Multimodal Graph Contrastive Learning for Multimedia-Based Recommendation

Kang Liu, Feng Xue, Dan Guo, *Member, IEEE*, Peijie Sun, Shengsheng Qian, *Member, IEEE*, and Richang Hong, *Member, IEEE*

***Abstract*—Multimedia-based recommendation is a challenging task that requires not only learning collaborative signals from user-item interaction, but also capturing modality-specific user interest clues from complex multimedia content. Though significant progress on this challenge has been made, we argue that current solutions remain limited by multimodal noise contamination. Specifically, a considerable proportion of multimedia content is irrelevant to the user preference, such as the background, overall layout, and brightness of images; the word order and semantic-free words in titles; etc. We take this irrelevant information as noise contamination to discover user preferences. Moreover, most recent research has been conducted by graph learning. This means that noise is diffused into the user and item representations with the message propagation; the contamination influence is further amplified. To tackle this problem, we develop a novel framework named Multimodal Graph Contrastive Learning (MGCL), which captures collaborative signals from interactions and uses visual and textual modalities to respectively extract modality-specific user preference clues. The key idea of MGCL involves two aspects: First, to alleviate noise contamination during graph learning, we construct three parallel graph convolution networks to independently generate three types of user and item representations, containing collaborative signals, visual preference clues, and textual preference clues. Second, to eliminate as much preference-independent noisy information as possible from the generated representations, we incorporate sufficient self-supervised signals into the model optimization with the help of contrastive learning, thus enhancing the expressiveness of the user and item representations. Extensive experiments validate the effectiveness and scalability of MGCL at https://github.com/hfutmars/MGCL.**

***Index Terms*—Recommender system, collaborative filtering, multimodal user preference, graph convolution network, contrastive learning.**

I. Introduction

MULTIMEDIA-BASED recommendation (MMRec) algorithms play a crucial role in many online services, such as short-video sharing platforms [1], [2], e-commerce [3], [4], social media [5], [6], and food-related applications [7], [8]. Unlike traditional recommendation methods (mainly collaborative filtering [9], short for CF) that only learn collaborative signals (i.e., behavioral similarity reflected by interaction data), mainstream MMRec methods explore additional clues pertinent to user preferences from multiple modalities (image, text, audio, etc.), which we refer to as multimodal preference clues. Typically, the paradigm of mainstream MMRec methods involves two steps. First, pre-trained neural networks are employed to extract deep features from items' multimodal content. Then, the features are incorporated into the recommendation framework to capture both collaborative signals and multimodal preference clues.

According to the utilization of modality information, existing MMRec works can be roughly divided into two categories: single-modality and multimodal models. Early MMRec efforts are mostly single-modality models, employing either descriptive images or textual content to enrich the item representations. Considering the obvious difference in the distribution of user preferences across different modalities, most subsequent MMRec works are multimodal models, which leverage multiple modalities simultaneously to profile item properties and model user preferences more fully. Section II introduces these methods. In recent years, graph convolution network (GCN), as a powerful neural network for graph data, has been widely applied in MMRec research and achieved outstanding recommendation performance. In principle, compared to traditional recommendation methods, GCNs capture more interactions (both direct and indirect) through neighbor aggregation, which provides more supervised signals for discover clues pertinent to user interests hidden in different modality content.

Manuscript received 5 July 2022; revised 17 January 2023; accepted 23 February 2023. Date of publication 17 March 2023; date of current version 15 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62272143, in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2022-054, in part by the Anhui Provincial Major Science and Technology Project under Grant 202203a05020025, in part by the Seventh Special Support Plan for Innovation and Entrepreneurship in Anhui Province, and in part by the Fellowship of China Postdoctoral Science Foundation under Grant 2022TQ0178. The associate editor coordinating the review of this manuscript and approving it for publication was Dr Liang Lin. *(Corresponding author: Feng Xue.)*

Kang Liu, Dan Guo, and Richang Hong are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: kangliu1225@gmail.com; guodan@hfut.edu.cn; hongrc.hfut@gmail.com).

Feng Xue is with the Key Laboratory of Knowledge Engineering with Big Data of Ministry of Education, China, the Intelligent Interconnected Systems Laboratory of Anhui Province, and the School of Software, Hefei University of Technology, Hefei 230601, China (e-mail: feng.xue@hfut.edu.cn).

Peijie Sun is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: sun.hfut@gmail.com).

Shengsheng Qian is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100040, China (e-mail: shengsheng.qian@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TMM.2023.3251108

Fig. 1. Illustration of the noise contamination in visual modality, where *Target item* is similar to *Pos item*, rather than *Neg item*. In *Example 1*, the noise is mainly the {background, people} in the images; in *Example 2*, the noise is mainly the {overall layout, people} in the images; in *Example 3*, the noise is mainly the {background, people} in the images. Noise-dominated images lead to incorrect similarity distribution, whereas preference-dominated interactions result in correct distribution.

However, we argue that existing MMRec works ignore the problem of noise contamination when modeling user preferences with multimodal information. More precisely, a considerable proportion of the modality content that users interact with is noise unrelated to their interests. Typically, user purchase behavior mainly depends on the visual and textual content related to the product, independent of the background, overall layout, and brightness of images; people in the images; word order; and words unrelated to the product in titles. Fig. 1 gives an example of noise contamination in the visual modality, which can easily lead to incorrect similarity measurements between products. From this view, the modality-specific feature of an item is impure for the user-oriented recommendation task, because it features a large amount of preference-irrelevant noise. Worse still, in most GCN-based MMRec methods, such impure multimodal features are continuously propagated to other nodes during the recursive graph convolutions, which intensifies the noise contamination. In addition, these methods usually treat collaborative signals and modality-specific preference clues as a whole—that is, the ID embeddings (i.e., vector mapped by ID) and modality-specific features are merged to perform message passing between nodes. As such, the collaborative signals captured in the ID embeddings are also affected by noise.

To tackle this problem, we need to eliminate the noise contamination in collaborative signals and multimodal preference clues. First, for collaborative signals, the simplest yet effective approach is to construct an additional GCN module dedicated to propagating ID embeddings, which prevents the explicit incorporation of multimodal information into ID embeddings, thus obtaining cleaner collaborative signals. Second, for multimodal user preferences, it is difficult to eliminate preference-irrelevant noise in multimodal content due to the complexity of multimodal content and the variability of user preferences. Therefore, our goal is to reduce the negative impact of noisy data on modeling multimodal user preferences, rather than to identify and remove preference-irrelevant parts from multimodal content. For this purpose, in this study, we propose to leverage cleaner collaborative signals to guide the modeling of multimodal user preferences. The specific strategy is to adopt contrastive learning [10] to maximize the mutual information between the captured collaborative signals and multimodal preferences, which drives the learning of multimodal representations toward capturing clues that are consistent with cleaner collaborative signals. Essentially, contrastive learning can be seen as representation alignment, which pulls together items with similar interaction behaviors and pushes away the dissimilar items, thus eliminating false similarities caused by multimodal noise to some extent.

Overall, we develop a new MMRec method, multimodal graph contrastive learning (MGCL), which is equipped with two main components, including *separated graph learning module* and *model training module*. The former constructs three parallel GCNs to process ID embeddings, visual features, and textual features, thus capturing collaborative signals, visual preference clues, and textual preference clues separately. Such a schema obtains purer collaborative signals by isolating the flow of multimodal information into the ID embeddings. The latter focuses on model optimization, which can be divided into two aspects. First, this module sets independent optimization targets for capturing collaborative signals and multimodal preference cues. Second, to filter out as much preference-independent multimodal noise as possible in representation learning, this module additionally sets a node self-discrimination task and constructs a contrastive loss, which amounts to using sufficient self-supervised signals to assist the model in mining as many preference-relevant clues as possible from complex multimedia content. We conduct extensive experiments on three public datasets to demonstrate the state-of-the-art performance of MGCL. Specifically, MGCL achieves a maximum of 6.37% and an average of 4.4% in improvement compared with the strongest baseline. Further comparison experiments validate the effectiveness of each component in MGCL and its excellent scalability.

We summarize the contributions of this paper as follows:

- We highlight the problem of multimodal noise contamination in existing works for MMRec, which weakens collaborative signal capturing and multimodal user preference modeling.
- We propose MGCL, a new GCN-based MMRec framework, which alleviates the problem of multimodal noise contamination and enhances the representations of users and items through a separated graph learning schema and contrastive learning.
- We conduct extensive experiments on three public datasets to validate the effectiveness and scalability of the proposed MGCL. We release the complete codes and data of MGCL at *https://github.com/hfutmars/MGCL*.

II. Related Work

In this section, we review the recommender methods that are most relevant to our work, including GCN-based recommendation methods and MMRec methods.

