# Latent Structure Mining with Contrastive Modality Fusion for Multimedia Recommendation

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, *Member, IEEE,* Mengqi Zhang,
Shu Wu, *Senior Member, IEEE,* and Liang Wang, *Fellow, IEEE*

**Abstract**—Multimedia contents are of predominance in the modern Web era. Recent years have witnessed growing research interests in multimedia recommendation, which aims to predict whether a user will interact with an item with multimodal contents. Most previous studies focus on modeling user-item interactions with multimodal features included as side information. However, this scheme is not well-designed for multimedia recommendation. Firstly, only *collaborative* item-item relationships are implicitly modeled through high-order item-user-item co-occurrences. Considering that items are associated with rich contents in multiple modalities, we argue that the latent *semantic* item-item structures underlying these multimodal contents could be beneficial for learning better item representations and assist the recommender models to comprehensively discover candidate items. Secondly, although previous studies consider multiple modalities, their ways of fusing multiple modalities by linear combination or concatenation is insufficient to fully capture content information of items and item relationships. To address these deficiencies, we propose a latent structure MIning with ContRastive mOdality fusion model, which we term MICRO for brevity. To be specific, we devise a novel modality-aware structure learning module, which learns item-item relationships for each modality. Based on the learned modality-aware latent item relationships, we perform graph convolutions to explicitly inject item affinities into modality-aware item representations. Additionally, we design a novel multimodal contrastive framework to facilitate item-level multimodal fusion by mining both modality-shared and modality-specific information. Finally, the item representations are plugged into existing collaborative filtering methods to make accurate recommendation. Extensive experiments on three real-world datasets demonstrate the superiority of our method over state-of-arts and rationalize the design choice of our work.

**Index Terms**—Multimedia Recommendation, Graph Structure Learning, Contrastive Learning.

## 1 Introduction

With the rapid development of Internet, information overload has become an increasingly crucial challenge. Nowadays, users are easily accessible to large amounts of online information represented in multiple modalities, including images, texts, videos, etc. For example, visual appearances and textual descriptions play important roles for selecting products online; visual covers and textual tags allow users to find interesting items from a large amount of instant videos. Therefore, multimedia recommendation, which aims to predict whether a user will interact with an item with multimodal contents, has attracted a lot of research interests and has been successfully applied to many online applications, such as e-commerce, instant video platforms, and social media platforms.

Collaborative Filtering (CF), as one of the most prevalent techniques in personalized recommendation, has been widely studied previously. Focusing on exploiting abundant user-item interactions, CF methods group users according to their historical interactions, by encoding users and items into low-dimensional dense vectors and making recommendation based on these embeddings [2–4]. Following traditional CF frameworks, early work on multimedia recommendation like VBPR [5], DeepStyle [6], and ACF [7] incorporates multimodal features as side information in addition to the learned dense vectors of items, so as to group users based on both historical interactions and item contents. Park et al. [8] propose to explicitly capture the information hidden in also-viewed products, i.e., a list of products that have also been viewed by users who have viewed a target product. The also-viewed relationship can be regarded as a special kind of co-interacted relationship. Lee and Abu-El-Haija [9] propose to minimize the similarity of the content vectors of co-watched items, which exploits co-interacted item-item

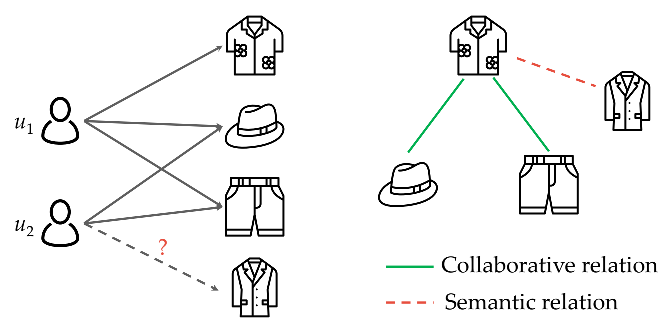


Fig. 1. A toy example of recommendation with two types of item relations. In this paper, we argue that semantic structures mined from multimodal features are helpful for comprehensively discovering candidate items supplementary to collaborative signals in traditional work.

- *This paper is an extended version of LATTICE [1], which has been published in the Proceedings of the 29th ACM International Conference on Multimedia.*
- *J. Zhang, Q. Liu, M. Zhang, S. Wu, and L. Wang are with the Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences.*
- *Y. Zhu is with the Department of Computer Science, University of California at Los Angeles.*
- *Corresponding author: Shu Wu (shu.wu@nlpr.ia.ac.cn).*

*Manuscript received April 19, 2005; revised August 26, 2015.*

relationships through item-user-item occurrences.

Inspired by the recent surge of graph neural networks [10, 11], Wang et al. [12] propose to model user-item relationships as bipartite graphs. The first-order connectivities in user-item graphs indicate the interaction history. The second-order connectivities reveal collaborative relations that similar users (or items) who have co-interacted with the same items (or users). These graph-based recommender systems [12–14] inject high-order connectivities into the embeddings to learn better representations. Recently, many attempts have been made to integrate multimodal contents into graph-based recommendation systems. MMGCN [15] constructs modality-specific user-item interaction graphs to model user preferences specific to each modality. Following MMGCN, GRCN [16] utilizes multimodal features to refine user-item interaction graphs by identifying false-positive feedbacks and pruning the corresponding noisy edges. HUIGN [17] constructs a co-interacted item graph, where the edge corresponds to the item pair consumed by the same users. By conducting hierarchical GNNs on the co-interacted item graph, HUIGN can mine users' intents at different levels.

Despite their effectiveness, previous attempts suffer from two limitations. Firstly, existing work fails to comprehensively model item-item relationships, which have been proved to be important in recommender systems [18]. Specifically, only **collaborative** relations are considered through high-order item-user-item co-occurrences [8, 9, 17]. However, **semantic** relations which reflect the content of items, are not explicitly modeled. Taking Figure 1 as an example, existing methods will recommend the shirt () for $u_2$ according to collaborative relations, since shirts (), hats (), and pants () all interacted with $u_1$. However, previous work may not be able to recommend coats () to $u_2$, which are semantically (visually in this example) similar to shirts. Considering that items are associated with rich multimodal content features in multimedia recommendation, there exist a wealth of semantic relations underlying multimodal contents, which would assist the recommender models to comprehensively discover candidate items.

Secondly, most previous attempts disregard the item-level multimodal fusion. Early work [5, 6, 19] only focuses on uni-modal information; other work on multimedia recommendation [15, 16] conducts multimodal fusion by simple linear combination or concatenation, the inductive bias behind which is that all items share the same fusion mechanism (e.g., the same combination weights). However, users usually focus on different modalities when browsing different items. For example, one may pay more attention to the visual modality when selecting clothes, while focusing more on textual information when picking books. To this end, we conduct item-level multimodal fusion, allowing the model to utilize the most important parts of different items in a flexible manner and therefore learn better item representations. Specifically, we propose to mine latent semantic item-item relationships underlying multimodal features of items and conduct item-level multimodal fusion based on the learned structures.

As shown in Figure 2, the proposed MICRO consists of four key components. Firstly, we develop a novel modality-aware structure learning layer, which learns modality-aware item structures from content features of each modality. Secondly, we perform graph convolutions on the learned modality-aware latent graphs to explicitly consider item relationships of each modality individually. Thirdly, we devise a novel multimodal contrastive framework to consider both modality-shared and modality-specific information. Finally, the resulting enhanced item representations are infused with item relationships in multiple modalities, which will be added into the output item embeddings of CF models to make recommendations.

Our model enjoys two additional benefits. Firstly, MICRO can alleviate the cold-start problem. Previous graph-based multimedia recommendation methods face cold-start problems where long-tailed items are only interacted with few users or even never interacted with users. Since previous methods utilize multimodal content features based on the user-item interaction graph, those long-tailed items will become isolated nodes in that graph, which will reduce the effectiveness of multimodal information. Our work, on the contrary, can alleviate the cold-start problem in two ways: (1) we mine latent item-item structures and the long-tailed items will get similar user feedbacks from their learned neighbors; (2) the multimodal contrastive framework serves as an auxiliary training signal that helps learn better item representations involved with relation information. The second benefit is that, MICRO can serve as a flexible play-and-plug module. Unlike previous attempts which utilize multimodal features based on dedicated user-item aggregation strategies, MICRO separates the usage of multimodal features with the usage of user-item interactions and is thus agnostic to downstream CF methods.

