Learning the User's Deeper Preferences for Multi-modal Recommendation Systems

FEI LEI, ZHONGQI CAO, YUNING YANG, YIBO DING, and CONG ZHANG, Beijing University of Technology, China

Recommendation system plays an important role in the rapid development of micro-video sharing platform. Micro-video has rich modal features, such as visual, audio, and text. It is of great significance to carry out personalized recommendation by integrating multi-modal features. However, most of the current multi-modal recommendation systems can only enrich the feature representation on the item side, while it leads to poor learning of user preferences. To solve this problem, we propose a novel module named **Learning the User's Deeper Preferences (LUDP)**, which constructs the item-item modal similarity graph and user preference graph in each modality to explore the learning of item and user representation. Specifically, we construct item-item similar modalities graph using multi-modal features, the item ID embedding is propagated and aggregated on the graph to learn the latent structural information of items; The user preference graph is constructed through the historical interaction between the user and item, on which the multi-modal features are aggregated as the user's preference for the modal. Finally, combining the two parts as auxiliary information enhances the user and item representation learned from the collaborative signals to learn deeper user preferences. Through a large number of experiments on two public datasets (TikTok, Movielens), our model is proved to be superior to the most advanced multi-modal recommendation methods.

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

Additional Key Words and Phrases: Graph convolutional networks, multimodal recommendation

ACM Reference format:
Fei Lei, Zhongqi Cao, Yuning Yang, Yibo Ding, and Cong Zhang. 2023. Learning the User's Deeper Preferences for Multi-modal Recommendation Systems. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 3s, Article 138 (February 2023), 18 pages.
https://doi.org/10.1145/3573010

1 INTRODUCTION

In recent years, micro-videos take up a lot of time in people's online life. The success of micro-video sharing platforms such as TikTok and Kwai is inseparable from the recommendation system mechanism behind them. The most classical recommendation algorithm in the early days is **Collaborative Filtering (CF)**, which predicts the preferences of similar users based on the historical

Fei Lei and Zhongqi Cao contributed equally to this research.

Authors' addresses: F. Lei, Z. Cao, Y. Yang, Y. Ding, and C. Zhang, Beijing University of Technology, 100 Pingleyuan, Chaoyang Qu, Beijing, China; emails: leifei@bjut.edu.cn, caozhongqi@emails.bjut.edu.cn, yangyuning@emails.bjut.edu.cn, Dingyb@emails.bjut.edu.cn, zc73080@emails.bjut.edu.cn.

1551-6857/2023/02-ART138 $15.00
https://doi.org/10.1145/3573010

interaction records of user and item. However, the existence of data sparsity brings many limitations to the algorithm. **Matrix factorization (MF)** algorithms [5, 8, 23] are based on embedding user ID to model user preferences, which solve the problem of data sparsity. In addition, there are a number of methods to solve the problem of cold starts [24, 37, 42], which are based on the content or modal of items. With the development of deep learning, recent studies have applied **graph convolution networks (GCN)** in the recommendation system. According to user-item historical interactions, Xiang Wang et al. [34] model a bipartite graph, and verified that graph convolutional networks can effectively capture higher-order information between user and item to learn better representation. LightGCN [16] simplifies NGCF model [34] by removing the operation of feature transformation and nonlinear activation, which learned a better representation of item and user. Moreover, recommendation system based on knowledge graph [9, 28] has also attracted wide attention.

Despite the great success of these methods, there is still a long way to establish a good representation in micro-video recommendation. Micro-video mainly involves complex modal features such as visual, audio and text, so it is very important to capture user preferences in each modal. For this kind of multimedia related micro-video recommendation, it is crucial to study the user preference information of items and the user preference. **Multi-modal graph convolution network (MMGCN)** [39] models modal features with users' interests and builds multiple interaction graphs to enhance the representation of user preference in different modalities by multiple modal features of the items. Although MMGCN has inspired us in modeling user interest, it only enhances the representation of user interest in different modalities based on the interaction information between the user and the item, ignoring the similarity information between the modalities. Wei et al. [38] designed the graph refinement layer and the graph convolution layer, used the neighbor path mechanism to refine the prototype network, and learned the high-quality collaborative embedding of user and item. Guo et al. [10] proposed a bipartite graph structure, where they calculated the similarity of the attribute features of items and built an item-item graph to alleviate the common cold start problem in recommendation.

According to the current research, most of them use graph convolutional networks to capture the higher-order collaboration signals between users and item. We find that the item representation can be learned better by mining structural information between item through fine-grained modal similarity information. At the same time, inspired by MMGCN, users' interest preferences for different modalities are different, and such a study is more meaningful to capture users' deeper preferences. As shown in Figure 1, assuming that user u_1 prefers visual effects and background music, while user u_2 is more attracted by text information and background music, considering that users with similar historical behavior have the same interest preferences, therefore, according to the synergistic signal, u_2 will probably click i_3; and according to the fine-grained modal similarity information, the i_4, i_5, and i_6 with higher modal similarity to i_3 can be mined. Considering that u_2 is more interested in background music and text, i_5 and i_6, which are more similar to i_3 in audio modality and text modality, can be recommended to u_2, so that better representations of user and item can be learned through fine-grained modal similarity features and fine-grained user preferences.

In this paper, we aim at investigating how to use the GCN propagation mechanism to deal with similar modalities and learn users' preferences on different modalities, then explore the deeper preferences of users over different modalities. Our proposed LUDP framework consists of three designs: (1) On the item side, we build similarity modal graphs by using modal features to obtain richer similar items on different modalities and design the attention layer to obtain the weights of different modalities. (2) On the user side, we capture users' preferences on different modalities, the modal features are aggregated over the user preference graph to obtain users' preference information. The preference features are used to concatenate with the user's ID embedding to learn

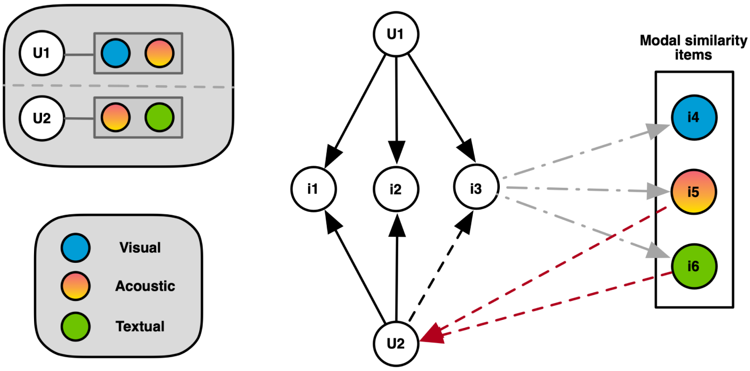


Fig. 1. Schematic illustration of recommending candidate items using fine-grained modal similarity of items and fine-grained user preferences.

the user's preferences for different modalities. (3) The ID embedding of users and items are aggregated and propagated on the bipartite graph as the primary user and item representation through collaborative signaling. The similar modal features extracted from the item side are concatenated with the item ID embedding as the final representation of the item, and the user interest features extracted from the user side are concatenated with the user ID embedding as the final representation of the user. As a result, from the multi-modal perspective, mining more items of similar modalities and extracting useful signals in similar modalities by users' preferences for modalities can enhance the representation of users and items.