A. GCN-Based Recommendation Methods

As a deep neural network tailored for graph-structured data, GCN [11], [12] has attracted considerable attention in recent years and has been widely used in recommender systems [13]. The general paradigm of GCN is iteratively collecting collaborative signals from neighbor nodes and integrating them into the representation of the target node. NGCF [14] is a fusion framework of GCN and matrix factorization (MF) [15], which performs graph convolutions on a user-item interaction graph to explicitly encode high-order collaborative signals. Some researchers have found that the nonlinear components in GCNs are not helpful in capturing collaborative signals [16], [17]. Therefore, a lightweight GCN framework, LightGCN [18], is proposed, which removes the nonlinear activation functions and weight transformation matrices from the classical GCN structure, achieving a significant improvement. In addition to collaborative signals, GCN can also capture popularity features. JMPGCF [19] captures multi-grained popularity features simultaneously using the popularity-aware Graph Laplacian Norm to better match the varying sensitivity of users to popularity. More recently, some researchers have tried to incorporate contrastive learning [20], [21] into the GCN-based recommender methods. Specifically, as a self-supervised graph learning framework, SGL [22] generates multiple views through dropout operations and conducts contrastive learning between them, thus enhancing the model robustness. In addition, EGLN [23] maximizes the mutual information between the node embeddings and global structural features to achieve mutual enhancement between them.

B. MMRec Methods

Existing mainstream MMRec methods [24], [25] follow a hybrid paradigm based on both multimedia content and CF [9]. Depending on the use of modality information in the multimedia content, MMRec methods can be roughly divided into three categories, as follows:

1) Visual Modality: Visual-modality-based MMRec models are usually suitable for recommendation scenarios that highly rely on visual content. VBPR [26] is a representative approach that uses the features obtained by a pre-trained Convolution Neural Network (CNN) to enrich item representations. DUIF [27] enriches the user representation with additional user-related features on top of VBPR. VPOI [28] also utilizes similar pre-extracted visual features and incorporates them into the PMF [29] framework for Point-of-Interest (POI) check-in recommendation. To further improve the utilization of visual features, some subsequent studies [30], [31] have started to incorporate a CNN module into the recommendation framework and participate in parameter updating, but such an end-to-end schema greatly increases the model's computational burden. Other researchers have tried to use visual information to discover more preference-related signals. For example, ACF [32] feeds component-level visual content into the attention networks to learn the weight of user preferences for historical items. In addition, NPR [33] further extracts spatial and topical information associated with item images to better model user preferences in the visual modality.

2) Textual Modality: Common textual modality information includes descriptive titles and user reviews. Most MMRec methods based solely on the textual modality usually focus on processing review information to achieve the rating prediction task (i.e., explicit feedback). ConvMF [34], an early work, is a fusion framework of CNN and PMF, wherein a CNN module is used to extract review features and enrich the item representations. DeepCoNN [35] is a classical two-tower model that uses two parallel CNN modules to model user and item review documents, respectively, and employs a Factorization Machine (FM) for rating prediction. Some subsequent works have emerged to improve DeepCoNN, such as using an attention mechanism to calculate the importance of each word in a review [36], or each review in a document [37], or incorporating interaction information [38]. Recently, many researchers have explored the application of GCN to textual-modality-based MMRec. For example, RMG [39] uses attention-based GCNs to aggregate review features for all users and items, thus obtaining their representations. RGCL [40] is a recently proposed approach that uses a multi-relational GCN to generate node representations, involves review features as edges on the graph for message propagation, and uses contrastive learning to further enhance the model performance.

3) Multiple Modalities: Typically, users exhibit different interests in different modality information, so it is a better idea to utilize multiple modalities simultaneously. CKE [41] is a representative work that incorporates visual features, textual features, and knowledge graphs as a whole into the item representations, and uses MF as the overall recommendation framework. Unlike CKE, MMGCN [42] constructs multiple GCN modules to process different modalities (visual, text, and audio) separately to model modality-specific user preferences. Build on MMGCN, MGAT [43] constructs an attention network to calculate the importance of different modalities, so as to model more fine-grained multimodal user preferences. Distinct from these methods, GRCN [44] does not incorporate multimodal features into the representations in the graph convolution process, but leverages multimodal information to refine the user-item interaction graph, which can alleviate the negative impact of irrelevant multimodal information to a certain extent. Recently, some researchers have begun to explore user intents hidden in multimodal information. For example, HUIGN [1] constructs hierarchical GCNs and designs two types of neighbor aggregations to generate representations of different user intents. In addition, a recently proposed work, DMRL [45], employs a disentanglement technique to explicitly mine multiple user intents and allocate them to different chunked representations. Notably, Du et al. argued that there are spurious correlations (similar to the multimodal noise in this paper) between multimodal content and user preferences; therefore, they proposed InvRL [46] to eliminates such spurious correlations with counterfactual inference. However, we argue that

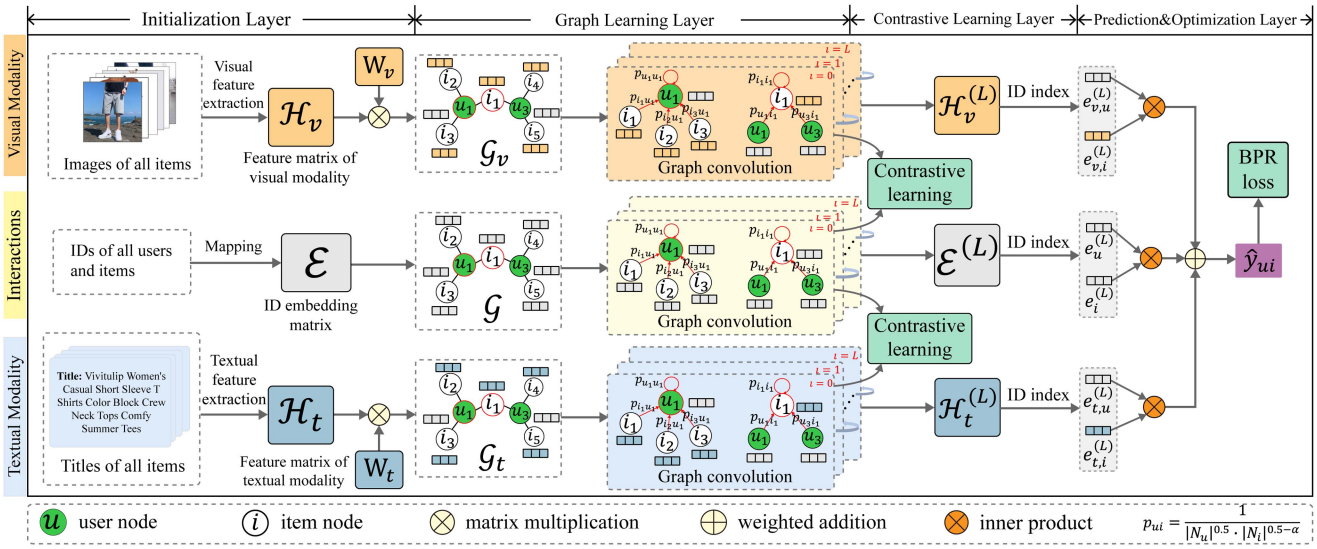

Fig. 2. Illustration of the proposed MGCL, which separately models the visual preferences clues, collaborative signals, and textual preference clues. In addition, the contrastive learning layer is used to achieve the representation alignment between collaborative signals and modality-specific preferences.

although this counterfactual schema can reduce spurious correlations in multimodal content, it ignores the problem that collaborative signals will also be contaminated by these spurious correlations.

III. Preliminary

Here, we first describe our task. Then, we construct two kinds of graphs, user-item interaction graph and multimodal interaction graph, which are used as information carriers to capture collaborative signals and multimodal user preference clues, respectively.

Our Task: Given the user-item interactions and the multimodal contents of items in a dataset, our task is to construct a model of representation learning that can predict the likelihood that a user would interact with an item.

User-Item Interaction Graph $\mathcal{G}$: Following the conventional graph learning methods in recommendation work, we transform the implicit feedback data into a bipartite user-item interaction graph $\mathcal{G} = \{(u, r_{ui}, i)|u \in \mathcal{U}, i \in \mathcal{I}\}$, where $\mathcal{U}$ and $\mathcal{I}$ are the user and item sets, respectively, and $r_{ui} = 1$ denotes that there is an interaction between user u and item i, otherwise $r_{ui} = 0$.

Multimodal Interaction Graph $\mathcal{G}_m$: In most recommendation scenarios, each item contains the information on multiple modalities such as images, titles, and reviews. We use pre-trained deep neural networks to process the multimodal contents of all items, obtaining deep feature matrices, $\mathcal{H}_m \in \mathbb{R}^{|\mathcal{I}| \times d_m}$, where $m \in \{v, t\}$ is the modality indicator, v and t denote visual and textual modalities, respectively, $|\mathcal{I}|$ is the number of all items, and d_m denotes the feature length of modality m. Note that, we can further enhance $\mathcal{H}_m$ by preemptively removing a small amount of noise (e.g., background, word order, and semantic-free words) that can be identified and removed by advanced techniques in Computer Vision (CV) and Natural Language Processing (NLP), which is left for future research. We construct the graphs $\mathcal{G}_m$, which has exactly the same structure as the interaction graph $\mathcal{G}$, to store multimodal features. We term it as multimodal interaction graph.

