism to co-learn the semantic information of image and text modality, respectively, and then concat them. Besides, the proportions of various modals are also different in some models. MML [40] designs an attention layer based on id information and is assisted by visual and text information. In MCPTR [34], each modal occupies the same position, and the self-attention mechanism determines the fusion weight. By comparison, HCGCN [38] pays more attention to the visual and text information of the item itself.

*2.2.2 Fine-gained Attention.* The multimodal data contains both global and fine-grained features, such as the tone of an audio recording or the pattern on a piece of clothing. Since coarse-grained fusion is often invasive and irreversible [27], it will damage the original modality's information and degrade the recommendation performance. Therefore, some works consider fine-grained fusion, which selectively fuses fine-grained feature information between different modalities.

Fine-grained fusion is significant in the fashion recommendation scenario. POG [5] is a sizeable online clothing RS based on transformer architecture. In the encoder, it excavates the deep features belonging to the collocation scheme in fashion images through multi-layer attention, which continuously realizes fine-grained integration. Compared with

POG, NOR [25] applies both encoder-decoder transformer architecture, which contains fine-grained self-attention structures. It can generate the corresponding scheme description according to collocation information. Besides, to increase interpretability, EFRM [18] also designs a Semantic Extraction Network (SEN) to extract the local features, and finally fuses the two features with fine-grained attention preference. VECF [6] performs image segmentation to integrate image features of each patch with other modalities. UVCAN [26] makes image segmentation of video screenshots like VECF and fuse image patches with id information and text information through the attention mechanism, respectively. MM-Rec [52] first extracts the region of interest from the image of news through the target detection algorithm Mask-RCNN, and then fuses POI with news content using co-attention.

Some models design unique internal structures for better fine-grained fusion. For instance, MKGformer [8] achieves fine-grained fusion by sharing some QKV parameters and a related perceptual fusion module. MGAT [44] uses a gated attention mechanism to focus on the user's local preferences. MARIO [21] predicts user preferences by considering the individual impact of each modality on each interaction. So the model designs a modality-aware attention mechanism to identify the influence of various modality on each interaction and do point multiplication for different modalities.

#### 2.2.3 *Combined Attention.*

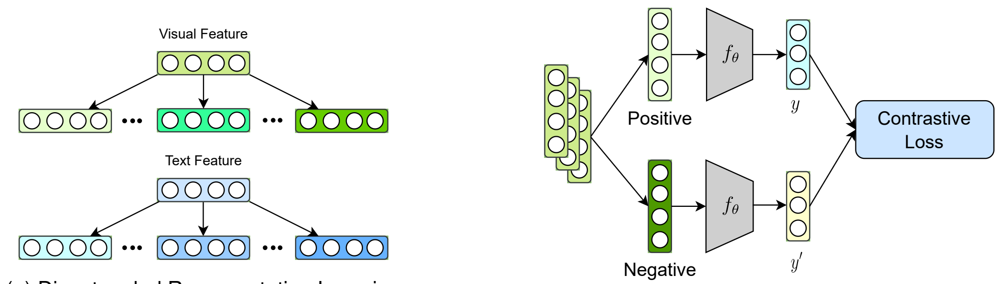
Based on fine-grained fusion, some models design combined fusion structures, hoping that the fusion of fine-grained features can also preserve the aggregation of global information. NOVA [27] introduces side information to sequential recommendation. It points out that directly fusing different modal features with vanilla attention usually brings little effect or even degrades performance. So it proposes a non-invasive attention mechanism with two branches, id embedding in a separated one to preserve interactive information in the fusion process. NRPA [30] offers a personalized attention network, which considers user preferences represented by user comments. It uses personalized word-level attention to select more important words in comments for each user/item, and passes the comment information attention layer through fine-grained and coarse-grained fusion in turn. VLSNR [14] is another application of sequential recommendation - news recommendation. It can model users' temporary and long-term interests and realize fine-grained and coarse-grained fusion by multi-head attention and GRU network. MARank [57] designs a Multi-order Attention layer, which combines the Attention and Resnet into a unified structure to fuse information.

