A Multimedia Graph Collaborative Filter

JINGZHOU SUN\(^1\), HE CHANG\(^1\), WENXUAN ZHAO\(^1\), YOUJIAN YU\(^2\), LIFANG YANG\(^1\), AND XIANGLIN HUANG\(^1\)

\(^1\)Key Laboratory of Convergent Media and Intelligent Technology, Communication University of China, Ministry of Education, Beijing 100024, China
\(^2\)School of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300384, China

Corresponding author: Youjian Yu (yyj@tcu.edu.cn)

This work was funded by the scheme—Research on Key Technologies of Mobile Digital Reading Service, the scheme is in the major project—R&D and Application of Mobile Digital Reading Service Technologies, which is in The National Key Research and Development Program—R&D and Application Demonstration of Common Key Technologies in Modern Service Industry (Grant No. 2019YFB1406101). This work was supported by the Key Laboratory of Convergent Media and Intelligent Technology, Communication University of China, Ministry of Education.

ABSTRACT The Multimedia recommendation has always been an active research field in the area of personalized recommendation, and the core of multimedia recommenders is learning multimodal representations for the users and items. Traditional multimedia recommenders first extract multimodal (such as visual, acoustic, and textual) features by pre-trained networks, and then incorporate ID embeddings with the features to enrich the representations of both users and items. However, these methods only employ multimodal information of the directly interacting items. Recent graph-based efforts utilize high-order connectivities and the message pass mechanism of the graph. The multimodal information of high-hop items is propagated along with the high-order connectivities in a graph and aggregated to enrich the representations of users and items, such as adopting a parallel graph structure to model user preferences on different modalities. However, the users’ preferences for different modalities are unknown. The bipartite graph structure for the entire modalities should differ from the single modality. In this work, we devise a new multimedia recommendation framework, a Multimedia Graph Collaborative Filter (MGCF). In MGCF, a light graph framework with a fusion component is developed to integrate and distill the useful multimodal information. Moreover, the attention mechanism is adopted to enhance the importance of aggregated information. Extensive experiments are conducted on two public datasets: Tiktok and MovieLens. The results show that our model outperforms several state-of-art multimedia recommenders. Further analysis demonstrates the importance of modeling multimodal information for better user and item representations. The code will be open soon.

INDEX TERMS Multimedia recommendation, collaborative filter, graph neural network, deep learning.

I. INTRODUCTION

The multimedia recommendation has always been an active research field in personalized recommendation and [1]–[9] have been applied to much multimedia content sharing systems [10], such as TikTok, Snap, etc. Learning vector representations for users and items play a critical role in multimedia recommenders. Typically, The traditional multimedia recommenders directly incorporate ID embeddings and the pre-trained multimodal information (visual, acoustic, and textual) for users’ and items’ vector representations. For instance, VBPR(Visual Bayesian Personalized Ranking from Implicit Feedback) [11] devises an ID embedding and learnable multimodal preference vector for the user and then uses the inner-product of the ID embedding and the multimodal features of the item to predict the interaction between users and items. ACF [12] argues that the implicit feedback does not reflect the real user preference on the item or modality and introduces a novel attention mechanism of Collaborative Filters to address the challenging item- and component-level implicit feedback in the multimedia recommendation. The previous development of multimedia recommenders obviously enriches the items’ representation and model users’ multimodal preferences by incorporating multimodal information. However, the learning effect still largely depends upon the user-item interaction. It does not
utilize the high-order connectivities and in fact, the message pass mechanism could better enhance the representations of users and items. [13], [14].

Recent research efforts have approved the reasonability of the graph-based personalized recommendation approaches [15]–[22]. Typically, the graph-based approaches adopt the multilayer graph convolutional network (GCN) [23] to enrich the users’ or items’ representations model along with the high-order connectivities. In the process of the GCN operation, the neighbors are used to enrich the current user or item’s representation, which is similar to item-level aggregation in ACF (Attentive collaborative filtering). After multiple operations, the high-hop users and items’ information will also be injected into representations of the current users and items. In graph-based multimedia recommendation, GraphCAR (Content-aware Multimedia Recommendation with Graph Autoencoder) [42] propose a content-aware multimedia recommendation model with graph autoencoder, which directly adopts raw modal features of items. MMGCN (Graph Convolution Network for Personalized Recommendation of Micro video) [18] devised a parallel graph network, in which users’ modal preferences are modeled on three separated child modal graph networks (textual, acoustic, and visual). DualGNN (Dual Graph Neural Network for Multimedia Recommendation) [43] and HUIGN (Hierarchical User Intent Graph Network for Multimedia Recommendation) [10] respectively adopt user co-occurrence graph and item co-interact graph to model user preference on different modalities. However, similar to MMGCN, the preference of users are modeled in several graph networks. MGAT [17] argues that the structure of MMGCN is very complex and neglects the different importance of the passed messages. Instead, a lighter parallel attentive gated graph network should be reconstructed.

Though all the aforementioned work has achieved satisfactory performance, we argue that the graph-based multimedia recommenders in which the same bipartite graph for each modality is constructed, are not reasonable enough after a detailed analysis of this two papers [46], [47]. We agree that the interaction between users and items indicates the user’s preference for the entire item’s modalities. However, inside each item, users’ preferences on different modalities are unknown, and it does not mean that the user is interested in every single modality. Hence, the bipartite graph for each modality may be different from the entire bipartite graph. For example, as shown in Fig. 1, the user prefers visual and acoustic background music while seems not interested in the textual content.