IV. Methodology

In this section, we elaborate on our proposed MGCL, as shown in Fig. 2. MGCL is equipped with two main modules: (a) Separated Graph Learning Module, which performs message propagation on the user-item interaction graph $\mathcal{G}$ and multimodal interaction graphs $\mathcal{G}_v$ and $\mathcal{G}_t$ to generate multiple types of representations for all user and item nodes; and (b) Model Training Module, which alternatively optimizes the recommendation task and contrastive learning task to capture collaborative signals and multimodal preference clues.

A. Separated Graph Learning Module

To capture collaborative signals and multimodal preference clues and prevent multimodal noise from contaminating collaborative signals, we develop a GCN module, termed a *separated graph learning module*. This module performs message propagation on the three interaction graphs ($\mathcal{G}$, $\mathcal{G}_v$, and $\mathcal{G}_t$) separately to generate three types of representations of users and items.

1) Feature Initialization: Before performing message propagation, we need to initialize all user and item nodes on the three graphs. Because the user nodes on all graphs and the item nodes on $\mathcal{G}$ carry only ID information, we initialize the feature representations of these nodes by mapping their IDs to low-dimensional dense vectors (ID embeddings) according to CF-based recommendation methods:

$$\mathcal{E} = \{e_{u_1}^{(0)}, \ldots, e_{u_{|\mathcal{U}|}}^{(0)}, e_{i_1}^{(0)}, \ldots, e_{i_{|\mathcal{I}|}}^{(0)}, e_{v,u_1}^{(0)}, \ldots, e_{v,u_{|\mathcal{U}|}}^{(0)}, e_{t,u_1}^{(0)}, \ldots, e_{t,u_{|\mathcal{U}|}}^{(0)}\}, \quad (1)$$

where $\mathcal{E} \in \mathbb{R}^{(3|\mathcal{U}|+|\mathcal{I}|)\times d}$, $|\mathcal{U}|$ and $|\mathcal{I}|$ are the number of users and items, respectively, and d is the embedding length.

For item nodes on $\mathcal{G}_v$ and $\mathcal{G}_t$, we initialize them using their corresponding visual features and textual features, respectively, as follows:

$$\mathcal{H}_v = \{h_{v,i_1}, \ldots, h_{v,i_{|\mathcal{I}|}}\}, \quad \mathcal{H}_t = \{h_{t,i_1}, \ldots, h_{t,i_{|\mathcal{I}|}}\}, \tag{2}$$

where $\mathcal{H}_v \in \mathbb{R}^{|\mathcal{I}|\times d_v}$, $\mathcal{H}_t \in \mathbb{R}^{|\mathcal{I}|\times d_t}$, and d_v and d_t are the lengths of the visual and textual features, respectively.

To align the lengths of all node features, we use a weight transformation matrix to downscale the features for each node in $\mathcal{H}_v$ and $\mathcal{H}_t$. We formulate such transformation for the visual feature of node i_1 as follows:

$$e_{v,i_1}^{(0)} = W_v^T h_{v,i_1}, \tag{3}$$

where $W_v \in \mathbb{R}^{d_v \times d}$ is a visual weight transformation matrix. Similarly, we can obtain the i_1's down-scaled representation $e_{t,i_1}^{(0)}$ through a textual weight transformation matrix $W_t \in \mathbb{R}^{d_t \times d}$.

By performing this dimensional transformation on all item features in $\mathcal{H}_v$ and $\mathcal{H}_t$, we obtain the new visual and textual feature matrices on $\mathcal{G}_v$ and $\mathcal{G}_t$ as follows:

$$\mathcal{E}_v = \left\{e_{v,i_1}^{(0)}, \ldots, e_{v,i_{|\mathcal{I}|}}^{(0)}\right\}, \quad \mathcal{E}_t = \left\{e_{t,i_1}^{(0)}, \ldots, e_{t,i_{|\mathcal{I}|}}^{(0)}\right\}. \tag{4}$$

2) Graph Learning on $\mathcal{G}$: To capture clean collaborative signals and incorporate them into the representations of users and items, we perform message propagation iteratively on $\mathcal{G}$ for each node. Taking user node u_1 as example, the representation of u_1 obtained after $(l-1)$ graph convolutions can be formulated as follows:

$$e_{u_1}^{(l)} = \mathrm{AGG}\left(e_{u_1}^{(l-1)}, \left\{e_i^{(l-1)} : i \in N_{u_1}\right\}\right), \tag{5}$$

where $e_i^{(0)}$ is initialized in (1), N_{u_1} is the neighbors of u_1 in $\mathcal{G}$, and $\mathrm{AGG}(\cdot)$ denotes the neighbor aggregation operation, which could be flexibly implemented using different functions, such as weighted summation, max (or mean) pooling, concatenation, and nonlinear networks. Here, we employ a light variant of neighbor aggregation, which removes the nonlinear components (activation function and weight transformation matrix) from the traditional GCNs. The specific representation generation for u_1 is formulated as follows:

$$e_{u_1}^{(l)} = \sum_{i \in N_{u_1} \cup u_1} \frac{1}{|N_{u_1}|^{0.5}|N_i|^{0.5-\alpha}} \cdot e_i^{(l-1)}, \tag{6}$$

where $|N_u|$ and $|N_i|$ denotes the size of N_u and N_i, respectively. Additionally, following recent graph learning based method [19], we fine-tune the classical Graph Laplacian Norm $\frac{1}{|N_{i_1}|^{0.5}|N_u|^{0.5}}$ to the form of $\frac{1}{|N_{i_1}|^{0.5}|N_u|^{0.5-\alpha}}$, which is referred to as the popularity-aware norm, where α is a hyper-parameter used to control the model's sensitivity to popularity information. Performing a similar graph convolution operation as in (6) on all user and item nodes in $\mathcal{G}$, we can obtain a set of user and item representations:

$$\left\{e_{u_1}^{(l)}, \ldots, e_{u_{|\mathcal{U}|}}^{(l)}, e_{i_1}^{(l)}, \ldots, e_{i_{|\mathcal{I}|}}^{(l)}\right\}. \tag{7}$$

3) Graph Learning on $\mathcal{G}_v$ and $\mathcal{G}_t$: To capture multimodal preference clues and incorporate them into user and item representations, we here also use lightweight GCNs to perform message propagation over multimodal interaction graphs ($\mathcal{G}_v$ and $\mathcal{G}_t$). In addition, we set a low GCN depth to prevent noise in the multimodal content from being over-propagated. Specifically, for the visual modality v, the representation generation of user u_1 at layer k is formulated as follows:

$$e_{v,u_1}^{(k)} = \sum_{i \in N_{u_1} \cup u_1} \frac{1}{|N_{u_1}|^{0.5}|N_i|^{0.5-\alpha}} \cdot e_{v,i}^{(k-1)}, \tag{8}$$

where $e_{v,i}^{(0)}$ is initialized in (4). Similarly, for textual modality, we can use the graph convolution operation in (8) to obtain u_1's representation $e_{t,u_1}^{(k)}$.

By performing these above operations on all user and item nodes in both $\mathcal{G}_v$ and $\mathcal{G}_t$, we can obtain two additional sets of user and item representations:

$$\left\{\left\{e_{m,u_1}^{(k)}, \ldots, e_{m,u_{|\mathcal{U}|}}^{(k)}, e_{m,i_1}^{(k)}, \ldots, e_{m,i_{|\mathcal{I}|}}^{(k)}\right\} | m \in \{v,t\}\right\}. \tag{9}$$

4) Prediction: Having obtained three types of representations of all users and items (*cf.* (7) and 9), we now use them to reconstruct historical interactions, i.e., to predict the interaction likelihood between any user-item pair. For simplicity, given user u_1 and item i_1, we present the specific prediction process for their interaction probabilities.

First, we consider the results of collaborative signals for interaction prediction. Specifically, we find the corresponding feature representations of u_1 and i_1 in the representation matrix of (7), and use their inner product as the interaction probability, as follows:

$$\hat{y}_{u_1 i_1}^C = \left(e_{u_1}^{(l)}\right)^T \cdot e_{i_1}^{(l)}. \tag{10}$$

Next, we perform prediction using the captured multimodal preference clues. We find the representations of u_1 and i_1 in the visual and textual modalities, respectively, from the representation matrix of (9), and also prediction their interaction probability using inner product as the core operation as follows:

$$\hat{y}_{u_1 i_1}^M = \lambda_m \cdot \sum_{m \in \{v,t\}} \left(e_{m,u_1}^{(k)}\right)^T \cdot e_{m,i_1}^{(k)}, \tag{11}$$

where λ_m is a hyper-parameter to control the contribution of multimodal feature to user preference prediction.