#### 2.2.4 *Other Fusion Methods.*

In addition to fusing the multimodal information through attention weights, some works apply some simple methods, including average pooling, concat operations [60], and Gating mechanism [27]. Nevertheless, they rarely appear alone and often in combination with the graph and attention mechanisms, as mentioned above. Existing work [27] has shown that simple interactions, if appropriately used, will not damage the recommendation effect, and can reduce the complexity of the model. Besides, some early models adopt structures such as RNN [14] and LSTM [16], attempting to model user temporal preferences through multimodal information. However, with the development of deep learning techniques like attention mechanisms and CNN, they have been used less in these years. Some models fuse the multimodal feature through linear and nonlinear layers. Lv et.al. [35] set a linear layer at the place to fuse the textual and visual features. In MMT-Net [23], three context invariants of restaurant data are artificially marked, and interaction is carried out through a three-layer MLP network.

### 2.3 Filtration

As multimodal data differs from user interaction data, it contains much information unrelated to user preferences. For example, as shown in Figure 2(c), the interaction between movie 3 and user 1 is noisy, which should be removed. Filtering out noisy data in multimodal recommendation tasks can usually improve the recommendation performance. It is worth




(a) Disentangled Representation Learning

(b) Contrastive Learning

Fig. 3. The illustration to two types of feature enhancement.

noting that noise can exist in interaction graph or multimodal feature itself, so filtration can be embedded in the bridge and fusion, respectively.

Some models use image processing to denoise. For example, VECF [6] and UVCAN [26] perform image segmentation to remove noise from the image so that they can better model the user's personalized interests. MM-Rec [52] uses a target detection algorithm to select the significant margin of the image.

In addition, many structures based on graph neural networks are also used for denoising. Due to the sparsity of user-item interaction and noise of item features, the representation of users and items learned through graph aggregation inherently contain noise. FREEDOM [63] designs a degree-sensitive edge pruning method to denoise the user-item interaction graph. GRCN [55] detects whether the user accidentally interacts with a noisy item. Unlike the traditional GCN model, GRCN can adaptively adjust the structure of the interaction graph during training to identify and prune wrong interaction information. PMGCRN [20] also takes user interactions with uninterested items into account, but unlike GRCN, it handles mismatched interactions with an active attention mechanism to correct users' wrong preferences. Besides, MEGCF [31] focuses on the mismatch problem between multimodal feature extraction and user interest modeling. It firstly constructs a multimodal user-item graph, then uses sentiment information from comment data to fine-grained weight neighbor aggregation in the GCN module to filter information.

## 3 MULTIMODAL FEATURE ENHANCEMENT

Different modality representations of the same object have unique and common semantic information. Therefore, the recommendation performance and generalization of MRS can be significantly improved if the unique and common characteristics can be distinguished. Recently, to solve this problem, some models are equipped with **Disentangled Representation Learning (DRL)** and **Contrastive Learning (CL)** to carry out feature enhancement based on interaction, as shown in Figure 3.

### 3.1 Disentangled Representation Learning

The features of different modalities have various importance to the user's preference on a particular factor of the target item in RS. However, the representations of different factors in each modality are often entangled, so many researchers have introduced decomposition learning techniques to dig out the meticulous factors in user preference, such as DICER [61], MacridVAE [36], CDR [4]. Besides, the multimodal recommendation is discovering helpful information formed by various hidden factors from multimodal data, which are highly entangled in complex ways. MDR [49] proposes a multimodal disentangled recommendation that can learn well-disentangled representations carrying complementary and standard information from different modalities. DMRL [28] considers the different contributions of various modality features for each disentanglement factor to capture user preferences. Furthermore, PAMD [15] designs a disentangled encoder to extract their modality-common features while preserving their modality-specific features automatically. Besides, contrastive learning guarantees the consistency and gap between separated modal representations. Compared with MacridVAE, SEM-MacridVAE [48] considers item semantic information when learning disentangled representations from user behaviors.