This work aims to learn the representations of users and items with multimodal information by GCN. To avoid the drawback of the unreasonable graph structure of each modality, we adopt a single bipartite graph. For a better representation, a fusion component is proposed to integrate and distill the useful multimodal information, and the attention mechanism is introduced to weigh the importance of different neighbors.

We devise a new multimedia recommendation framework, a Multimedia Graph Collaborative Filter (aka MGCF). In MGCF, the multimodal information is first fused with the item’s ID embedding by the fusion component, and the multimodal information of high-hop items will then be propagated along with the high-order connectivities in the graph and aggregated to enrich the representations of the users and the items. Meanwhile, the attention mechanism is adopted to enhance the importance of aggregated information. Extensive experiments are conducted on two public datasets: Tiktok and MovieLens. The results show that our model outperforms several state-of-art multimedia recommenders. Further analysis discloses the importance of modeling multimodal collaborative filter signals for better user and item representations. Our key contributions are summarized as follows:

- We devise a generic light framework for the multimedia recommendation referred to as MGCF. It adopts a single graph that reflects the user’s actual preference, to learn the representations of users and items.
- We design a fusion component to integrate and distill the useful multimodal information. The attention mechanism is introduced to weigh the importance of different neighbors.
- We perform extensive experiments on two datasets to verify the rationality and effectiveness of MGCF. We will release the code and parameter settings upon acceptance.

II. RELATED WORK
In this section, we introduce some work related to our method, including multimodal personalized recommendation and graph convolution networks.
A. MULTIMODAL PERSONALIZED RECOMMENDATION

In modern recommendation systems, collaborative filtering is usually used for the recommendation of personalized tasks. The CF-based method [1]–[3], [24], [37], [38] captures users’ historical feedback and learns low dimensional vectors to represent users’ preferences. However, when the CF-based method encounters excessively sparse interaction, it will lead to poor performance. To solve this problem, researchers developed a hybrid method to introduce multimodal information into feedback and representation [11], [39], [40], [46], [47]. VBPR proposes an extensible factor decomposition model, which integrates visual modality features into the user’s preference representation to predict user preferences. Then the researchers tried to combine the visual, acoustic, and textual modalities to represent the user’s comprehensive preferences by adopting the graph convolutional network.

GraphCAR [42] proposes a graph autoencoder to model user preference from multimedia content. MMGCN [18] and MGAT [17] combine the visual, acoustic, and textual modalities in three separated graph networks. DualGNN [43] adopts user-item bipartite graph and user co-occurrence graph to improve the representation of users on different modalities. Different from DualGNN, HUIGN [10], [43] learns multi-level user preference on different modalities from co-interacted items.

B. GRAPH CONVOLUTION NETWORKS

Graph convolution networks have been widely used in various applications due to their effectiveness and simplicity. With convolutional graph operations, the local structure information of nodes can be encoded into its representation through message passing and aggregation mechanisms. The graph-based recommendation method has also been widely used. GraphCAR [42] uses an encoder based on graph networks to combine user-item interaction with user attributes and multimedia content. Neural Graph Collaborative Filtering (NGCF) [13] embeds the user-item graph into the collaborative filtering framework based on graph neural network and enriches the embedded data with its high-order connectivity to obtain collaborative signals. Light Graph Convolution Network (LightGCN) [25] simplifies the NGCF framework and discards redundant parameters, making the model lighter, easier to train, and more effective. The multimedia content contains rich multimodal information and is utilized in many fields [41]. The Multimodal Graph Convolution Network (MMGCN) generates user and project representations in specific modes so that the characteristic representations of user preferences can be modeled more clearly. The Multimodal Graph Attention Network (MGAT) introduces a gated attention mechanism based on MMGCN to distinguish user preferences under different modes to capture hidden interaction modes. DualGNN and HUIGN respectively adopt user co-occurrence graph and item co-interact graph to model user preference on different modalities. GRCN [44] introduces prototype learning to denoises the information from the neighbors and refine the structure of interaction graph.

III. PRELIMINARIES

In this article, we use italic capital letters to represent sets (e.g., $\mathcal{V}$). Besides, we use $id$ to represent the user preference and $m \in \mathcal{M} = \{v, a, t\}$ as the modality indicator, where $v, a$, and $t$ represent visual, acoustic, and textual modalities. Moreover, we use $N_u$ to represent the item that has interaction with the user $u$. Similarly, $N_i$ is the representation of the user who has interaction with the item $i$. Some key concepts are introduced as follows.

A. USER-ITEM BIPARTITE GRAPH

We set the historical interaction between user and item as an edge on a graph, so we get a bipartite graph $G = (\mathcal{V}, \mathcal{E})$ to represent them all, where the node-set $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$, and $\mathcal{E} = \{(u, i) | u \in \mathcal{U}, i \in \mathcal{I}\}$ is the edge set, to represent the interaction between user $u$ and item $i$. It is worth noting that we formalized $G$ as an undirected graph.

