SURVEY

Multimodal Recommender Systems: A Survey

QIDONG LIU, Xi'an Jiaotong University, Xi'an, Shaanxi, China

JIAXI HU, City University of Hong Kong, Hong Kong, Hong Kong

YUTIAN XIAO, City University of Hong Kong, Hong Kong, Hong Kong

XIANGYU ZHAO, City University of Hong Kong, Hong Kong, Hong Kong

JINGTONG GAO, City University of Hong Kong, Hong Kong, Hong Kong

WANYU WANG, City University of Hong Kong, Hong Kong, Hong Kong

View all

Published: 10 October 2024
Online AM: 10 September 2024
Accepted: 22 August 2024
Revised: 08 June 2024
Received: 26 June 2023

Multimodal Recommender Systems: A Survey

QIDONG LIU, Xi'an Jiaotong University, Xi'an, China and City University of Hong Kong, Hong Kong, Hong Kong
JIAXI HU, City University of Hong Kong, Hong Kong, Hong Kong
YUTIAN XIAO, City University of Hong Kong, Hong Kong, Hong Kong
XIANGYU ZHAO, City University of Hong Kong, Hong Kong, Hong Kong
JINGTONG GAO, City University of Hong Kong, Hong Kong, Hong Kong
WANYU WANG, City University of Hong Kong, Hong Kong, Hong Kong
QING LI, The Hong Kong Polytechnic University, Hong Kong, Hong Kong
JILIANG TANG, Michigan State University, East Lansing, United States

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 videos, news, and and so on, 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 article, 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 four categories, i.e., **Modality Encoder**, **Feature Interaction**, **Feature Enhancement**, and **Model Optimization**. Besides, 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 article. To access more details of the surveyed articles, such as implementation code, we open source a repository.[1]

[1]https://github.com/Applied-Machine-Learning-Lab/Awesome-Multimodal-Recommender-Systems

Q. Liu, J. Hu, and Y. Xiao contributed equally to this research.

This research was partially supported by Research Impact Fund (No.R1015-23), APRC - CityU New Research Initiatives (No.9610565, Start-up Grant for New Faculty of CityU), CityU - HKIDS Early Career Research Grant (No.9360163), Hong Kong ITC Innovation and Technology Fund Midstream Research Programme for Universities Project (No.ITS/034/22MS), Hong Kong Environmental and Conservation Fund (No. 88/2022), and SIRG - CityU Strategic Interdisciplinary Research Grant (No.7020046), Huawei (Huawei Innovation Research Program), Tencent (CCF-Tencent Open Fund, Tencent Rhino-Bird Focused Research Program), Ant Group (CCF-Ant Research Fund, Ant Group Research Fund), Alibaba (CCF-Alimama Tech Kangaroo Fund (No. 2024002)), CCF-BaiChuan-Ebtech Foundation Model Fund, and Kuaishou.

Authors' Contact Information: Qidong Liu, Xi'an Jiaotong University, Xi'an, China and City University of Hong Kong, Hong Kong, Hong Kong; e-mail: liuqidong@stu.xjtu.edu.cn; Jiaxi Hu, City University of Hong Kong, Hong Kong, Hong Kong; e-mail: jiaxihu2-c@my.cityu.edu.hk; Yutian Xiao, City University of Hong Kong, Hong Kong, Hong Kong; e-mail: yutiaxiao2-c@my.cityu.edu.hk; Xiangyu Zhao (Corresponding author), City University of Hong Kong, Hong Kong, Hong Kong; e-mail: xianzhao@cityu.edu.hk; Jingtong Gao, City University of Hong Kong, Hong Kong, Hong Kong; e-mail: jt.g@my.cityu.edu.hk; Wanyu Wang, City University of Hong Kong, Hong Kong, Hong Kong; e-mail: wanyuwang4-c@my.cityu.edu.hk; Qing Li, The Hong Kong Polytechnic University, Hong Kong, Hong Kong; e-mail: qing-prof.li@polyu.edu.hk; Jiliang Tang, Michigan State University, East Lansing, State of Michigan, United States; e-mail: tangjili@msu.edu.

ACM 0360-0300/2024/10-ART26
https://doi.org/10.1145/3695461

CCS Concepts: • **Information systems** → **Recommender systems**; **Multimedia and multimodal retrieval;**

Additional Key Words and Phrases: Recommender systems, multi-modal, multi-media

ACM Reference Format:
Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal Recommender Systems: A Survey. *ACM Comput. Surv.* 57, 2, Article 26 (October 2024), 17 pages. https://doi.org/10.1145/3695461

1 Introduction

With the advancement of the internet, many multimedia online services are emerging, such as fashion recommendation [9] and so on. These multimedia applications give a chance to push the RS towards the path of understanding recommended items, which is very beneficial. On the one hand, understanding can help RS make use of abundant multimodal information of items to alleviate the problems of data sparsity [9]. On the other hand, it assists the RS in knowing about the user's preference more deeply from a semantic level. Considering the prevalence of multimedia services, **multimodal recommender system** (**MRS**) is promising to become a general pattern of RS in the future. Therefore, more research has focused on MRS recently, and a review to survey and categorize them is urgently needed.