Our contribution can be summarized as follows:

- We design the item-item modal similarity graph from the modal perspective. This similarity graph can effectively tap the potential relationships among items to enhance the representation of items.
- We propose LUDP, a novel framework, which mines potential relationships between items in similar modalities and learns deeper user preferences based on the features of different modalities of items that users have interacted with historically.
- Conducting extensive experiments on two public datasets, we demonstrate the effectiveness of our proposed method. The results show that our method outperforms the baseline methods.

2 RELATED WORK

2.1 Collaborative Filtering

The approaches are based on CF personalized recommendations through similarity analysis of user-item history behaviors [18, 27, 29]. These methods transform each user-item pair into a suitable embedding representation, modeling through historical interactions and learning the representation of users and items. After that, the user's preference for items is predicted based on the inner product of the user and item embeddings. Most traditional CF methods [23, 29] take the interaction matrix between the user and item as input to learn the potential representation. These methods only can capture the first-order connectivity of collaborative signals, so the embedding learned is poor and cannot well represent user preferences. In recent years, due to the advanced effectiveness of graph neural networks in modeling topologies, researchers have constructed bipartite graphs from user-item interactions to learn their embeddings. Many graph neural network based models, such as GC-MC [2], NGCF [34], LightGCN [16], and so on, have achieved good results. Most of these models differ from the way of propagating and aggregating information. For example, NGCF learns better embedding representation by capturing higher-order collaborative signals through graph convolution operations. LightGCN [16] simplifies on this basis [34] and constructs a linear graph convolution model with state-of-the-art performance. In addition, there are many works [19, 32, 43] that have introduced attention mechanism into graph convolution as a new aggregation method, and this is a good improvement in advancing performance. These models demonstrate the ability of graph models to learn better representations in recommendations. In our model, we draw on advanced graph models to learn higher-order collaborative signals. We also introduce multimodal features and design item similarity graphs and user preference graphs to enhance the embedding representation.

2.2 Graph Convolution Networks

In recommender systems, user-item interactions can be naturally constructed as bipartite graphs, and GCN is highly expressive in solving graph-related tasks. For the moment, GCN-based models [2, 16, 34, 41] were widely used in recommendation systems and have been shown to be effective in mining higher-order information in interaction signals. PinSAGE [41] is the first industrial-grade GCN-based recommendation system that uses the mechanism of aggregation and propagation to learn the embedding of user and item. Since then, graph-based approaches have received a lot of attentions in the field of recommendation. NGCF [34] explicitly modeled higher-order connectivity between user items by using graph convolutional networks to enhance embedding. LightGCN [16] proposed a simpler propagation, aggregation method, and achieved the most advanced performance. Inspired by the GCN propagation aggregation mechanism, MMGCN [39] trained modal features and user preferences in the interaction graph, which improved the embedding of users' different modal preferences. He et al. [13] learned user-specific preference embedding and item-specific attribute embedding through graph representation learning techniques. Ma et al. [26] studied the potential factors affecting neighbor information transmission from the propagation process of GCN to disentangle them, and proposed the **disentangled Graph Convolutional Network (DisenGCN)** to learn disentangled node representations. Guo et al. [10] proposed a Dual Graph enhanced Embedding Neural Network to solve the problem of item feature sparsity and user behavior sparsity. Liu et al. [25] proposed the IMP-GCN model, which uses higher-order item-item similarity to capture more useful neighbor nodes and solve the over-smoothing problem caused by the convolution of multilayer graphs. In contrast, we construct the item-item graph with the displayed fine-grained modal features, which is conducive to learning a better representation of the item based on mining the underlying structural information of the item.

2.3 Multi-modal Personalized Recommendation

Multi-modal recommendation systems take the large amount of multimedia content information of items into account, which have been successfully applied to each area in the recommendation systems [6, 15, 39]. Most of the approaches enhance the information on the item side by modal features. For example, He et al. [14] consider that people tend to pay a lot of attention to the picture information of items when shopping, so they extend MF by extracting visual features from images to improve recommendation performance. ACF [3] is based on the attention mechanism in the item and component layers for handling recommendation tasks in the multimedia domain. Apart from that, most researchers focus on designing frameworks for extracting better modal features. A recommendation model based on users' multimodal preferences was proposed by Xu et al. [40]. It works on capturing high-level conceptual information and exploring the connection between text and visuals between users and items. CITING [4] combined image features and text features to develop a feature-aware matrix factorization framework. In recent studies, many researchers have

Table 1. Notations Used in This Paper

Notations	Frequency
$\mathbf{U}$	user set
$\mathbf{I}$	item set
$\mathbf{Y}$	the 0-1 matrix
e_u	ID embedding representation of u
e_i	ID embedding representation of u
$m \in \mathbf{M} = \{v, a, t\}$	visual, acoustic, and textual modalities
$\mathbf{E}_u$	the user representation on user-item interaction graph
$\mathbf{E}_i$	the item representation on user-item interaction graph
$\mathbf{G}_{\mathbf{II}}$	KNN-graph
$\mathbf{H}$	the final item representations on item-item similarity graph
$\mathbf{P}$	the final fused modal preferences
e_u^*	the final representations of user
e_i^*	the final representations of item
y_{ui}^*	prediction score
$\mathcal{L}$	loss function

connected multimodal information with GCN to solve the problems of data sparsity as well as cold start in recommender systems [12, 30]. MGAT [31] can propagate information over the interaction graph of different modalities while adaptively capturing user preferences for different modalities by using a gated attention mechanism. Wei et al. [36] designed the HUIG model to learn multi-level user intent from the collaborative interaction patterns of products to obtain high-quality user and product representations and further improve recommendation performance. Dual-GCN [33] exploits correlations between users to mine each user's specific fusion modality. Compared to these approaches, we focus more on exploring item-to-item relationships, learning better item representations, and also exploring each user's preferences for specific modalities to the extent that deeper user preferences are captured.