B. EMBEDDING

The primary purpose of the embedding layer is to map different sparse high-dimensional vectors to lower dense vectors of the same dimension, thus realizing the operation of the relationship between various features and it can also prevent excessive space occupation caused by excessive sparsity. For example, we have $1 \times d$ as the raw feature vector, and through the embedding layer $(d, d')$, in which $d'$ is much smaller than $d$, we receive a $1 \times d'$ vector as the final feature representation.

C. HIGH-ORDER CONNECTIVITY

User-item connectivity can be formulated as the relationship hidden in the user behavior data. Knowing that the user’s first-order connectivity consists of his/her historical interacting items, which reflects the user’s preference directly, we can use the user group of the item to describe the item features. Then we obtain the high-order connections among nodes by conducting random walks on the graph. For example, the path $u_1 - i_1 - u_2 - i_4$ indicates that $u_1$ might have a similar preference with $u_2$ due to their common choice of adopting $i_1$; in addition, we can recommend $i_4$ to $u_1$ according to the mechanism of collaborative filtering.

D. AGGREGATION MECHANISM

After $l$ layers information propagation, each node catches the information within its $l − hop$ neighbors through the aggregation mechanism. The existing aggregation mechanism includes average-pooling, max-pooling, and attention mechanisms.

IV. METHODOLOGY

In this section, we will present the MGCF model. As illustrated in Fig. 2, the overall model consists of three components, namely the fusion layer, the aggregation layer, and the
prediction layer. The fusion layer proposes the ID embedding of user and item and fusion component to model user preferences on special multimodal information, which users prefer. By applying the attention and message-passing mechanism in the aggregation layer, we integrate multimodal information from neighbors to enhance the representations of the users and items. Lastly, the prediction layer estimates the affinity scores of a user-item pair based on the final representation.

A. FUSION LAYER

1) EMBEDDING

In the mainstream recommender models [3], [13], [18], [24], [25], the user and the item are usually projected as an embedding vector $e_u \in \mathbb{R}^d$ and $e_i \in \mathbb{R}^d$ through their unique ID information, where $d$ denotes the embedding size. [26], [27] Moreover, beyond the ID information, each item $i$ has particular characteristics in different modalities. For simplicity, we use $e_{mi} \in \mathbb{R}^d$ to denote the features in the $m$-th modality of item $i$, where $m \in M$ represent the visual, acoustic, and textual modality. All of these embeddings are as follows:

$$E = [e_u, e_i, e_{mi}|u \in \mathcal{U}, i \in \mathcal{I}, m \in \mathcal{M}] \quad (1)$$

2) MODALITY FUSION

Following the prior recommender models on multimodal representation, we first project $e_{mi}$ onto the same latent space as $e_i$ to quantify the influence of each modality:

$$\hat{e}_{mi} = \tanh(W_{1,m}e_{mi}) \quad (2)$$

where $W_{1,m} \in \mathbb{R}^{d \times d_a}$ is a trainable transformation matrix to obtain the features of each modality in ID embedding space, and $d$ is the transformation size. The representation of different modalities hence can be compared and combined in the same latent space.

Intuitively, the features of each modality are separately fed into a parallel aggregation layer to obtain the representation of users/items. In contrast, we adopt in the MGCF framework a fusion function $f(\cdot)$ to integrate the different modal information and item information into a unified hyperplane before feeding it to a single aggregation layer, which leads to a lighter framework and models user preferences on the special part of the entire item. Different modality fusion methods may have different effects on the performance of our model. We implement $f(\cdot)$ via:

- **Average Fusion**: Basically, we first get the three modality features’ representation in the same space from the MLP layer and then use the additive average to represent the multimodal feature:

$$f_{avg}(m, i) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \hat{e}_{mi} + e_i \quad (3)$$

- **Raw Fusion**: We concatenate the raw modality features first and then pass the result through an MLP layer to represent the multimodal feature:

$$f_{rf}(m, i) = \tanh(\hat{W}_2(e_{vi} \parallel e_{ai} \parallel e_{ri})) + e_i \quad (4)$$

where $\parallel$ is the concatenation operation and $\hat{W}_2 \in \mathbb{R}^{d \times (d_a + d_b + d)}$ is the transformation matrix.

- **Align Fusion**: We try to align the three modal features through the MLP layer first and concatenate them through an MLP layer to represent the multimodal feature:

$$f_{af}(m, i) = W_2(\hat{e}_{vi} \parallel \hat{e}_{ai} \parallel \hat{e}_{ri}) + e_i \quad (5)$$

where $\parallel$ is a concatenation operation and $W_2 \in \mathbb{R}^{d \times 3d_a}$ is the transformation matrix.

The impacts of these modality fusion methods on the MGCF performance will be discussed in the next section.

B. AGGREGATION LAYER

Unlike other recommendation models, graph-based methods with multimodal exploit the user-item interaction graph and stack multiple aggregation layers for each modality to refine the representation. To lighten these graph-based models and have better performance, we devised a light bipartite graph with the aggregated collaborative filter signal.