In summary, the main contribution of this work is threefold.

- We highlight the importance of explicitly exploiting item relationships and explicitly consider item-level multimodal fusion in multimedia recommendation.
- We propose a novel method to mine latent item relations and conduct item-level multimodal fusion based on the mined structures, which consider both modality-shared and modality-specific information.
- We perform extensive experiments on three public datasets. Notably, our method outperforms the state-of-the-art methods by 20% on average in terms of different metrics, validating the effectiveness of our proposed model.

To foster reproducible research, our code is made publicly available at https://github.com/CRIPAC-DIG/MICRO.

## 2 Preliminaries

In this section, we first formulate the multimedia recommendation problem. Then, to motivate our model design, we use two simple and intuitive experiments, from item and user perspectives respectively, to show that users tend to buy semantically similar items. That is, semantic item-item relationships are helpful for comprehensively discovering candidate items.

### 2.1 Problem Definition

Let $\mathcal{U}, \mathcal{I}\ (|\mathcal{I}| = N)$ denote the set of users and items, respectively. Each user $u \in \mathcal{U}$ is associated with a set

TABLE 1
Average semantic similarity of all items and co-interacted items.

| Dataset | Modality | All Items | Co-interacted Items |
|---|---|---|---|
| Clothing | Visual | 0.2239 | 0.3958 |
| | Textual | 0.4206 | 0.5830 |
| Sports | Visual | 0.2184 | 0.3547 |
| | Textual | 0.3895 | 0.5423 |
| Baby | Visual | 0.2240 | 0.3534 |
| | Textual | 0.4413 | 0.5405 |

TABLE 2
The proportion (%) of users buying semantically similar items with respect to different $k$.

| Dataset | Modality | $k = 5$ | $k = 10$ | $k = 15$ | $k = 20$ |
|---|---|---|---|---|---|
| Clothing | Visual | 46.88 | 51.34 | 54.22 | 56.33 |
| | Textual | 54.90 | 60.21 | 63.41 | 65.57 |
| Sports | Visual | 42.18 | 45.58 | 47.71 | 49.39 |
| | Textual | 53.56 | 58.48 | 61.64 | 63.90 |
| Baby | Visual | 44.37 | 48.17 | 50.87 | 53.36 |
| | Textual | 55.25 | 59.58 | 62.45 | 64.89 |

of items $\mathcal{I}^u$ with positive feedbacks which indicate the preference score $y_{ui} = 1$ for $i \in \mathcal{I}^u$. $\mathbf{x}_u, \mathbf{x}_i \in \mathbb{R}^d$ is the input ID embedding of $u$ and $i$, respectively, where $d$ is the embedding dimension. Besides user-item interactions, multimodal features are offered as content information of items. We denote the modality features of item $i$ as $\mathbf{e}_i^m \in \mathbb{R}^{d_m}$, where $d_m$ denotes the dimension of the features, $m \in \mathcal{M}$ is the modality, and $\mathcal{M}$ is the set of modalities. The purpose of multimedia recommendation is to accurately predict users' preferences by ranking items for each user according to predicted preference scores $\hat{y}_{ui}$. In this paper, we consider visual and textual modalities denoted by $\mathcal{M} = \{\mathrm{v}, \mathrm{t}\}$. Please kindly note that our method is not fixed to the two modalities and multiple modalities can be involved.

## 2.2 Pilot Studies

Firstly, from the item perspective, we conduct an experiment to show that co-interacted items are much more semantically similar. We compute the cosine similarity between all items as the baseline and compute the similarity between co-interacted items. The averages are summarized in Table 1. We can observe that co-interacted items (items bought by the same user) are much more similar, which demonstrates that users tend to buy semantically similar items.

Secondly, from the user perspective, we count the proportion of users buying semantically similar items. We intuitively define $i_1$ and $i_2$ are semantically similar if $i_1$ is among the $k$ items most similar to $i_2$, or $i_2$ is among the $k$ items most similar to $i_1$, where a smaller $k$ means a smaller range. Table 2 reports the proportion of users buying semantically similar items with respect to different $k$. We can observe that even with a small $k$, the majority of users tend to buy semantically similar items.

# 3 The Proposed Method

In this section, we introduce our model in detail. As illustrated in Figure 2, there are four main components in our proposed framework: (1) a modality-aware graph structure learning layer that learns item graph structures from content features of each modality, (2) graph convolutional layers that learn the modality-aware item embeddings by injecting item-item affinities based on the learned graph structures, (3) an contrastive multimodal fusion framework to promote item-level multimodal fusion by considering both modality-shared and modality-specific information, and (4) downstream CF methods.

## 3.1 Modality-aware Latent Structure Mining

Multimodal features provide rich and meaningful content information of items, while existing methods only utilize multimodal features as side information for each item, ignoring the important *semantic* relationships of items underlying features. In this section, we introduce how to discover the underlying latent graph structure of item graphs in order to learn better item representations.

To be specific, we first construct initial $k$-Nearest-Neighbor ($k$NN) modality-aware item graphs $\widetilde{\mathbf{S}}^m$ by utilizing raw multimodal features. After that, we learn the latent graph structures $\widetilde{\mathbf{A}}^m$ from transformed multimodal features. Finally, we combine the learned structures with the initial structures by a skip connection.

### 3.1.1 Constructing Initial Modality-aware Graphs

We first construct initial $k$NN modality-aware graph $\mathbf{S}^m$ by using raw features for each modality $m$. Based on the hypothesis that similar items are more likely to interact than dissimilar items [20], we quantify the *semantic* relationship between two items by their similarity. Common options for node similarity measurement include cosine similarity [21], kernel-based functions [22], and attention mechanisms [23]. Our method is agnostic to similarity measurements, and we opt to the simple and parameter-free cosine similarity in this paper. The similarity matrix $\mathbf{S}^m \in \mathbb{R}^{N \times N}$ is computed by

$$S_{ij}^m = \frac{(\mathbf{e}_i^m)^\top \mathbf{e}_j^m}{\|\mathbf{e}_i^m\| \|\mathbf{e}_j^m\|}. \tag{1}$$

Typically, the graph adjacency matrix is supposed to be non-negative but $S_{ij}$ ranges between $[-1, 1]$. Thus, we suppress its negative entries to zeros. Moreover, common graph structures are much sparser other than a fully-connected graph, which is computationally demanding and might introduce noisy, unimportant edges [23]. We conduct $k$NN sparsification [24] on the dense graph: for each item $i$, we only keep edges with the top-$k$ confidence scores:

$$\widehat{S}_{ij}^m = \begin{cases} S_{ij}^m, & S_{ij}^m \in \text{top-}k(S_{i,:}^m), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $S_{i,:}^m$ denotes the $i$-row of $\mathbf{S}$, and $\widehat{\mathbf{S}}^m$ is the resulting sparsified, directed graph adjacency matrix. To alleviate the exploding or vanishing gradient problem [10], we normalize the adjacency matrix as:

$$\widetilde{\mathbf{S}}^m = (\mathbf{D}^m)^{-\frac{1}{2}} \widehat{\mathbf{S}}^m (\mathbf{D}^m)^{-\frac{1}{2}}, \tag{3}$$

Fig. 2. The overall framework of our proposed MICRO model. Firstly, we develop a novel modality-aware structure learning layer to mine the modality-aware latent item-item semantic relationships from multimodal features. Secondly, we employ graph convolutions on the learned modality-aware graphs to explicitly model item relationships of each modality individually. Thirdly, we devise a novel contrastive multimodal fusion framework to adaptively capture item relationships shared between multiple modalities in a self-supervised manner. Finally, the resulting item representations are infused with item relationships in multiple modalities, which will be added into the output item embeddings of CF models to make recommendation.

where $D^m \in \mathbb{R}^{N \times N}$ is the diagonal degree matrix of $\tilde{S}^m$ and $D^m_{ii} = \sum_j \tilde{S}^m_{ij}$.

#### 3.1.2 Learning Latent Modality-aware Graphs

Although we have obtained the modality-aware initial graph structures $\tilde{S}^m$ by utilizing raw multimodal features, they may not be ideal for the recommendation task. This is because the raw multimodal features are often noisy or even incomplete due to the inevitably error-prone data measurement or collection. To this end, we propose to dynamically learn the graph structures by the transformed multimodal features and combine the learned structures with initial ones.