In general, the recommender system focuses on collaborative or side information, which refers to the identifier (abbreviated to id) and tabular features of items, such as genera and published year. By comparison, in an MRS, multimodal features, such as image, audio, and text, play a vital role. For simplicity, we define the MRS as *the recommender system for the items with multimodal features*. In the following subsections, we will introduce the general procedures and our taxonomy 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 Representation**, **Feature Interaction**, and **Recommendation**. We take the movie recommendation as an example to illustrate the general procedures as follows:

Raw Feature Representation. Each movie possesses two types of features: tabular features that describe its important characteristics using numerical values or classifications (such as genera or year), and multimodal features that depict the movie across various modalities of representation (such as poster images and textual introduction). To handle the tabular features, general recommender systems often adopt an embedding layer to transform sparse discrete features into dense vectors [18]. Specifically, the embedding layer treats each feature field as a discrete set and maps it to a fixed-length representation. However, multimodal features often have varying formats and cannot share such an embedding layer. Therefore, the multimodal features are often fed into different modality encoders to extract comprehensive representations. The modality encoders are often general architectures used in other fields, such as ViT [12] for images and Bert [11] 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 [65]. Therefore, in this procedure, MRS seeks to fuse and interact multimodal representations

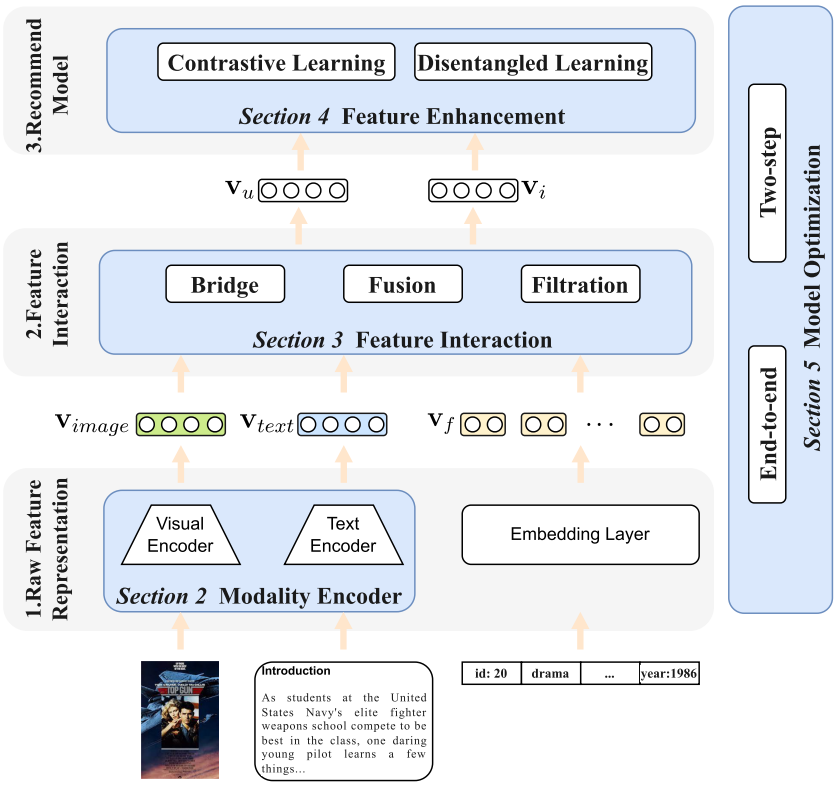

Fig. 1. The general procedures of multimodal recommender system.

$\mathbf{v}_f, \mathbf{v}_{image}$, and $\mathbf{v}_{text}$ to get a unified item and user representations, which are often used to get the recommendation list [13, 38].

Recommendation. After the second procedure, we get the representations of user and item, denoted as $\mathbf{v}_u$ and $\mathbf{v}_i$. The general recommendation models absorb these two representations and give the recommendation probabilities for different items [22]. However, the problem of data sparsity always degrades recommendation performance. Therefore, many research studies [34, 40] propose to enhance the representations by incorporating multimodal information.

1.2 Taxonomy

Multimodal features bring the chance to alleviate the problem of data sparsity. However, due to the complexity and heterogeneity of the multimodal features, there are some challenges for each procedure and the whole process of the MRS. Such challenges block the benefit of multimodal features and even have adverse effects on RS. Therefore, to make the best of multimodal information, existing research focuses on facing one or some of these challenges. Then, the major challenges and corresponding solutions are listed according to the procedure of MRS.

— ***Challenge 1****: For the raw multimodal inputs, e.g., images or texts, how to get representations from the complex modality features.* The embedding layer has been widely adopted in general RS to learn the representation from raw features. Nevertheless, it is hard to get representations from complex images or texts. For example, extracting useful information from the raw poster or introduction of movies is difficult but vital evidently. Thanks to the advancements

in computer vision and natural language processing, many encoders can be borrowed by MRS to get representations, such as ViT [12] for images. In this article, we denote all encoders that handle multimodal features as **Modality Encoder** (in Section 2).

— ***Challenge 2****: For the feature interaction procedure, how to fuse the modality features in different semantic spaces and get various preferences for each modality.* The heterogeneous nature of multimodal features causes difficulty in learning item representation and user preference for MRS. Case in point, the features of the movie poster and introduction are in totally distinct representation space, which hinders the imputation of the user's preference. To face this challenge, many works design the model by extracting the relationships between users, items, and modalities. These works are categorized into **Feature Interaction** technique (in Section 3).

— ***Challenge 3****: For the recommendation procedure, how to get comprehensive representations for recommendation models under the data-sparse condition.* Though the multimodal features enrich the information of items, the sparsity problem still exists because of the small volume of interaction records in RS. As an example, the sparsity of one typical movie recommendation dataset, Movielens-1M,[2] is beyond 95%. Compared with general RS, multimodal features can be utilized further to enhance the representation of the user and item. We denote this line of work as **Feature Enhancement** (in Section 4).