1) INFORMATION AGGREGATION

In MGCF, we adopt an aggregated representation instead of the neighbor’s original feature by the attention mechanism. For an user (or item) node in a bipartite graph $G$, we formalize the representation being propagated from its neighbors as follows:

$$e_{Nu} = \text{LeakyReLU}(\sum_{i \in N_u} W_3((f_u(u, i)e_i)e_u)) \quad (6)$$

where $N_u \{i(u, i) \in G\}$ denotes the neighbors of the target node $u$; $W_3 \in \mathbb{R}^{d' \times d}$ is the trainable weight matrix to distill useful knowledge, where $d'$ is the size of this weight matrix; $e_u \in \mathbb{R}^d$ and $e_i \in \mathbb{R}^d$ is the $d$-dimensional representation of the target node $u$ and the connected nodes with multimodal information; $f_u(u, i)$ is the attention score indicating the contributions of neighbors nodes. In contrast with aggregation only connected neighbors in [25], we aggregate the target node and its neighbors by the element inner product. The attention mechanism, to be introduced in the section, warrants the effectiveness of information from neighbor nodes.

2) NEIGHBOR-AWARE ATTENTION

There are several methods of attention mechanism existing [17], [21], [28]–[30] in Graph attention networks. Inspired by the attention mechanism of MGAT, we incorporate the attention mechanism (without the gate mechanism) into our model to learn the varying weight of each neighbor. We implement $f_u(\cdot)$ using the following two types of the attention mechanisms:

- **Concatenation Attention** which concatenates each target node and its neighbors:

$$f_{a,c} = \frac{\exp(\text{LeakyReLU}(\hat{W}_4(e_u \parallel \hat{e}_r)))}{\sum_{e \in N_u} \exp(\text{LeakyReLU}(\hat{W}_4(e_u \parallel \hat{e}_r)))} \quad (7)$$
where $\hat{W}_4 \in \mathbb{R}^{d \times 2d}$ is a trainable weight matrix, $\|\|$ is the concatenation operation, and LeakyReLU($\cdot$) is used as a nonlinear activation function.

- **Inner Product Attention** which conducts an inner product operation between each target node and its neighbors:

$$ f_{u,i,p} = \frac{\exp(\text{LeakyReLU}(W_4(e_u \odot \hat{e}_l))))}{\sum_{l \in N_u} \exp(\text{LeakyReLU}(W_4(e_u \odot \hat{e}_l)))} \quad (8) $$

where $W_4 \in \mathbb{R}^{d \times d}$ is a trainable weight matrix; $\odot$ is the inner product operation.

The final attention scores are used to aggregate the neighbor node representation to enhance the performance on information propagation. The impact of these attention mechanisms on the MGCF performance will be discussed in section IV.

C. PREDICTION LAYER

In this section, we stack more aggregation layers to explore the higher-order connectivity among nodes in the bipartite user-item graph. More formally, the representation for the $l$-th aggregation layer is recursively formulated as:

$$ e_u^{(l)} = \text{LeakyReLU}(W_3 e_{N_u}^{(l-1)} + e_u^{(l-1)}) \quad (9) $$

where $e_u^{(l-1)}$ is the representation from the previous $(l-1)$-hop aggregation layers, containing the information propagated $(l-1)$ steps, and $e_u^{(0)}$ is initialized embedding. As a result, $e_u^{(l)}$ characterizes the user preferences on different modal features of items, which is obtained from a light neural network.

We carry out the same operation on items. After stacking $L$ aggregation layers, we produce the representations of each node as follows:

$$ e_u^* = e_u^0 \parallel \cdots \parallel e_u^L \quad (10) $$

Finally, we conduct the inner product of the user and item representation to estimate the user’s preference towards the target item. The user’s preference is formulated as:

$$ y_{u,i} = e_u^* e_i^T \quad (12) $$

D. OPTIMIZATION

Among the mainstream optimization methods [31]–[36], we choose Bayesian Personalized Ranking (BPR), which is a well-known pairwise ranking optimization framework, to optimize the model parameters. We formulate this as follows:

$$ L = \sum_{(u,i,j) \in \mathcal{O}} -\ln(\delta(y_{ui} - y_{uj})) + \lambda \|\theta\|_2^2 \quad (13) $$

where $\mathcal{O} = \{(u,i,j) | (u,i) \in R^+, (u,j) \in R^- \}$ is the training dataset; $R^+$ is the dataset that contains observed interaction between user $u$ and item $i$, while $R^-$ is the unobserved interactions; $\delta()$ is the sigmoid function; $\lambda$ is the decay factor and $\theta$ is the parameters used in the model.

E. COMPLEXITY ANALYSIS

In this section, we analyze the complexity of the MGCF, and compare it with MMGCN and MGAT. Suppose the number of nodes and edges in the user-item bipartite graph are $|\mathcal{V}|$ and $|\mathcal{E}|$, respectively. Let $\mathcal{T}$ denotes the number of triplets in the training set. The complexity mainly comes from two parts of the model:

- **Graph Convolution:** The complexity of graph convolution of MMGCN is $O(3L|\mathcal{E}|d + 3L|\mathcal{V}|d^2)$, since MMGCN constructs three sub-graphs for different modalities and uses the feature transformation in graph convolution. MGAT not only uses three sub-graphs...
like MMGCN, but also conducts the gate and attention mechanism in each graph, the complexity of MGAT is $O(3L|E|d^3 + 3L|V|d^3)$. The proposed MGCF is lighter than MGAT since it only uses one graph and an attention mechanism, the complexity is $O(L|E|d^2 + L|V|d^2)$.