Firstly, we transform raw modality features into high-level features $\mathbf{e}_i^m$:

$$\mathbf{e}_i^m = W_m e_i^m + b_m, \tag{4}$$

where $W_m \in \mathbb{R}^{d \times d_m}$ and $b_m \in \mathbb{R}^d$ denote the trainable transformation matrix and the bias vector, respectively. We dynamically infer the graph structures utilizing $\mathbf{e}_i^m$, repeat the graph learning process described in Eqs. (1, 2, 3) and obtain the adjacency matrix $\hat{A}^m$.

Although the initial graph could be noisy, it still carries rich and useful information regarding item graph structures. Also, drastic change of adjacency matrix will lead to unstable training. To keep rich information of initial item graph and stabilize the training process, we add a skip connection that combines the learned graph with the initial graph:

$$A^m = \lambda \tilde{S}^m + (1-\lambda)\hat{A}^m, \tag{5}$$

where $\lambda \in (0, 1)$ is the coefficient of skip connection that controls the amount of information from the initial structure. The obtained $A^m$ is the final graph adjacency matrix representing latent structures for modality $m$.

It is worth mentioning that both $\tilde{S}^m$ and $\hat{A}^m$ are sparsified and normalized matrices, thus the final adjacency matrix $A^m$ is also sparsified and normalized, which is computationally efficient and stabilizes gradient backpropagation.

### 3.2 Item Affinity Learning with Graph Convolutions

After obtaining the modality-aware latent structures, we perform graph convolution operations to learn better item representations by injecting item-item affinities into the embedding process. Graph convolutions can be treated as message propagation and aggregation. Through propagating the item representations from its neighbors, one item can aggregate information within the first-order neighborhood. Furthermore, by stacking multiple graph convolutional layers, the high-order item-item relationships can be captured.

Following Wu et al. [25] and He et al. [14], we employ simple message propagation and aggregation without feature transformation and non-linear activations which is effective and computationally efficient. In the $l$-th layer, the message passing and aggregation could be formulated as:

$$H^m_{(l)} = A^m H^m_{(l-1)}, \tag{6}$$

where $H^m_{(l)} \in \mathbb{R}^{N \times d}$ is the $l$-th layer item embedding matrix of modality $m$, the $i$-th row of which denotes the embedding vector of item $i$. For all modalities $m \in \mathcal{M}$, we use the same item ID embedding matrix to initialize the input embedding matrix $H^m_{(0)}$. We utilize ID embedding vector of items as input representations rather than multimodal features, since we employ graph convolutions to directly capture item-item affinities and multimodal features are used to bridge semantic relationships. After stacking $L$ layers, $H^m_{(L)}$ encodes the high-order item-item relationships of modality $m$.

### 3.3 Multimodal Fusion with Contrastive Learning

Multiple modalities convey both complementary and supplementary information [26]. To this end, we first utilize contrastive learning to extract modality-shared representations and then deploy an orthogonality constraint to extract modality-specific representations.

#### 3.3.1 Mining Modality-Shared Information

For simplicity, we omit the subscript $(L)$ and use $h_i^m$ hereafter to denote the $i$-th row of $H_{(L)}^m$, which represents the embedding of item $i$ in modality $m$. The importance of each modality corresponding to item $i$ can be formulated as follows:

$$\alpha_i^m = \mathrm{softmax}\left(q_1^\top \tanh(W h_i^m + b)\right), \tag{7}$$

where $q_1 \in \mathbb{R}^d$ denotes attention vector and $W \in \mathbb{R}^{d \times d}, b \in \mathbb{R}^d$ denote the weight matrix and the bias vector, respectively. Note that these parameters are shared for all modalities. Then, the modality-common embedding of item $i$ can be represented as:

$$h_{c,i} = \sum_{m=1}^{|\mathcal{M}|} \alpha_i^m h_i^m. \tag{8}$$

We devise a novel self-supervised auxiliary task to adaptively distill the shared information from multiple modalities. Existing contrastive learning frameworks [27] seek to maximize the agreement among differently augmented views of the same data examples, which has been proven to be effective in multi-view representation learning [28, 29] and multimodal tasks [30, 31]. In this work, since multiple modality-aware graphs are involved, we propose to construct self-supervision signals by maximizing the agreement between item representations under individual modalities and the fused multimodal representations. In this way, the fused multimodal representations can adaptively capture item-item relationships shared between multiple modalities in a self-supervised manner. The resulting contrastive loss can be mathematically expressed as:

$$\mathcal{L}_C = -\frac{1}{|\mathcal{I}|}\frac{1}{|\mathcal{M}|}\sum_{i \in \mathcal{I}}\sum_{m \in \mathcal{M}} I(h_i^m, h_{c,i}), \tag{9}$$

where $I(\cdot,\cdot)$ denotes the mutual information which quantifies the agreement between two representations, which is implemented by the InfoNCE estimator [27]. Specifically, for $I(h_i^m, h_{c,i})$, we set $(h_i^m, h_{c,i})$ as positive samples, while all other item embeddings in an individual modality $(h_i^m, h_j^m)$ and the fused multimodal embeddings $(h_i^m, h_{c,j})_{(j \neq i)}$ are considered as negatives:

$$I(h_i^m, h_{c,i}) = \log \frac{e^{\theta(h_i^m, h_{c,i})/\tau}}{e^{\theta(h_i^m, h_{c,i})/\tau} + \sum_{j \neq i} e^{\theta(h_i^m, h_{c,j})/\tau} + e^{\theta(h_i^m, h_j^m)/\tau}}, \tag{10}$$

where $\tau \in \mathbb{R}$ is a temperature parameter and $\theta(\cdot,\cdot)$ is the critic function which is a simple cosine similarity function in this work.

The proposed objective also conceptually relates to contrastive knowledge distillation [32], where several teacher models (representations under different individual modalities) and one student model (the modality-shared representations) are employed. By forcing the embeddings between several teachers and a student to be close, these modality-shared representations adaptively collect information from all modality-aware item relations. Additionally, the multimodal contrastive framework serves as a self-supervised auxiliary task, where the external self-supervision signals are introduced to learn better item representations involved with relation information from multiple modalities, which would further alleviate the cold-start problem.

#### 3.3.2 Mining Modality-Specific Information

Multiple modalities usually convey both complementary and supplementary information [26]. Our contrastive learning framework could adaptively extract the modality-shared supplementary information from all modalities. The distinctive characteristics held by each modality are also important for fully understanding item relationships. Previous work [33–35] notices that the modality-specific information could complement the modality-shared features captured in the invariant space and provides comprehensive multimodal representations.

To this end, we propose to mine modality-specific representations. Specifically, the modality-specific representations $h_{s,i}^m$ of each modality $m$ are obtained by subtracting the modality-common representations $h_{c,i}$ from the modality representation $h_i^m$:

$$h_{s,i}^v = h_i^v - h_{c,i}, \tag{11}$$

$$h_{s,i}^t = h_i^t - h_{c,i}. \tag{12}$$

To ensure that the modality-specific representations of different modalities do not encode shared information of each other, we employ a soft orthogonality constraint:

$$\mathcal{L}_S = \frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}} \| {h_{s,i}^v}^\top h_{s,i}^t \|^2. \tag{13}$$

Then, we employ a simplified attention module to fuse the modality-shared and modality-specific representations. The importance of each representation corresponding to item $i$ can be formulated as:

$$\beta_{c,i}, \beta_{s,i}^v, \beta_{s,i}^t = \mathrm{softmax}\left(q_2^\top \left[h_{c,i}, h_{s,i}^v, h_{s,i}^t\right]\right), \tag{14}$$

where $q_2 \in \mathbb{R}^d$ denotes the attention vector. Then, the final fused multimodal representation of item $i$ can be represented as:

$$h_i = \beta_{c,i} h_{c,i} + \beta_{s,i}^v h_{s,i}^v + \beta_{s,i}^t h_{s,i}^t. \tag{15}$$

### 3.4 Incorporating with Collaborative Filtering Methods

Unlike previous attempts which utilize multimodal features based on ad-hoc user-item aggregation strategies, MICRO separates the usage of multimodal features with the usage of user-item interactions and is agnostic to downstream CF methods. Specifically, we learn item representations from mined item relations and then combine them with downstream CF methods that model user-item interactions. It is flexible and could be served as a play-and-plug module for any CF methods.