3 PROBLEM STATEMENT

Before discussing our proposed model in detail, we first define the basic concept of recommendation task. The symbols used in this paper are shown in Table 1.

Definition 1 (Interaction Information). In general, some interactive information such as users' clicks and likes on an item represent users' potential preferences for the item. There are a set of m users $\mathbf{U} = \{u_1, u_2, \dots, u_m\}$ and a set of n items $\mathbf{I} = \{i_1, i_2, \dots, i_n\}$. Define the scoring matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$ based on whether there is an interaction between the user and the item, where $y_{ui} = 1$ denotes an observed interaction between user u and item i, otherwise $y_{ui} = 0$. $e_u, e_i \in \mathbb{R}^d$ is the input ID embedding representation of u and i, respectively, where d is the embedding size.

Definition 2 (Modal Information). Micro-videos contain multiple modal information — visual, acoustic, and textual features. We use $e_i^m \in \mathbb{R}^{d_m}$ as the modality indicator, where d_m denotes the modality features sizes, $m \in \mathbf{M} = \{v, a, t\}$ is the visual, acoustic, and textual modalities, respectively. Note that these modal features are pre-trained. In our model, we fine-tune these modal features to the point of learning a modal representation that better matches our recommendation task.

Definition 3 (User-item Bigraph Network). Construct a bipartite graph structure based on the interaction information between the user and the item. The bipartite graph $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ is

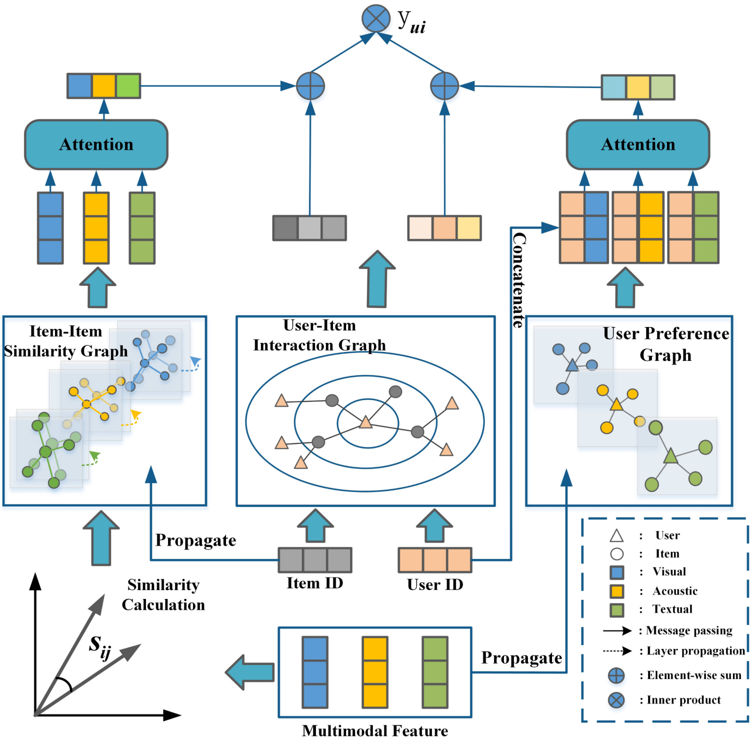


Fig. 2. Schematic illustration of our proposed framework. It includes the construction of item-item similarity graphs, user-item interaction graphs, and user preference graphs.

defined, where $\mathcal{V}$ denotes the vertex (user or item) set, and $\mathcal{E}$ denotes the edge (relation) set. The representation of nodes in the graph is the ID embedding of the user and item.

4 METHODS

Our overall framework of LUDP, which Figure 2 shows, consists of three components: (1) user-item interaction graph. Learning the representation of each node from interaction data. (2) item-item similarity graph. Mining the potential structure of an item with similar modalities. (3) user preference graph. Learning user's deeper preference for different modalities. Finally, the enhanced representations of user and item are predicted by dot product operation.

4.1 User-Item Interaction Graph

Existing GCN models [16, 34] have been shown to effectively capture the higher-order signals between user and item. Therefore, create a user-item bipartite graph $\mathbf{G} = \{(u,i)|u \in \mathbf{U}, i \in \mathbf{I}\}$ through the user-item interactions matrix $\mathbf{Y}$. There is an observed interaction between user u and item i if an edge $y_{ui} = 1$, otherwise $y_{ui} = 0$.

It is well known that interactive items provide direct user preferences. The information of users and items is aggregated and propagated on the interaction graph $\mathbf{G}$, which facilitates the capture of higher-order collaborative signals. The state-of-the-art GCN models such as GCN [22], NGCF [34], and LightGCN [16] have better performance in information propagation. To make the training easier and more efficient, we choose LightGCN as the main model to obtain user and item embeddings with higher-order collaboration signals. Therefore, the propagation rules on the interaction graph are defined as follows:

$$e_u^{(l+1)} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} e_i^{(l)},$$
$$e_i^{(l+1)} = \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i|}\sqrt{|\mathcal{N}_u|}} e_u^{(l)}, \tag{1}$$

where $\mathcal{N}_i$, $\mathcal{N}_u$ denote the neighbors of item and user, respectively. l is the number of layers of the network.

Following [16], we sum the node embedding representations obtained at each layer as the final embedding representations on user-item interaction graph:

$$\mathbf{E}_u = \sum_{l=0}^{L} e_u^{(l)},$$
$$\mathbf{E}_i = \sum_{l=0}^{L} e_i^{(l)}, \tag{2}$$

Note that the representation obtained from the collaboration signal is the main representation. Besides, fine-grained modal information is introduced to enhance the representation of users and items.