### 3.2 Contrastive Learning

Unlike DRL, cotrastive learning methods enhance the representation by data augmentation, which is also helpful to handle the sparsity problem. Many works in MRS have introduced CL loss functions, mainly for the modality alignment and enhancing the deep feature information between positive and negative samples.

MCPTR [34] proposes a novel CL loss, which makes the different modal representations of the same item have semantic similarity. In addition, GHMFC [46] constructs two contrastive learning modules, based on the entity embedding representations derived from the graph neural network. The two CL loss functions are in two directions, *i.e.,* text to image and image to text. Cross-CBR [37] proposes a CL loss to align the graph representation from the bundle view and item view. While, MICRO [59] focus on shared modal information and specific modal information. In CMCKG [2], entity embeddings are obtained from both descriptive attributes and structural link information through knowledge graphs for contrastive loss. In HCGCN [38], to enforce visual and textual item features mapped into the same semantic space, it refers to CLIP [41] that adopts contrastive learning and maximizes the similarity of correct visual-textual pair in a batch. Besides, it sets up weights for different CL loss functions.

Because the core of contrastive learning is to mine the relationship between positive and negative samples, many models adopt data augment methods to construct positive samples in recommendation scenarios. MGMC [62] designs a graph enhancement to augment the samples and introduces meta-learning to increase model generalization. MML [40] is a sequential recommendation model that expands the training data by constructing a subset of the user's historical purchase item sequence. LHBPMR [38] selects items with similar preferences from the graph convolution to construct positive samples. MMGCL [54] constructs positive samples by modal edge loss and modal masking. And Victor [24] firstly constructs samples through Chinese semantics. Combo-Fashion [64] is a bundle fashion recommendation model so it contruct negtive and positive fashion matching scheme. Most of the existing models consider removing information that does not belong to user preferences in multimodal data. By comparison, QRec [56] starts from the opposite point of view, adding uniform noise to the multimodal information as a positive sample to improve the generalization of the model. Besides, though UMPR [53] does not have an explicit CL loss function, it also constructs a loss function that describes the difference between visual positive and negative samples.

## 4 MODEL OPTIMIZATION

Unlike traditional recommendation tasks, due to the existence of multimodal information, the computational requirements for model training are greatly increased when multimodal encoders and recommendation models are trained together. Therefore, the multimodal recommendation model can be divided into two categories during training: **End-to-end training** and **Two-step training**. As shown in Figure 4(a), End-to-end training can update the parameters of all layers in the model with each gradient obtained through backpropagation. While the two-step training includes the first stage of pretraining multimodal encoders and the second stage of task-oriented optimization, which is illustrated in Figure 4(b).

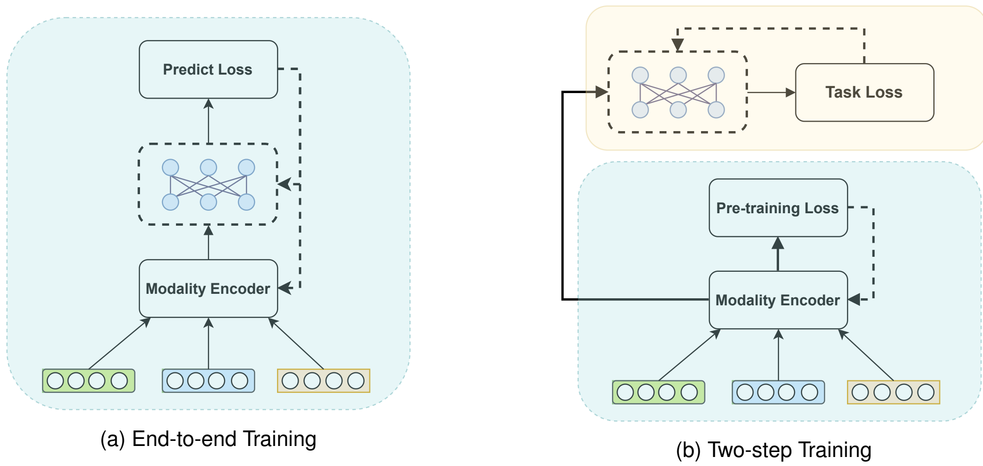



Fig. 4. The illustration to two types of training schemes.