— ***Challenge 4:*** *For the whole process of MRS, how to optimize the lightweight recommendation models and parameterized modality encoder.* Taking a movie recommendation model [29] as an example, the parameter scale of its text encoder for movie introduction is about 110M, while the basic RS accounts for less than 10M. To solve the optimization problem, some MRS works propose novel techniques, which are clustered into **Model Optimization** (in Section 5).

Based on the four challenges mentioned above, we organize the rest of this article according to the corresponding technical solutions, i.e., **Modality Encoder**, **Feature Interaction**, **Feature Enhancement**, and **Model Optimization**. As far as we know, this survey is totally different from the existing two MRS surveys. One review [10] organized the research following the different modalities in real applications. The other latest survey [80] paid more attention to the RS itself while ignoring the characteristics of MRS. By comparison, our survey organizes the description concerning various types of techniques, especially for multimodal, which may help readers better understand the general MRS architecture. Also, we try to collect all recent works to help readers know about the recent advancements in this field.

2 Modality Encoder

The multimodal features of items are critical in constructing more specific user interests and enhancing model interpretability. In the context of recommendation tasks, item id information is typically represented by dense vectors obtained through a trainable embedding table. However, the multimodal features of items require corresponding encoders to obtain dense feature representations to extract more comprehensive information. In this section, in addition to pre-published feature vectors, we provide a brief introduction to commonly used encoders for three types of multimodal features, i.e., **Visual**, **Textual**, and **Other modalities**, and provide detailed encoder information for existing MRS models in Table 1.

— **Visual Encoder:** Visual feature is one vital modality in MRS, such as the poster for movie recommendation and clothing image for fashion recommendation. Most early MRS use a

[2] https://grouplens.org/datasets/movielens/

Table 1. Category for Modality Encoder

Modality	Category	Related Works
Visual Encoder	CNN	[5–7, 14, 20, 26, 30, 31, 33–37, 45, 61, 76, 84]
	ResNet	[1, 16, 17, 38, 39, 46, 47, 52, 58, 59, 65, 66]
	Transformer	[8, 13]
Textual Encoder	Word2vec	[2, 14, 36, 52, 59, 61, 65]
	RNN	[30, 68]
	CNN	[5, 62]
	Sentence-transformer	[20, 26, 45, 63, 64, 73, 74, 81, 84]
	Bert	[1, 8, 13, 17, 29, 32–34, 38, 39, 41, 46, 47, 58, 66]
Other Modality Encoder	Published Feature	[51, 53, 59, 63–65, 70–74, 83]

CNN-based pre-trained model as an image encoder. It compresses and extracts features through convolution and pooling operations from raw pixel information. For example, MMGCN [65] adopts ResNet [15] to extract information from the visual information of the news. POG [6] uses VGG as an image encoder for cloth pictures. Recently, visual pre-training models based on the transformer [54] architecture have achieved better performance, and some MRS models [8, 13] have started to use ViT [12] to extract visual features. So, we mainly categorize the visual encoder into CNN-based, ResNet-based, and Transformer-based.

— **Textual Encoder:** Textual information, such as item descriptions, often contain more semantic information than images, making them more suitable for enhancing user interest modeling [82]. Some MRS models [14, 52] typically utilize GloVe [48] or Word2Vec [44] to extract text features. In addition, some models also use encoders such as Text-CNN [21]. With the development of natural language models, text encoders have gradually been standardized to Bert [11]. In conclusion, we categorize textual encoders into five lines, i.e., Bert-based, RNN-based, CNN-based, Sentence-transformer-based, and Word2vec-based.

— **Other Modality Encoder:** For acoustic and video data, they are often converted into text or visual information before input into the textual or visual encoders. Besides, some methods directly use the pre-published feature vectors contained in original datasets to model the acoustic and video modalities.

3 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 various feature spaces of different modalities into common space, finally elevating the performance and generalization of the 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. For readability, we also categorize the existing works based on their interaction type in Table 2.

3.1 Bridge

Here bridge refers to the construction of a multimodal information transfer channel. It focuses on capturing the inter-relationship between users and items considering multimodal information. The difference between multimedia and traditional recommendation is that the items contain rich

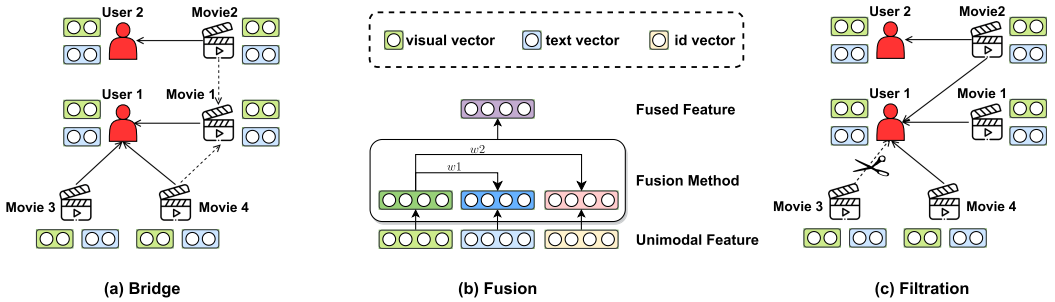

Fig. 2. The illustration of three types of feature interaction.