4.2 Item-Item Similarity Graph

The bipartite graph can only reflect the interaction between user and item. While the mining of the potential semantic relationship of items is also important in multi-modal recommendation. These pre-trained modal features are distributed differently in space, while similar features are distributed more intensively in space. Therefore, clustering method can be used to classify these modal features. As a result, we build item-item similarity graph based on modal features. We first calculate the similarity using cosine similarity [35] between each of the two items at the same modal. The modality similarity matrix $S^m \in \mathbb{R}^{n \times n}$ formulate as follows:

$$S^m(i,j) = \frac{e_i^m \cdot (e_j^m)^T}{||e_i^m|| \cdot ||e_j^m||} \tag{3}$$

Secondly, we build KNN graphs based on the similarity of each of two items on the same modality. Specifically, by setting a pre-defined k, the first k items that are more similar to the target item are taken to determine whether there are edges among them. KNN-graph $\mathbf{G_{II}}$ is defined as:

$$\mathbf{G}_{\mathbf{II}}^m(i,j) = \begin{cases} S_{ij}^m, & S_{ij}^m \in top-k(S_i^m), \\ 0, & otherise. \end{cases} \tag{4}$$

where S_i^m denotes all values in row i of the similarity matrix S.

After constructing similar graphs $\mathbf{G_{II}}$ on different modalities, the graph convolution operation is performed. We use the ID embedding of the item as the node representation instead of the modal feature. In this way, it is possible to effectively learn the potential connections among items on different modalities to enrich the representation of items. In addition, we are based on multi layer graph convolutional networks to capture the relationship between higher-order terms and items. Since the representation of the node contains only the ID embedding of the item, we adopt the idea of [16] and omit the feature transformation and nonlinear activation operations. The message passing and aggregation are represented as follows:

$$h_i^{(l+1)} = \sum_{j \in \mathcal{N}_i} \frac{1}{\sqrt{\mathcal{N}_i}} h_j^{(l)} \tag{5}$$

where $h^{(l)}$ denotes the node representation of layer l in the modal graph $\mathcal{G}_{II}$. $\mathcal{N}_i$ is the neighbor node representation of item i. The node representations are aggregated and propagated with information on each modal graph to obtain the final representation $\mathbf{h}^m$. Since different modalities convey different amounts of information, therefore, we use an attention mechanism [20] to fuse the item representations $\mathbf{h}^m$ learned in different modalities. The importance of the different modes is calculated as follows:

$$w_i^m = q^T tanh(\mathbf{W}_m \mathbf{h}_i^m + b) \tag{6}$$

where $q \in \mathbb{R}_m^d$ is attention vector; $\mathbf{W}_m \in \mathbb{R}^{d_m \times d_m}$, $b \in \mathbb{R}^{d_m}$ is the trainable weight matrix and bias vector, respectively. After this, we use the softmax function for normalization to obtain the weight score:

$$\alpha_i^m = \frac{exp(w_i^m)}{\sum_{m \in \mathcal{M}} exp(w_i^m)} \tag{7}$$

Finally, the merged item i representation is as follows:

$$\mathbf{H}_i = \sum_{m \in \mathcal{M}} \alpha_i^m \mathbf{h}_i^m \tag{8}$$

4.3 User Preference Graph

On the item side, instead of directly incorporating the features of the modalities into the network as side information, we construct a modal structure graph to capture similar items on different modalities for the purpose of enhancing item representation. [1] states that the fusion of multiple modalities can lead to more robust predictions as well as capture complementary information that is not visible in unimodality. The user's historical interaction information can be considered as the user's potential preferences. Inspired by the aggregation mechanism of GCN, we aggregate modal features as user's preferences on the interaction graph. Technically, we can integrate the modality-aware information from the history items to create a preference representation of user u as :

$$p^m = f(\mathcal{N}_u) \tag{9}$$

where $\mathcal{N}_u = \{i|(u,i) \in \mathbf{G}\}$ denotes the neighbors of user u; p^m is the preference representation of user u; and $f(\cdot)$ is the aggregator function to characterize each first-order connection (u,i). We apply an averaging pooling operation to aggregate the modal features and use $LeakyReLU(\cdot)$ for nonlinear transformation, $f(\cdot)$ as follows:

$$p^m = LeakyReLU\left(\frac{1}{|\mathcal{N}_u|} \sum_{j \in \mathcal{N}_u} \mathbf{W}_1 e_i^m\right) \tag{10}$$

where $e_i^m \in \mathbb{R}^{d_m}$ is the size of item i in modality m; $\mathbf{W}_1 \in \mathbb{R}^{d'_m \times d_m}$ is the trainable weight matrix, which is used to extract useful information from the modalities, where d'_m is the transformation size.

Based on the fact that users' preferences for different modalities are different, we map the aggregated user preference representations from modalities to the space of user ID embedding and concatenate with them:

$$\hat{p}^m = LeakyReLU(\mathbf{W}_2 p^m) + e_u \tag{11}$$

where $\mathbf{W}_2 \in \mathbb{R}^{d \times d'_m}$.

As with the item side, we also fuse user preference in different modalities through the attention mechanism. The final fused modal preferences is as follows:

$$\mathbf{P} = \sum_{m \in \mathcal{M}} \beta^m \hat{p}^m \tag{12}$$

where β^m is the significance score of user preference to each modality m.

4.4 Model Prediction

We use the potential modal structure and user preference of the learned item to add to the representation obtained from the collaboration graph respectively, to enhance the representation of the item and the user. The final representations of user u and items i are as follows:

$$e_u^* = \mathbf{E}_u + \mathbf{P}, \quad e_i^* = \mathbf{E}_i + \mathbf{H} \tag{13}$$

Finally, we conduct inner product of user and item representations to predict their matching score:

$$y_{ui}^* = {e_u^*}^T e_i^* \tag{14}$$

4.5 Optimization

We opt for **Bayesian Personalized Ranking (BPR)** loss to optimize the model parameters, which considers that user interactive items should be assigned with higher prediction scores than the unobserved items:

$$\mathcal{L}_{BPR} = \sum_{(u,i,j) \in O} -ln\, \sigma(y_{ui}^* - y_{uj}^*) + \lambda ||\Theta||_2^2 \tag{15}$$

where $O \in \{(u,i,j)|(u,i) \in \mathbf{Y}^+, (u,j) \in \mathbf{Y}^-\}$ is a set of triples for training dataset; $\mathbf{Y}^+$ is the observed interaction, while $\mathbf{Y}^-$ is the unobserved interactions; $\sigma(\cdot)$ is the sigmoid function; λ is the regularization weight and Θ is the parameters of the model.

5 EXPERIMENTS

In this section, we present our experimental results in detail, and answer the following research questions: **RQ**1: How does our model perform compared with the state-of-the-art recommended method? **RQ**2: How does the multi-modal feature affect the performance of the recommendation in our model? And can our model learn the underlying structural information of items from item-item graphs of similar modalities and learn the users' deeper preferences? How does the effects of attention layer? **RQ**3: How does the depth of network and parameter settings affect the effect?