- **BPR Loss**: The inner product is the only option conducted in the prediction layer for all three models. The time cost of a training epoch is $O(Td)$.

### V. EXPERIMENTS

In this section, we will detail our experiments, including experimental settings, performance comparisons, case studies of MGCF, and complexity analysis.

#### A. EXPERIMENT SETTINGS

1) **DATASETS**

As the micro-video contains multi-modalities — frames, soundtracks, and descriptions [17], [18], following MGAT, we conduct several experiments on the two publicly accessible designed for the micro-video recommendation, namely MovieLens and TikTok, to evaluate the performance of MGCF. More importantly, each of these datasets contains rich user-item interaction records and plenty of multimodal features, in which all the multimodal features are extracted as feature vectors by pre-trained deep neural networks, such as ResNet, VGGish, and Sentence2Vector. The details of these datasets are summarized in Table 2.

2) **BASELINES**

To evaluate our proposed model, we compared it with several state-of-the-art baselines. We briefly divided these baselines into three groups: 1) deep-learning-based methods (i.e., VBPR, ACF), 2) pure graph-based collaborative filtering methods (i.e., NGCF, LightGCN, GraphSAGE) and 3) multi-media algorithms (i.e., GraphCAR, DualGNN, GRCN, HUIGN, MMGCN, MGAT).

- **VBPR** [11]: It adds visual features into the item’s representation and then uses Matrix Factorization to learn the representation of users and items based on their previous interactions. In our model, we fuse the multimodal features of the micro video into one feature vector and then integrate it with user-ID to predict the interaction between users and items.
- **ACF** [12]: This method uses item-level attention to deal with the implicit feedback of media recommendations. We then use a similar attention mechanism to distinguish preferences in different modalities.
- **GraphCAR** [42]: It uses a graph autoencoder to combine informative multimedia content with user-item interaction. The preference scores for each user are generated from user-item interaction, user attributes, and multimedia content.
- **GraphSAGE** [45]: This is a graph-based model in which the hidden data is represented by the information aggregated from neighbor nodes. In our work, we combine all three modality features with each node to gain their complete representation.
- **NGCF** [13]: The NGCF method integrates collaborative signals into embedding explicitly. Then it merges the information that passes from multiple levels of neighbors to represent the high-order feature interactions in a graph. In our experiment, we use the combination of the multimodal features and the item-ID as the item representation, based on the process of embedding in the NGCF framework.
- **LightGCN** [25]: The LightGCN framework simplifies the NGCF framework by discarding the redundant parameters, making the model lighter, easier to train, and more effective. Based on its lightweight design, we add multimodal features and fuse them to improve our model.
- **MMGCN** [18]: This model enriches the user-side representation by constructing a bipartite graph based on the user-item interactions for each modality and then performs multi-modal information merging for users and items, respectively. To simplify this method, we merely fuse the user’s preference representation and the multi-modal representation of the item into one graph.
- **DualGNN** [43]: This is a GCN-based multi-media recommendation method, which adopts a user-item bipartite graph and user co-occurrence graph to learn user preference on different modalities.
- **HUIGN** [10]: This model learns the multi-level user intents from the co-interacted items to enrich the representation of users and items. By graph convolution operations, it models coarser-grained user intents.
- **GRCN** [44]: This is a GCN-based framework, which can adaptively discover and prune noisy edges in the user-item bipartite graph. By introducing a prototypical network, it models the user preference from meaningful nodes.
- **MGAT** [17]: It uses a gated mechanism and the attention mechanism to distinguish user preferences among different modes. We use a similar attention mechanism to distinguish the importance of different modes.

3) **EVALUATION METRICS AND PARAMETER SETTINGS**

We divided each dataset randomly into three portions as the training, validation, and test sets, with a number ratio of 8:1:1. The validation set is used to discover the optimal hyperparameters for performance evaluation in our experiments. We then use the widely-used metrics: Precision@K, Recall@K, and NDCG@K to measure the top-K performance. These are commonly used evaluation indicators in the field of recommendation systems, in which Precision@K stands for the proportion of correctly predicted relevant results in all returned results, Recall@K stands for the proportion of correctly predicted relevant results in all relevant results, and NDCG@K is an evaluation index considering the return order. For all models, We set $K=10$ and then report the average scores in the test set. We randomly initialize the parameters with the
Gaussian distribution, use the LeakyReLU as the activation function, and optimize the model with SGD to train our model. We set the batch size as 1024, the latent feature dimension $[32, 64, 128, 256]$ the learning rate $[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$, and the regularizer 0.2. The results are shown in Table 3-Table 6. The other baselines use the hyper-parameters as used in the original papers. To ensure fairness, we use multimodal features in all baselines.

### B. PERFORMANCE COMPARISON

We carry out a performance comparison between MGCF and the most advanced recommendation algorithms on the Tiktok and MovieLens datasets. The results are summarized in Table 3, from which we have the following observations:

- MGCF performs better than all the baseline models, which demonstrates that the design of our model is reasonable. Compared to VBPR and ACF, our model uses high-order connectivity to facilitate the representation learning process rather than the traditional collaborative filtering methods. Compared to the GNN-based models like GraphSAGE and NGCF, the improvement brought by MGCF is attributed to modality fusion and the attention mechanism, which are intended to capture the user preferences on the special part of the entire item effectively.