Finally, we sum these results as the final interaction probability as follows:

$$\hat{y}_{u_1 i_1} = \hat{y}_{u_1 i_1}^C + \hat{y}_{u_1 i_1}^M. \tag{12}$$

B. Model Training Module

To optimize the user and item representations generated by the *separated graph learning module* so that they match user preferences well, a model training module is proposed here. This module constructs two types of optimization targets, including BPR loss and contrastive loss, as follows.

1) BPR Loss: For the recommendation task in MGCL, we apply a widely used Bayesian Personalized Ranking (BPR) loss [47] as the basic objective function, which assumes that users prefer interacted items to non-interacted ones. Specifically, we set independent optimization targets for the different preferences, as follows.

First, to incorporate the CF signals hidden in the interaction data into the representations in (7), we design the following loss function:

$$\mathcal{L}_c = \sum_{(u,i,j)\in\mathcal{O}} -ln\sigma\left(\left(e_u^{(l)}\right)^T \cdot e_i^{(l)} - \left(e_u^{(l)}\right)^T \cdot e_j^{(l)}\right), \quad (13)$$

where $\mathcal{O} = \{(u,i,j)|i \in N_u, j \notin N_u\}$ is the triplet training data, N_u denotes the interacted item set for user u (or denotes the neighbors of user node u), and $\sigma(\cdot)$ is the sigmoid function.

Next, to employ the interaction data to discover user interest clues in multimodal content and incorporate them into the representations in (9), we design the following loss function for modality m:

$$\mathcal{L}_m = \sum_{(u,i,j)\in\mathcal{O}} -ln\sigma\left(\left(e_{m,u}^{(k)}\right)^T \cdot e_{m,i}^{(k)} - \left(e_{m,u}^{(k)}\right)^T \cdot e_{m,j}^{(k)}\right), \quad (14)$$

Finally, we jointly optimize $\mathcal{L}_c$ and $\mathcal{L}_m$ as follows:

$$\mathcal{L} = \mathcal{L}_c + \sum_{m\in\{v,t\}} \mathcal{L}_m + \lambda(||\mathcal{H}_1||_2^2 + ||\mathcal{H}_2||_2^2). \quad (15)$$

To prevent overfitting, L_2 regularization is applied on $\mathcal{H}_1$ and $\mathcal{H}_2$, which are controlled by λ, where $\mathcal{H}_1$ and $\mathcal{H}_2$ are the initialized ID embedding matrix (*cf.* (1)) and the down-scaled multimodal feature matrices (*cf.* (4)), respectively.

2) Contrastive Loss: Distinct from the recommendation task, a node self-discrimination task is proposed here to implement contrastive learning between collaborative signals and multimodal preference clues. The motivation for this design is to utilize sufficient self-supervised signals to assist the model in mining as many clues pertinent to user preferences as possible from complex multimodal content, which is equivalent to alleviating the multimodal noise contamination on representations.

Specifically, this task assumes that the similarity between representations of the same node over different preferences should be higher than the similarity between representations of different nodes over the same preference. Building on this, we first follow SimCLR [48] to construct a contrastive loss for user nodes, as follows:

$$\mathcal{L}_{\text{user}}^m = \sum_{u\in\mathcal{U}'} -\log\left(\frac{\exp\left(\cos\left(e_u^{(0)}, e_{m,u}^{(0)}\right)/\tau\right)}{\sum_{j\in\mathcal{U}'}\exp\left(\cos\left(e_u^{(0)}, e_{m,j}^{(0)}\right)/\tau\right)}\right), \quad (16)$$

where $\cos(\cdot)$ is a cosine similarity function, τ is the temperature hyper-parameter, $\mathcal{U}'$ denotes a set of user nodes in a batch of training data, and m is a modality indicator.

We can also obtain a contrastive loss $\mathcal{L}_{item}^m$ for item nodes in the similar way. By considering both user nodes and item nodes in the visual and textual modalities, we obtain the final contrastive loss as follows:

$$\mathcal{L}_{\text{CL}} = \sum_{m\in\{v,t\}} (\mathcal{L}_{\text{user}}^m + \mathcal{L}_{\text{item}}^m). \quad (17)$$

Note that (17) does not consider contrastive learning across modalities because user preference distribution on different modalities is differential and therefore, there is no need to enforce their consistency. We validate the effectiveness of contrastive loss in Sections V-C and V-D.

3) Alternative Training: MGCL is equipped with two different optimization objectives, including BPR loss $\mathcal{L}$ and contrastive loss $\mathcal{L}_{CL}$. Therefore, to ensure that they are fully optimized, we train them alternatively in the manner described in Algorithm 1.

Algorithm 1: MODEL TRAINING.

Input: Dataset $\mathbf{D}$; multimodal feature matrices for all items $\{\mathcal{H}_m|m \in \{v,t\}\}$; learning rate lr; optimal graph convolution layers l and k; embedding size d; coefficient of L_2 regularization λ.

Output: Updated ID embeddings $\mathcal{E}$ and weight transformation matrices $\{W_m|m \in \{v,t\}\}$.

```
Initialize E and {W_m|m ∈ {v,t}} ← Xavier initializer.
for each epoch do
  for each mini-batch do
    Generate the node representations that contain collaborative signals ← Equations (6).
    Generate the node representations that contain multimodal preference clues ← Equations (8).
    Update the parameters E and {W_m|m ∈ {v,t}} ← Equation (15).
    Update the parameters E and {W_m|m ∈ {v,t}} ← Equation (17).
  end for
end for
```

C. Complexity Analysis

Here we analyze the time complexity of MGCL in the complete training phase. Given the number of edges on the user-item interaction graph $\mathcal{G}$ and multimodal interaction graphs $\mathcal{G}_v$ and $\mathcal{G}_t$ is $|\mathcal{E}|$, d is the embedding size, s is the number of training epochs, B is the batch size of the training data, and L_1 and L_2 denotes the GCN depth on $\mathcal{G}$ and $\{\mathcal{G}_v, \mathcal{G}_t\}$, respectively. The time complexity comes mainly from three operations, including (1) graph convolution, (2) BPR loss calculation, and (3) contrastive loss calculation.

- First, for the process of graph convolution, the complexity of MGCL on $\mathcal{G}$, $\mathcal{G}_v$, and $\mathcal{G}_t$ are $\mathcal{O}\left(2|\mathcal{E}|L_1\frac{ds|\mathcal{E}|}{B}\right)$, $\mathcal{O}\left(2|\mathcal{E}|L_2\frac{ds|\mathcal{E}|}{B}\right)$, and $\mathcal{O}\left(2|\mathcal{E}|L_2\frac{ds|\mathcal{E}|}{B}\right)$, respectively. From this, the complexity of graph convolutions in the complete training process is $\mathcal{O}\left(2|\mathcal{E}|\left(L_1+2L_2\right)\frac{ds|\mathcal{E}|}{B}\right)$.
- Next, for the calculation of BPR loss, the core operation in prediction function is inner product and its complexity

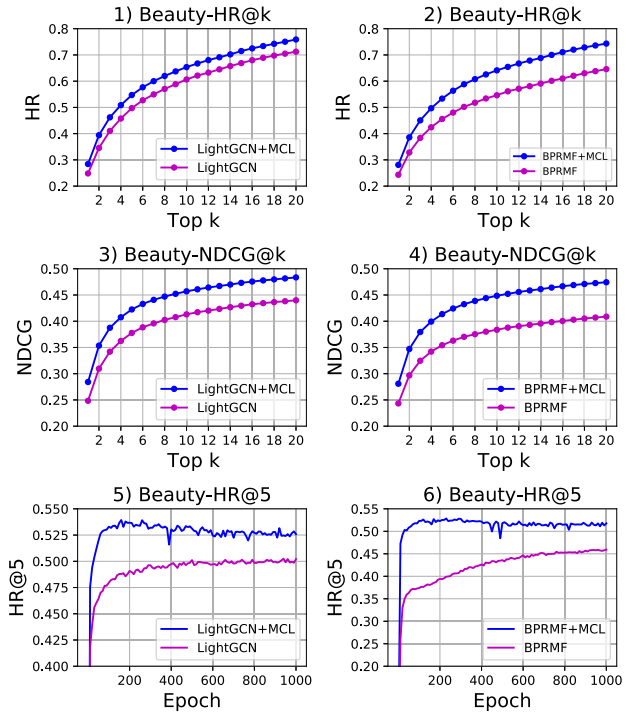

Fig. 3. Effect of MCL on BPRMF and LightGCN.

is $\mathcal{O}(d)$. MGCL needs to perform preference predictions on the graphs $\mathcal{G}$, $\mathcal{G}_v$, and $\mathcal{G}_t$, and the corresponding time complexity is $\mathcal{O}\left(6\frac{ds|\mathcal{E}|}{B}\right)$.

