DualGNN: Dual Graph Neural Network for Multimedia Recommendation

Qifan Wang, Yinwei Wei, *Member, IEEE*, Jianhua Yin, *Member, IEEE*, Jianlong Wu, Xuemeng Song, *Member, IEEE*, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*—One of the important factors affecting micro-video recommender systems is to model the multi-modal user preference on the micro-video. Despite the remarkable performance of prior arts, they are still limited by fusing the user preference derived from different modalities in a unified manner, ignoring the users tend to place different emphasis on different modalities. Furthermore, modality-missing is ubiquity and unavoidable in the micro-video recommendation, some modalities information of micro-videos are lacked in many cases, which negatively affects the multi-modal fusion operations. To overcome these disadvantages, we propose a novel framework for the micro-video recommendation, dubbed Dual Graph Neural Network (DualGNN), upon the user-microvideo bipartite and user co-occurrence graphs, which leverages the correlation between users to collaboratively mine the particular fusion pattern for each user. Specifically, we first introduce a single-modal representation learning module, which performs graph operations on the user-microvideo graph in each modality to capture single-modal user preferences on different modalities. And then, we devise a multi-modal representation learning module to explicitly model the user's attentions over different modalities and inductively learn the multi-modal user preference. Finally, we propose a prediction module to rank the potential micro-videos for users. Extensive experiments on two public datasets demonstrate the significant superiority of our DualGNN over state-of-the-arts methods.**

***Index Terms*—Micro-video recommender systems, graph neural network, multi-modal fusion, representation learning.**

I. Introduction

With the rise of social media, micro-videos have become ubiquitous in our daily life. Facing the overload of micro-videos on the sharing platforms (*e.g.,* Tiktok and Kwai), the service providers are troubled by locating the interesting micro-videos for users. To tackle this drawback, they develop the micro-video recommender system, which aims at discovering the users' tastes and accordingly ranking the candidate micro-videos.

For this purpose, one common solution is incorporating the content information into the collaborative filtering scheme, so as to model the user preference with the content information [1]–[3]. For example, Liu *et al.* [4] proposed to extract the semantic information from reviews, and combine them with the matrix factorization technique. Generally, the approaches could be classified into two categories: modality-agnostic methods and modality-aware methods. For the modality-agnostic methods, they ignore the difference of user preference among the multiple modalities and treat the multi-modal user preference as a whole. For instance, Liu *et al.* [5] proposed a User-Video Co-Attention Network (UVCAN), which concatenates multi-modal features of users and micro-videos, and uses the attention mechanism to learn multi-modal representations for them. However, such methods forgo the distinction between different modalities, therefore, some modality-aware approaches are proposed more recently. For example, Wei *et al.* [6] proposed a Multi-Modal Graph Convolutional Network (MMGCN), which learns the single-modal user preferences and concatenates them to represent the multi-modal user preference on the micro-video. Despite the remarkable performance achieved by previous studies, we argue that there are two challenges still remained, which harm the multi-modal user preference modeling and cause the sub-optimal performance.

- It is presupposed that the multi-modal fusion pattern for all users follows a pre-defined and unified manner. However, this assumption is too strict, since the users tend to place different emphasis on different modalities and have their opinions on integrally choosing the multi-modal instances. Taking Figure 1 as an example, User 1 and User 2 prefer the high-quality frame, while User 3 and User 4 are probably attracted by the beautiful soundtrack. Moreover, a unified fusion operation (*e.g.,* element-wise mean) makes each user's single-modal representations equally aggregated. In the contrast, the personalized fusion pattern for each user should give higher proportions to visual and audio modalities for User 1, User 2, and give higher proportions to audio and textual modalities for User 3 and User 4. Therefore, how to define the particular fusion function for each user is our first challenge.
- Modality-missing is ubiquity and unavoidable in the micro-video recommendation. For the micro-video, its

Manuscript received 11 August 2021; revised 26 October 2021; accepted 15 December 2021. Date of publication 24 December 2021; date of current version 12 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62172261 and 61802231 and in part by the Shandong Provincial Natural Science Foundation under Grant ZR2019QF001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wen-Huang Cheng. *(Corresponding authors: Yinwei Wei; Jianhua Yin.)*

Qifan Wang, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie are with the College of Computer Science and Technology, Shandong University, Qingdao 266237, China (e-mail: wqf@mail.sdu.edu.cn; jhyin@sdu.edu.cn; jlwu1992@sdu.edu.cn; sxmustc@gmail.com; nieliqiang@gmail.com).

Yinwei Wei is with the School of Computing, National University of Singapore, Singapore 119077 (e-mail: weiyinwei@hotmail.com).

Digital Object Identifier 10.1109/TMM.2021.3138298

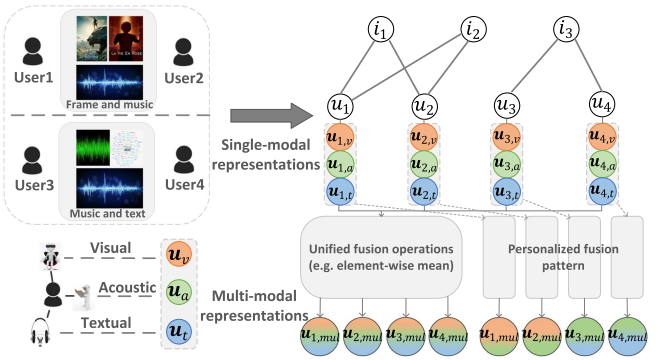


Fig. 1. Illustration of the difference between the unified operations (*e.g.,* element-wise mean) and the personalized fusion pattern in multi-modal user preference representation learning, where $\boldsymbol{u}_v, \boldsymbol{u}_a, \boldsymbol{u}_t$ denote single-modal representations of user u. And $\boldsymbol{u}_{mul}$ denotes the multi-modal representation user u, and i denotes micro-video.

development is powered by its easy-to-operate and instant sharing. However, the loose requirement for posting is a double-edged sword, which inevitably leads to information missing [7]. Modality-missing not only perturbs the single-modal user preference representation, but negatively affects the multi-modal fusion operations. Particularly, the unified fusion pattern is more difficult to adapt to the uncertain multi-modal inputs. Therefore, how to make the multi-modal fusion pattern be robustness facing the modality-missing cases is our second challenge.