- The performance of the graph-based model is better than the CF-based model on both datasets. It proves that using message-passing to inject the information of the neighbors into the head node representation can improve the representation of the micro videos. Moreover, we find that VBPR performs better than GraphSAGE and NGCF with the MovieLens dataset. The reason primarily lies in the difference in data types: most videos in MovieLens are much longer than those in Tiktok and the raw features in long videos are much more complicated than in shorter ones. What’s more, the relationships among the modalities can be highly implicated in long videos.

- The performance of lightGCN is better than that of MGAT on MovieLens. The conclusion does not hold when it comes to Tiktok. It shows that the performance benefits from the lighter graph structure when the dataset is less sparse. MGCF no outperforms lightGCN, which shares the same bipartite graph. It proves that, compared to only ID embedding being used, an activate function is more necessary for multimodal features.

- The performance of lightGCN is better than that of MGAT on MovieLens. The conclusion does not hold when it comes to Tiktok. It shows that the performance benefits from the lighter graph structure when the dataset is less sparse. MGCF no outperforms lightGCN, which shares the same bipartite graph. It proves that, compared to only ID embedding being used, an activate function is more necessary for multimodal features.

- As we can see from the results, there is a gap between the performance of our proposed model and GraphCAR. The gap is caused by the fusion component. Instead of raw modal features, we integrate and distill the different modal features of items before feeding them into the user-item bipartite graph. Compared with the DualGNN, HUIGN, and GRCN, we introduce the attention mechanism to weigh the importance of different neighbors. The result shows that the performance of our proposed model is better.

### C. CASE STUDIES FOR MGCF

In this section, we present the case studies for MGCF to investigate the factors that may have effects on the model performance.
TABLE 4. Effect of the number of modalities.

|           | TikTok | MovieLens |
|-----------|--------|-----------|
| Model     | Precision | Recall | NDCG | Precision | Recall | NDCG |
| MGCF_v   | 0.1245  | 0.5940  | 0.3786 | 0.1291 | 0.5474 | 0.3302 |
| MGCF_a   | 0.1240  | 0.5912  | 0.3736 | 0.1247 | 0.5299 | 0.3144 |
| MGCF_t   | 0.1223  | 0.5855  | 0.3729 | 0.1210 | 0.5146 | 0.3055 |
| MGCF_u   | 0.1265  | 0.5938  | 0.3801 | 0.1300 | 0.5510 | 0.3316 |
| MGCF_cat | 0.1234  | 0.5837  | 0.3716 | 0.1294 | 0.5484 | 0.3314 |
| MGCF_vat | 0.1238  | 0.5908  | 0.3746 | 0.1291 | 0.5477 | 0.3288 |

TABLE 5. Effect of attention mechanism.

|           | TikTok | MovieLens |
|-----------|--------|-----------|
| Model     | Precision | Recall | NDCG | Precision | Recall | NDCG |
| MGCF_v   | 0.1202  | 0.5435  | 0.3445 | 0.1257 | 0.5334 | 0.3155 |
| MGCF_cat | 0.1277  | 0.6058  | 0.3834 | 0.1307 | 0.5552 | 0.3361 |
| MGCF_ip  | 0.1282  | 0.6079  | 0.3870 | 0.1332 | 0.5622 | 0.3415 |

TABLE 6. Effect of the modality combination mode.

|           | TikTok | MovieLens |
|-----------|--------|-----------|
| Model     | Precision | Recall | NDCG | Precision | Recall | NDCG |
| MGCF_avg | 0.1282  | 0.6079  | 0.3870 | 0.1332 | 0.5622 | 0.3415 |
| MGCF_tav | 0.1300  | 0.6143  | 0.3992 | 0.1331 | 0.5631 | 0.3439 |
| MGCF_stav | 0.1308  | 0.6152  | 0.4050 | 0.1334 | 0.5643 | 0.3457 |
| MGCF_al   | 0.1308  | 0.6197  | 0.3987 | 0.1342 | 0.5654 | 0.3448 |

1) EFFECT OF THE NUMBER OF MODALITIES
We designed six groups of experiments to measure the influence of the number of modalities on our model, including three single-modality and three dual-modality. The results are summarized in Table 4, from which we observe the following:

- In the single-modality comparison, the visual modality performs the best, the auditory modality performance is the second, and the text modality effect is the weakest, which is also in line with the common fact, that is, when users choose multimedia content, they mainly focus on the visual quality and sound effect of the media, while the text is often considered as auxiliary information and is given less attention. In conclusion, different performances in a single modality verify that users have different preferences for different modalities, so it is essential to capture information from all modalities, to gain better performance.

- Following the previous conclusion, the result of the dual-modality combination is also reasonable, and the combinations with visual modality are more dominant. However, this does not indicate that the text modality should not be taken into account since all the single-modality or dual-modality experimental results are not as good as the original MGCF model.