- Third, for the calculation of contrastive loss, the complexity for each training batch is $\mathcal{O}(4(B+1)d)$. Thus, the complexity for a complete training process is $\mathcal{O}\left(4(B+1)ds\frac{|\mathcal{E}|}{B}\right)$.

Overall, the training complexity of MGCL is $\mathcal{O}\left((2|\mathcal{E}|L_1+2|\mathcal{E}|L_2+6+4(B+1))\frac{ds|\mathcal{E}|}{B}\right)$, whereas the complexity of the most efficient GCN-based recommendation method (LightGCN) is $\mathcal{O}((2|\mathcal{E}|L_1+2)\frac{ds|\mathcal{E}|}{B})$. Note that L_1 is larger than L_2 (we set $L_1=3$ and $L_2=1$ on the Beauty dataset). The complexity of MGCL is about 1.3 times that of LightGCN on the Beauty dataset. In practice, taking the Beauty dataset as an example, each training epoch of MGCL and LightGCN takes 4.8 s and 3.9 s, respectively, and each testing process of MGCL and LightGCN takes 27 s and 25.6 s, respectively, which is acceptable because MGCL converges much faster (*cf.* Fig. 3).

V. Experiment

In this section, we conduct experiments to evaluate our proposed MGCL. We aim to answer the following five main research questions:

- **RQ1:** How does MGCL perform compared with the state-of-the-art baselines?
- **RQ2:** Are key components (multimodal features, contrastive learning module, iterative training strategy, etc) in MGCL helpful?
- **RQ3:** Can collaborative signal capturing and multimodal preference modeling be enhanced simultaneously by using the contrastive learning module?
- **RQ4:** How is the scalability of MGCL (or can MGCL be flexibly applied to other methods)?
- **RQ5:** Does MGCL have the ability to alleviate the multimodal noise contamination?

TABLE I
Statistics of the Datasets, Where # V and # T Denote the Length of Visual and Textual Features, Respectively, and Density is Calculated by Using $\#Interaction/(\#User \times \#Item)$

Dataset	# User	# Item	# Interaction	Density	# V	# T
Beauty	15,576	8,678	139,318	0.00103	4096	300
Art	25,165	9,324	201,427	0.00086	4096	300
Taobao	12,539	8,735	83,648	0.00076	4096	-

A. Experimental Settings

1) Datasets: Since our work aims to study multimedia-based recommendations, we choose two public datasets **Beauty** and **Arts_crafts_and_Sewing** (short for **Art**)[1] introduced by [49]; both of them contain images and titles. Besides, we select another dataset released in the Tianchi competition, **Taobao**.[2] This dataset provides visual content only. To ensure the quality of these three datasets, we apply the 5-core setting, that is, retaining that all users and items have at least five interactions. We present the details of these three datasets in Table I. Following the widely used setting [26], [50], [51], we apply the *leave-one-out* method for evaluation [47], where one item is randomly sampled for each user to form the test set, another item to form the validation set, and the rest of the interaction data as the training set.

2) Evaluate Metrics: As the focus of this work is ranking and recommending top-k items to users, we select two widely used protocols [50], [51], [52]: *Hit Ratio* (HR) and *Normalized Discounted Cumulative Gain* (NDCG) to evaluate model performance. We compute the average HR@k and NDCG@k for each user in the test set, where k is the size of recommendation list.

3) Baselines: To evaluate the overall performance of our proposed MGCL, we compare it with three types of baselines. The first is classical MF-based methods: BPRMF and SVD++.

- **BPRMF** [15]: This is a classical MF model optimized by BPR loss.
- **SVD++** [53]: Based on BPRMF, this model incorporates historical interacted items into the user embeddings. SVD++ can also be viewed as a one-layer GCN that only passes messages for user nodes.

The second is graph learning based CF methods: NGCF, LightGCN, and SGL.

- **NGCF** [14]: This model employs GCN to generate the user and item representations and concatenates the embeddings obtained at each layer as the final representation.

[1] http://deepyeti.ucsd.edu/jianmo/amazon/index.html
[2] https://tianchi.aliyun.com/competition/entrance/231506/information

TABLE II
OVERALL PERFORMANCE COMPARISON

Models	Beauty				Art				Taobao			
	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10
(1) BPRMF	0.4274	0.3343	0.5173	0.3634	0.6333	0.5597	0.7052	0.5829	0.3215	0.2465	0.4049	0.2733
(2) SVD++	0.4584	0.3592	0.5520	0.3895	0.6530	0.5627	0.7425	0.5916	0.3374	0.2523	0.4293	0.2819
(3) NGCF	0.4853	0.3776	0.5820	0.4089	0.6742	0.5882	0.7541	0.6141	0.3575	0.2658	0.4593	0.2986
(4) LightGCN	0.5002	0.3807	0.6063	0.4152	0.6814	0.5886	0.7639	0.6153	0.3848	0.2840	0.4893	0.3176
(5) SGL	0.5018	0.3828	0.6090	0.4176	0.6835	0.5912	0.7692	0.6179	0.3858	0.2842	0.4914	0.3216
(6) VBPR	0.4722	0.3665	0.5670	0.3973	0.6699	0.5830	0.7464	0.6078	0.3464	0.2639	0.4364	0.2928
(7) MMGCN	0.4934	0.3714	0.6067	0.4081	0.6769	0.5643	0.7702	0.5945	0.3649	0.2709	0.4695	0.3047
(8) MGAT	0.5010	0.3781	0.6152	0.4152	0.6825	0.5784	0.7699	0.6067	0.3783	0.2820	0.4882	0.3175
(9) GRCN	0.5087	0.3910	0.6204	0.4272	0.6905	0.5937	0.7743	0.6208	0.3865	0.2861	0.4996	0.3225
(10) InvRL	0.5130	0.3955	0.6097	0.4268	0.6965	0.5986	0.7748	0.6237	0.3913	0.2926	0.4897	0.3244
(11) DMRL	0.5230	0.4075	0.6215	0.4396	0.6972	0.6008	0.7775	0.6272	0.3757	0.2876	0.4594	0.3146
MGCL	**0.5564**	**0.4322**	**0.6592**	**0.4656**	**0.7095**	**0.6121**	**0.7905**	**0.6359**	**0.4106**	**0.3064**	**0.5217**	**0.3423**
%Imp.	6.37%	6.06%	6.07%	5.91%	1.76%	1.88%	1.67%	1.39%	4.93%	4.72%	6.53%	5.52%

- **LightGCN** [18]: This is a light variant of NGCF, which removes the nonlinear networks from traditional graph convolution layers and accumulates the embeddings obtained at each layer as the final representation.
- **SGL** [22]: This model constructs three views for the user-item interaction graph and maximizes the mutual information between them using the contrastive learning approach.

The third is MMRec methods, including VBPR, MMGCN, MGAT, GRCN, InvRL, and DMRL. MMGCN, MGAT, GRCN, and InvRL are GCN-based MMRec methods.

- **VBPR** [26]: This model incorporates visual features into the item representations on the basis of BPRMF.
- **MMGCN** [42]: This model considers the features of multiple modalities and applies GCNs to propagate them on the user-item graph.
- **MGAT** [43]: Building on MMGCN, this model constructs attention networks to compute dynamic weights for different modalities to model more fine-grained multimodal user preferences.
- **GRCN** [44]: This model utilizes multimodal features to refine the user-item interaction graph (or to weight neighbor aggregation), and the final representation of the target node are the concatenations of the output of GCN and the original multimodal features.
- **InvRL** [46]: This model employs UltraGCN [54] as the base recommendation module and applies counterfactual inference techniques to enhance the multimodal representations of users and items.
- **DMRL** [46]: This model applies a disentanglement technique to multimodal representation learning and uses an attention mechanism to assign weights to the disentangled representations. Note that DMRL is not a GCN-based model.

4) Hyper-Parameter Settings: For all methods, we set the embedding size and batch size to 64 and 2048, respectively. We tune the learning rate in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and search the coefficient of L_2 regularization in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. For GCN-based methods, we tune the number of graph convolution layers in $\{1, 2, 3, 4, 5, 6\}$. Besides, we use the Xavier initializer [55] to initialize the embeddings and weight transformation matrices for all models.