### 4.1 End-to-end Training

Since multimodal recommender systems use pictures, texts, audio and other multimedia information, some common encoders in other fields, such as Vit [13], Resnet [17] , Bert [12], are often adopted when processing these multimodal data. The parameters of these pretrained models are often very huge. For example, the number of parameters of Vit-Base [13] reaches 86M, which is a great challenge for computing resources. To solve this problem, most MRS adopt pretrained encoders directly and only train the recommendation model in an end-to-end pattern. NOVA [27] and VLSNR [14] use a pretrained encoder to encode images and text features, then embed the resulting multimodal feature vectors through the model and recommends for users. They shows that introducing multimodal data without updating encoder parameters can also improve the recommendation performance. MCPTR [34] fine-tunes the encoder's parameters with only 100 epochs by recommendation and contrastive loss.

Some end-to-end methods also aim to reduce the amount of computation while improving the recommendation performance. They often decrease the number of parameters required to be updated when training. For instance, MKG-former [8] is a multi-layer transformer structure where many attention layer parameters are shared to reduce computation. FREEDOM [63] is designed to freeze some parameters of graph structure, dramatically reducing memory costs, and achieving a denoising effect to improve the recommendation performance.

### 4.2 Two-step Training

Compared with end-to-end pattern, two-stage training scheme can target downstream tasks better, but it requests much higher computing resources. Thus, few MRS adopts two-step training. PMGT [33] proposes a pretrained graph transformer referring to Bert's structure. It learns item representations with two objectives: graph structure reconstruction and masked node feature reconstruction. In POG [5], it trains a pretrained transformer to learn the fashion matching knowledge, and then recommends for users through a cloth generation model. Besides, it is common in sequential recommendation tasks where it is difficult to train the model in an end-to-end scheme. For example, in the pretraining stage, MML [40] first trains the meta-learner through meta-learning to increase model generalization, then trains the item embedding generator

Table 2. Summary of the MRS datasets.

| Data | Field | Modality | Scale | Link |
|---|---|---|---|---|
| Tiktok | Micro-video | V,T,M,A | 726K+ | https://paperswithcode.com/dataset/tiktok-dataset |
| Kwai | Micro-video | V,T,M | 1 million+ | https://zenodo.org/record/4023390#.Y9YZ6XZBw7c |
| Movielens + IMDB | Movie | V,T | 100k~25m | https://grouplens.org/datasets/movielens/ |
| Douban | Movie,Book,Music | V,T | 1 million+ | https://github.com/FengZhu-Joey/GA-DTCDR/tree/main/Data |
| Yelp | POI | V,T,POI | 1 million+ | https://www.yelp.com/dataset |
| Amazon | E-commerce | V,T | 100 million+ | https://cseweb.ucsd.edu/ jmcauley/datasets.html#amazon_reviews |
| Book-Crossings | Book | V,T | 1 million+ | http://www2.informatik.uni-freiburg.de/ cziegler/BX/ |
| Amazon Books | Book | V,T | 3 million | https://jmcauley.ucsd.edu/data/amazon/ |
| Amazon Fashion | Fashion | V,T | 1 million | https://jmcauley.ucsd.edu/data/amazon/ |
| POG | Fashion | V,T | 1 million+ | https://drive.google.com/drive/folders/1xFdx5xuNXHGsUVG2VIohFTXf9S7G5veq |
| Tianmao | Fashion | V,T | 8 million+ | https://tianchi.aliyun.com/dataset/43 |
| Taobao | Fashion | V,T | 1 million+ | https://tianchi.aliyun.com/dataset/52 |
| Tianchi News | News | T | 3 million+ | https://tianchi.aliyun.com/competition/entrance/531842/introduction |
| MIND | News | V,T | 15 million+ | https://msnews.github.io/ |
| Last.FM | Music | V,T,A | 186 k+ | https://www.heywhale.com/mw/dataset/5cfe0526e727f8002c36b9d9/content |
| MSD | Music | T,A | 48 million+ | http://millionsongdataset.com/challenge/ |

[1] 'V','T','M','A' indicates the visual data, the textual data, the video data, and acoustic data respectively.

in the second stage. Besides, TESM [39] and Victor [24] pretrain a well-designed graph neural network and a video transformer, respectively.