5.1 Experiment Setup

5.1.1 Datasets. We evaluate the performance of our proposed model on two publicly available datasets: TikTok and MovieLens. The statistics of datasets are in Table 2.

- **TikTok**[1]**:** It contains user ID, Item ID, and their interactions (e.g., finish, like). In addition, this dataset also provides multimodal features for each micro-video: visual features, audio features, and text features. Moreover, the text features are extracted from the subtitles of micro-videos given by users.
- **MovieLens**[2]**:** This dataset is widely used in recommendation systems. In order to construct a dataset suitable for multimodal recommendations, we collected some modality features of movies from the MovieLens-10M dataset. For example, text features such as the title and

[1] http://ai-lab-challenge.bytedance.com/tce/vc/.
[2] https://grouplens.org/datasets/movielens/.

Table 2. Dataset Statistics

Dataset	#Interactions	#Items	#Users	Sparsity	V	A	T
TikTok	1493532	34756	18855	99.77%	128	256	128
MovieLens	896006	10698	23018	99.64%	2048	256	768

V, A, and T Denote the Dimensions of Visual, Acoustic, and Textual Modalities, Respectively.

plot description of a movie are extracted by Bert [7]; the corresponding posters were also collected and visual features were extracted by Resnet [11]; for acoustic modality, we split the audio message from the trailer and adopt VGGish [17] to learn the acoustic features.

5.1.2 Baselines. To evaluate the performance of our model, we compare LUDP with several state-of-the-art methods. These baseline methods are mainly classified into CF-based and GCN-based methods.

- **MF** [27]**:** This method uses the **Bayesian personalized ranking (BPR)** loss to optimize Matrix Factorization, which exploits the user-item direct interactions only as the target value of interaction function.
- **VBPR** [14]: The model stitches visual features with ID embeddings as representations of items and uses matrix factorization framework to reconstruct the history of user-item interactions. In this experiment, we introduce multimodal features while adhering to its ideas.
- **ACF** [3]**:** This model introduces two attention modules to handle implicit feedback at the item level and component level. We consider each modality as a component of an item to explore the user preferences of a particular modality and the characteristics of the item.
- **NGCF** [34]**:** The method constructs a bipartite graph using the historical interaction information of users and items, then obtains the collaboration signal and the higher-order connection signal by the graph convolution operation.
- **MMGCN** [39]**:** It is one of the state-of-the-art multimodal recommendation methods, which learn user preferences for specific models. Performing graph convolution operation in each modality to obtain a representation of the user and the item.
- **LightGCN** [16]**:** LightGCN is the best collaborative filtering model based on graph networks. It achieves better performance with lighter graph operation by removing feature changes and nonlinear activation.
- **GRCN** [38]**:** This is one of the state-of-the-art multimodal recommendation methods. It identifies false-positive feedback by the graph thinning layer and graph convolution layer. And the corresponding noise edge in the interaction diagram is trimmed.
- **LATTICE** [42]**:** It is also one of the state-of-the-art multimodal recommendation methods. It explores the potential structural information of items by constructing a modal similarity graph of item using modal features and letting the item embedding propagate on this graph. Finally, the learned item representation is used to complement the synergistic signal modeled by the traditional CF methods.

5.1.3 Evaluation Metrics. We randomly split the dataset into training, validation, and testing in the ratio of 8:1:1. We also treat the user items that pairs interact with as positive samples and the random selection of items that the user does not interact with as negative samples. For each user in the validation and the testing sets, we treated all items who did not click before as the negative samples. We compared the recommended top-K list $R(u)$ to the ground truth list $T(u)$ for each user u and used three commonly used evaluation metrics: Precision @K, Recall @K, and NDCG

@K. The larger the value of the three measures, the better the recommendation effect. Here we set $K = 10$ and report the averaged metrics for all users.

- **Precision@K:** This evaluation metric calculates the proportion of ground truth among items recommended to users:
$$Precision@K = \frac{R(u) \cap T(u)}{R(u)} \tag{16}$$
- **Recall@K:** This is the most commonly used evaluation metric for recommendation tasks. It denotes the percentage of correctly predicted items in the list recommended to the user in the ground truth list $T(u)$. The formula is as follows:
$$Recall@K = \frac{R(u) \cap T(u)}{T(u)} \tag{17}$$
- **NDCG@K: Normalized Discounted Cumulative Gain (NDCG)** is a commonly used ranking evaluation metric. It means the higher the rank, the higher the score. It is formally defined as:
$$DCG_K = \sum_{i=1}^{K} \frac{2^{r_{ui}} - 1}{log_2(1+i)}, NDCG@K = \frac{DCG_K}{IDCG_K} \tag{18}$$
where $IDCG$ is ideal DCG, and r_{ui} represents interaction between the user u and item i.

5.1.4 Hyper-parameter Setting. We implemented our method using PyTorch. In the model learning process, we randomly initialize model parameters with a Gaussian distribution and set the embedding dimension d fixed to 64, and optimize the model with the Adam [21] optimizer. The batch size was searched in 128, 256, 512, 1024 and the learning rate was searched in 0.0001, 0.0005, 0.001, 0.005, 0.01. The $L2$ normalization coefficient was tuned in 0, 0.00001, 0.0001, 0.001, 0.01, 0.1. The k values in item-item similarity graph was set in 5, 10, 20, 50. In addition, we stop training if Recall@10 on the validation set does not increase for 10 successive epochs to avoid over fitting.