B. Overall Comparison (RQ1)

To evaluate the superiority of the proposed MGCL, we compare it with all baselines *w.r.t.* {HR,NDCG}@{5,10} on the three datasets. As shown in Table II, we obtain the following findings:

- SVD++ and GCN-based CF methods (Table II, Models 3-5) consistently outperform BPRMF, which illustrates the importance of capturing higher-order collaborative signals. LightGCN achieves a significant improvement over NGCF in all cases, which is attributed to the fact that lightweight GCN is more suitable for capturing collaborative signals.
- The MMRec baselines (Table II, Models 6-9) generally outperform SVD++ and BPRMF, which demonstrates that modeling additional multimodal user preferences can improve the quality of recommendation. MGAT is slightly stronger than MMGCN, which is due to the attention networks in MGAT that learn more fine-grained weights for the different modalities.
- Notably, most GCN-based MMRec baselines (Table II, Models 6-8) are weaker than GCN-based CF baselines (LightGCN and SGL) *w.r.t.* NDCG on the three datasets. This result illustrates the presence of multimodal noise irrelevant to user preferences, and directly utilizing the complete multimodal features can contaminate the user and item representations, thus weakening the ranking preference modeling.
- GRCN, as a GCN-based MMRec baseline, consistently outperforms other baselines on both HR and NDCG, which is attributed to the fact that GRCN weakens the role of multimodal features in message propagation, thus alleviating the problem of multimodal noise contamination in the captured collaborative signals. This result also demonstrates that the separated graph learning schema in MGCL is also effective in capturing high-quality collaborative signals.
- InvRL always outperforms other baselines, mainly due to the fact that the backbone used in InvRL is the advanced

TABLE III
ABLATION STUDY. NOTE THAT, VARIANTS **w/o T** AND **w/o M&CL** DO NOT CONDUCT THE EXPERIMENTS ON THE TAOBAO DATASET DUE TO THE LACK OF TEXTUAL DATA IN THAT DATASETS

Models	Beauty		Art		Taobao	
	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5
MGCL	**0.5564**	**0.4322**	**0.7095**	**0.6121**	**0.4106**	**0.3064**
w/o CL	0.5415	0.4213	0.6992	0.5944	0.3992	0.2982
$\%Imp.$	-2.68%	-2.52%	-1.45%	-2.89%	-2.78%	-2.68%
w/o AT	0.5474	0.4261	0.7027	0.5958	0.3868	0.2900
$\%Imp.$	-1.62%	-1.41%	-0.96%	-2.66%	-5.80%	-5.35%
w/o V	0.5478	0.4248	0.7047	0.6029	0.3929	0.2942
$\%Imp.$	-1.55%	-1.71%	-0.68%	-1.50%	-4.31%	-3.98%
w/o T	0.5338	0.4131	0.6938	0.5996	-	-
$\%Imp.$	-4.06%	-4.42%	-2.21%	-2.04%	-%	-%
w/o M&CL	0.5149	0.3989	0.6827	0.5852	-	-
$\%Imp.$	-7.46%	-7.70%	-3.78%	-4.39%	-%	-%

UltraGCN and the additional use of counterfactual inference techniques further enhances the user and item representations. In addition, DMRL outperforms InvRL on the Beauty and Art datasets, partly due to the effectiveness of the disentanglement strategy in DMRL and partly due to the differences in how they fuse multimodal features—that is, InvRL fuses multiple modalities as a whole, whereas DMRL treats them separately. This result demonstrates that the separate manner is better at modeling user preferences across modalities. On the Taobao dataset, DMRL performs poorly, probably because the sparsest Taobao dataset is highly dependent on higher-order collaborative signals, whereas DMRL does not adopt the graph learning schema and thus lacks important higher-order collaborative signals.

- Our proposed MGCL yields the best performance on the three datasets. Specifically, MGCL achieves the mean improvement of 4.4% compared to the best performance of the baselines, which demonstrates the effectiveness of MGCL. We attribute this result to two main reasons, including (1) the separated graph learning schema in MGCL ensures that the collaborative signals are protected from multimodal noise contamination, and (2) the contrastive learning module in MGCL alleviates noise contamination in multimodal user preference modeling—that is, both collaborative signals and multimodal preference clues captured by MGCL are superior to those captured by the baselines.

C. Ablation Studies of Main Modules (RQ2)

We conduct ablation experiments to validate the effectiveness of each component in MGCL. We set up the following variants: **w/o AT**, which jointly trains recommendation and contrastive learning modules instead of the alternative training strategy; **w/o CL**, which removes the contrastive learning module from MGCL; **w/o V** and **w/o T**, which remove visual and textual features from MGCL, respectively; and **w/o M&CL**, which removes both multimodal features and the contrastive learning module from MGCL.

TABLE IV
EFFECTS OF CONTRASTIVE LEARNING ON THE COLLABORATIVE SIGNAL CAPTURING AND MULTIMODAL PREFERENCE MODELING

Models	Beauty		Art		Taobao	
	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5
$\mathbf{C}_{\text{w/o CL}}$	0.5092	0.3973	0.6905	0.6015	0.3914	0.2921
C	**0.5190**	**0.4056**	**0.6971**	**0.6061**	**0.3942**	**0.2944**
$\%Imp.$	1.92%	2.09%	0.96%	0.76%	0.72%	0.79%
$\mathbf{M}_{\text{w/o CL}}$	0.5210	0.4073	0.6681	0.5787	0.3602	0.2708
M	**0.5300**	**0.4126**	**0.6766**	**0.5843**	**0.3662**	**0.2751**
$\%Imp.$	1.73%	1.30%	1.27%	0.96%	1.67%	1.59%

Table III records the performance of these variants *w.r.t.* {HR, NDCG}@5 on the three datasets, generating the following findings:

- w/o CL is weaker than MGCL on the three datasets. This result verifies the effectiveness of the contrastive learning module in MGCL, which is further investigated in Section V-D. Additionally, w/o AT generally underperforms MGCL, which illustrates that the alternative training strategy can better optimize the recommendation and node self-discrimination tasks.
- An interesting phenomenon is that w/o V slightly underperforms MGCL on the Beauty and Art datasets, whereas w/o T is significantly weaker than MGCL. This result demonstrates that the contribution of different modalities to triggering user interest is significantly different. For the Beauty and Art datasets, the textual features play a more crucial role in user preference modeling than the visual features. Note that w/o V significantly underperforms MGCL on the Taobao dataset, which is attributed to the fact that w/o V on the Taobao dataset removes both the visual features and contrastive learning module from MGCL.
- w/o M&CL significantly underperforms w/o V, w/o T, and w/o CL on the Beauty and Art datasets, which demonstrates the importance of simultaneously using multimodal features and contrastive learning module for modeling multimodal user preferences.

D. Ablation Studies of Contrastive Learning (RQ3)

To investigate the effect of the contrastive learning module in MGCL on the capture of collaborative signals and multimodal preferences, we set up the following variants: **C**, which uses only the embeddings learned on the user-item interaction graph to predict interaction likelihood, i.e., $\sum_{m\in M}(e_{m,u}^{(k)})^T \cdot e_{m,i}^{(k)}$ are removed from (12); $\mathbf{C}_{w/o\,\text{CL}}$, which removes the contrastive learning module from **C**; **M**, which uses only the embeddings learned on the multimodal interaction graph to predict interaction likelihood, i.e., $(e_u^{(l)})^T \cdot e_i^{(l)}$ is removed from (12); and $\mathbf{M}_{w/o\,\text{CL}}$ removes the contrastive learning module from **M**.

Table IV records the performance of these variants on the three datasets *w.r.t.* {HR,NDCG}@5. We find that **C** and **M** consistently outperform $\text{C}_{w/o\,\text{CL}}$ and $\text{M}_{w/o\,\text{CL}}$, respectively, suggesting that the captured collaborative signals and multimodal preference clues can be further enhanced by performing contrastive

TABLE V
EFFECT OF MCL ON BPRMF AND LIGHTGCN

Methods	HR			NDCG		
	k=5	k=10	k=20	k=5	k=10	k=20
BPRMF	0.4558	0.5463	0.6459	0.3543	0.3837	0.4088
BPRMF+MCL	**0.5286**	**0.6263**	**0.7284**	**0.4108**	**0.4425**	**0.4684**
$\%Imp.$	15.97%	14.64%	12.77%	15.95%	15.32%	14.58%
LightGCN	0.5002	0.6063	0.7178	0.3777	0.4131	0.4399
LightGCN+MCL	**0.5394**	**0.6464**	**0.7490**	**0.4117**	**0.4463**	**0.4723**
$\%Imp.$	7.84%	6.61%	4.35%	9.00%	8.04%	7.37%

learning between them. An interesting phenomenon is that **M** outperforms **C** on the Beauty dataset, whereas M underperforms **C** on the Art and Taobao datasets, which is consistent with the results in Table III—that is, the multimodal features give the most improvements on the Beauty dataset. This result suggests that the importance of collaborative signals and multimodal preference clues for modeling overall user preferences varies from scenario to scenario.

E. Scalability of MGCL (RQ4)

In this part, we investigate the scalability of MGCL. Specifically, we believe that the key components in MGCL can be combined as a model-agnostic framework and applied to other MF-based recommender methods seamlessly. We term this framework as Multimodal Contrastive Learning (MCL), which includes the prediction function (cf. (12)), objective function (cf. (15)), contrastive learning module (cf. (17)), and alternative training strategy (cf. Algorithm 1). Here we incorporate MCL into BPRMF and LightGCN to obtain two model variants **BPRMF+MCL** and **LightGCN+MCL**.