- The performance of three modalities is better than either dual-modality combination or single-modality, indicating that more modality information is beneficial to the MGCF performance. Furthermore, in all combinations, the model including the visual modality information performs better than any other without it, which reveals that vision information has a greater impact on user preferences.

2) EFFECT OF ATTENTION MECHANISM
We test the performance of the model with and without the attention mechanism. The comparison results are summarized in Table 5, and we have the following observations:

- Obviously, the attention mechanism makes a great deal of contribution to the model. Models using it work much better than those which do not use it, demonstrating the necessity and effectiveness of the attention mechanisms in GNN-based models.

- The MGCF_ip and MGCF_cat make use of the attention mechanism by the inner product and concatenation operation between $x_i$ and $x_j$, respectively. The result shows that the MGCF_ip performs slightly better than MGCF_cat and we argue that using more modality information and parameters may make it easier to overfit the model.

3) EFFECT OF THE MODALITY COMBINATION MODE
Several different experiments on modal combinatorial relationships are designed to test the effect of the combinatorial relationship on model accuracy. The comparison results are summarized in Table 6, the methodology and observations are as follows:

- We add up the features of the three modalities together to average and add the item embedding vector to represent the final features of the item. The outcome is not as good as we expected. Therefore, adding the modality features without nonlinear transformation is the first
choice of our model. However, this method may have some problems because it does not take into account the combinatorial relationships between different modalities and has no focus on any of them. This is also the reason that we introduce other fusion methods.

- From the statistics in the next two options, we can conclude that the combination of the item features may affect the outcome. In the Tiktok dataset, all the options that apply concatenation perform better than the average option. The raw option is almost equivalent to the eqnarray option, and the all option performs better than the raw option. We think that the combination with item ID before going into the MLP layer has a more comprehensive expression of the item features. In the meantime, in the MovieLens dataset, the results follow the patterns observed in the Tiktok dataset. The eqnarray option demonstrates its superiority over the raw option and the all option performs the best.

In conclusion, the fusion mechanism is necessary for enhancing the model performance. The raw option combines the features of different modalities into a whole representation and this is more reasonable than avg. The reason that the eqnarray option is better than the raw option is that the former takes the alignment information between modalities into account, which reflects the influence of the potential connection between modalities on our model, while the operation of simple concatenation might confuse the relationship between modality features. Finally, the all option shows that MLP and fusion have different emphases. Similar to the component-level attention in the ACF model, taking the item ID feature into MLP with the modality features in all option is equivalent to learning the representation of the item again.

VI. CONCLUSION

In this research, we proposed a multimedia recommendation model that mainly focuses on integrating multimedia information to model the representations of the user and items based on user-item interactions. Though the performance improves continuously, we argue that the latest state-of-art parallel graph framework with the same graph structure is undesirable since user preferences on different modalities are unknown upon implicit feedback. The entire learned repre-
sentations would be limited. We devised a light graph-based multimedia recommendation algorithm, MGCF. In MGCF, a fusion component is used to integrate and distill the useful multimodal information, which models better representations with multimodal information based on multiple GCN operations. Moreover, we adopted the attention mechanism to weigh the importance of different neighbors. We conducted extensive experiments on two benchmark datasets to justify the superiority of our proposed model. Empirical results showed that MGCF exhibited substantial improvements over the state-of-the-art baselines. In future work, we will try to exploit methods for capturing more precise user preferences on different modalities. Moreover, we will utilize causality to promote the recommendation performance for multimedia.

REFERENCES

[1] Y. Koren, “Factorization meets the neighborhood: A multifaceted collaborative filtering model,” in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Y. Li, B. Liu, and S. Sarawagi, Eds, Las Vegas, NV, USA, Aug. 2008, pp. 426–434.

[2] S. Rendle, “Factorization machines,” in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, Eds, IEEE Computer Society, Dec. 2010, pp. 995–1000.

[3] S. Rendle, J. Herlocker, L. Laﬀerty, and J. Konstan, “The netﬂix prize problem formulation,” in *Proc. 25th ACM SIGKDD Conf. Data Mining (SIGKDD)*, San Diego, CA, USA, Aug. 2019, pp. 1–9.

[4] T. Joachims, “Learning to rank from example queries: An analysis of margin-based ranking models,” in *Proc. 24th ACM SIGKDD Conf. Data Mining (SIGKDD)*, San Diego, CA, USA, Aug. 2018, pp. 1–10.

[5] N. Dalvi and D. Suciu, “On the expressiveness of query languages for information retrieval,” in *Proc. 14th ACM SIGMOD Conf. Management Data (SIGMOD)*, New York, NY, USA, Jun. 2005, pp. 590–601.

[6] J. B. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, J. Ramapuram, J. Togelius, and M. Blanke, “Graph neural networks for graphs,” in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, T. Wang, R. Ahmed, and L. Zettlemoyer, Eds., New Orleans, LA, USA, Jun. 2017, pp. 1–20.

[7] S. Wang, J. Qiu, Q. Meng, Y. Dong, J. Zhang, H. Yang, M. Ding, K. Wang, and J. Tang, “GCG: Graph contrastive coding for graph neural network pre-training,” in *Proc. 26th ACM Conf. Knowl. Discovery Data Mining, KDD*, Beijing, China, Jul. 2020, pp. 1101–1110.