To remedy the challenges, we propose to design an inductive model, which is able to capture the specific multi-modal fusion pattern for each user. To this end, we follow the fact that users who have the same behavior tend to own similar preferences and thus the co-occurrence users (i.e., users who have browsed some of the same videos) reveal their similar preferences with respect to both single-modal and multi-modal cases. Therefore, we introduce to model the co-occurrence relationship between users to collaboratively mine the particular fusion pattern for each user. For this purpose, we build a novel framework, termed Dual Graph Neural Network (DualGNN), upon the user-microvideo bipartite graph and the user co-occurrence graph. In particular, we first simplify the graph-based model on the multimedia recommendation and devise a new single-modal preference learning module, which performs the graph operations on the user-microvideo graph in each modality to capture single-modal user preferences on different modalities. And then, we design a multi-modal representation learning module to represent the multi-modal user preference. Different from existing works that perform a unified operation (*e.g.,* element-wise mean) on multi-modal information, we disentangle the learning process into the information construction and aggregation operations, in order to explicitly model the user's attentions over different modalities and inductively learn the multi-modal user preference. More specifically, we organize the co-occurrence users as a graph structure and initialize each user node with its single-modal representations weighted by the individual attention on each modality. Furthermore, by iteratively conducting the graph operations, we capture the co-occurrence relationship and inject it into the user nodes, facilitating to learn the particular multi-modal user representation for each user. Finally, a prediction module could be used to rank the potential micro-videos for users by measuring the similarity of each user and micro-video pair. To demonstrate the effectiveness of our proposed method, we conduct extensive experiments on two publicly accessible datasets. The experimental results show that DualGNN outperforms state-of-the-art baselines, such as MMGCN, LR-GCCF [8], and LightGCN [9]. To summarize, our contributions are three-fold:

- We argue the issue that the users tend to place different emphasis on different modalities, and highlight the significance of learning particular fusion pattern for each user in multi-modal fusion.
- We propose a new framework DualGNN, which captures the user's single-modal preferences on different modalities, then explicitly modeling the user's attentions over different modalities, and inductively learns the multi-modal user preference. Further, the learned multi-modal fusion pattern could reduce the negative influence of the modality-missing cases.
- We perform extensive experiments on two public datasets to demonstrate the effectiveness of our approach. Our codes are available in https://github.com/wqf321/dualgnn

The remainder of this paper is organized as follows. We introduce related work in section II, and give the description of the problem statement of DualGNN in Section III. In section IV, the detail of the proposed model is presented. In section V, we setup the experiments and present the results with analysis. Finally, we conclude our work in Section VI.

II. Related Work

A. Micro-Video Personalized Recommendation

Recommender systems play a pivotal role in current micro-video sharing platforms, which attracts a large number of researchers to design efficient recommendation models [10], [11]. These models parameterize users and micro-videos as embeddings, and learn the embedding parameters by reconstructing historical user-microvideo interactions. For example, He *et al.* [12] proposed a Visual Bayesian Personalized Ranking model (VBPR) model based on CF-framework. It leverages visual features to enrich the ID embeddings of micro-videos, and learns better representation of users. Liu *et al.* [5] proposed a novel framework User-Video Co-Attention Network (UVCAN), in order to learn multi-modal information from both user and microvideo side using attention mechanism. Jiang *et al.* [1] developed a parallel temporal mask network, which is able to learn multiple temporal information for micro-video recommendation. However, most of these works only leverage the multi-modal information as a whole or even ignore the multi-modal information. Wei *et al.* [6] developed a new method MMGCN which employs information propagation on the modality-aware bipartite user-microvideo graph, in order to obtain better user representations based on multiple modalities micro-video content information.

However, they simply average users' multi-modal representations as a result, which is insufficient to model users' preferences for different modalities, and cannot deal with the modality-missing problem. Different from the existing studies, we propose to model a personalized multi-modal fusion pattern for each user, and iteratively conduct the graph operations on the user co-occurrence graph, so as to inductively learn the multi-modal user preference and solve the modality-missing problem.

B. GNN-Based Personalized Recommendation

Deep learning models have been widely used in structured data, such as audio, image, and text [13]–[15]. However, it is difficult to define an intuitive computing framework for graph data due to its complexity. Towards this end, Kipf *et al.* [16] proposed a graph convolutional network to propagate neighbor information on the graph. Due to its effectiveness and simplicity, Graph Neural Networks (GNN) have been widely used in the computer vision [17], [18], information retrieval, and recommendation [3], [19]. For instance, Ji *et al.* [20] transformed the irregular superpixel information to a structured feature representation, and utilized the graph neural network to interact the context of superpixel nodes for saliency detection. For user preference modeling, Berg *et al.* [21] utilized the graph convolutional operation on the bipartite interaction graph to generate the user preference representation by aggregating its neighbor micro-videos' features. Recent years, a new method Neural Graph Collaborative Filtering (NGCF) [22] is proposed to explicitly integrate the collaborative signals into the embedding process. Further, He *et al.* [9] simplified the design of GCN by abandoning the use of feature transformation and nonlinear activation, in order to let it be more concise and appropriate for the recommendation task, and proposed a strong model named LightGCN. Recently, Liu *et al.* [23] proposed an IMP-GCN model to exploits high-order neighbors from the same sub-graph, and designed an unsupervised sub-graph generation module, which can effectively identify users with common interests by exploiting both user feature and graph structure.

While these GCN-based recommendation models almost simply apply graph convolution network propagation on the bipartite interaction graph, ignoring user's multi-modal preference for micro-videos, thus we design the multi-modal representation learning module to learn user's multi-modal preference.

C. Multi-Modal Fusion

Multi-modal fusion has gained much attention of many researchers due to the benefit it provides for various multimedia analysis tasks. Traditional multi-modal fusion mainly consist of early, late, and hybrid fusion approaches. In the early fusion schemes, features from different modalities are integrated as a whole and input to the network. For example, Couprie *et al.* [24] presents an early fusion strategy via a simple concatenation of RGB and depth channels before feeding into a segmentation network. Hu *et al.* [25] proposed a novel early fusion architecture based on attention mechanism, known as ACNet, which selectively gathers valuable features from RGB and depth branches. In contrast, late fusion methods merge data after a separate full processing, and the individual modalities can be processed by powerful targeted approaches. Wei *et al.* [26] augmented feature vectors by the cooperative nets in each modalities and fed into an attention net, followed by a late fusion over the prediction results from different modalities. The hybrid fusion strategy is proposed to combine the strengths of early fusion and late fusion as an alternative method. Wang *et al.* [27] proposed to combine early and late fusion together, and designed a universal hybrid fusion framework that can effectively overcome the shortcomings of the late fusion scheme.

However, most of the traditional multi-modal fusion approaches are unified, which is unfavorable to describe users' personalized multi-modal preferences in recommendation tasks. Considering that it is unrealistic to learn the particular multi-modal fusion function separately for each user, we proposed to explicitly model the user's attentions over different modalities, and inductively learn the multi-modal user preference.

III. Preliminary

Given a set of N users $u \in \mathcal{U}$ and a set of M micro-videos $i \in \mathcal{I}$. Let $\mathcal{P}^+ = \{p_{ui} | u \in \mathcal{U}, i \in \mathcal{I}\}$ be the observed interactions, where p_{ui} denotes the edge between user u and micro-video i. We organize the user-microvideo bipartite graph as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ according to their historical interactions. Whereinto, $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$ denotes the node set of all users and micro-videos, and the edge set $\mathcal{E} = \mathcal{P}^+$ represents observed user-microvideo interactions. Beyond the interactions, each micro-video contains multi-modal content information. And, we use $m \in \mathcal{M} = \{v, a, t\}$ as the modality indicator, where v, a, and t refer to the visual, acoustic, and textual modalities, respectively. The representations of the input user u of each modality are randomly initialized as $\boldsymbol{u}_m^{(0)} \in \mathbb{R}^d$, and the representations of the input micro-video i are pre-processed by multi-modal features as $\boldsymbol{i}_m^{(0)} \in \mathbb{R}^d$, where d denotes the dimension of user and micro-video representations.