## 5 APPLICATION AND RESOURCES

In this section, we enumerate typical MRS applications and provide each application scenario's dataset. In addition, We introduce several open-source multimodal recommendation frameworks in the second subsection.

### 5.1 Application and Corresponding Dataset

Nowadays, when users browse the online shopping platform, they will receive a large amount of multimodal information about items, which will influence users' behavior imperceptibly. For example, In **fashion recommendation scenarios**, Users are often tempted to buy something they do not need because of the image of the cloth. In **movie recommendation scenarios**, users will only click on the content if they are attracted to the movie poster and title. Making full use of these multimodal data can improve the recommendation effect of the model. Besides, the quality of multimodal data directly affects the recommendation effect. So, as shown in Table 2, we summarize several popularly used datasets for MRS based on application scenarios. This will guide the researchers to obtain these MRS datasets conveniently. Anyone who wishes to use these datasets can refer to the corresponding citations and websites for more details.

### 5.2 Open Source Architecture

**MMRec**[1]: MMRec is a MultiModal Recommendation toolbox based on PyTorch. It integrates more than ten outstanding multimodal recommendation system models, such as MMGCN [51].

[1] https://github.com/enoche/MMRec

**Cornac**[2] [42]: Cornac is a comparative framework for multimodal recommender systems. It derives the whole experimental procedures for MRS, *i.e.,* data, models, metrics and experiment. Besides, cornac is highly compatible with mainstream deep learning frameworks such as TensorFlow and PyTorch.

## 6 CHALLENGES

To inspire the researchers who want to devote to this field, we list several existing challenges for promising researches:

- **A Universal Solution.** It is worth noting that though some methods for different stages in a model are proposed [24], there is no up-to-date universal solution with the combinations of these techniques provided.
- **Model Interpretability**: The complexity of multimodal models can make it difficult to understand and interpret the recommendations generated by the system, which can limit the trust and transparency of the system. Though few pioneers [6, 18] refer to it, but it is still need to be explored.
- **Computational Complexity**: MRS require large amounts of data and computational resources, making it challenging to scale to large datasets and populations. The complexity of multimodal data and models can increase the computational cost and time required for recommendation generation, making it challenging for real-time applications.
- **General MRS dataset.** Currently, the dataset of MRS is still limited, and the modalities covered are not extensive enough. In addition, the quality and availability of data for different modalities may vary, which can affect the accuracy and reliability of the recommendations.

## 7 CONCLUSION

Multimodal recommender system is becoming one of the leading research directions in recommender systems, benefitting from its aggregation advantage on different modalities. In this paper, we introduce taxonomies of multimodal recommendation models, *i.e.,* feature interaction, feature enhancement and model optimization based on challenges faced in different modeling stages. We also summarize the dataset and open-source codes to facilitate researchers. To conclude, we believe that by summarizing these taxonomies, techniques and resources, our survey will provide guidance and stimulate further research by raising awareness of different methods and featuring new topics.

## REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *TPAMI* (2018).

[2] Xianshuai Cao, Yuliang Shi, Jihu Wang, Han Yu, Xinjun Wang, and Zhongmin Yan. 2022. Cross-modal Knowledge Graph Contrastive Learning for Machine Learning Method Recommendation. In *Proc. of ACM MM*.

[3] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Bundle recommendation with graph convolutional networks. In *Proc. of SIGIR*.

[4] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. 2021. Curriculum Disentangled Recommendation with Noisy Multi-feedback. *Proc. of NeurIPS* (2021).

[5] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. POG: personalized outfit generation for fashion recommendation at Alibaba iFashion. In *Proc. of KDD*.

[6] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proc. of SIGIR*.

[7] Xi Chen, Yangsiyi Lu, Yuehai Wang, and Jianyi Yang. 2021. CMBF: Cross-Modal-Based Fusion Recommendation Algorithm. *Sensors* (2021).