Firstly, we represent user preference by aggregating the semantic item-item relation information of interacted items:

$$h_u = \frac{1}{|\mathcal{N}_u|}\sum_{i \in \mathcal{N}_u} h_i. \tag{16}$$

We denote the output user and item embeddings from CF methods as $\tilde{x}_u, \tilde{x}_i \in \mathbb{R}^d$, respectively. Finally, the user-item preference scores are obtained by taking inner product of enhanced user embeddings and item embeddings:

$$\hat{y}_{ui} = \left(\bar{e}_u + \frac{h_u}{\|h_u\|}\right)^{\top}\left(\bar{e}_i + \frac{h_i}{\|h_i\|}\right). \tag{17}$$

### 3.5 Optimization

We adopt the Bayesian Personalized Ranking (BPR) loss [36] to compute pair-wise rankings, which encourage the prediction of an observed entry to be higher than its unobserved counterparts:

$$L_{BPR} = \sum_{u \in U}\sum_{i \in I_u}\sum_{j \notin I_u} -\ln \sigma(\hat{y}_{ui} - \hat{y}_{uj}), \tag{18}$$

where $I_u$ indicates the observed items associated with user $u$ and $(u, i, j)$ denotes the pairwise training triples where $i \in I_u$ is the positive item and $j \notin I_u$ is the negative item sampled from unobserved interactions. $\sigma(\cdot)$ is the sigmoid function.

The overall objective function can be formulated as:

$$L = L_{BPR} + \lambda_C L_C + \lambda_S L_S, \tag{19}$$

where $\lambda_C, \lambda_S$ are hyperparameters to control the effect of the contrastive auxiliary task and the orthogonality constraint, respectively.

## 4 Experiments

In this section, we conduct experiments on three widely used real-world datasets to answer the following four research questions:

- RQ1: How does our model perform compared with the state-of-the-art multimedia recommendation methods and other CF methods in both warm-start and cold-start settings?
- RQ2: How do the structure mining and contrastive learning modules contribute to the model performance?
- RQ3: How sensitive is our model under the perturbation of several key hyperparameters?
- RQ4: How does each modality contribute to the final representations?

### 4.1 Experimental Settings

#### 4.1.1 Datasets

We conduct experiments on three categories of widely used Amazon datasets introduced by McAuley et al. [37]: (a) Clothing, Shoes and Jewelry, (b) Sports and Outdoors, and (c) Baby, which we refer to as Clothing, Sports, and Baby in brevity. The statistics of these three datasets are summarized in Table 3. The three datasets include both visual and textual modalities. We use the 4,096-dimensional visual features that have been extracted and published. For the textual modality, we extract textual embeddings by concatenating the title, descriptions, categories, and brand of each item and utilize sentence-transformers [38] to obtain 1,024-dimensional sentence embeddings.

TABLE 3
Statistics of the datasets

| Dataset[1] | # Users | # Items | # Interactions | Density |
|---|---|---|---|---|
| Clothing | 39,387 | 23,033 | 237,488 | 0.00026 |
| Sports | 35,598 | 18,357 | 256,308 | 0.00039 |
| Baby | 19,445 | 7,050 | 139,110 | 0.00101 |

[1] Datasets can be accessed at http://jmcauley.ucsd.edu/data/amazon/links.html.

#### 4.1.2 Baselines

To evaluate the effectiveness of our proposed model, we compare it with several state-of-the-art recommendation models. These baselines fall into two groups: CF methods (i.e., ItemKNN, MF, NGCF, LightGCN, SGL) and deep content-aware recommendation models (i.e., VBPR, MMGCN, GRCN).

- ItemKNN [39] computes the similarity between the items, and compute the similarity between a basket of items and a candidate recommender item.
- MF [36] optimizes Matrix Factorization using the Bayesian personalized ranking (BPR) loss, which exploits the user-item direct interactions only as the target value of interaction function.
- NGCF [12] explicitly models user-item interactions by a bipartite graph. By leveraging graph convolutional operations, it allows the embeddings of users and items to interact with each other to harvest the collaborative signals as well as high-order connectivity signals.
- LightGCN [14] argues the unnecessarily complicated design of GCNs (i.e., feature transformation and nonlinear activation) for recommendation systems and proposes a light model which only consists of two essential components: light graph convolution and layer combination.
- SGL [40] generates multiple views of a node and maximizing the agreement between different views of the same node.
- VBPR [5] integrates the visual features and ID embeddings of each item as its representation based upon the BPR model and feeds them into the matrix factorization framework. In our experiments, we concatenate multimodal features as content information to predict the interactions between users and items.
- MMGCN [15] is one of the state-of-the-art multimodal recommendation methods, which constructs modal-specific graphs and refines modal-specific representations for users and items. It aggregates all model-specific representations to obtain the representations of users or items for prediction.
- GRCN [16] is also one of the state-of-the-arts multimodal recommendation methods. It refines user-item interaction graph by identifying the false-positive feedback and prunes the corresponding noisy edges in the interaction graph.

#### 4.1.3 Evaluation Protocols

We conduct experiments in both warm-start and cold-start settings.

Warm-start setting. For each dataset, we select 80% of historical interactions of each user to constitute the training

TABLE 4
Performance comparison of our MICRO with different baselines in terms of Recall@20 (R@20), Precision@20 (P@20), and NDCG@20. The best performance is highlighted in bold and the second is highlighted by underlines. Improvement indicates the relative improvement of MICRO compared to the best baseline in percentage. All improvements are significant with p-value 0.05.

| Model | Clothing | | | Sports | | | Baby | | |
|---|---|---|---|---|---|---|---|---|---|
| | R@20 | P@20 | NDCG@20 | R@20 | P@20 | NDCG@20 | R@20 | P@20 | NDCG@20 |
| ItemKNN | 0.0280 | 0.0014 | 0.0131 | 0.0410 | 0.0022 | 0.0212 | 0.0317 | 0.0017 | 0.0152 |
| MF | 0.0191 | 0.0010 | 0.0088 | 0.0430 | 0.0023 | 0.0202 | 0.0440 | 0.0024 | 0.0200 |
| NGCF | 0.0387 | 0.0020 | 0.0168 | 0.0728 | 0.0038 | 0.0332 | 0.0591 | 0.0032 | 0.0261 |
| LightGCN | 0.0470 | 0.0024 | 0.0215 | 0.0803 | 0.0042 | 0.0377 | 0.0698 | 0.0037 | 0.0319 |
| SGL | 0.0598 | 0.0030 | 0.0268 | 0.0905 | 0.0047 | 0.0412 | 0.0745 | 0.0040 | 0.0328 |
| VBPR | 0.0481 | 0.0024 | 0.0205 | 0.0582 | 0.0031 | 0.0265 | 0.0486 | 0.0026 | 0.0213 |
| MMGCN | 0.0501 | 0.0024 | 0.0221 | 0.0638 | 0.0034 | 0.0279 | 0.0640 | 0.0032 | 0.0284 |
| GRCN | 0.0631 | 0.0032 | 0.0276 | 0.0833 | 0.0044 | 0.0377 | 0.0754 | 0.0040 | 0.0336 |
| MICRO | 0.0824 | 0.0042 | 0.0371 | 0.1005 | 0.0052 | 0.0467 | 0.0898 | 0.0047 | 0.0407 |
| Improvement | 30.5% | 31.3% | 34.4% | 11.0% | 10.6% | 13.3% | 19.1% | 17.5% | 21.1% |

set, 10% for validation, and the remaining 10% for the test set. For each observed user-item interaction, we treat it as a positive pair and then conduct the negative sampling strategy to pair them with one negative item that the user does not interact before.

Cold-start setting. We remove all user-item interaction pairs associated with a randomly selected 20% item set from the training set. We further divide the half of the items (10%) into the validation set and half (10%) into the test set. In other words, these items are entirely unseen in the training set.

We adopt three widely-used metrics to evaluate the performance of preference ranking: Recall@k, NDCG@k, and Precision@k. By default, we set k = 20 and report the averaged metrics for all users in the test set.

#### 4.1.4 Implementation Details

We implemente our method in PyTorch [41] and set the embedding dimension d fixed to 64 for all models for fair comparison. We optimize all models with the Adam [42] optimizer, where the batch size is fixed at 1024. We use the Xavier initializer [43] to initialize the model parameters. The optimal hyper-parameters are determined via grid search on the validation set: the learning rate is set to $0.0005$, the coefficient of $\ell_2$ normalization is set to $10^{-4}$. The k of kNN sparsification is set to 10, the of skip connection is set to $0.7$, and the temperature parameter is set to $0.5$. Besides, we stop training if Recall@20 on the validation set does not increase for 10 successive epochs to avoid overfitting.