Table 2. Category for Feature Interaction

Interaction	Goal	Category	Related Works
Bridge	Capture inter-relationship between users and items	User-item Graph	[33, 46, 53, 59, 63–65, 70, 72, 83]
		Item-item Graph	[3, 17, 39, 43, 45, 72–74, 74, 84]
		KG	[2, 8, 35, 52, 55, 58, 62]
Fusion	Combine various preference to modalities	Coarse-grained Attention	[7, 38, 40, 47, 63, 72, 84]
		Fine-grained Attention	[6, 8, 14, 20, 20, 28, 34, 39, 46, 53, 67], [5, 16, 24, 25, 30, 31, 35, 66, 85]
		Combined Attention	[13, 17, 32, 36]
		Other Fusion Methods	[1, 4, 29, 41, 41, 61, 68, 76, 83]
Filtration	Filter out noisy data	Filtration	[26, 37, 51, 52, 69, 71, 79, 81]

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 networks can enhance user representation through information exchange between users and items and further capture the user's preference for different modal information. Figure 2(a) gives an example: many models 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.

3.1.1 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 [65] 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 [71] improves the performance of recommendations by adaptively modifying the graph's structure during model training to delete incorrect interaction data (users clicked uninterested 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 [59] utilizes the correlation between users to learn user preferences based on the bipartite and user co-occurrence graph. Also, MMGCL [70] 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 [53] 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 contained in user behaviors.

3.1.2 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 [73] constructs an item-item graph for each modality based on the user-item bipartite graph. It aggregates them to obtain the latent item graphs. MICRO [74] 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 [45] 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 [39] 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 the user-item interaction, user-bundle interaction, and bundle-item affiliation into a heterogeneous graph, using graph convolution to extract fine-gained features. Cross-CBR [43] builds the user-bundle graph, the user-item diagram, and the item-bundle graph, using **contrastive learning** (**CL**) to align them from the bundle and item views.

3.1.3 Knowledge Graph (KG). KGs 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 [52] is the first model to introduce a KG into the multimodal recommendation. MKGAT proposes a multimodal graph attention technique to model multimodal KG 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 KG. SI-MKR [62] proposes an enhanced multimodal recommendation method based on alternate training and the KG representation based on MKR [55]. Besides, most MRSs ignore the problem of data type diversity. SI-MKR solves it by adding user and item attribute information from the KG. By comparison, MMKGV [35] adopts a graph attention network for information dissemination and information aggregation on a KG, which combines multimodal information and uses the triplet reasoning relationship of the KG. CMCKG [2] treats information from descriptive attributes and structural connections as two modals and learns node representation by maximizing consistency between these two views.

3.2 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 [27]. Compared with bridge, fusion is more concerned about the multimodal intra-relationships of items. To be specific, it aims at combining various preferences into modalities. Since the inter- and intra-relationships are vital to learning comprehensive representations, many MRS models [46, 53] 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.

3.2.1 Coarse-grained Attention. Some models apply attention mechanisms to fuse information from multiple modalities at a coarse-grained level. For example, UVCAN [38] 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. 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 concatenate them. Besides, the proportions of various modals are also different in some models. MML [47] designs an attention layer based on id information and is assisted by visual and text information. Liu et al. [40] point out that each modal occupies the same position, and the self-attention mechanism determines the fusion weight. By comparison, HCGCN [45] pays more attention to the visual and text information of the item itself. Besides, MGCN [72] proposes behavior-aware attention to combining the modalities with distinct importance. Zhou et al. [84] highlight the difference between identity and other modalities and then design multi-level attention for fusion.

3.2.2 Fine-gained Attention. The multimodal data contains both global and fine-grained features, such as the tone of the audio recording or the pattern of clothing. Since coarse-grained fusion is often invasive and irreversible [32], it will damage the original modality information and degrade the recommendation performance. Therefore, some works consider fine-grained fusion, which selectively fuses fine-grained 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. The encoder 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 [30] 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 [16] 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 [31] conducts image segmentation of video screenshots like VECF and fuse image patches with id information and text information through the attention mechanism, respectively. MM-Rec [66] 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. MINER [25], DMIN [67], and SUM [28] all build interest representations of different aspects of the user by the multimodal information.

Some other 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 [53] uses a gated attention mechanism to focus on the user's local preferences. MARIO [20] predicts user preferences by considering the individual impact of each modality on each interaction. So, it designs a modality-aware attention mechanism to identify the influence of various modalities on each interaction and conducts point multiplication for different modalities.

3.2.3 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 [32] introduces side information to the 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 separate ones to preserve interactive information in the fusion process.

NRPA [36] 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 to the attention layer through fine-grained and coarse-grained fusion in turn. VLSNR [13] is another application of sequential recommendation, i.e., 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. Moreover, MMSR [17] employs dual attention to preserve modalities' temporal order for the sequential recommendation.

3.2.4 Other Fusion Methods. In addition to fusing the multimodal information through attention weights, some works apply other simple methods, including concat operations [76], and gating mechanism [32]. Nevertheless, they rarely appear alone and often in combination with the graph and attention mechanisms, as mentioned above. Existing work [32] 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 [13], attempting to model user temporal preferences through multimodal information. The other models fuse the multimodal feature through linear and nonlinear layers. Lv et al. [41] 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 MLPs. Recently, more fabricated architectures have been developed for fusion, such as the mixture of expert [1] and MLP mixer [29].