[8] Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion. *arXiv preprint arXiv:2205.02357* (2022).

[9] Yashar Deldjoo, Fatemeh Nazary, Arnau Ramisa, Julian Mcauley, Giovanni Pellegrini, Alejandro Bellogin, and Tommaso Di Noia. 2022. A review of modern fashion recommender systems. *arXiv preprint arXiv:2202.02757* (2022).

[2]https://github.com/PreferredAI/cornac

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

[11] Yashar Deldjoo, Markus Schedl, and Peter Knees. 2021. Content-driven music recommendation: Evolution, state of the art, and challenges. *arXiv preprint arXiv:2107.11803* (2021).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] 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. *arXiv preprint arXiv:2010.11929* (2020).

[14] Songhao Han, Wei Huang, and Xiaotian Luan. 2022. VLSNR: Vision-Linguistics Coordination Time Sequence-aware News Recommendation. *arXiv preprint arXiv:2210.02946* (2022).

[15] Tengyue Han, Pengfei Wang, Shaozhang Niu, and Chenliang Li. 2022. Modality Matches Modality: Pretraining Modality-Disentangled Item Representations for Recommendation. In *Proceedings of the ACM Web Conference 2022*.

[16] Casper Hansen, Christian Hansen, Lucas Maystre, Rishabh Mehrotra, Brian Brost, Federico Tomasi, and Mounia Lalmas. 2020. Contextual and sequential user embeddings for large-scale music recommendation. In *Proc. of RecSys*.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*.

[18] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W Zheng, and Qi Liu. 2019. Explainable fashion recommendation: A semantic attribute region guided approach. *arXiv preprint arXiv:1905.12862* (2019).

[19] Umair Javed, Kamran Shaukat, Ibrahim A Hameed, Farhat Iqbal, Talha Mahboob Alam, and Suhuai Luo. 2021. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)* (2021).

[20] Xiangen Jia, Yihong Dong, Feng Zhu, Yu Xin, and Jiangbo Qian. 2022. Preference-corrected multimodal graph convolutional recommendation network. *Applied Intelligence* (2022).

[21] Taeri Kim, Yeon-Chang Lee, Kijung Shin, and Sang-Wook Kim. 2022. MARIO: Modality-Aware Attention and Modality-Preserving Decoders for Multimedia Recommendation. In *Proc. of CIKM*.

[22] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* (2009).

[23] Adit Krishnan, Mahashweta Das, Mangesh Bendre, Hao Yang, and Hari Sundaram. 2020. Transfer learning via contextual invariants for one-to-many cross-domain recommendation. In *Proc. of SIGIR*.

[24] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. 2021. Understanding chinese video and language via contrastive multimodal pre-training. In *Proc. of ACM MM*.

[25] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten De Rijke. 2019. Explainable outfit recommendation with joint outfit matching and comment generation. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[26] Bo Liu. 2022. Implicit semantic-based personalized micro-videos recommendation. *arXiv preprint arXiv:2205.03297* (2022).

[27] Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. 2021. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proc. of AAAI*.

[28] Fan Liu, Zhiyong Cheng, Huilin Chen, Anan Liu, Liqiang Nie, and Mohan Kankanhalli. 2022. Disentangled Multimodal Representation Learning for Recommendation. *arXiv preprint arXiv:2203.05406* (2022).

[29] Huizhi Liu, Chen Li, and Lihua Tian. 2022. Multi-modal Graph Attention Network for Video Recommendation. In *Proc. of CCET*.

[30] Hongtao Liu, Fangzhao Wu, Wenjun Wang, Xianchen Wang, Pengfei Jiao, Chuhan Wu, and Xing Xie. 2019. NRPA: neural recommendation with personalized attention. In *Proc. of SIGIR*.

[31] Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. 2022. MEGCF: Multimodal entity graph collaborative filtering for personalized recommendation. *TOIS* (2022).

[32] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-video co-attention network for personalized micro-video recommendation. In *Proc. of WWW*.

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

[34] Zhuang Liu, Yunpu Ma, Matthias Schubert, Yuanxin Ouyang, and Zhang Xiong. 2022. Multi-Modal Contrastive Pre-training for Recommendation. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*.