Table V and Fig. 3 record their performance on the Beauty dataset, generating the following findings:

- For BPRMF, its HR and NDCG improve by an average of 16.51% and 16.56%, respectively, after the incorporation of the MCL module. LightGCN also achieves significant improvements after incorporating the MCL module. These results demonstrate the remarkable scalability of the MCL module. Essentially, the gains come mainly from the additional modeling of multimodal preferences, and the further enhancement of the representations by contrastive learning.
- According to Fig. 3 (5)–(6), it is clear that LightGCN and BPRMF converge faster after incorporating the MCL module. Specifically, the optimal training epochs for LightGCN+MCL and BPRMF+MCL are 160 and 120, respectively, whereas those for LightGCN and BPRMF are 660 and 960, respectively. This indicates that our proposed MCL module can significantly reduce the training cost of the model, meanwhile ensuring the higher accuracy of recommendation. We attribute this result to two main reasons: (a) the additional utilization of multimodal features in MCL, which compensates for the difficulty of user preference learning due to sparse interaction data, and (b) the contrastive learning task in MCL introduces more additional self-supervised signals, which can further enhance the representations of users and items, thus speeding up the model convergence. In addition, the improvements of NDCG are relatively greater than that of HR, which indicates that the incorporation of the MCL module improves the modeling of ranking preferences, i.e., the recommendation list generated by the model not only makes it easier to find items that match the users' preferences, but also ranks the items closer to the top of the list.

F. Case Study (RQ5)

To investigate the effectiveness of our proposed MGCL in eliminating noise contamination, we design a case study on the Taobao dataset, as shown in Fig. 4. For each example in the figure, the item on the left is not similar to the target item, but its noise content is similar to that of the target item, whereas the item on the right is truly similar to the target item, but its noise content is different from that of the target item. We select MMGCN and MGCL to calculate the similarities between items (we normalize these similarities to facilitate comparison). We generate the following findings:

- In Fig. 4(a) and 4(b), MMGCN incorrectly predicts that the target items are more similar to the left items, which may be attributed to the fact that they have similar backgrounds and layouts and the presence of people. In contrast, MGCL correctly predicts that the target items are more similar with the right items with high confidence, although they have many dissimilarities in content. These results demonstrate that MGCL can effectively reduce the negative impact of preference-irrelevant content on user preference modeling.
- Fig. 4(c) is a difficult example in which i_{105}, i_{413}, and i_{110} are products in the same category but with differences in style. More precisely, i_{105} and i_{413} are extremely similar in visual content but different in style (i_{105} is a dress and i_{413} is a long dress), whereas i_{105} and i_{110} are very different in noisy content (the background and the presence of people) but the same in style (they are both long dresses). Clearly, i_{105} and i_{110} are more similar in terms of user preference. Unexpectedly, both MMGCN and MGCL incorrectly predict that i_{105} and i_{413} are more similar, but MGCL is less confident in this prediction than MMGCN, which also suggests that MGCL can alleviate multimodal noise contamination to some extent.

G. Hyper-Parameter Studies

Here, as shown in Table VI, we first study the depth of graph convolution layers L. Then, as shown in Figs. 5, we analyze the effect of hyper-parameters λ.

1) Effect of the Number of Graph Convolution Layers: To investigate the effect of the number of graph convolution layers, we search the layers in {1, 2, 3, 4, 5, 6}. As shown in Table VI, we find that the model performance gradually improves as the number of graph convolution layers increases on the three datasets, which demonstrates that capturing high-order collaborative relations can better model user preference. However, the

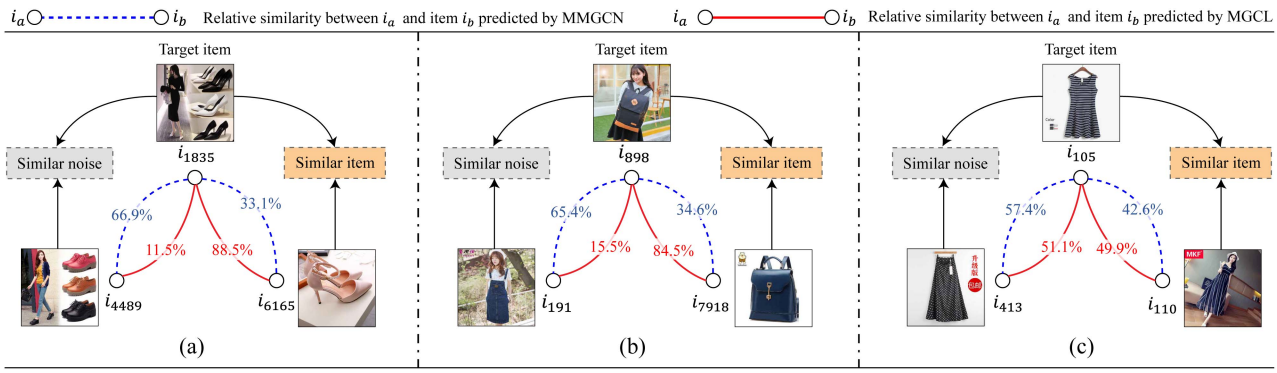

Fig. 4. Effect of MGCL on eliminating the noise contamination, where i_{1835} denotes an item with the ID of 1835. (a), (b), and (c) represent the similarity predictions in the three difficult cases, respectively.

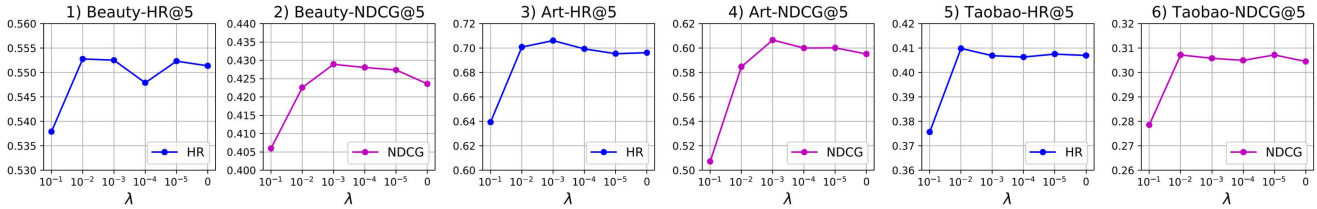

Fig. 5. Performance of MGCL *w.r.t.* different L_2 regularization coefficient λ on the three datasets.

TABLE VI
EFFECT OF THE NUMBER OF GRAPH CONVOLUTION LAYERS ON MGCL

Layers	Beauty		Art		Taobao	
	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5
1	0.5476	0.4216	0.6989	0.6017	0.3866	0.2890
2	0.5516	0.4287	0.7057	0.6084	0.4010	0.3015
3	0.5546	0.4305	0.7066	0.6085	0.4039	0.3039
4	**0.5564**	**0.4322**	0.7070	0.6105	0.4069	0.3054
5	0.5540	0.4298	**0.7095**	**0.6121**	**0.4106**	**0.3064**
6	0.5501	0.4270	0.7055	0.6101	0.4035	0.3019

performance of MGCL begins to decrease as the layers continue to go deeper, which is attributed to the fact that too deeper graph convolution layers are prone to the issue of over-smoothing. To be more specific, the optimal layers on the three datasets are 4, 5, and 5, respectively. The reason for this is that Art and Taobao datasets are sparser than the Beauty dataset. Therefore, on the Art and Taobao datasets, the model needs to perform more neighbor aggregation to obtain enough collaborative signals.

2) Effect of L_2 Regularization Coefficient λ: MGCL achieves the best results on the Beauty, Art, and Taobao datasets when $\lambda = 10^{-3}$, $\lambda = 10^{-3}$, and $\lambda = 10^{-2}$, respectively. Specifically, we observe that there is no significant degradation in the performance of MGCL when the λ is small (even when the λ is set to 0). This result illustrates that the proposed MGCL has high stability against overfitting. When λ is greater than 10^{-2}, the model performance drops sharply on the three datasets, probably because the overly large L_2 regularization coefficient limits the learning ability of the model. Therefore, when applying MGCL to a new recommendation scenario (i.e., a new dataset), we recommend choosing a smaller L_2 regularization coefficient.

VI. CONCLUSION AND FUTURE WORK

In this work, we propose a novel MMRec model, MGCL. Specifically, we first construct three parallel GCNs to capture collaborative signals, visual preference clues, and textual preference clues, respectively. Then, we employ contrastive learning to eliminate noise contamination in multimodal user preference modeling. Next, to ensure sufficient optimization of the recommendation task and contrastive learning task, we adopt an alternative training strategy to optimize them. Finally, we conduct extensive experiments to validate the effectiveness of MGCL.