### 4.2 Performance Comparison (RQ1)

We start by comparing the performance of all methods, and then explore how the our method alleviate the cold-start problem. In this subsection, we combine MICRO with LightGCN as a downstream CF method, and will also conduct experiments with different CF methods in Section 4.3.

#### 4.2.1 Overall Performance

Table 4 reports the performance comparison results, from which we can observe:

- Our method MICRO significantly outperforms both CF methods and content-aware methods, verifying the effectiveness of our methods. Specifically, MICRO improves over the strongest baselines in terms of Recall@20 by 24.1%, 18.6%, and 18.3% in Clothing, Sports, and Baby, respectively. This indicates that our proposed method is well-designed for multimedia recommendation by discovering underlying item-item relationships and is able to conduct fine-grained multimodal fusion through the contrastive auxiliary task.
- Compared with CF methods, content-aware methods yield better overall performance, which indicates that multimodal features provide rich content information about items, and can boost recommendation accuracy. Even without utilizing item content information, the self-supervised method SGL achieves competitive performance on the three datasets and even outperforms the powerful content-aware method GRCN, which demonstrates that the auxiliary self-supervised task can also improve node representation learning.
- Additionally, existing content-aware recommendation models are highly dependent on the representativeness of multimodal features and thus obtain fluctuating performance over different datasets. For the Clothing dataset where visual features are very important in revealing item attributes [5, 6], VBPR, MMGCN, and GRCN outperform all CF methods. For the other two datasets where multimodal features may not directly reveal item attributes, content-aware methods obtain relatively small improvements. The performance of VBPR and MMGCN is even inferior to the CF method LightGCN. Different from existing content-aware methods, we discover the latent item relationships underlying multimodal features instead of directly using them as side information. The latent item relationships are less dependent on the representativeness of multimodal features and thus we are able to obtain robust performance.

#### 4.2.2 Performance in the Cold-start Setting

The cold-start problem remains a prominent challenge in recommendation systems [44]. Multimodal features of items provide rich content information, which can be exploited to alleviate the cold-start problem. We conduct cold-start experiments and compare with representative baselines. MICRO w/o. fusion is the simplified variant of MICRO, which discards multimodal fusion described in Section 3.3

TABLE 5
Performance of our proposed MICRO on top of different downstream collaborative  ltering (CF) methods.

| Model | Clothing | | | Sports | | | Baby | | |
|---|---|---|---|---|---|---|---|---|---|
| | R@20 | P@20 | NDCG@20 | R@20 | P@20 | NDCG@20 | R@20 | P@20 | NDCG@20 |
| MF | 0.0191 | 0.0010 | 0.0088 | 0.0430 | 0.0023 | 0.0202 | 0.0440 | 0.0024 | 0.0200 |
| MF+feats | 0.0456 | 0.0023 | 0.0197 | 0.0889 | 0.0047 | 0.0403 | 0.0701 | 0.0037 | 0.0306 |
| MICRO/feats | 0.0729 | 0.0037 | 0.0323 | 0.0889 | 0.0047 | 0.0403 | 0.0840 | 0.0044 | 0.0376 |
| MICRO w/o. fusion | 0.0758 | 0.0038 | 0.0339 | 0.0940 | 0.0050 | 0.0436 | 0.0827 | 0.0044 | 0.0366 |
| MICRO w/o. speci c | 0.0785 | 0.0039 | 0.0351 | 0.0968 | 0.0051 | 0.0450 | 0.0845 | 0.0045 | 0.0384 |
| MICRO w/o. preference | 0.0727 | 0.0036 | 0.0321 | 0.0883 | 0.0047 | 0.0408 | 0.0752 | 0.0040 | 0.0336 |
| MICRO | 0.0797 | 0.0040 | 0.0355 | 0.0971 | 0.0051 | 0.0450 | 0.0854 | 0.0045 | 0.0385 |
| NGCF | 0.0387 | 0.0020 | 0.0168 | 0.0728 | 0.0038 | 0.0332 | 0.0591 | 0.0032 | 0.0261 |
| NGCF+feats | 0.0436 | 0.0022 | 0.0190 | 0.0748 | 0.0040 | 0.0344 | 0.0660 | 0.0035 | 0.0295 |
| MICRO/feats | 0.0676 | 0.0034 | 0.0297 | 0.0932 | 0.0049 | 0.0422 | 0.0799 | 0.0042 | 0.0356 |
| MICRO w/o. fusion | 0.0639 | 0.0032 | 0.0288 | 0.0900 | 0.0048 | 0.0408 | 0.0766 | 0.0041 | 0.0340 |
| MICRO w/o. speci c | 0.0735 | 0.0037 | 0.0333 | 0.0957 | 0.0049 | 0.0438 | 0.0799 | 0.0042 | 0.0360 |
| MICRO w/o. preference | 0.0639 | 0.0032 | 0.0282 | 0.0888 | 0.0040 | 0.0047 | 0.0785 | 0.0041 | 0.0337 |
| MICRO | 0.0743 | 0.0038 | 0.0336 | 0.0962 | 0.0051 | 0.0440 | 0.0805 | 0.0042 | 0.0355 |
| LightGCN | 0.0470 | 0.0024 | 0.0215 | 0.0803 | 0.0042 | 0.0377 | 0.0698 | 0.0037 | 0.0319 |
| LightGCN+feats | 0.0477 | 0.0024 | 0.0208 | 0.0754 | 0.0040 | 0.0350 | 0.0793 | 0.0042 | 0.0344 |
| MICRO/feats | 0.0736 | 0.0037 | 0.0331 | 0.0945 | 0.0050 | 0.0433 | 0.0892 | 0.0047 | 0.0404 |
| MICRO w/o. fusion | 0.0729 | 0.0037 | 0.0331 | 0.0925 | 0.0049 | 0.0428 | 0.0849 | 0.0045 | 0.0377 |
| MICRO w/o. speci c | 0.0803 | 0.0042 | 0.0368 | 0.0981 | 0.0051 | 0.0461 | 0.0880 | 0.0046 | 0.0404 |
| MICRO w/o. preference | 0.0796 | 0.0040 | 0.0358 | 0.0992 | 0.0051 | 0.0461 | 0.0890 | 0.0047 | 0.0401 |
| MICRO | 0.0824 | 0.0042 | 0.0371 | 0.1005 | 0.0052 | 0.0467 | 0.0898 | 0.0047 | 0.0407 |

Fig. 3. Performance comparison of our method with different baselines in the cold-start setting.

and only utilizes the BPR loss in Eq. (18). Figure 3 reports the results of performance, from which we can observe:

Both MICRO w/o. fusion and MICRO can alleviate the cold-start problem and outperform all baselines on three datasets. They learn item graphs from multimodal features, along which cold-start items will get similar feedbacks from relevant neighbors through neighborhood aggregation of graph convolutions.

Additionally, MICRO outperforms MICRO w/o. fusion on three datasets. In MICRO, the multimodal contrastive framework serves as a self-supervised auxiliary task. The self-supervision signals are constructed by maximizing the agreement between item representations under individual modalities and the multimodal fused representations to learn better item representations which encode item relationships from multiple modalities. In this way, the cold-start problem would be further alleviated.

CF methods MF and LightGCN obtain poor performance under the cold-start setting in general, primarily because they only leverage users' feedbacks to predict the interactions between users and items. Although these methods may work well for items with suf cient feedbacks, they cannot help in the cold-start setting, since no user-item interaction is available to update the representations of cold-start items.

The content-aware model VBPR outperforms CF methods in general, which indicates that the content information provided by multimodal features bene ts recommendation for cold-start items. In particular, content information can help bridge the gap between the existing items to cold-start items. However, some graph-based content-aware methods such as GRCN, although perform well in the warm-start setting, obtain poor performance in the cold-start setting. GRCN utilizes multimodal features on user-item interaction bipartite graphs, which is also heavily dependent on user-item interactions. For cold-start items, they never interact with users and become isolated nodes in the user-item graphs, leading to deteriorated performance.