5.2 Performance Comparison (RQ1)

We compare the performance of LUDP with the baseline models. Table 3 shows the experimental results, and we have the following observations:

- Our method LUDP achieves the best performance among all baseline methods and significantly outperforms other multi-modal recommendations. More precisely, LUDP improves over the strongest baselines by 4.07%, 5.33%, and 5.32% in terms of Precision@10, Recall@10, and NDCG@10 on TikTok; as well as improves over the strongest baselines by 6.16%, 6.77%, and 8.44% in terms of Precision@10, Recall@10, and NDCG@10 on MoiveLens. This demonstrates the effectiveness of our model in multi-modal recommendation. Mining potential relationships between items through multi-modal features and capturing user preferences for a single modality enable better learning of item and user representations.
- Compared with CF methods, GCN methods have the better performance overall. This shows that the graph convolution operation can efficiently aggregate and propagate information to learn higher-order collaborative signals.
- In addition, the content-aware approach has a better performance overall compared to the CF approach. For example, in the TikTok dataset, MMGCN is close to the powerful LightGCN. GRCN is even better than LATTICE, which may be caused by the fact that this dataset contains too many false positive interactions, which are better handled. In the MovieLens dataset, the difference between the content-aware approach and the CF approach is not

Table 3. Comparison Results between our Model and the Baselines

Model	TikTok			MovieLens		
	Precision@10	Recall@10	NDCG@10	Precision@10	Recall@10	NDCG@10
MF	0.0306	0.0466	0.0432	0.0813	0.1511	0.1369
VBPR	0.0339	0.0526	0.0484	0.0876	0.1641	0.1561
ACF	0.0298	0.0433	0.0414	0.0792	0.1408	0.1221
NGCF	0.0332	0.0513	0.0481	0.1021	0.1804	0.1676
MMGCN	0.0352	0.0536	0.0493	0.1194	0.1917	0.1759
LightGCN	0.0357	0.0561	0.0501	0.1207	0.1978	0.1793
GRCN	0.0368	0.0577	0.0526	0.1232	0.1948	0.1843
LATTICE	0.0361	0.0582	0.0512	0.1266	0.2054	0.1872
LUDP	**0.0383**	**0.0613**	**0.0554**	**0.1344**	**0.2193**	**0.2030**
%Improv.	4.07%	5.33%	5.32%	6.16%	6.77%	8.44%

Bold scores are the best in each column, while underlined scores are the second best. %Improv. denotes the relative performance improvement of LUDP over the second best.

Table 4. Performances Comparison over Different Modalities

Dataset	Model	Precision@10	Recall@10	NDCG@10
TikTok	Visual	0.0377	0.0602	0.0549
	Acoustic	0.0368	0.0591	0.0542
	Textual	0.0362	0.0584	0.0531
	All	**0.0383**	**0.0613**	**0.0554**
MovieLens	Visual	0.1298	0.2097	0.1932
	Acoustic	0.1189	0.1901	0.1744
	Textual	0.1285	0.1988	0.1842
	All	**0.1344**	**0.2193**	**0.2030**

much, which is likely caused by the fact that the modal feature may be contained by noise and the content-aware approach failed to learn better modal information.

5.3 Ablation Study of LUDP(RQ2)

5.3.1 Effects of Modalities. To explore the effects of different modalities, we compare the impact on different modalities over two datasets. The results of the performance comparison are shown in Table 4. We have the following observations:

- As expected, the multi-modal features have better expressiveness compared to single-modal features. This indicates that rich modal features are more beneficial to explore the potential features between items to improve the representation of items. At the same time, different modal features have different effects on the results, and exploring the relationship between users and different modalities can effectively capture users' preferences in different modalities.
- As can be seen from Table 4, for the comparison between single modes, it is not difficult to find that visual features have the greatest promoting effect on the results. While visual features are more important than other modal features, it may be because visual information is easier to attract users' attention when they watch a micro-video again.

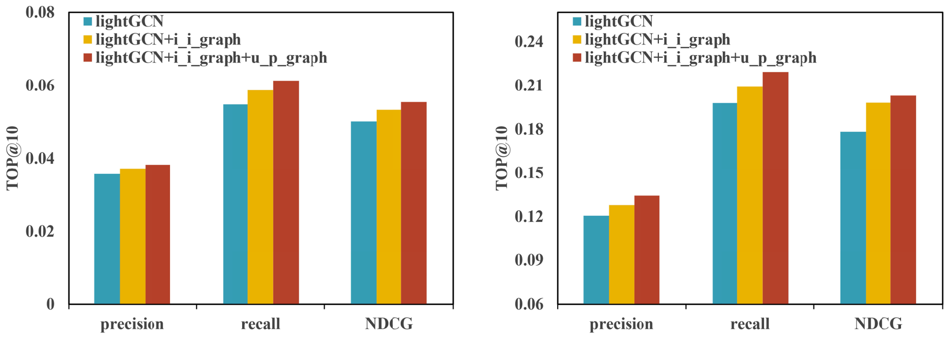
- Moreover, in the TikTok dataset, text features are less expressive, while in the MovieLens dataset, audio features are worse in performance. This is reasonable because users in the TikTok dataset pay more attention to background music, which is the reason why micro-videos are currently popular. Since users spend less time in each micro-video, they tend to pay less attention to the text features of the item. In the MovieLens dataset, however, the text describes the storyline, which is highly relevant to the content, and users may watch the video based on the storyline, which makes the audio features less important.
- Last but not least, we found that the result of single modal were better than the baseline models with hybrid multi-modal results in the TikTok dataset. This may be caused by two reasons: One is that the distribution of this dataset is relatively dense, so that users have little difference in preference for different modes. In addition, the similarity between more items is stronger after the modal similarity diagram is passed, so that single mode also has a good effect. The other is that we have made modal enhancement representations based on LightGCN, and LightGCN model has better learning ability than MMGCN.

5.3.2 Effects of Item-Item Similarity Graph and User Preference Graph. In this set of experiments, we investigate whether the similarity graph constructed by modalities can explore the potential relationship between items, and at the same time, analyze whether the designed user preference graph can explore users' preferences for different modalities from items.

Our model augments the representations of item and user with item-item similarity graphs and user preference graphs. Therefore, we choose to add only the item-side similarity graph module to explore whether the modal information can be used to learn the potential features of the item. Note that we use lightGCN as the convolution method of the interaction graph to capture the collaboration signal, as shown in Figure 3. When we construct the similarity graph using only the modal features to explore the potential features of the item, it is obvious that the effect is highly improved. This indicates that the design of item-item similarity graph can enhance the item representations learned by the collaborative signal, and learn the potential relationships between item and item under the same modality.