As mentioned before, we aim to construct the co-occurrence relationship between users. Thus, we record the number of micro-videos that have been both interacted by each user pair in a common way. We use a matrix $\boldsymbol{C} \in \mathbb{R}^{N \times N}$ to represent each user's co-occurrence times with other users, where N denotes the size of the user set. Whereinto, each entry of $\boldsymbol{C}$, denoted $C_{u,u'}$, is the number of co-occurrences between u and u'. Further, we use set $\mathcal{C}_u = \{C_{u,u^1}, C_{u,u^2}, \ldots, C_{u,u^{N-1}}\}$ to denote the co-occurrence times of all co-occurrence users with user u.

IV. Method

In this section, we address the aforementioned challenges of modeling the users' multi-modal preferences representation. We begin with a brief overview of our framework and then elaborate on its components.

As illustrated in Figure 2, the DualGNN framework consists of three components: 1) a single-modal representation learning module, which performs the graph network operations to capture the modal-specific user preference and micro-video representation on each modality user-microvideo bipartite graph;

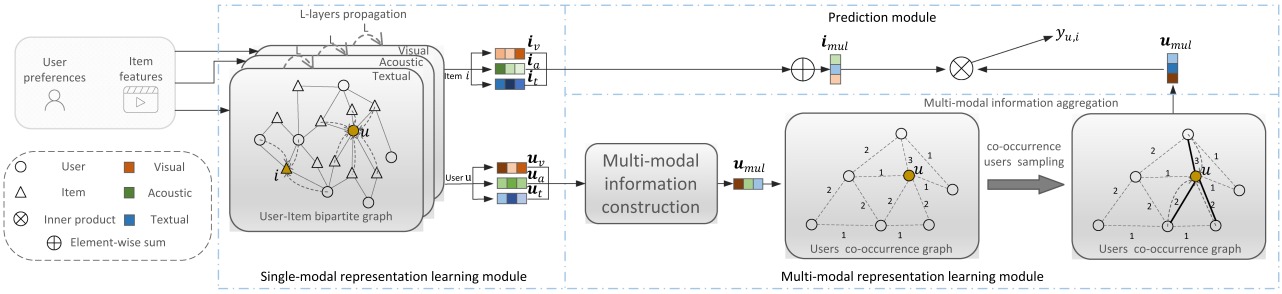


Fig. 2. The overall framework of our proposed DualGNN. It consists of the single-modal representation learning module which captures the single-modal user preference on each modality user-microvideo bipartite graph, the multi-modal representation learning module that explicitly models the user's tastes on different modalities and inductively learns the multi-modal user preference, and the prediction module to estimate the user's preference towards the target micro-video.

2) a multi-modal representation learning module that explicitly models the user's taste on different modalities and inductively learns the multi-modal user preference; and 3) a prediction module that ranks the potential micro-videos for users by measuring the similarity of each user and micro-video pair.

A. Single-Modal Representation Learning Module

Following the settings in MMGCN, we aim at performing the graph convolutional operations on the single-modal bipartite graph to learn the user preference in each modality (a.k.a. single-modal representation). However, we are inspired by the arguments that the feature transformation and nonlinear activation of common GCN have no positive impact on the effectiveness of collaborative filtering [9], and accordingly simplify the graph convolutional operations for the multimedia recommendation. Specifically, we discard the self-loop information propagation of each node and purely model its collaborative signal. Then, the feature transformation is also ignored to reduce the cost and facilitate the model optimization. Therefore, at the $(l+1)$-th layer, the operation could be formulated as:

$$\begin{aligned} \boldsymbol{u}_m^{(l+1)} &= \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \boldsymbol{i}_m^{(l)}, \\ \boldsymbol{i}_m^{(l+1)} &= \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i|}\sqrt{|\mathcal{N}_u|}} \boldsymbol{u}_m^{(l)}, \end{aligned} \tag{1}$$

where $\mathcal{N}_u$ and $\mathcal{N}_i$ represent the neighbors of u and i in the bipartite graph, respectively. In addition, $\boldsymbol{u}_m^{(l)} \in \mathbb{R}^d$ and $\boldsymbol{i}_m^{(l)} \in \mathbb{R}^d$ denote the representations of user and micro-video learned from the previous layer in each modality, respectively. We use the symmetric normalization $\frac{1}{\sqrt{|\mathcal{N}_i|}\sqrt{|\mathcal{N}_u|}}$ to avoid the scale of representations increasing with graph convolution operations.

By iteratively conducting the above operations, the users and micro-videos obtain their collaborative signal from each layer. After L layers propagation, we combine them with the features of each node to form the desired single-modal representations of each user and micro-video, formally,

$$\boldsymbol{u}_m = \sum_{l=0}^{L} \boldsymbol{u}_m^{(l)} \quad , \quad \boldsymbol{i}_m = \sum_{l=0}^{L} \boldsymbol{i}_m^{(l)}. \tag{2}$$

As a result, the informative signals are encoded into the single-modal user and micro-video representations. The same operations are adopted to the bipartite graph of each modality. After propagating on different modalities bipartite graphs, we gain the representations of the users and micro-videos in each modality.

B. Multi-Modal Representation Learning Module

To inductively learn the particular multi-modal fusion pattern for each user, we disentangle the learning process into the information construction and aggregation operations.

1) Multi-modal information construction: Inspired by previous methods on fusing multi-modal information [6], we also present several construction methods for the multi-modal information construction. At first, we initialize a parameter set $\{\alpha_{u,v} = 1, \alpha_{u,a} = 1, \alpha_{u,t} = 1\}$ for each user, where $\alpha_{u,v}, \alpha_{u,a}$, and $\alpha_{u,t}$ denote the user's preference for visual, acoustic, and textual modalities, respectively.

a) Attentively concatenation construction: An intuitive method to construct the multi-modal information is to concatenate each single-modal representation of a user as:

$$\begin{cases} \boldsymbol{h}_u = \alpha_{u,v}\boldsymbol{u}_v || \alpha_{u,a}\boldsymbol{u}_a || \alpha_{u,t}\boldsymbol{u}_t, \\ \boldsymbol{u}_{mul} = \boldsymbol{W}_m \boldsymbol{h}_u + \boldsymbol{b}_m, \end{cases} \tag{3}$$

where $||$ denotes the concatenation operation, $\boldsymbol{W}_m \in \mathbb{R}^{d \times 3\,d}$ and $\boldsymbol{b}_m \in \mathbb{R}^d$ denote linear transformation matrix and bias, respectively. u_{mul} denotes the constructed multi-modal representation of u. In this way, each single-modal representation could be intactly considered in constructing the multi-modal information.

b) Attentively sum construction: Inspired by the fact that the element-wise sum can preserve most features of the multi-modal information [28], we could attentively integrate the single-modal preference of users as:

$$\boldsymbol{u}_{mul} = \alpha_{u,v}\boldsymbol{u}_v + \alpha_{u,a}\boldsymbol{u}_a + \alpha_{u,t}\boldsymbol{u}_t. \tag{4}$$

c) Attentively maximum construction: For user's single modal representations, we select the maximum value of each dimension as the user's multi-modal preference representation. Such operation can be formally defined as:

$$\boldsymbol{u}_{mul} = max(\alpha_{u,v}\boldsymbol{u}_v, \alpha_{u,a}\boldsymbol{u}_a, \alpha_{u,t}\boldsymbol{u}_t). \tag{5}$$

This operation is based on the assumption that the most prominent single-modal representation of each user carries the richest information for its multi-modal representation.