### 4.3 Ablation Studies (RQ2)

In this subsection, we combine MICRO with three common-used CF methods, i.e., MF, NGCF, and LightGCN to validate the effectiveness and flexibility of our proposed method. For each CF method, we compare it with the following variants:

- CF+feats does not consider latent item-item relationships and directly uses transformed multimodal features to replace the item representations learned from item graphs in Eq. (17).
- MICRO/feats uses multimodal features as the input initial item embeddings of graph convolutions instead of ID embeddings.
- MICRO w/o. fusion discards the modality fusion in Section 3.3 and only utilizes the BPR loss in Eq. (18), which is equivalent to LATTICE [1].
- MICRO w/o. specifc : discards the modality-specific information mining module in Section 3.3.2. Specifically, it only utilizes modality-shared representation $h_i^c$ in Eq. (8) as the final multimodal item representation and $L_{BPR}; L_C$ in Eq. (18).
- MICRO w/o. preference : ignores the user preference $h_u$ in Eq. (17) and producing user-item score with $\hat{y}_{ui} = (\tilde{e}_u)^{\top} \left(\tilde{e}_i + \frac{h_i}{\|h_i\|}\right)$.

Table 5 summarizes the performance, from which we have the following observations:

- MICRO significantly and consistently outperforms all original CF methods and CF+feats variants on three datasets, obtaining up to 68.8% improvements over the CF+feats variants, verifying the flexibility of our plug-in paradigm.
- Even without the contrastive auxiliary task, MICRO w/o. fusion obtains significant improvements over CF+feats, indicating the effectiveness of discovering latent item-item relationships from multimodal features. Furthermore, the improvements between MICRO and MICRO w/o. fusion show the importance of fine-grained multimodal fusion, through which we can capture item relationships shared between modalities adaptively.
- Based on the learned item graph structures, MICRO/feats employs graph convolutions on multimodal features. Our proposed method MICRO utilizes the same learned structures but employ graph convolutions on item ID embeddings, which aims to directly model item affinities. The improvements between them validate the effectiveness of explicitly modeling item affinities where multimodal features are only used to bridge semantic relationships between items.
- Multiple modalities convey both complementary and supplementary information. The modality-specific information could complement the modality-shared features captured in the invariant space and provides comprehensive multimodal representations. The improvements between MICRO and MICRO w/o. specific indicates the specific information could also boost recommendation.
- The improvements between MICRO and MICRO w/o. preference indicates the effectiveness of adding user preference representations by aggregating the multimodal representation of history items. Different from the aggregating operation in CF methods which encodes collaborative signals, we aim to encode semantic item-item relationships conveyed by multimodal content into the user preference representation.

Fig. 4. Performance comparison of various hyperparameters $k$ and .

### 4.4 Sensitivity Analysis (RQ3)

Since the graph structure learning layer and the contrastive auxiliary task play pivotal roles in our method, in this subsection, we conduct sensitivity analysis with different hyper-parameters on graph structure learning layers and the contrastive auxiliary task. Firstly, we investigate performance of MICRO-LightGCN with respect to different $k$ value of the $k$-NN sparsification operation since $k$ is important which determines the number of neighbors of each item, and controls the amount of information propagated from neighbors. Secondly, we discuss how the skip connection coefficient affects the performance which controls the amount of information from the initial graph structures. Finally, we explore how the auxiliary task magnitude $_c$ and $_s$ affects the performance.

#### 4.4.1 Impact of Varied $k$ Values

Figures 4(a)(c)(e) present the results of performance comparison. $k = 0$ means no item relationships are included and the model is degenerated to LightGCN. We have the following observations:

- Our method gains significant improvement between $k = 0$ and $k = 5$, which validates the rationality of item

relationships mined from multimodal features. Even if only a small part of the neighbors are included, we can obtain better item representations by aggregating meaningful and important information from the neighbors, which boost the recommendation performance.
- Furthermore, the performance first improves as $k$ increases, which verifies the effectiveness of information aggregation along item-item graphs since more neighbors bring more meaningful information that helps to make more accurate recommendations.
- The trend, however, declines when $k$ continues to increase, since there may exist many unimportant neighbors that will inevitably introduce noise to information propagation. This demonstrates the necessity of conducting $k$NN sparsification on the learned dense graph.

#### 4.4.2 Impact of Varied Coefficients

Figures 4(b)(d)(f) present the performance comparison. $= 0$ means only consider the graph structure learned by the transformed multimodal features, and $= 1$ means we only consider the initial structure generated by the raw multimodal features. We have the following observations:

- When we set $= 0$, the model obtains poor performance. It only learns graph structure from the transformed features, completely updating the adjacency matrix every time, ignoring the rich and useful information of raw features and resulting in fluctuated training process.
- The performance first grows as becomes larger, validating the importance of initial structures constructed by raw multimodal features. However, it begins to deteriorate when continues to increase, since raw features are often noisy due to the inevitably error-prone data measurement or collection process. Learning the graph structures dynamically can reduce noise.
- Overall, there are no apparent sharp rises and falls, indicating that our method is not that sensitive to the selection of . Notably, all models surpass the baselines (c.f. Table 4), proving the effectiveness of item graphs.

#### 4.4.3 Impact of Varied Coefficients

We investigate how the coefficients of the contrastive auxiliary task $_c$ and the orthogonality constraint $_s$ affect the performance. Figures 5(a)(c)(e) report the performance. We can observe that:

- With the increase of $_c$ and $_s$, the performances on all datasets first rise and is always better than $_c = 0$ and $_s = 0$. The primary recommendation task achieves decent gains when jointly optimized with the two auxiliary tasks even with a small $_c$ and $_s$.
- However, it begins to decline when $_c$ and $_s$ continue to increase. A small can promote the primary task, while a larger one would mislead it. The benefits brought by the self-supervised task and orthogonality constraint could be easily neutralized and the recommendation task is sensitive to the magnitude of them.

#### 4.4.4 Impact of Varied Layer Number $L$

In order to investigate the effect of multiple graph convolution layers and high-order information, we search the number

Fig. 5. Performance comparison of various hyperparameters and $L$.

of layers $L$ in the range of $\{0, 1, 2, 3, 4, 5\}$. Figures 6 report the performance. We can observe that:

- When $L$ increases from 0 to 1, the performance increases significantly on all datasets, indicating that the item-item relationships can effectively boost recommendation.
- The best performed hop varies from different datasets. Specifically, MICRO achieves the best performance with $L = 1$ in Clothing, $L = 2$ in Sports and Baby. Applying a too deep architecture might introduce noisy, unimportant item relationships to the representation learning.
- When varying the number of layers, MICRO consistently and significantly outperforms baselines on all datasets. It again verifies the effectiveness of item-item relationships.

### 4.5 Investigation of the Contribution of Each Modality (RQ4)

In this subsection, we aim to explore the contribution of each modality. Table 6 reports the performance comparison over different modalities. We observe that the performances of utilizing multiple modalities are better than that of ones within the single modality, demonstrating that incorporating the information from multiple modalities facilitates comprehensive understanding of items. Additionally, textual modality contributes more than visual modality in general. It is reasonable since textual modality provide more fine-grained

Fig. 6. Performance comparison of varied hyperparameters $L$.

TABLE 6
Performance comparison over different modalities.

| Dataset | Model | R@20 | P@20 | NDCG@20 |
|---|---|---|---|---|
| Clothing | Visual | 0.0626 | 0.0032 | 0.0277 |
| | Textual | 0.0765 | 0.0038 | 0.0349 |
| | Both | 0.0824 | 0.0042 | 0.0371 |
| Sports | Visual | 0.0868 | 0.0046 | 0.0409 |
| | Textual | 0.0940 | 0.0049 | 0.0435 |
| | Both | 0.1005 | 0.0052 | 0.0467 |
| Baby | Visual | 0.0768 | 0.0040 | 0.0339 |
| | Textual | 0.0835 | 0.0043 | 0.0380 |
| | Both | 0.0898 | 0.0047 | 0.0407 |

information which directly reveals the titles, categories and descriptions of items while visual modality only provides coarse-grained visual appearances.

# 5 Related Work

## 5.1 Multimedia Recommendation

Collaborative ltering (CF) has achieved great success in recommendation systems, which leverage users' feedbacks (such as clicks and purchases) to predict the preference of users and make recommendations. However, CF-based methods suffer from sparse data with limited user-item interactions and rarely accessed items. To address the problem of data sparsity, it is important to exploit other information besides user-item interactions. Multimedia recommendation systems consider massive multimedia content information of items, which have been successfully applied to many applications, such as e-commerce, instant video platforms and social media platforms [37, 45–47].