3.3 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 noting that noise can exist in the 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 [31] perform image segmentation to remove noise from the image so that they can better model the user's personalized interests. MM-Rec [66] 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 interactions and the noise of item features, the representation of users and items learned through graph aggregation inherently contain noise. FREEDOM [81] designs a degree-sensitive edge pruning method to denoise the user-item interaction graph. GRCN [71] detects whether the user accidentally interacts with a noisy item. Unlike the GCN model, GRCN can adaptively adjust the structure of the interaction graph during training to identify and prune wrong interaction information. PMGCRN [19] 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 [37] focuses on the mismatch problem between multimodal feature extraction and user interest modeling. It first constructs a multimodal user-item graph and then uses sentiment information from comment data to fine-grained weight neighbor aggregation in the GCN module to filter noise. MAGAE [69] is a model designed to handle uncertainty in multi-modal data. Besides, Shang et al. [51] first find out the problem of modality imbalance and propose to filter out the items with sensitive modalities. Then, MCLN [26] integrates a novel counterfactual framework to eliminate the noise.

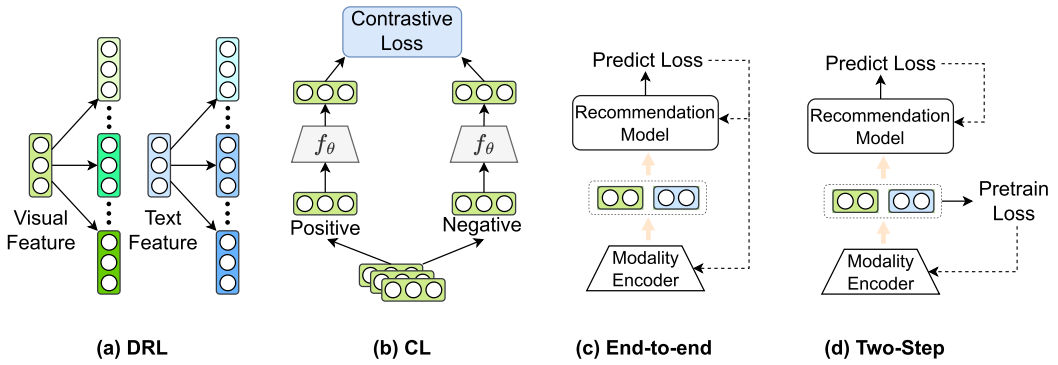

Fig. 3. The illustration of feature enhancement and model optimization.

4 Feature Enhancement

Different modality representations of the same object have unique or 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 have been equipped with **Disentangled Representation Learning (DRL)** and **CL** to carry out feature enhancement based on interaction, as shown in Figure 3(a) and (b).

4.1 DRL

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 [77], MacridVAE [42], and CDR [4]. Besides, the multimodal recommendation is to discover helpful information formed by various hidden factors from multimodal data, which are highly entangled in complex ways. MDR [61] proposes a multimodal disentangled recommendation that can learn well-disentangled representations carrying complementary and standard information from different modalities. DMRL [34] considers the different contributions of various modality features for each disentanglement factor to capture user preferences. Furthermore, PAMD [14] designs a disentangled encoder to extract modality-common features while preserving modality-specific features automatically. Besides, the designed CL guarantees the consistency and gap between separated modal representations. Compared with MacridVAE, SEM-MacridVAE [60] considers item semantic information when learning disentangled representations from user behaviors.

4.2 CL

Unlike DRL, CL methods enhance the representation by data augmentation, which is also helpful in handling the sparsity problem. Besides, Many works have introduced CL loss functions mainly for modality alignment.

Liu et al. [40] propose a novel CL loss, which makes the different modal representations of the same item have semantic similarity. In addition, GHMFC [58] constructs two CL 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 [43] proposes a CL loss to align the graph representation from the bundle view and item view. MICRO [74] focuses on both shared modal information and specific modal information. In CMCKG [2], entity embeddings are obtained from both descriptive attributes and structural link information through

KGs for contrastive loss. In HCGCN [45], to enforce visual and textual item features mapped into the same semantic space, it refers to CLIP [49] that adopts CL and maximizes the similarity of correct visual-textual pair in a batch. Furthermore, Zhou et al. [83] and Wei et al. [63] both design the modality-based contrastive loss for alignment and digging out the inter-relationships between distinct modalities.

Because the core of CL is to mine the relationship between positive and negative samples, many models adopt data augmentation methods to construct positive samples in recommendation scenarios. MGMC [78] designs a graph enhancement to augment the samples and introduces meta-learning to increase model generalization. MML [47] is a sequential recommendation model that expands the training data by constructing a subset of the user's historical purchase item sequence. LHBPMR [45] selects items with similar preferences from the graph convolution to construct positive samples. MMGCL [70] constructs positive samples by modal edge dropping and modal masking. Also, Victor [24] firstly constructs samples through Chinese semantics. Combo-Fashion [85] is a bundle fashion recommendation model, so it constructs negative and positive fashion matching schemes. Most of the existing models consider removing information that does not belong to user preferences in multimodal data. By comparison, UMPR [68] directly constructs a loss that describes the difference between visual positive and negative samples.

5 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 RS are trained together. Therefore, the MRS can be divided into two categories during training: **End-to-end training** and **Two-step training**. As shown in Figure 3(c), End-to-end training can update the parameters of all layers in the model with each gradient obtained through backpropagation. By comparison, 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 3(d).