2) Multi-modal information aggregation: Following the fact that users who have interacted with the same micro-videos are generally close to each other in the multi-modal preference. We argue that the personalized fusion pattern of each user is hidden in such user co-occurrence relationship.

However, the co-occurrence times between a user and his/her co-occurrence users is not consistent. More specially, the user may have a large number of co-occurrence times with a small group of users, while most users only co-occur a few times with the user. Thus, we argue that only users with a certain number of co-occurrences have more similar multi-modal preferences. For this purpose, we propose a Top-K sampling strategy for the user co-occurrence graph construction as follows.

Top-K sampling strategy: First, we define the sampled graph as $\mathcal{G}_U = \{\mathcal{U}, \mathcal{A}\}$, and $\mathcal{A} = \{(u, u')|u, u' \in \mathcal{U}\}$ reflects the node pairs between u and u'. We sample the top-K frequent users for each user from the user co-occurrence matrix $\boldsymbol{C}$, and an edge in graph $\mathcal{G}_U$ could be defined as $q_{uu'}$. If $q_{uu'} = 1$, it indicates that $C_{u,u'}$ belongs to the top-K values of $\mathcal{C}_u$, otherwise $q_{uu'} \neq 1$. We will discuss the effect of K on the model performance in the experiment section.

Thereafter, based on the sampled graph $\mathcal{G}_U$, we design two aggregation methods to derive each user's fusion pattern in the user co-occurrence graph as follows.

Mean aggregation: This method simply averages the representations of each user's neighbor nodes as the aggregated information, and updates the user's representation as:

$$\boldsymbol{u}_{mul}^{(l'+1)} = \boldsymbol{u}_{mul}^{(l')} + \sum_{u' \in \mathcal{N}_{u,c}} \frac{1}{|\mathcal{N}_{u,c}|} \boldsymbol{u}_{mul}^{\prime(l')}, \tag{6}$$

where l' is the number of GCN layers, and $\mathcal{N}_{u,c}$ denotes user u's neighbor nodes in the user co-occurrence graph.

Softmax weighted aggregation: In order to enhance the impact of neighbor users who have more co-occurrence times, we use the softmax function to compute each user's aggregation weight:

$$\boldsymbol{u}_{mul}(l'+1) = \boldsymbol{u}_{mul}(l') + \sum_{u' \in \mathcal{N}_{u,c}} \frac{\exp(C_{u,u'})}{\sum_{u' \in \mathcal{N}_{u,c}} \exp(C_{u,u'})} \boldsymbol{u}'_{mul}(l'). \tag{7}$$

After L' layers propagation, each user's personalized fusion pattern could be mined from its neighbor nodes in the user co-occurrence graph. Noticing that we don't design the symmetric multi-modal representation learning module for micro-videos, and the reasons can be summarized in the following two aspects. For one side, micro-videos' features are more objective than the users' preferences to characterize the instance. For another, the micro-videos' fusion patterns are probably consistent with their exposure accessed by users. More specifically, since the micro-videos are collected from the same platform, we believe that they expose to users in a unified manner, which causes the micro-video's same fusion pattern.

C. Prediction Module

After the propagation of previous modules, we capture the representations of nodes as:

$$\begin{cases} \boldsymbol{u}^* = \boldsymbol{u}_{mul}^{(L')}, \\ \boldsymbol{i}^* = \boldsymbol{i}_v + \boldsymbol{i}_a + \boldsymbol{i}_t, \end{cases} \tag{8}$$

where $\boldsymbol{u}^*$ and $\boldsymbol{i}^*$ denote the final representation of user u and micro-video i, respectively. $\boldsymbol{u}_{mul}^{(L')}$ denotes the user's representation after multi-modal information aggregation. Finally, we compute the inner product between user and micro-video representations as:

$$y_{u,i} = \boldsymbol{u}*^{\mathsf{T}} \boldsymbol{i}*, \tag{9}$$

where the output $y_{u,i}$ denotes the user's preference towards the target micro-video. A high score suggests that the user prefers the micro-video and vice versa.

D. Optimization

To optimize the model parameters, we adopt the Bayesian Personalized Ranking [29] loss to rank user-microvideo pairs. Thus, we construct a triplet of one user u, one observed micro-video i, and one unobserved micro-video j, formally as:

$$\mathcal{R} = \{(u, i, j)|(u, i) \in \mathcal{E}, (u, j) \notin \mathcal{E}\}, \tag{10}$$

where $\mathcal{R}$ is a triplet set for training. And we formulate the objective function as:

$$\mathcal{L} = \sum_{(u,i,j) \in \mathcal{R}} - \ln \mu(y_{u,i} - y_{u,j}) + \lambda \left|\left|\theta\right|\right|_2, \tag{11}$$

where $\mu(\cdot)$, λ, and θ denote the $sigmod$ function, regularization weight, and parameters of models, respectively.

E. Model Complexity

In this section, we analyze the complexity of the proposed DualGNN, and compare it with MMGCN and LightGCN. Suppose the number of nodes and edges in the user-microvideo interaction graph are $|\mathcal{V}|$ and $|\mathcal{E}|$, respectively. Let $|\mathcal{U}|$ denotes the number of nodes in the user co-occurrence graph, and $\mathcal{T}$ is the number of triplets in the training set. The complexity mainly comes from two parts:

Graph Convolution: The complexity of graph convolution of LightGCN is $O(L|\mathcal{E}|d)$. Since DualGNN constructs three sub-graphs for different modalities, the complexity of DualGNN on sub-graphs is $O(3L|\mathcal{E}|d)$. Considering the multi-modal information aggregation of DualGNN, the complexity of graph convolution of DualGNN is $O(3L|\mathcal{E}|d + L'|\mathcal{U}|Kd)$. MMGCN not only constructs three sub-graphs for different modalities, but also uses the feature transformation in graph convolution. Thus, the complexity of graph convolution of MMGCN is $O(3L|\mathcal{E}|d + 3L|\mathcal{V}|d^2)$.

BPR Loss: For all these models, only the inner product is conducted in the prediction layer, for which the time cost of the whole training epoch is $O(\mathcal{T}d)$.

We summarize the time complexity in training among DualGNN, MMGCN, and LightGCN in Table I.

TABLE I
THE COMPARISON OF ANALYTICAL TIME COMPLEXITY AMONG DUALGNN, MMGCN, AND LIGHTGCN

Model	DualGNN	MMGCN	LightGCN
Graph Convolution	$O(3L\|\mathcal{E}\|d + L'\|\mathcal{U}\|Kd)$	$O(3L\|\mathcal{E}\|d + 3L\|\mathcal{V}\|d^2)$	$O(L\|\mathcal{E}\|d)$
Loss	$O(\mathcal{T}d)$	$O(\mathcal{T}d)$	$O(\mathcal{T}d)$

TABLE II
BASIC STATISTICS OF THE DATASETS. NOTE THAT **V**, **A**, AND **T** DENOTE DIMENSIONS OF VISUAL, ACOUSTIC, AND TEXTUAL MODALITIES, RESPECTIVELY

Dataset	Interactions	Micro-videos	Users	Sparsity	**V**	**A**	**T**
MovieLens	1,239,508	5,986	55,485	99.63%	2,048	128	100
Tiktok	726,065	66,456	32,309	99.96%	128	128	128

V. EXPERIMENTS

In this section, we conducted experiments to demonstrate the effectiveness of our proposed model and answer the following questions:

- RQ1: How do our model and state-of-the-art methods perform on the real-world datasets?
- RQ2: How do different components (e.g. multi-modal information construction methods, the size of sampled co-occurrence users) of our model and hyper-parameters affect DualGNN?
- RQ3: How does DualGNN affect the representation of multi-modal user preference?
- RQ4: Could DualGNN work well in the modality-missing cases?