For example, VBPR [5] extends matrix factorization by incorporating visual features extracted from product images to improve the performance. DVBPR [48] attempts to jointly train the image representation as well as the parameters in a recommender model. Sherlock [19] incorporates categorical information for recommendation based on visual features. DeepStyle [6] disentangles category information from visual representations for learning style features of items and sensing preferences of users. ACF [7] introduces an item-level and component-level attention model for inferring the underlying users' preferences encoded in the implicit user feedbacks. VECF [49] models users' various attentions on different image regions and reviews. MV-RNN [50] uses multimodal features for sequential recommendation in a recurrent framework. Recently, Graph Neural Networks (GNNs) have been introduced into recommendation systems [12, 13, 51] and especially multimodal recommendation systems [15, 16, 52]. MMGCN [15] constructs modal-speci c graph and conducts graph convolutional operations, to capture the modal-speci c user preference and distills the item representations simultaneously. In this way, the learned user representation can re ect the users' speci c interests on items. Following MMGCN, GRCN [16] focuses on adaptively re ning the structure of interaction graph to discover and prune potential false-positive edges. There are several prior studies [8, 9, 17] that propose to explore collaborative item relationships through high-order item-user-item co-occurrences. For example, HUIGN [17] constructs a co-interacted item graph which exhibits users' intents at different levels. It aims to learn multi-level user intents from the co-interacted patterns of items and further enhance the recommendation performance. PAMD [35] takes the modality-speci c information which could complement the modality-shared features into consideration.

The above methods directly utilize multimodal features as side information of each item and disregard ne-grained multimodal fusion. In our model, we step further by discovering semantic item-item relationships from multimodal features, and conduct ne-grained multimodal fusion to inject complementary item-item relationships from multiple modalities into the item representations.

## 5.2 Deep Graph Structure Learning

GNNs have shown great power on analyzing graph-structured data and have been widely employed for graph analytical tasks across a variety of domains, including node classi cation [10, 53], link prediction [54], information retrieval [55, 56], etc. However, most GNN methods are highly sensitive to the quality of graph structures and usually require a perfect graph structure that are hard to construct in real-world applications [57]. Since GNNs recursively aggregate information from neighborhoods of one node to compute its node embedding, such an iterative mechanism has cascading effects — small noise in a graph will be propagated to neighboring nodes, affecting the embeddings of many others. Additionally, there also exist many real-world applications where initial graph structures are not available. Recently, considerable literature has arisen around the central theme of Graph Structure Learning (GSL), which targets at jointly learning an optimized graph structure and corresponding representations. There are three categories of GSL methods: metric learning [21–23], probabilistic modeling [57–59], and direct optimization approaches [60–62].

For example, IDGL [23] casts the graph learning problem into a similarity metric learning problem and leverage adaptive graph regularization for controlling the quality of the learned graph; DGM [63] predicts a probabilistic graph, allowing a discrete graph to be sampled accordingly in order to be used in any graph convolutional operator. NeuralSparse [58] considers the graph sparsi cation task by removing task-irrelevant edges. It utilizes a deep neural network to learn $k$-neighbor subgraphs by selecting at most $k$ neighbors for each node in the graph. We kindly refer to [64] for a recent overview of approaches for graph structure learning.

In personalized recommendation, although user-item interactions can be formulated as a bipartite graph naturally, item-item relations remain rarely explored. To model item relationships explicitly, we employ metric learning approaches to represent edge weights as a distance measure between two end nodes, which ts for multimedia recommendation since rich content information can be included to measure the semantic relationship between two items.

### 5.3 Contrastive Learning

Self-supervised learning is an emerging technique to learn representations by self-de ned supervision signals generated from raw data without relying on annotated labels. Contrastive learning (CL) has become a dominant branch of self-supervised learning, which targets at obtaining robust and discriminative representations by grouping positive samples closer and negative samples far from each other. For visual data, negative samples can be generated using a multiple-stage augmentation pipeline [27, 65, 66], consisting of color jitter, random ip, cropping, resizing, rotation, color distortion, etc. The latest advances extend self-supervised learning to graph representation learning. Velickovic et al. [67] introduce an objective function measuring the Mutual Information (MI) between global graph embeddings and local node embeddings. GraphCL [68] and GRACE [53] propose a node-level contrastive objective to simplify previous work. Furthermore, Zhu et al. [69] propose a contrastive method with adaptive augmentation that incorporates various priors for topological and semantic aspects of the graph. Generally, most CL work differs from each other in terms of the generation of negative samples and contrastive objectives.

There also exist several works combining self-supervised learning with collaborative ltering [40, 70], session-based recommendation [71], social recommendation [72, 73] and multimedia recommendation [74, 75]. Wu et al. [40] introduce self-supervised auxiliary task into collaborative ltering and improve both accuracy and robustness of GNNs for recommendation. Yao et al. [70] utilize self-supervised learning to learn better latent relationship of item features for large-scale item recommendations. Zhou et al. [71] utilize contrastive learning to learn the correlations among attribute, item, subsequence, and sequence. Wei et al.[75] aim to maximize the mutual information between item content and collaborative signals to alleviate the cold-start problem.

In this work, since multiple modality-aware graphs are involved, the individual modality-aware item representations and multimodal fused representations are natural positive pairs. We utilize contrastive learning to maximize the agreement between item representations under an individual modality and the multimodal fused representations. In this way, the fused multimodal representations can adaptively capture item-item relationships shared between multiple modalities in a self-supervised manner.

## 6 Conclusion

In this paper, we have proposed the latent structure mining method (MICRO) for multimodal recommendation, which leverages graph structure learning to discover latent item relationships underlying multimodal features and devises a novel contrastive framework to fuse multimodal item relationships. In particular, we rst develop a modality-aware structure learning layer and graph convolutions to inject modality-aware item relationships into item representations. Furthermore, we propose a novel multimodal contrastive framework to adaptively capture item-item relationships shared between multiple modalities in a self-supervised manner. Finally, the resulting enhanced item representations are infused with item relationships in multiple modalities, which will be added into the output item embeddings of CF models to make recommendations. Empirical results on three public datasets have demonstrated the effectiveness of our proposed model.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (62141608, 62236010, and 62206291).

## References


[1] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, “Mining latent structures for multimedia recommendation,” in ACM Multimedia, 2021, pp. 3872–3880. 1, 9

[2] C. C. Aggarwal, Recommender Systems: The Textbook. Springer, 2016. 1

[3] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural Collaborative Filtering,” in WWW, 2017, pp. 173–182.

[4] X. Su and T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” Adv. Artif. Intell. , pp. 421 425:1–421 425, 2009. 1

[5] R. He and J. McAuley, “VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback,” in AAAI , 2016, pp. 144–150. 1, 2, 6, 7, 11

[6] Q. Liu, S. Wu, and L. Wang, “DeepStyle: Learning User Preferences for Visual Recommendation,” in SIGIR, 2017, pp. 841–844. 1, 2, 7, 11

[7] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, “Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention,” in SIGIR, 2017, pp. 335–344. 1, 11

[8] C. Park, D. Kim, J. Oh, and H. Yu, “Do" also-viewed" products help user rating prediction?” in WWW, 2017, pp. 1113–1122. 1, 2, 11

[9] J. Lee and S. Abu-El-Haija, “Large-scale content-only video recommendation,” in ICCVW, 2017, pp. 987–995. 1, 2, 11

[10] T. N. Kipf and M. Welling, “Semi-Supervised Classi cation with Graph Convolutional Networks,” in ICLR, 2017. 2, 3, 11

[11] P. Velicković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” in ICLR, 2018. 2

[12] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, “Neural Graph Collaborative Filtering,” in SIGIR, 2019, pp. 165–174. 2, 6, 11

[13] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, “Session-based Recommendation with Graph Neural Networks,” in AAAI , 2019, pp. 346–353. 11

[14] X. He, K. Deng, X. Wang, Y. Li, Y.-D. Zhang, and M. Wang, “LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation,” in SIGIR, 2020, pp. 639–648. 2, 4, 6

[15] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, “MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video,” in ACM Multimedia, 2019, pp. 1437–1445. 2, 6, 11

[16] Y. Wei, X. Wang, L. Nie, X. He, and T.-S. Chua, “Graph-Re ned Convolutional Network for Multimedia Recommendation with Implicit Feedback,” in ACM Multimedia, 2020, pp. 3451–3459. 2, 6, 11