[35] Junmei Lv, Bin Song, Jie Guo, Xiaojiang Du, and Mohsen Guizani. 2019. Interest-related item similarity model based on multimodal data for top-N recommendation. *IEEE Access* (2019).

[36] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. *Proc. of NeurIPS* (2019).

[37] Yunshan Ma, Yingzhi He, An Zhang, Xiang Wang, and Tat-Seng Chua. 2022. CrossCBR: Cross-view Contrastive Learning for Bundle Recommendation. *arXiv preprint arXiv:2206.00242* (2022).

[38] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning Hybrid Behavior Patterns for Multimedia Recommendation. In *Proc. of ACM MM*.

[39] Juan Ni, Zhenhua Huang, Yang Hu, and Chen Lin. 2022. A two-stage embedding model for recommendation with multimodal auxiliary information. *Information Sciences* (2022).

[40] 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 *Proc. of CIKM*.

[41] 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 *Proc. of ICML*.

[42] Aghiles Salah, Quoc-Tuan Truong, and Hady W Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *Journal of Machine Learning Research* (2020).

[43] 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 *Proc. of CIKM*.

[44] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. MGAT: multimodal graph attention network for recommendation. *Information Processing & Management* (2020).

[45] Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2019. Multi-task feature learning for knowledge graph enhanced recommendation. In *Proc. of WWW*.

[46] Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022. Multimodal Entity Linking with Gated Hierarchical Fusion and Contrastive Training. In *Proc. of SIGIR*.

[47] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* (2021).

[48] Xin Wang, Hong Chen, Yuwei Zhou, Jianxin Ma, and Wenwu Zhu. 2023. Disentangled Representation Learning for Recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[49] Xin Wang, Hong Chen, and Wenwu Zhu. 2021. Multimodal disentangled representation for recommendation. In *Proc. of ICME*.

[50] Yuequn Wang, Liyan Dong, Hao Zhang, Xintao Ma, Yongli Li, and Minghui Sun. 2020. An enhanced multi-modal recommendation based on alternate training with knowledge graph representation. *Ieee Access* (2020).

[51] 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 *Proc. of ACM MM*.

[52] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Mm-rec: multimodal news recommendation. *arXiv preprint arXiv:2104.07407* (2021).

[53] Cai Xu, Ziyu Guan, Wei Zhao, Quanzhou Wu, Meng Yan, Long Chen, and Qiguang Miao. 2020. Recommendation by users' multimodal preferences for smart city applications. *IEEE Transactions on Industrial Informatics* (2020).

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

[55] Wei Yinwei, Wang Xiang, Nie Liqiang, He Xiangnan, and Chua Tat-Seng. 2021. GRCN: Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. *arXiv preprint arXiv:2111.02036* (2021).

[56] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proc. of SIGIR*.

[57] Lu Yu, Chuxu Zhang, Shangsong Liang, and Xiangliang Zhang. 2019. Multi-order attentive ranking model for sequential recommendation. In *Proc. of AAAI*.

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

[59] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent Structure Mining with Contrastive Modality Fusion for Multimedia Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[60] Xiaoyan Zhang, Haihua Luo, Bowei Chen, and Guibing Guo. 2020. Multi-view visual Bayesian personalized ranking for restaurant recommendation. *Applied Intelligence* (2020).

[61] Yin Zhang, Ziwei Zhu, Yun He, and James Caverlee. 2020. Content-collaborative disentanglement representation learning for enhanced recommendation. In *Fourteenth ACM Conference on Recommender Systems*.

[62] Feng Zhao and Donglin Wang. 2021. Multimodal graph meta contrastive learning. In *Proc. of CIKM*.

[63] Xin Zhou. 2022. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. *arXiv preprint arXiv:2211.06924* (2022).

[64] Chenxu Zhu, Peng Du, Weinan Zhang, Yong Yu, and Yang Cao. 2022. Combo-Fashion: Fashion Clothes Matching CTR Prediction with Item History. In *Proc. of KDD*.


Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009