5.1 End-to-end Training

Since MRSs use pictures, texts, audio and other multimedia information, some common encoders in other fields, such as Vit [12], Resnet [15], and Bert [11], 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 [12] 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 [32] and VLSNR [13] 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 show that introducing multimodal data without updating encoder parameters can also improve the recommendation performance. Liu et al. [40] propose to fine-tune the encoder's parameters with only 100 epochs by recommendation and contrastive loss. In particular, MG [79] eliminates the noise contained in multimodal inputs by a gradient strategy. Recently, some researchers [56, 75] only utilize multimodal features of items by pretrained encoders, which can free MRS from the limitation of item identity in a specific dataset.

Some end-to-end methods also aim at reducing the amount of computation while improving the recommendation performance. They often decrease the number of parameters required to be updated when training. For instance, MKGformer [8] is a multi-layer transformer structure where many attention layer parameters are shared to reduce computation. FREEDOM [81] is designed to freeze some parameters of graph structure, dramatically reducing memory costs, and achieving a denoising effect.

5.2 Two-step Training

Compared with the end-to-end pattern, the two-stage training scheme can target downstream tasks better, but it requests much higher computing resources. Thus, few MRS adopt two-step training. PMGT [39] 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 pretrains a transformer to learn the fashion matching knowledge, and then recommends for users through a cloth generation model. Besides, it is common in sequential recommendation, where it is difficult to train the model in an end-to-end scheme. For example, in the pretraining stage, MML [47] first trains the meta-learner through meta-learning to increase model generalization, then trains the item embedding generator in the second stage. Besides, TESM [46] and Victor [24] pretrain a well-designed graph neural network and a video transformer, respectively. Recently, some more advanced techniques have been adapted for higher training efficiency, such as knowledge distillation and prompt tuning. As for the former one, SGFD [33] distills a lighter modality encoder from a pretrained modality encoder, when finetuning for the recommendation task. Also, PromptMM [64] proposes a pretrain-prompt scheme to achieve easier finetuning and higher task adaptability.

6 Applications and Resources

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. Though most researchers aim at proposing a general MRS model that can be applied to all applications, it is better to design a unique model for some typical applications. For example, in fashion recommendation applications, users are often tempted to buy clothing because of the style of the clothing that they do not need. POG [5] proposes to utilize the image and title of the clothing item to predict whether the style is compatible and preferred by the user. Besides, the content of items is important for News recommendation, so the textual feature is highlighted in the related researches [13, 25]. Normally, the general RS can also be adapted to these applications; however, they show markedly inferior performance compared with MRS models [5, 25, 29, 85].

Dataset is one necessary resource to research MRS, especially for these typical applications. Therefore, to ease access to such vital resources, we summarize several popular datasets for MRS according to its applications in Table 3. 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. In terms of the evaluation metrics for MRS models, they are often the same as the general RS, such as hit rate and normalized discounted cumulative gain.

As mentioned before, MRS often consists of several architecture components, which causes technical difficulty in implementation for practical systems. For example, the modality encoders with extensive parameters are difficult to deploy. Furthermore, it is hard to devise a unique training pipeline for MRS models because many branches exist, such as different interaction modules. For data pre-processing, the settings of data split and filtration vary, which leads to challenges in reproduction. To face these issues, two open-source benchmarks are helpful:

- **MMRec**[3]: MMRec is a multimodal recommendation toolbox based on PyTorch. It integrates more than ten outstanding multimodal recommendation system models, such as MMGCN [65].
- **Cornac**[4] [50]: Cornac is a comparative framework for MRSs. It derives the whole experimental procedures for MRS, i.e., data, models, metrics, and experiment. Besides, cornac

[3] https://github.com/enoche/MMRec
[4] https://github.com/PreferredAI/cornac

Table 3. 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' indicate the visual data, textual data, video data, and acoustic data, respectively.

is highly compatible with mainstream deep learning frameworks such as TensorFlow and PyTorch.

7 Challenges and Future Directions

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

— **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, 16] refer to it, it still needs to be explored.
— **Computational Complexity.** MRS requires large amounts of computational resources due to the parameter-intensive modality encoder, making it challenging to train the models on large datasets and populations. Also, the complexity of multimodal data and models can increase the inference cost and time required for recommendation generation, making it challenging for real-time applications.
— **Risk of Overfitting.** Due to the sparsity of the MRS data and informative representation obtained from the fabricated modality encoders, the MRS models are inclined to suffer from overfitting.
— **Privacy.** Though multimodal information can benefit recommender systems by alleviating data sparsity, it also increases the risk of privacy leakage. How to protect individual privacy under the condition of affluent multimodal information is also a great challenge for the researchers.
— **Large General MRS Dataset.** Currently, the scale of the MRS dataset 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. Therefore, a large high-quality MRS dataset with abundant modalities is urgently needed.

— **Incomplete and Biased Data.** In real-world applications, the multimodal data is often incomplete or biased. For example, one specific modality may be missed during the data collection. Besides, the practical interactions are often skewed by popularity. Addressing these two data challenges will accelerate the applications of MRS to industrial.

Besides, we also give out several future directions as follows:

— **Cross-modal Representation Learning.** Developing better methods for cross-modal representation learning, such as transfer learning and large-scale pre-training encoder is expected to lead to more effective and efficient MRS.

— **Integration with Other Technologies.** An integration with technologies such as augmented reality and virtual reality is expected to enhance user experience and provide new opportunities for multimodal recommendation.

— **Utilization of Multimodal Large Language Models (MMLLM).** The MLLM have shown brilliant understanding and reasoning abilities. Thus, the adaptation of MLLM to MRS is relatively promising [57].