At the beginning, we presented settings of datasets, baselines, evaluation metrics, and parameters, and then answered the above four questions.

A. Datasets

As the micro-video contains rich multimedia information - frames, sound tracks, and descriptions, we performed experiments on two datasets designed for the micro-video recommendation (i.e., Movielens[1] and Tiktok[2]). The statistics of datasets are summarized in Table II.

Movielens: This dataset is widely used in the personalized recommendation. Beyond the user and micro-video interaction records, this dataset contains the videos' trailers, titles, and descriptions [6]. The keyframes and audio tracks are extracted from the trailers so that we can leverage the visual and acoustic features. Consistent with MMGCN, we used ResNet [30], VGGish [31], and Sentance2Vector [32] to extract the visual, acoustic, and textual features from the frames, audio tracks, and descriptions, respectively.

Tiktok: This dataset is released by Tiktok, which contains users, micro-videos with duration of 3-15 seconds and their historical interactions. In addition, the multi-modal features (i.e. visual, acoustic and textual) are extracted and published without providing the raw data.

For each dataset, we used the ratio 8:1:1 to randomly split the historical interactions of each user and constituted the training set, validation set, and testing set. In the training set, for each user-microvideo pair, we randomly sampled one micro-video, which the user has not interacted with, to construct the triple for optimizing the model. The validation set and testing set are respectively used to tune the hyper-parameters and evaluate the performance in the experiments.

B. Baselines

To demonstrate the effectiveness of our proposed method, we compared it with the following methods:

VBPR [12]: This is a benchmark model in the multimedia recommendation. It incorporates the content information into the collaborative filtering framework. Specially, we concatenated the multi-modal features of the micro-video into a single feature vector to infer interactions between users and micro-videos.

MMGCN [6]: This is a GCN-based framework designed for the micro-video recommendation, in which the single-modal user preference can be learned. It refines the micro-videos' representations with the learned users' single-modal preferences by performing the GCN propagation on the user-microvideo graph in each modality.

STAR-GCN [33]: This is a multi-block graph encoder-decoder framework for graph convolutional matrix completion. The graph encoder generates node representations by encoding semantic graph structures, and the decoder aims to recover the input node embeddings. In our experiments, we leveraged the recovered node embeddings to compute ranking scores.

LR-GCCF [8]: This is a general GCN-based Collaborative Filtering (CF) model for recommendation, which uses a simple linear embedding propagation for each layer, and adopts a residual based network structure which concatenates the embeddings obtained at each layer to predict the user preference for micro-videos.

LightGCN [9]: This is a light yet effective model by including the most essential ingredients of GCN for recommendation. This model adopts the simple weighted sum aggregator as graph convolution operation and combines the embeddings obtained at each layer to form the final representation of each user.

C. Evaluation Metrics

For each user in the validation and testing sets, we treated all micro-videos which she/he did not consume before as the negative samples. With the trained model, we scored the interactions of the user and micro-video pairs and ranked them in the descending order. Consistent with most recommendation works, we adopted Recall@P and Normalized Discounted Cumulative Gain (NDCG@P) to evaluate the performance of proposed model, where @P denotes top-P candidate micro-videos of each user. Specially, we set P =1, 5, 10 and reported the average values of the above metrics for all users during the testing phase.

D. Parameter Settings

We used Xavier [34] to initialize parameters, and optimized DualGNN with Adam [35], based on the default mini-batch size

[1][Online]. Available: https://movielens.org/.
[2][Online]. Available: https://www.tiktok.com/.

TABLE III
PERFORMANCE COMPARISON BETWEEN OUR MODEL AND THE BASELINES OVER TWO DATASETS

Model	Movielens						Tiktok					
	Recall@10	NDCG@10	Recall@5	NDCG@5	Recall@1	NDCG@1	Recall@10	NDCG@10	Recall@5	NDCG@5	Recall@1	NDCG@1
VBPR	0.2180	0.2389	0.1324	0.2028	0.0362	0.0997	0.0428	0.0426	0.0267	0.0325	0.0084	0.0149
MMGCN	0.2305	0.2529	0.1506	0.2140	0.0451	0.1102	0.0786	0.0710	0.0475	0.0540	0.0134	0.0236
STAR-GCN	0.2413	0.2653	0.1598	0.2260	0.0496	0.1193	0.0833	0.0748	0.0500	0.0565	0.0148	0.0248
LR-GCCF	0.2445	0.2720	0.1596	0.2325	0.0497	0.1259	0.0779	0.0732	0.0477	0.0552	0.0129	0.0227
LightGCN	0.2658	0.2941	0.1792	0.2553	0.0595	0.1421	0.1208	0.1081	0.0763	0.0840	0.0234	0.0383
DualGNN	**0.2822**	**0.3131**	**0.1911**	**0.2745**	**0.0661**	**0.1579**	**0.1318**	**0.1151**	**0.0842**	**0.0901**	**0.0250**	**0.0403**
%Improv.	**6.17%**	**6.46%**	**6.64%**	**7.52%**	**11.09%**	**11.12%**	**9.11%**	**6.48%**	**10.35%**	**7.26%**	**6.84%**	**5.22%**

as 1024. We set K=40 and K=50 for the number of users for multi-modal information aggregation on the Movielens dataset and the Tiktok dataset, respectively. We searched the learning rate in {0.1, 0.01, 0.001, 0.0001, 0.00001} and the regularization weight in {0.1, 0.01, 0.001, 0.0001, 0.00001}. Since multi-modal features in the Movielens dataset have different dimensions, we unified the dimension to 64 as the input of all models. For the Tiktok dataset, we used the dimension 128 for all models to ensure a fair comparison. Besides, we chose the attentively sum construction and the softmax weighted aggregation (in section 2.3) for DualGNN, and set $L = 2$, $L' = 1$ as the number of GCN layers of the single-modal representation learning module and the multi-modal representation learning module, respectively. Our implementation is based on the Pytorch[3] API.

E. Performances Comparison (RQ1)

The comparative results are summarized in Table III. Note that "%Improv." denotes the improvement of DualGNN's result compared to the best result of baselines, and the best result of the baselines is indicated by "__". From the results, we have the following observations:

(1) All GCN-based models achieve better results than the benchmark model VBPR. It demonstrates that the graph convolutional operations benefit the representation learning in the recommendation, and it do make sense to use the GCN-based framework to learn the multi-modal user representation.

(2) DualGNN obviously outperforms MMGCN in all metrics. The reason has two aspects: on the one hand, DualGNN reduces noise in the GCN propagation by adopting the simple sum aggregator and abandoning the use of self-loop and feature transformation; on the other hand, DualGNN models personalized fusion pattern for each user, while MMGCN simply averages user's each single-modal representation as his/her multi-modal preference.