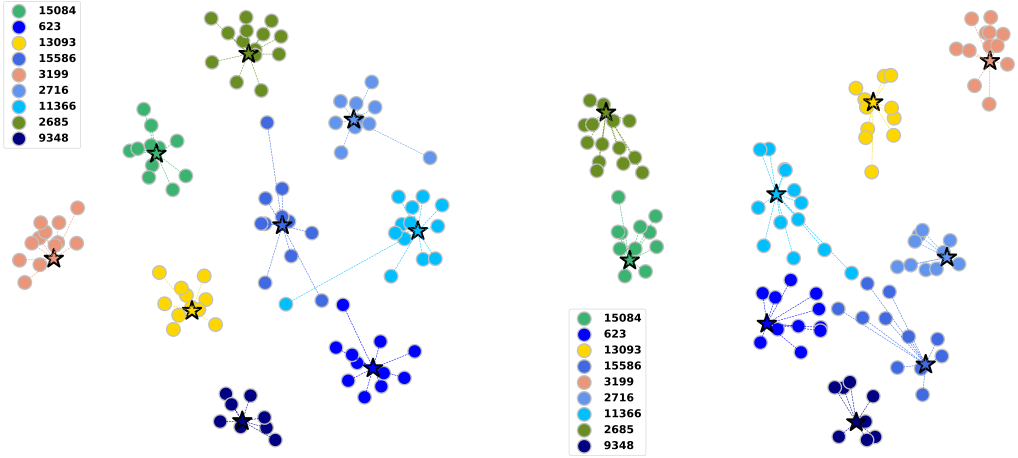
Meanwhile, we believe that there is still a need to learn user-side preferences for modalities, so as to better filter out some noise in redundant modalities. For example, users may be interested in items with similar modalities, but many items with similar modalities often contain more noise, and a large number of recommendations will often lead to user resentment. Therefore, we introduce the user interest graph to solve this problem. In Figure 3, we can see that the effect of adding the modal interest graph has increased significantly again, although not as much as the improvement of adding the item-item similarity graph, but it is easy to see that considering users' interest in different modalities improves the representation of items while enhancing the user representation to achieve the optimal recommendation effect.

5.3.3 Visualization. To show that our model learns better embedding, visualization of the learned t-SNE transformed representations derived from LightGCN and LUDP is to be compared. As shown in Figure 4, we randomly took some users to show the best visual representation. The pentagram in the figure represents the user, and the circle connected to it represents the item that the user interacts with. From this figure, we can see that both LightGCN and LUDP models cluster well in two dimensions. Moreover, LightGCN's cluster effect is more evenly distributed, while in our model LUDP, the user interacts items more closely and similar users tend to cluster more together. It may be caused by our design of item-item modal similarity graph, where users can dig out potential items with modal similarity, resulting in the representation of users also becoming close.



(a) Tiktok (b) Movielens

Fig. 3. Performance comparison of variants of LUDP.



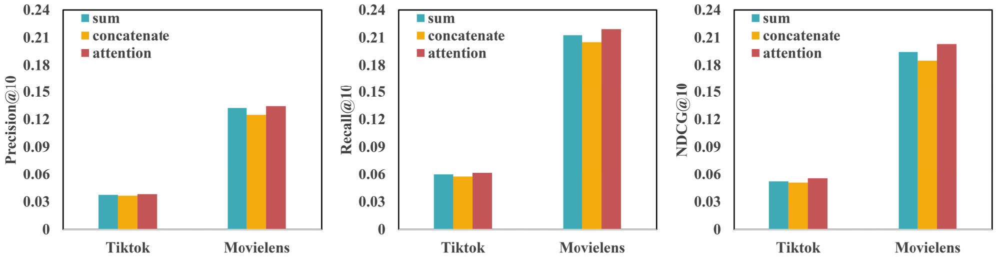
(a) LightGCN (b) LUDP

Fig. 4. Visualization of the learned t-SNE transformed representations derived from LightGCN and LUDP.

5.3.4 Effects of Attention Layer. In our model, we train different weights for both the item and user side multimodal fusion. To achieve the effect of attention on different modalities, in this section we evaluate different fusion methods for multiple modalities, such as sum, concatenate, and attention. These fusion methods are evaluated at Recall@10, as shown in Figure 5; we have the following findings: It can be seen from the figure that the fusion approach of Attention performs best in both datasets. This may be because the different modalities are independent of each other and have different effects on item and user representation enhancement. Next, the fusion of fully connected layers is the worst, while the direct summation fusion approach is only a little worse than Attention. We speculate that the fusion method of splicing and then passing through the fully connected layers will make training more difficult and lead to poorer results. In contrast, direct summation does not train more parameters, and it considers that each modality has the same degree of influence.

5.4 Network Depth and Setting of Hyperparameters(RQ3)

5.4.1 Network Depth Analysis. To investigate whether our model can learn information from different embedding layers, we superimposed deeper layers to vary the depth of the model. In our



(a) Precision@10 (b) Recall@10 (c) NDCG@10

Fig. 5. Performance comparison of variants of embeddings fusion.

Table 5. Performance Comparison with Different Number of Layers

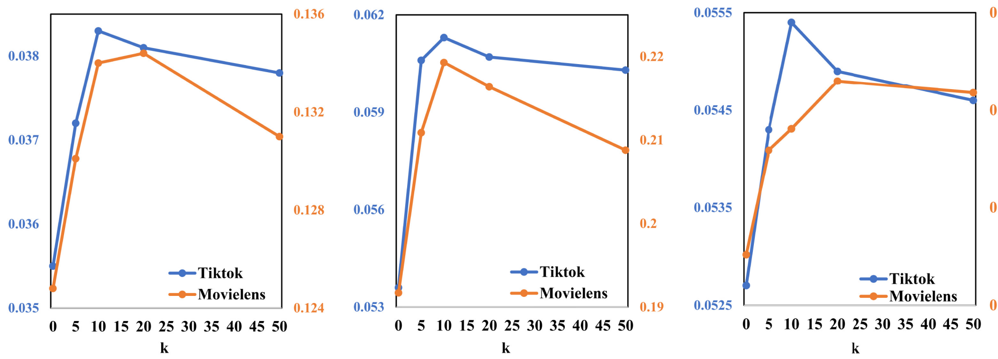
Dataset	Layer	Precision10	Recall@10	NDCG10
TikTok	One	**0.0383**	**0.0613**	**0.0554**
	Two	0.0376	0.0602	0.0547
	Three	0.0366	0.0588	0.0532
Movielens	One	0.1288	0.2015	0.1892
	Two	**0.1344**	**0.2193**	**0.2030**
	Three	0.1298	0.2088	0.1927

model, both the user-item interaction graph and the multimodal similarity graph use layer overlay. However, since the user-item interaction graph is a baseline graph, we fixed it to two layers to explore the effect of multimodal similarity graphs with different layers. As shown in Table 5, we found that more layers of overlay did not improve the performance on the TikTok dataset, which may be caused by overfitting due to the sparsity of the data. In contrast, on the MovieLens dataset, the best performance is achieved by stacking two layers, and the performance may decrease by continuing to increase the number of layers. It may be that the deeper structure may introduce noise to the representation learning and lead to performance degradation.