[8] 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, Sep. 2020, Art. no. 102277.

[9] Y. Wei, X. Wang, L. Nie, X. He, and R. Hong, “MMPGCN: Multi-modal graph convolution network for personalized recommendation of micro-video,” in *Proc. 27th ACM Int. Conf. Multimedia*, L. Amsaleg, B. He, M. A. Larson, G. Nevatia, H. Huang, C. Ngo, and W. T. Ooi, Eds, Nice, France, Oct. 2019, pp. 1437–1445.

[10] F. Moni, M. M. Bronstein, and X. Bresson, “Geometric matrix completion with recurrent multi-graph neural networks,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3697–3707.

[11] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, “Gated graph sequence neural networks,” in *Proc. 4th Int. Conf. Learn. Represent.*, Y. Bengio and J. Yamins, Eds., San Juan, Puerto Rico, 2016.

[12] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, “Gated graph sequence neural networks,” in *Proc. 4th Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds, San Juan, Puerto Rico, May 2016.

[13] M. Niepert, M. Ahmed, and K. Kutzkov, “Learning convolutional neural networks for graphs,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2014–2023.

[14] Y. Jiang, H. Ma, Y. Liu, and Z. Li, “Attentional social recommendation system with graph convolutional network,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8, doi: 10.1109/IJCNN52387.2021.9534075.

[15] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. Chua, “Neural collaborative filtering,” in *Proc. 26th Int. Conf. World Wide Web*, R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, Eds, Perth, WA, Australia, Apr. 2017, pp. 173–182.

[16] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, “LightGCN: Simplifying and powering graph convolution network for recommendation,” in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds, Virtual Event, China, Jul. 2020, pp. 639–648.

[17] L. Wang, Y. Li, and S. Lazebnik, “Learning deep structure-preserving image-text embeddings,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 191–198.

[18] Y. He, 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, Jun. 11, 2021, doi: 10.1109/TMM.2021.3088307.

[19] R. He and J. J. McAuley, “VBPR: Visual Bayesian personalized ranking from implicit feedback,” in *Proc. 50th Conf. Artif. Intell.*, D. Schuurmans and S. Teller, Eds, Phoenix, AR, USA: AAAI Press, Feb. 2016, pp. 144–150.

[20] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, “Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention,” in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White, Eds, Tokyo, Japan, Aug. 2017, pp. 335–344.

[21] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, “Neural graph collaborative filtering,” in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, and F. Scholer, Eds, Paris, France, Jul. 2019, pp. 165–174.

[22] W. L. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds, Long Beach, CA, USA, 2017, pp. 1024–1034.
[36] Y. Tay, L. Anh Tuan, and S. C. Hui, “Latent relational metric learning via memory-based attention for collaborative ranking,” in Proc. World Wide Web Conf. World Wide Web - WWW, 2018, pp. 729–739.

[37] O. Barkan, N. Koenigstein, E. Yogev, and O. Katz, “CB2CF: A neural multiview content-to-collaborative filtering model for completely cold item recommendations,” in Proc. 13th ACM Conf. Recommender Syst., T. Bogers, A. Said, P. Brusilovsky, and D. Tikk, Eds. Copenhagen, Denmark, Sep. 2019, pp. 228–236.

[38] R. Mu, “A survey of recommender systems based on deep learning,” IEEE Access, vol. 6, pp. 69009–69022, 2018, doi: 10.1109/ACCESS.2018.2880197.

[39] X. Wang, X. He, Y. Cao, M. Liu, and T. Chua, “KGAT: Knowledge graph attention network for recommendation,” in Proc. 25th ACM Int. Conf. Knowl. Discovery Data Mining, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, Eds. Anchorage, AK, USA, 2019, pp. 950–958.

[40] S. M. Al-Ghuribi and S. A. Mohd Noah, “Multi-criteria review-based recommender system–The state of the art,” IEEE Access, vol. 7, pp. 169446–169468, 2019, doi: 10.1109/ACCESS.2019.2954861.

[41] X.-Y. Du, Y. Yang, L. Yang, F.-M. Shen, Z.-G. Qin, and J.-H. Tang, “Captioning videos using large-scale image corpus,” J. Comput. Sci. Technol., vol. 32, no. 3, pp. 480–493, May 2017.

[42] Q. Xu, F. Shen, L. Liu, and H. T. Shen, “GraphCAR: Content-aware multimedia recommendation with graph autoencoder,” in Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Jun. 2018, pp. 981–984.

[43] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, “DualGNN: Dual graph neural network for multimedia recommendation,” IEEE Trans. Multimedia, early access, Dec. 24, 2022, doi: 10.1109/TMM.2021.3138298.

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

[45] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 1–11.

[46] W. Guan, H. Wen, X. Song, C.-H. Yeh, X. Chang, and L. Nie, “Multimodal compatibility modeling via exploring the consistent and complementary correlations,” in Proc. 29th ACM Int. Conf. Multimedia, Oct. 2021, pp. 2299–2307.

[47] W. Guan, X. Song, T. Gan, J. Lin, X. Chang, and L. Nie, “Cooperation learning from multiple social networks: Consistent and complementary perspectives,” IEEE Trans. Cybern., vol. 51, no. 9, pp. 4501–4514, Sep. 2021.