(3) The proposed DualGNN framework consistently outperforms state-of-the-art baselines in all cases, which could demonstrate its effectiveness for the micro-video recommendation. We attribute the effectiveness of the DualGNN framework to model each user's different preferences for different modalities, and learn the multi-modal user preference in an inductive manner. Thus the DualGNN framework could make multi-modal users preference representation learning more accurately, and increase the accuracy of the top-k recommendation.

[3][Online]. Available: https://pytorch.org/.

TABLE IV
EFFECT OF COMPONENTS IN THE MULTI-MODAL REPRESENTATION LEARNING MODULE

Model	Movielens		Tiktok	
	Recall@10	NDCG@10	Recall@10	NDCG@10
Dual-A-U	0.2467	0.2692	0.1026	0.0897
Dual-A	0.2660	0.2890	0.1136	0.0995
Dual-U	0.2766	0.3068	0.1242	0.1101
DualGNN	**0.2822**	**0.3131**	**0.1318**	**0.1151**

F. Ablation Study (RQ2)

In this section, we conducted several experiments to study the proposed DualGNN as follows.

1) Effect of the components in the multi-modal representation learning module: Dual-A-U: To demonstrate the effectiveness of the multi-modal representation learning module, this variant simply averages each single-modal representation without the user co-occurrence graph propagation.

Dual-A: To evaluate the multi-modal information construction, this variant simply averages each single-modal representation, and feeds the generated representation into the user co-occurrence graph.

Dual-U: This model leverages attentively weighted-sum construction to construct the multi-modal information. To evaluate the co-occurrence graph, in this variant, we neglected it and predicted the interactions of the user-microvideo pairs directly.

As illustrated in Table IV, we observed that: (1) Dual-A outperforms Dual-A-U in all cases, showing that multi-modal information aggregation based on the user co-occurrence graph could boost the expressiveness of the user multi-modal representation; (2) the performance of Dual-U is superior than Dual-A-U, which implies that attentively constructing multiple modalities information can better represent the multi-modal user preference; and (3) DualGNN achieves the best performance, which demonstrates the effectiveness of the multi-modal representation learning module.

2) Effect of single modality and multi-modal information construction methods: We constructed several single modality versions of DualGNN to verify the necessity of the information construction for different modalities. And **Dual/v**, **Dual/a**, and **Dual/t** denote the model only captures single-modal user preference of visual, acoustic, and textual in the whole process, respectively. Note that all of these models feed single-modal representations into the user co-occurrence graph. We compared DualGNN with all these variants and showed the results in Figure 3.

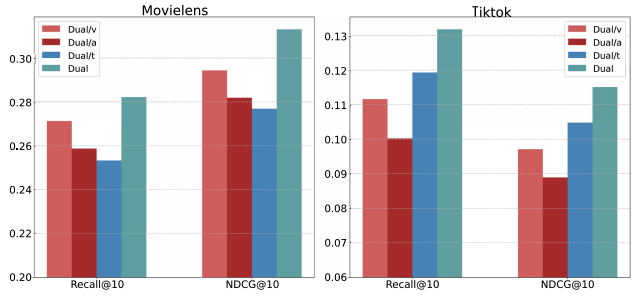


Fig. 3. Effect of single modality and multi-modal fusion.

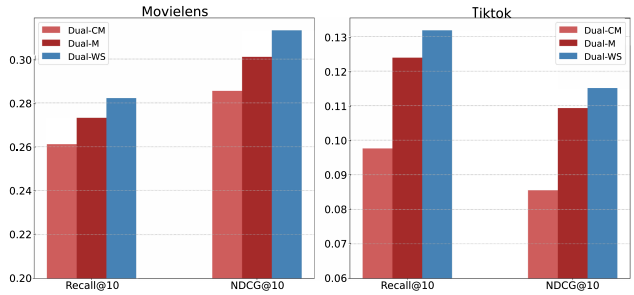


Fig. 4. Effect of different multi-modal information construction methods.

From the results in Figure 3, we have following observations: (1) Models with different single-modal features have different performance, which caused by the different representativeness of different modalities. Specifically, textual modality got the worst performance among single-modal variants in two datasets, while visual modality and acoustic modality got the best performance in Movielens and Tiktok, respectively; (2) **DualGNN** outperforms all variants in Figure 3. It demonstrates that fusing different modal's representation actually improves model's performance, thus the multi-modal representation learning module is indispensable.

For the construction methods mentioned in Section 2.3, we designed the corresponding model variants named as **Dual-CM**, **Dual-M**, and **Dual-WS**, which denote our model attentively using concatenation construction, maximum construction, and sum construction methods, respectively. The results of experiments are illustrated in Figure 4. Clearly, the attentively sum construction method outperforms than other methods, it implies that preserving most information of the multi-modal could capture the user's attentions over different modalities more comprehensively.

3) Effect of the sampled user co-occurrence number and aggregation methods: We conducted experiments to analyze the impact of the sampled user co-occurrence number and different aggregation methods. Specifically, we adopted the mean and softmax weighted aggregation methods described respectively in Eqn. 6 and Eqn. 7, namely **DualGNN-Mean** and **DualGNN-Softmax** to aggregate information from neighbor nodes in the user co-occurrence graph. The results are shown in Figure 5. From the results, we have the following observations: (1) When the sampled user co-occurrence number varies from 10 to 40, results of Recall@10 and NDCG@10 rise slowly. However, as the sampled user co-occurrence number increases further, the performance becomes worse. Since the co-occurrence times between each user and his/her co-occurrence users is not consistent, aggregating users with large count of co-occurrence times is more beneficial to model the multi-modal user preference. (2) The softmax weighted aggregation outperforms the mean aggregation on two datasets, which implies that giving more aggregation weight to neighbor users who have more co-occurrence times could boost the model's performance

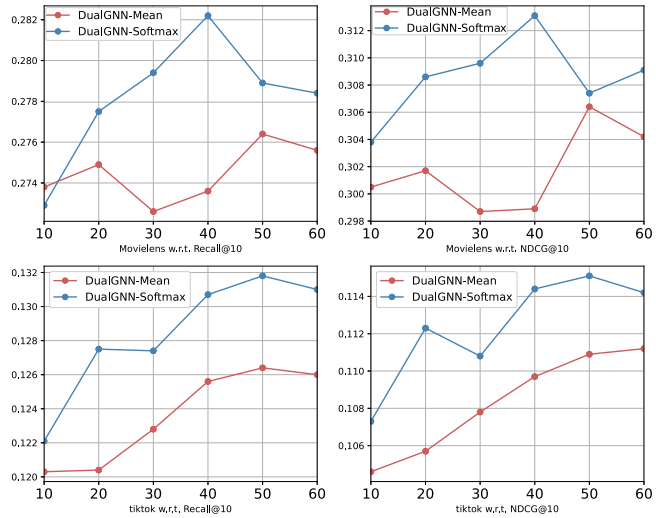


Fig. 5. Effect of the sampled user co-occurrence number and different information aggregation methods.

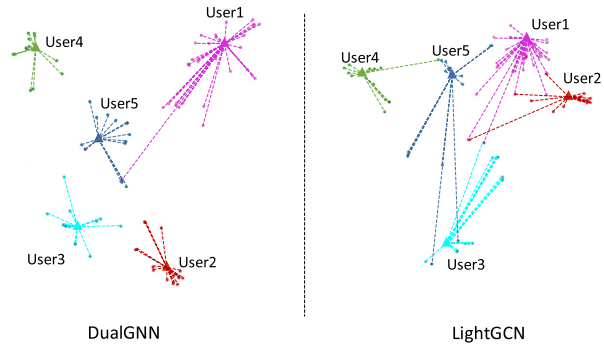


Fig. 6. Visualization of the learned t-SNE transformed representations comparison between LightGCN and DualGNN.