To the best of our knowledge, this is the first attempt to mitigate the problem of multimodal noise contamination in MMRec research by enhancing a graph learning schema and employing a contrastive learning technique. In the future, we need to further explore three main extensions. Specifically, (a) we would consider removing some of the easily distinguishable noise via advanced techniques in CV and NLP during the multimodal feature pre-processing to better address the noise contamination problem; (b) we would investigate how to more rationally combine causal inference techniques with multimodal user preference modeling in order to directly discover and eliminate preference-irrelevant noise information; and (c) we would explore how to leverage multimodal data to improve the interpretability of recommendations.

REFERENCES

[1] Y. Wei et al., "Hierarchical user intent graph network for multimedia recommendation," *IEEE Trans. Multimedia*, vol. 24, pp. 2701–2712, 2022.
[2] D. Cai, S. Qian, Q. Fang, and C. Xu, "Heterogeneous hierarchical feature aggregation network for personalized micro-video recommendation," *IEEE Trans. Multimedia*, vol. 24, pp. 805–818, 2022.

[3] K. Liu et al., "MEGCF: Multimodal entity graph collaborative filtering for personalized recommendation," *Trans. Inf. Syst.*, pp. 1–26, 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3544106

[4] X. Chen et al., "E-commerce storytelling recommendation using attentional domain-transfer network and adversarial pre-training," *IEEE Trans. Multimedia*, vol. 24, pp. 506–518, 2022.

[5] C. Yan, B. Gong, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1445–1451, Apr. 2021.

[6] L. Wu et al., "A hierarchical attention model for social contextual image recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 10, pp. 1854–1867, Oct. 2020.

[7] X. Gao et al., "Hierarchical attention network for visually-aware food recommendation," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1647–1659, Jun. 2020.

[8] W. Min, S. Jiang, and R. Jain, "Food recommendation: Framework, existing solutions, and challenges," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2659–2671, Oct. 2020.

[9] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide*, 2001, pp. 285–295.

[10] T. Yao et al., "Self-supervised learning for large-scale item recommendations," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 4321–4330.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–14.

[12] R. Ying et al., "Graph convolutional neural networks for web-scale recommender systems," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 974–983.

[13] S. Wang et al., "Graph learning based recommender systems: A review," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 4644–4652.

[14] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 165–174.

[15] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[16] F. Wu et al., "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn.*2019, pp. 6861–6871.

[17] K. Liu, F. Xue, and R. Hong, "RGCF: Refined graph convolution collaborative filtering with concise and expressive embedding," *Intell. Data Anal.*, vol. 26, no. 2, pp. 427–445, 2022.

[18] X. He et al., "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 639–648.

[19] K. Liu, F. Xue, X. He, D. Guo, and R. Hong, "Joint multi-grained popularity-aware graph convolution collaborative filtering for recommendation," *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 1, pp. 72–83, Feb. 2023.

[20] Y. Wei et al., "Contrastive learning for cold-start recommendation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 5382–5390.

[21] Y. Zhu et al., "Graph contrastive learning with adaptive augmentation," in *Proc. Web Conf.*, 2021, pp. 2069–2080.

[22] J. Wu et al., "Self-supervised graph learning for recommendation," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 726–735.

[23] Y. Yang, L. Wu, R. Hong, K. Zhang, and M. Wang, "Enhanced graph learning for collaborative filtering via mutual information maximization," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 71–80.

[24] L. Wu, X. He, X. Wang, K. Zhang, and M. Wang, "A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 05, 2022, doi: 10.1109/TKDE.2022.3145690.

[25] Q. Truong, A. Salah, and H. W. Lauw, "Multi-modal recommender systems: Hands-on exploration," in *Proc. 15th ACM Conf. Recommender Syst.*, 2021, pp. 834–837.

[26] R. He and J. J. McAuley, "VBPR: Visual Bayesian personalized ranking from implicit feedback," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 144–150.

[27] X. Geng, H. Zhang, J. Bian, and T.-S. Chua, "Learning image and user features for recommendation in social networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4274–4282.

[28] S. Wang et al., "What your images reveal: Exploiting visual contents for point-of-interest recommendation," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 391–400.

[29] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1257–1264.

[30] C. Lei, D. Liu, W. Li, Z. Zha, and H. Li, "Comparative deep learning of hybrid representations for image recommendations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2545–2553.

[31] W. C. Kang, C. Fang, Z. Wang, and J. McAuley, "Visually-aware fashion recommendation and design with generative image models," in *Proc. IEEE Int. Conf. Data Mining*, 2017, pp. 207–216.

[32] J. Chen et al., "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 335–344.

[33] W. Niu, J. Caverlee, and H. Lu, "Neural personalized ranking for image recommendation," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 423–431.

[34] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu, "Convolutional matrix factorization for document context-aware recommendation," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 233–240.

[35] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, 2017, pp. 425–434.

[36] D. Liu, J. Li, B. Du, J. Chang, and R. Gao, "DAML: Dual attention mutual learning between ratings and reviews for item recommendation," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 344–352.

[37] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural attentional rating regression with review-level explanations," in *Proc. World Wide Web Conf.*, 2018, pp. 1583–1592.

[38] H. Liu et al., "NRPA: Neural recommendation with personalized attention," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 1233–1236.

[39] C. Wu et al., "Reviews meet graphs: Enhancing user and item representations for recommendation with hierarchical attentive graph neural network," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 4884–4893.

[40] J. Shuai et al., "A review-aware graph contrastive learning framework for recommendation," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1283–1293.

[41] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y. Ma, "Collaborative knowledge base embedding for recommender systems," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 353–362.

[42] Y. Wei et al., "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1437–1445.

[43] Z. Tao et al., "MGAT: Multimodal graph attention network for recommendation," *Inf. Process. Manage.*, vol. 57, no. 5, 2020, Art. no. 102277.

[44] Y. Wei, X. Wang, L. Nie, X. He, and T. Chua, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3541–3549.

[45] F. Liu et al., "Disentangled multimodal representation learning for recommendation," *IEEE Trans. Multimedia*, early access, Oct. 26, 2022, doi: 10.1109/TMM.2022.3217449.

[46] X. Du, Z. Wu, F. Feng, X. He, and J. Tang, "Invariant representation learning for multimedia recommendation," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 619–628.

[47] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. UAI*, 2009, pp. 452–461.

[48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[49] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 188–197.

[50] X. He et al., "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.

[51] F. Xue et al., "Deep item-based collaborative filtering for top-n recommendation," *Trans. Inf. Syst.*, vol. 37, no. 3, pp. 1–25, 2019.

[52] X. He, Z. He, X. Du, and T. Chua, "Adversarial personalized ranking for recommendation," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 355–364.

[53] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov Data Mining*, 2008, pp. 426–434.

[54] K. Mao et al., "UltraGCN: Ultra simplification of graph convolutional networks for recommendation," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 1253–1262.
[55] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

Kang Liu received the B.S. degree from the China University of Geosciences, Wuhan, China, in 2017. He is currently working toward the M.S. and Ph.D. degrees with the Hefei University of Technology, Hefei, China. His research interests include recommender systems, data mining, and multimedia analysis.

Feng Xue received the Ph.D. degree from the Department of Computer Science, Hefei University of Technology (HFUT), Hefei, China. He is currently a Professor with HFUT. His research interests include artificial intelligence, multimedia analysis, and recommender systems.

Dan Guo (Member, IEEE) received the Ph.D. degree in system analysis and integration from the Huazhong University of Science and Technology, Wuhan, China, in 2010. She is currently a Professor with the School of Computer and Information, Hefei University of Technology, Hefei, China. Her research interests include computer vision, machine learning, and intelligent multimedia content analysis.

Peijie Sun received the Ph.D. degree from the Hefei University of Technology, Hefei, China, in 2022. He is currently a Postdoc with Tsinghua University, Beijing, China. He has authored or coauthored several papers in leading conferences and journals, which include, WWW, Proceeding 41st International ACM SIGIR Conference Research and Development in Information Retrieval, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTION ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and ACM TOIS. His research interests include data mining and recommender systems.

Shengsheng Qian (Member, IEEE) received the B.E. degree from the Jilin University, Changchun, China, in 2012, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His research interests include social media data mining and social event content analysis.

Richang Hong (Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. He is currently a Professor with the Hefei University of Technology, Hefei, China. From 2008 to 2010, he was a Research Fellow with the School of Computing, National University of Singapore, Singapore. He has co-authored more than 100 publications in his research areas, which include multimedia content analysis and social media. He was the recipient of the Best Paper Award in the ACM Multimedia 2010, Best Paper Award in the ACM ICMR 2015, and Honorable Mention of the IEEE Transactions on Multimedia Best Paper Award. He was an Associate Editor for the IEEE MULTIMEDIA MAGAZINE, *Information Sciences and Signal Processing*, Elsevier and the Technical Program Chair of the MMM 2016. He is a Member of ACM and the Executive Committee Member of the ACM SIGMM China Chapter.