8 Conclusion

MRS is becoming one of the leading directions in RS, benefitting from its aggregation advantage on different modalities. In this article, we introduce taxonomies of MRS, i.e., modality encoder, feature interaction, feature enhancement, and model optimization based on challenges faced in different modeling stages. We also summarize the dataset and open-source codes. At last, some challenges and future directions of MRS are proposed to inspire further research.

References

[1] Shuqing Bian, Xingyu Pan, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, and Ji-Rong Wen. 2023. Multi-modal mixture of experts represetation learning for sequential recommendation. In *Proceedings of the CIKM*.

[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 *Proceedings of the ACM MM*.

[3] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Bundle recommendation with graph convolutional networks. In *Proceedings of the 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. *Advances in Neural Information Processing Systems* 34 (2021), 26924–26936.

[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 *Proceedings of the 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 *Proceedings of the SIGIR*.

[7] Xi Chen, Yangsiyi Lu, Yuehai Wang, and Jianyi Yang. 2021. CMBF: Cross-modal-based fusion recommendation algorithm. *Sensors* 21, 16 (2021), 5275.

[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. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 904–915.

[9] Yashar Deldjoo, Fatemeh Nazary, Arnau Ramisa, Julian Mcauley, Giovanni Pellegrini, Alejandro Bellogin, and Tommaso Di Noia. 2023. A review of modern fashion recommender systems. *ACM Computing Surveys (CSUR)* 56, 4 (2023), 1–37.

[10] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–38.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Toutanova Kristina. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.
[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
[13] Songhao Han, Wei Huang, and Xiaotian Luan. 2022. VLSNR: Vision-linguistics coordination time sequence-aware news recommendation. arXiv:2210.02946. Retrieved from https://arxiv.org/abs/2210.02946
[14] Tengyue Han, Pengfei Wang, Shaozhang Niu, and Chenliang Li. 2022. Modality matches modality: Pretraining modality-disentangled item representations for recommendation. In *Proceedings of the WWW*.
[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the CVPR*.
[16] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W. Zheng, and Qi Liu. 2019. Explainable fashion recommendation: A semantic attribute region guided approach. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 4681–4688.
[17] Hengchang Hu, Wei Guo, Yong Liu, and Min-Yen Kan. 2023. Adaptive multi-modalities fusion in sequential recommendation systems. In *Proceedings of the CIKM*.
[18] 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)* 16, 3 (2021), 274–306.
[19] Xiangen Jia, Yihong Dong, Feng Zhu, Yu Xin, and Jiangbo Qian. 2022. Preference-corrected multimodal graph convolutional recommendation network. *Applied Intelligence* (2022), 1–16.
[20] Taeri Kim, Yeon-Chang Lee, Kijung Shin, and Sang-Wook Kim. 2022. MARIO: Modality-aware attention and modality-preserving decoders for multimedia recommendation. In *Proceedings of the CIKM*.
[21] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the EMNLP*. 1746–1751. DOI: https://doi.org/10.3115/v1/D14-1181
[22] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
[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 *Proceedings of the 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 *Proceedings of the ACM MM*.
[25] 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 *Proceedings of the ACL Findings*. 343–352.
[26] Shuaiyang Li, Dan Guo, Kang Liu, Richang Hong, and Feng Xue. 2023. Multimodal counterfactual learning network for multimedia-based recommendation. In *Proceedings of the SIGIR*.
[27] Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. IMF: Interactive multimodal fusion model for link prediction. In *Proceedings of the WWW*.
[28] Jianxun Lian, Iyad Batal, Zheng Liu, Akshay Soni, Eun Yong Kang, Yajun Wang, and Xing Xie. 2021. Multi-interest-aware user modeling for large-scale sequential recommendations. arXiv:2102.09211. Retrieved from https://arxiv.org/abs/2102.09211
[29] 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 *Proceedings of the WWW*.
[30] 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* 32, 8 (2019), 1502–1516.
[31] Bo Liu. 2022. Implicit semantic-based personalized micro-videos recommendation. arXiv:2205.03297. Retrieved from https://arxiv.org/abs/2205.03297
[32] 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 *Proceedings of the AAAI*.
[33] Fan Liu, Huilin Chen, Zhiyong Cheng, Liqiang Nie, and Mohan Kankanhalli. 2023. Semantic-guided feature distillation for multimodal recommendation. In *Proceedings of the ACM MM*.
[34] Fan Liu, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, and Mohan Kankanhalli. 2022. Disentangled multimodal representation learning for recommendation. *IEEE Transactions on Multimedia* 25 (2022), 7149–7159.
[35] Huizhi Liu, Chen Li, and Lihua Tian. 2022. Multi-modal graph attention network for video recommendation. In *Proceedings of the CCET*.