5.4.2 Setting of Hyperparameters K value. The construction of modal similarity graphs can effectively mine the potential structure of item. We adopt the K-NN sparsification operation to construct the sparse graphs, and set different values of k that determine the k item associated in a certain modality. Thus, the k-value determines the amount of information to be propagated. We tested the performance of our model on two datasets with different values of $k \in [5, 10, 20, 50]$, and Figure 6 represents the experimental results. We can find that the performance shows an increasing and then decreasing trend when k gradually increases. This is because when k is small, the relationship between items and items in a single modality is not fully utilized. When k becomes larger, some less similar item-item relationships (noise) may be introduced into the learning process, which affects the model performance.

6 CONCLUSION AND FUTURE WORK

In this work, we use modal information to construct modal similarity graphs and the item ID embeddings propagated and aggregated on the similarity graph to learn the potential structural information of items. At the same time, we construct user preference graphs based on interaction information, and let the modal information propagate on the user preference graphs to explore



(a) Precision@10 (b) Recall@10 (c) NDCG@10

Fig. 6. Performance impact of different k value settings.

users' preferences for different modalities. Finally, we use these learned user and item representations as auxiliary information to enhance the representations learned through collaborative signals on the bipartite graph. The validation results on two public datasets, TikTok and MovieLens, demonstrate the effectiveness of our proposed model.

In the future, we will focus on research in multimodal feature fusion. Multimodal recommendations require better modal features to be extracted, but these features often contain a lot of noise. The treatment of noise in the modalities is what we will study in the future.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[2] Rianne van den Berg, Thomas N. Kipf, and Max Welling. 2017. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263* (2017).

[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In *International ACM SIGIR Conference*. 335–344.

[4] Tao Chen, Xiangnan He, and Min Yen Kan. 2016. Context-aware image tweet modelling and recommendation. In *ACM on Multimedia Conference*. 1018–1027.

[5] Tianqi Chen, Weinan Zhang, Qiuxia Lu, Kailong Chen, and Yong Yu. 2012. SVDFeature: A toolkit for feature-based collaborative filtering. *Journal of Machine Learning Research* 13, 1 (2012), 3619–3622.

[6] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C. Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–28.

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

[8] Josef Feigl and Martin Bogdan. 2019. Neural networks for personalized item rankings. *Neurocomputing* 342 (2019), 60–65.

[9] Q. Guo, Z. Sun, J. Zhang, and Y. L. Theng. [n. d.]. Modeling heterogeneous influences for point-of-interest recommendation in location-based social networks. *Nanyang Technological University, Singapore* ([n. d.]).

[10] Wei Guo, Rong Su, Renhao Tan, Huifeng Guo, Yingxue Zhang, Zhirong Liu, Ruiming Tang, and Xiuqiang He. 2021. Dual graph enhanced embedding neural network for CTRPrediction. *arXiv preprint arXiv:2106.00314* (2021).

[11] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Li He, Hongxu Chen, Dingxian Wang, Shoaib Jameel, Philip Yu, and Guandong Xu. 2021. Click-through rate prediction with multi-modal hypergraphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 690–699.

[13] Ming He, Zekun Huang, and Han Wen. 2021. MPIA: Multiple preferences with item attributes for graph convolutional collaborative filtering. In *International Conference on Web Engineering*. Springer, 225–239.

[14] R. He and J. McAuley. 2015. VBPR: Visual Bayesian personalized ranking from implicit feedback. (2015).

[15] R. He and J. McAuley. 2016. *Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering*. International World Wide Web Conferences Steering Committee.

[16] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 639–648.

[17] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, and K. Wilson. 2016. CNN architectures for large-scale audio classification. *IEEE* (2016).

[18] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 263–272.

[19] Xiaowen Huang, Shengsheng Qian, Quan Fang, Jitao Sang, and Changsheng Xu. 2020. Meta-path augmented sequential recommendation with contextual co-attention network. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–24.

[20] H. Jie, S. Li, S. Gang, and S. Albanie. 2017. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP, 99 (2017).

[21] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *Computer Science* (2014).

[22] Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[23] Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[24] Xixun Lin, Jia Wu, Chuan Zhou, Shirui Pan, and Bin Wang. 2021. Task-adaptive neural process for user cold-start recommendation. (2021).

[25] Fan Liu, Zhiyong Cheng, Lei Zhu, Zan Gao, and Liqiang Nie. 2021. Interest-aware message-passing GCN for recommendation. In *Proceedings of the Web Conference 2021*. 1296–1305.

[26] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019. Disentangled graph convolutional networks. In *International Conference on Machine Learning*. PMLR, 4212–4221.

[27] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[28] B. V. Rossum and F. Frasincar. 2019. Augmenting LOD-based recommender systems using graph centrality measures. *Erasmus University Rotterdam,* Rotterdam, The Netherlands (2019).

[29] Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence* 2009 (2009).

[30] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1405–1414.

[31] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. MGAT: Multimodal graph attention network for recommendation. *Information Processing & Management* 57, 5 (2020), 102277.

[32] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

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

[34] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 165–174.

[35] Xiao Wang, Meiqi Zhu, Deyu Bo, Peng Cui, Chuan Shi, and Jian Pei. 2020. AM-GCN: Adaptive multi-channel graph convolutional networks. In *KDD'20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[36] Yinwei Wei, Xiang Wang, Xiangnan He, Liqiang Nie, Yong Rui, and Tat-Seng Chua. 2021. Hierarchical user intent graph network for multimedia recommendation. *IEEE Transactions on Multimedia* (2021).

[37] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5382–5390.

[38] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3541–3549.

[39] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.

[40] Cai Xu, Ziyu Guan, Wei Zhao, Quanzhou Wu, Meng Yan, Long Chen, and Qiguang Miao. 2020. Recommendation by users' multimodal preferences for smart city applications. *IEEE Transactions on Industrial Informatics* 17, 6 (2020), 4197–4205.

[41] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. *ACM* (2018).

[42] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. *arXiv preprint arXiv:2104.09036* (2021).

[43] Z. Zhao, Y. Yang, C. Li, and L. Nie. 2020. GuessUNeed: Recommending courses via neural attention network and course prerequisite relation embeddings. *ACM Transactions on Multimedia Computing Communications and Applications* 16, 4 (2020), 1–17.

Received 28 January 2022; revised 13 June 2022; accepted 20 November 2022