[17] Y. Wei, X. Wang, X. He, L. Nie, Y. Rui, and T.-S. Chua, “Hierarchical user intent graph network for multimedia recommendation,” TMM , 2021. 2, 11

[18] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, “Item-based Collaborative Filtering Recommendation Algorithms,” in WWW, 2001, pp. 285–295. 2

[19] R. He, C. Lin, J. Wang, and J. McAuley, “Sherlock: Sparse hierarchical embeddings for visually-aware one-class collaborative ltering,” in IJCAI, 2016. 2, 11

[20] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," Annu. Rev. Sociol, vol. 27, no. 1, pp. 415–444, 2003. 3
[21] X. Wang, M. Zhu, D. Bo, P. Cui, C. Shi, and J. Pei, "AM-GCN: Adaptive Multi-channel Graph Convolutional Networks," in KDD, 2020, pp. 1243–1253. 3, 11
[22] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive Graph Convolutional Neural Networks," in AAAI, 2018, pp. 3546–3553. 3
[23] Y. Chen, L. Wu, and M. Zaki, "Iterative Deep Graph Learning for Graph Neural Networks: Better and Robust Node Embeddings," in NeurIPS, 2020, pp. 19 314–19 326. 3, 11
[24] J. Chen, H.-r. Fang, and Y. Saad, "Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection," JMLR, vol. 10, no. 69, pp. 1989–2012, 2009. 3
[25] F. Wu, T. Zhang, A. H. de Souza Jr., C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying Graph Convolutional Networks," in ICML, 2019, pp. 6861–6871. 4
[26] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," TPAMI, vol. 41, no. 2, pp. 423–443, 2018. 4, 5
[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in ICML, vol. 119, 2020, pp. 10 709–10 719. 5, 12
[28] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in ECCV. Springer, 2020, pp. 776–794. 5
[29] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in CVPR, 2021, pp. 1182–1191. 5
[30] Y. Liu, Q. Fan, S. Zhang, H. Dong, T. Funkhouser, and L. Yi, "Contrastive multimodal fusion with tupleinfonce," in ICCV, 2021, pp. 754–763. 5
[31] S. Whitehead, H. Wu, H. Ji, R. Feris, and K. Saenko, "Separating skills and concepts for novel visual question answering," in CVPR, 2021, pp. 5632–5641. 5
[32] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in ICLR, 2019. 5
[33] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-speci c representations for multimodal sentiment analysis," in ACM Multimedia, 2020, pp. 1122–1131. 5
[34] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-speci c representations with self-supervised multi-task learning for multimodal sentiment analysis," in AAAI, vol. 35, no. 12, 2021, pp. 10 790–10 797.
[35] T. Han, P. Wang, S. Niu, and C. Li, "Modality matches modality: Pretraining modality-disentangled item representations for recommendation," in WWW, 2022, pp. 2058–2066. 5, 11
[36] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian Personalized Ranking from Implicit Feedback," in UAI, 2009, pp. 452–461. 6
[37] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-Based Recommendations on Styles and Substitutes," in SIGIR, 2015, pp. 43–52. 6, 11
[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in EMNLP, 2019, pp. 3980–3990. 6
[39] M. Deshpande and G. Karypis, "Item-based top-n recommendation algorithms," TOIS, vol. 22, no. 1, pp. 143–177, 2004. 6
[40] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, and X. Xie, "Self-supervised graph learning for recommendation," in SIGIR, 2021, pp. 726–735. 6, 12
[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in NeurIPS, 2019, pp. 8024–8035. 7
[42] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in ICLR, 2015. 7
[43] X. Glorot and Y. Bengio, "Understanding the Dif culty of Training Deep Feedforward Neural Networks," in AISTATS, 2010, pp. 249–256. 7
[44] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and Metrics for Cold-Start Recommendations," in SIGIR, 2002, pp. 253–260. 7
[45] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie, "Learning visual clothing style with heterogeneous dyadic co-occurrences," in ICCV, 2015, pp. 4642–4650. 11
[46] R. He and J. McAuley, "Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering," in WWW, 2016, pp. 507–517.
[47] Z. Cui, F. Yu, S. Wu, Q. Liu, and L. Wang, "Disentangled item representation for recommender systems," TIST, vol. 12, no. 2, Feb. 2021. 11
[48] W.-C. Kang, C. Fang, Z. Wang, and J. McAuley, "Visually-Aware Fashion Recommendation and Design with Generative Image Models," in ICDM, 2017, pp. 207–216. 11
[49] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation," in SIGIR, 2019, pp. 765–774. 11
[50] Q. Cui, S. Wu, Q. Liu, W. Zhong, and L. Wang, "Mv-rnn: A multi-view recurrent neural network for sequential recommendation," TKDE, vol. 32, no. 2, pp. 317–331, 2018. 11
[51] M. Zhang, S. Wu, M. Gao, X. Jiang, K. Xu, and L. Wang, "Personalized graph neural networks with attention mechanism for session-aware recommendation," TKDE, 2020. 11
[52] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T.-S. Chua, "Hierarchical fashion graph network for personalized out t recommendation," in SIGIR, 2020, pp. 159–168. 11
[53] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Deep Graph Contrastive Representation Learning," in GRL+@ICML, Jun. 2020. 11, 12
[54] M. Zhang and Y. Chen, "Link Prediction Based on Graph Neural Networks," in NeurIPS, 2018. 11
[55] Y. Zhang, J. Zhang, Z. Cui, S. Wu, and L. Wang, "A Graph-based Relevance Matching Model for Ad-hoc Retrieval," in AAAI, 2021. 11
[56] X. Yu, W. Xu, Z. Cui, S. Wu, and L. Wang, "Graph-based Hierarchical Relevance Matching Signals for Ad-hoc Retrieval," in WWW, 2021. 11
[57] L. Franceschi, M. Niepert, M. Pontil, and X. He, "Learning Discrete Structures for Graph Neural Networks," in ICML, 2019, pp. 1972–1982. 11
[58] C. Zheng, B. Zong, W. Cheng, D. Song, J. Ni, W. Yu, H. Chen, and W. Wang, "Robust Graph Representation Learning via Neural Sparsi cation," in ICML, 2020, pp. 11 458–11 468. 11
[59] D. Luo, W. Cheng, W. Yu, B. Zong, J. Ni, H. Chen, and X. Zhang, "Learning to Drop: Robust Graph Neural Network via Topological Denoising," in WSDM, 2021, pp. 779–787. 11
[60] L. Yang, Z. Kang, X. Cao, J. Di, B. Yang, and Y. Guo, "Topology Optimization based Graph Convolutional Network," in IJCAI, 2019, pp. 4054–4061. 11
[61] W. Jin, Y. Ma, X. Liu, X. Tang, S. Wang, and J. Tang, "Graph Structure Learning for Robust Graph Neural Networks," in KDD, 2020, pp. 66–74.
[62] X. Gao, W. Hu, and Z. Guo, "Exploring Structure-Adaptive Graph Learning for Robust Semi-Supervised Classi cation," in ICME, 2020, pp. 1–6. 11
[63] A. Kazi, L. Cosmo, N. Navab, and M. Bronstein, "Differentiable Graph Module (DGM) for Graph Convolutional Networks," arXiv.org, Feb. 2020. 11
[64] Y. Zhu, W. Xu, J. Zhang, Q. Liu, S. Wu, and L. Wang, "Deep Graph Structure Learning for Robust Representations: A Survey," arXiv.org, Mar. 2021. 11
[65] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning Representations by Maximizing Mutual Information Across Views," in NeurIPS, 2019, pp. 15 509–15 519. 12
[66] W. Falcon and K. Cho, "A Framework For Contrastive Self-Supervised Learning and Designing A New Approach," arXiv.org, Sep. 2020. 12
[67] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in ICLR, 2019. 12
[68] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in NeurIPS, vol. 33, 2020, pp. 5812–5823. 12
[69] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Graph Contrastive Learning with Adaptive Augmentation," in WWW, Apr. 2021, pp. 2069–2080. 12
[70] T. Yao, X. Yi, D. Z. Cheng, F. Yu, T. Chen, A. Menon, L. Hong, E. H. Chi, S. Tjoa, J. Kang et al., "Self-supervised learning for large-scale item recommendations," in CIKM, 2021, pp. 4321–4330. 12
[71] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization," in CIKM,