[36] Hongtao Liu, Fangzhao Wu, Wenjun Wang, Xianchen Wang, Pengfei Jiao, Chuhan Wu, and Xing Xie. 2019. NRPA: Neural recommendation with personalized attention. In *Proceedings of the SIGIR*.
[37] Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. 2023. MEGCF: Multimodal entity graph collaborative filtering for personalized recommendation. *ACM Transactions on Information Systems* 41, 2 (2023), 1–27.
[38] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-video co-attention network for personalized micro-video recommendation. In *Proceedings of the WWW*.
[39] 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 *Proceedings of the ACM MM*.
[40] Zhuang Liu, Yunpu Ma, Matthias Schubert, Yuanxin Ouyang, and Zhang Xiong. 2022. Multi-modal contrastive pre-training for recommendation. In *Proceedings of the ICMR*.
[41] 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* 7 (2019), 12809–12821.
[42] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. *Advances in Neural Information Processing Systems* 32 (2019).
[43] Yunshan Ma, Yingzhi He, An Zhang, Xiang Wang, and Tat-Seng Chua. 2022. Crosscbr: Cross-view contrastive learning for bundle recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1233–1241.
[44] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781. Retrieved from https://arxiv.org/abs/1301.3781
[45] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning hybrid behavior patterns for multimedia recommendation. In *Proceedings of the ACM MM*.
[46] Juan Ni, Zhenhua Huang, Yang Hu, and Chen Lin. 2022. A two-stage embedding model for recommendation with multimodal auxiliary information. *Information Sciences* 582 (2022), 22–37.
[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 *Proceedings of the CIKM*.
[48] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the EMNLP*. 1532–1543.
[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
[50] Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. 2020. Cornac: A comparative framework for multimodal recommender systems. *Journal of Machine Learning Research* 21, 95 (2020), 1–5.
[51] Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, Huimin Ma, and Yong Li. 2023. Enhancing adversarial robustness of multi-modal recommendation via modality balancing. In *Proceedings of the ACM MM*.
[52] 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 *Proceedings of the CIKM*.
[53] 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* 57, 5 (2020), 102277.
[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
[55] Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2019. Multi-task feature learning for knowledge graph enhanced recommendation. In *Proceedings of the WWW*.
[56] Jie Wang, Fajie Yuan, Mingyue Cheng, Joemon M. Jose, Chenyun Yu, Beibei Kong, Zhijin Wang, Bo Hu, and Zang Li. 2024. Transrec: Learning transferable recommendation from mixture-of-modality feedback. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 193–208.
[57] Maolin Wang, Yao Zhao, Jiajia Liu, Jingdong Chen, Chenyi Zhuang, Jinjie Gu, Ruocheng Guo, and Xiangyu Zhao. 2024. Large multimodal model compression via iterative efficient pruning and distillation. In *Companion Proceedings of the ACM Web Conference 2024*, 235–244.
[58] Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *Proceedings of the SIGIR*.
[59] 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).
[60] 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* 45, 1 (2023), 408–424.

[61] Xin Wang, Hong Chen, and Wenwu Zhu. 2021. Multimodal disentangled representation for recommendation. In *Proceedings of the ICME*.
[62] 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* 8 (2020), 213012–213026.
[63] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-modal self-supervised learning for recommendation. In *Proceedings of the WWW*.
[64] Wei Wei, Jiabin Tang, Lianghao Xia, Yangqin Jiang, and Chao Huang. 2024. PromptMM: Multi-modal knowledge distillation for recommendation with prompt-tuning. In *Proceedings of the WWW*.
[65] 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 *Proceedings of the ACM MM*.
[66] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. Mm-rec: Visiolinguistic model empowered multimodal news recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2560–2564.
[67] Zhibo Xiao, Luwei Yang, Wen Jiang, Yi Wei, Yi Hu, and Hao Wang. 2020. Deep multi-interest network for click-through rate prediction. In *Proceedings of the CIKM*. 2265–2268.
[68] 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* 17, 6 (2020), 4197–4205.
[69] Jing Yi and Zhenzhong Chen. 2021. Multi-modal variational graph auto-encoder for recommendation systems. *IEEE Transactions on Multimedia* 24 (2021), 1067–1079.
[70] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the SIGIR*.
[71] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3541–3549.
[72] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the ACM MM*.
[73] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the ACM MM*.
[74] 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* 35, 9 (2022), 9154–9167.
[75] Lingzi Zhang, Xin Zhou, and Zhiqi Shen. 2023. Multimodal pre-training framework for sequential recommendation via contrastive learning. arXiv:2303.11879. Retrieved from https://arxiv.org/abs/2303.11879
[76] Xiaoyan Zhang, Haihua Luo, Bowei Chen, and Guibing Guo. 2020. Multi-view visual Bayesian personalized ranking for restaurant recommendation. *Applied Intelligence* 50, 9 (2020), 2901–2915.
[77] Yin Zhang, Ziwei Zhu, Yun He, and James Caverlee. 2020. Content-collaborative disentanglement representation learning for enhanced recommendation. In *Proceedings of the RecSys*.
[78] Feng Zhao and Donglin Wang. 2021. Multimodal graph meta contrastive learning. In *Proceedings of the CIKM*.
[79] Shanshan Zhong, Zhongzhan Huang, Daifeng Li, Wushao Wen, Jinghui Qin, and Liang Lin. 2024. Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima. In *Proceedings of the WWW*.
[80] 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:2302.04473. Retrieved from https://arxiv.org/abs/2302.04473
[81] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 935–943.
[82] Xin Zhou. 2023. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*. 1–2.
[83] 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 *Proceedings of the WWW*.
[84] Yan Zhou, Jie Guo, Hao Sun, Bin Song, and Fei Richard Yu. 2023. Attention-guided multi-step fusion: A hierarchical fusion network for multimodal recommendation. In *Proceedings of the SIGIR*.
[85] Chenxu Zhu, Peng Du, Weinan Zhang, Yong Yu, and Yang Cao. 2022. Combo-Fashion: Fashion clothes matching CTR prediction with item history. In *Proceedings of the KDD*.

Received 26 June 2023; revised 8 June 2024; accepted 22 August 2024