G. Case Study (RQ3)

1) t-SNE Comparison: In this section, we provided a visualization experiments for the DualGNN framework.

We provided 5 users (represented by triangles in different colours) randomly selected from the Movielens datasets with all of their visited micro-videos, and we used the t-Distributed Stochastic Neighbor Embedding (t-SNE) in 2-dimension to exhibit representations of users and micro-videos. As illustrated in Figure 6, we visualized their representations, which are learned from LightGCN and DualGNN, respectively. We found that compared with LightGCN's distribution, DualGNN could represent the users more discriminately. Specially, there are several

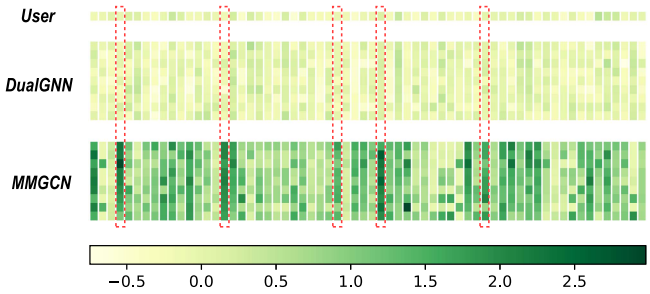


Fig. 7. Visualization of the co-occurrence users' multi-modal representations.

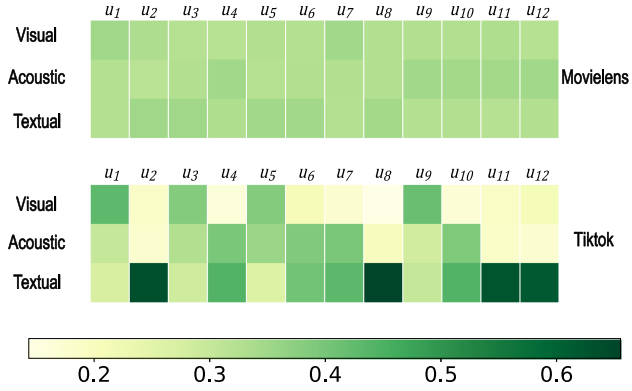


Fig. 8. Visualization of learned construction weight of users selected from two datasets.

micro-videos that are far away from their corresponding users in the result of LightGCN. The reason we suggest is that LightGCN ignores micro-videos' multi-modal features, and only using ID embeddings is not efficient to model user's multi-modal tastes. On the contrary, DualGNN could model user's multi-modal preference effectively.

2) Fusion Pattern Visualization: Considering that MMGCN integrates user's single-model preferences into multi-modal preference, we compared the multi-modal preference representation learned from DualGNN with which learned from MMGCN, in order to verify that DualGNN has indeed learned the user's personalized fusion pattern. We provided several co-occurrence users who have interacted with the same micro-videos from the Movielens dataset. The fusion pattern among the co-occurrence users could be computed as the similarity of the embedding between the selected user and his/her co-occurrence users.

As illustrated in Figure 7, we visualized their embeddings. Each selected user has 64 dim values, and we placed several red dashed boxes to show the fusion pattern difference between representations learned from DualGNN and MMGCN. It is obviously that the same dim values of the selected user and his/her co-occurrence users learned from DualGNN are more similar than those learned from MMGCN. The result demonstrates that our proposed model could better capture the user's multi-modal preference, by using the personalized fusion pattern to inductively fuse each single-modal preference of the user.

In addition, we selected 12 users (denoted as $u_1 \ldots u_{12}$ in Figure 8) from each dataset and showed their learned construction weights for different modalities in Figure 8. The discriminative weights of different modalities show that each user has personalized taste on different modalities. Specially, users' construction weights of different modalities is more discriminating on Tiktok than on Movielens. The reason is that Movielens is more dense than Tiktok, so the difference in user preferences for different modalities is likely to be less obvious.

H. Modality Missing Study (RQ4)

To study how DualGNN perform in the modality-missing cases, we mimicked the scenario in the training data. Specifically, we first select a certain number of micro-videos from all the micro-videos by the drop-ratio, then shuffle and divide them equally into three groups for different modalities, and then remove these micro-videos associated with the edges from

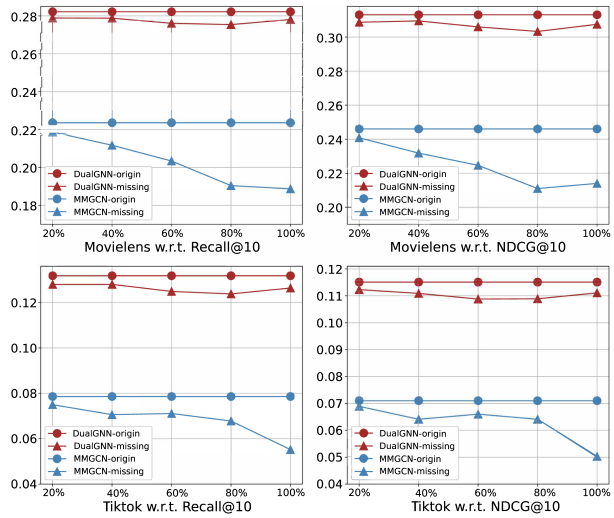


Fig. 9. Comparative experiment under the situation of modality-missing.

the interaction graph in each modality. In this way, when the drop-ratio grows to 100%, at least one modality information of each micro-video is missing. Then, we compared the performances of MMGCN and DualGNN in this setting, and results of experiments are shown in Figure 9.

Note that the horizontal axis refers to the proportion of micro-videos selected accounting for the total number of micro-videos. Clearly, DualGNN shows significant improvements over MMGCN in the modality-missing cases, and we found when the drop-ratio grows, the improvements are more significant. To be more specific, when drop-ratio is set as 100%, the decrease percentage of DualGNN is 1.45% and 4.12% in Movielens and Tiktok, respectively, and the decrease percentage of MMGCN is 15.61% and 29.7% in Movielens and Tiktok, respectively. Considering the lack of items' information in specific modality, the user-item bipartite graph could be perturbed by removing the corresponding edges, which hinders the representation learning of the user preference. Therefore, the reason why DualGNN has obvious improvement compared with MMGCN is that MMGCN could not supplement for the modality-missing information after learning single-modal user preference, while DualGNN could make up for it using the user co-occurrence graph. This promising

finding again verifies the significance of multi-modal information aggregation in solving the modality-missing problem.

VI. Conclusion

In this paper, we aim to solve the challenge of modeling users' preferences for multiple modalities and the modality-missing problem in the micro-video recommendation task. Therefore, we develop a novel model, named DualGNN, which consists two key modules, i.e., the single-modal representation learning module and the multi-modal representation learning module, in order to attentively construct the multi-modal information and inductively model users' multi-modal preferences. To the best of our knowledge, this work is the first attempt to consider how to define the particular fusion pattern for each user in the micro-video recommendation. However, we find that the model using the multi-modal features is easy to be overfiting, and we guess it may be caused by the naive feature pre-processing and aggregation methods. In the future, we expect to study on this question and give some proper solutions.

References

[1] H. Jiang, W. Wang, Y. Wei, Z. Gao, Y. Wang, and L. Nie, "What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 3487–3495.

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

[3] S. C. Hidayati, C.-C. Hsu, Y.-T. Chang, K.-L. Hua, J. Fu, and W.-H. Cheng, "What dress fits me best? fashion recommendation on the clothing style for personal body shape," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 438–446.

[4] H. Liu, Y. Guo, J. Yin, Z. Gao, and L. Nie, "Review polarity-wise recommender," 2021, *arXiv:2106.04155*.

[5] S. Liu, Z. Chen, H. Liu, and X. Hu, "User-video co-attention network for personalized micro-video recommendation," in *Proc. Int. Conf. World Wide Web Conf.*, 2019, pp. 3020–3026.

[6] 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 *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1437–1445.

[7] J. Zhang, L. Nie, X. Wang, X. He, X. Huang, and T. S. Chua, "Shorter-is-better: Venue category estimation from micro-video," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1415–1424.

[8] L. Chen, L. Wu, R. Hong, K. Zhang, and M. Wang, "Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 27–34.

[9] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 639–648.

[10] Y. Li, M. Liu, J. Yin, C. Cui, X.-S. Xu, and L. Nie, "Routing micro-videos via a temporal graph-guided recommendation system," in *Proc. 27th ACM Int. Conf. Multimedia*. Association for Computing Machinery, 2019, pp. 1464–1472.

[11] Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, and T.-S. Chua, "MGAT: Multimodal graph attention network for recommendation," *Inf. Process. Manage.*, vol. 57, no. 5, 2020, Art. no. 102277.

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

[13] L. Lo, H. X. Xie, H.-H. Shuai, and W.-H. Cheng, "Facial chirality: Using self-face reflection to learn discriminative features for facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2021, pp. 1–6.

[14] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "BeautyGlow: On-demand makeup transfer framework with reversible generative network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10042–10050.

[15] W. Ji, X. Li, F. Wu, Z. Pan, and Y. Zhuang, "Human-centric clothing segmentation via deformable semantic locality-preserving network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4837–4848, Dec. 2020.

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

[17] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 4, pp. 1–41, May 2022, doi: 10.1145/3447239.

[18] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "Au-assisted graph attention convolutional network for micro-expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2871–2880.

[19] Y. Wei, X. Wang, X. He, L. Nie, Y. Rui, and T.-S. Chua, "Hierarchical user intent graph network for multimedia recommendation," *IEEE Trans. Multimedia*, early access, 2021, doi: 10.1109/TMM.2021.3088307.

[20] W. Ji, X. Li, L. Wei, F. Wu, and Y. Zhuang, "Context-aware graph label propagation network for saliency detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8177–8186, 2020, doi: 10.1109/TIP.2020.3002083.

[21] R. v. d. Berg, T. N. Kipf, and M. Welling, "Graph convolutional matrix completion," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 1–9.

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

[23] F. Liu, Z. Cheng, L. Zhu, Z. Gao, and L. Nie, "Interest-aware message-passing GCN for recommendation," in *Proc. Int. Conf. World Wide Web Conf.*, 2021, pp. 1296–1305.

[24] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," 2013, *arXiv:1301.3572*.

[25] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1440–1444.

[26] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Trans. Image Process.*, vol. 29, pp. 1–14, 2019, doi: 10.1109/TIP.2019.2923608.

[27] Y. Wang, X. Xu, W. Yu, R. Xu, Z. Cao, and H. T. Shen, "Combine early and late fusion together: A hybrid fusion framework for image-text matching," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[28] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, and H. Mayer, "Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling," *Remote Sens.*, vol. 10, no. 1, pp. 52–65, 2018.

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

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

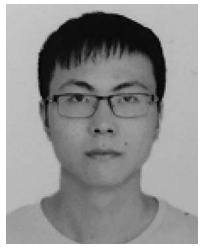
[31] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 131–135.

[32] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–16.

[33] J. Zhang, X. Shi, S. Zhao, and I. King, "Star-GCN: Stacked and reconstructed graph convolutional networks for recommender systems," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4264–4270.

[34] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–16.



Qifan Wang received the B.E. degree in control science and engineering in 2020 from Shandong University, Jinan, China, where he is currently working toward the M.S. degree with the School of Computer Science and Technology. His research interests include multimedia computing, information retrieval, and machine learning.

Yinwei Wei (Member, IEEE) received the M.S. degree from Tianjin University, Tianjin, China, and the Ph.D. degree from Shandong University, Jinan, China. He is currently a Research Fellow with the NExT, National University of Singapore, Singapore. He has authored or coauthored in top forums, such as ACM MM, IEEE TRANSACTIONS ON MULTIMEDIA, and TIP. His research interests include multimedia computing and recommendation. He was a PC Member of several conferences, such as MM, AAAI, and IJCAI, and a Reviewer of *TPAMI*, *TIP*, and IEEE TRANSACTIONS ON MULTIMEDIA.

Jianhua Yin (Member, IEEE) received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2017. He is currently an Assistant Professor with the School of Computer Science and Technology, Shandong University, Jinan, China. He has authored or coauthored several papers in the top venues, such as ACM SIGKDD, MM, SIGIR, and IEEE ICDE. His research interests mainly include data mining and machine learning applications. In addition, he was a PC Member for some leading international conferences, and an Invited Reviewer of top journals.

Jianlong Wu received the B.Eng. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2014, and the Ph.D. degree from Peking University, Beijing, China, in 2019. He is currently an Assistant Professor with the School of Computer Science and Technology, Shandong University, Jinan, China. He has authored or coauthored more than 20 research papers in top journals and conferences, such as TIP, ICML, NeurIPS, and ICCV. His research interests include computer vision and machine learning, and especially weakly supervised learning. He is a Senior Program Committee Member of IJCAI 2021, the Area Chair of ICPR 2020, and a Reviewer of many top journals and conferences, including *TPAMI*, IJCV, ICML, and ICCV.

Xuemeng Song (Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2012, and the Ph.D. degree from the School of Computing, National University of Singapore, Singapore, in 2016. She is currently an Associate Professor with Shandong University, Jinan, China. She has authored or coauthored several papers in the top venues, such as ACM SIGIR, MM and TOIS. Her research interests include the information retrieval and social network analysis. In addition, she was a reviewer of many top conferences and journals.

Liqiang Nie (Senior Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University, Xi'an, China, and the Ph.D. degree from from the National University of Singapore (NUS), Singapore. After Ph.D., he continued his research with NUS as a Research fellow for three and half years. He is currently a Professor with Shandong University, Jinan, China. Meanwhile, he is the Adjunct Dean with Shandong AI Institute. He has authored or coauthored more than 150 papers and received around 11,000 Google scholar citations. His research interests include multimedia analysis and search. He is an AE of information science, *IEEE TKDE*, IEEE TRANSACTIONS ON MULTIMEDIA, and ACM ToMM. He was the recipient of many awards, like SIGIR Best Paper Honorable mention in 2019, ACM MM Best Paper Finalist in 2019, SIGIR Best Student Paper in 2021, SIGMM Rising Star in 2020, TR35 China, and DAMO Academy Young Fellow in 2020.