Knowledge-Based Systems 311 (2025) 113035

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Graph attention contrastive learning with missing modality for multimodal recommendation

Wenqian Zhao, Kai Yang *, Peijin Ding, Ce Na, Wen Li

College of Information Engineering, Yangzhou University, Yangzhou 225127, Jiangsu, China

ARTICLE INFO

Keywords:
Multimodal recommendation
Missing modality
Contrastive learning
Graph neural network

ABSTRACT

Multimodal recommendation plays an important role in many online content-sharing platforms. Most existing reported approaches of multimodal recommendation employ user-interaction graphs or auxiliary graphs (e.g., user–user or item–item relation graphs) to augment user and/or item representations. However, real-world data suffer the problem of missing modality which affects recommendation performance. In this paper, we propose the Graph Attention Contrastive Learning with Missing Modality (MMGACL) model utilizing modality complementation and modality fusion of modality-aware user–item graphs to enhance recommendations. In particular, we construct user–item bipartite graphs for each modality and extract item subgraphs, leveraging contextual information to enhance item representations. Thereafter, we employ a bimodal attention mechanism to provide complementary information across modalities and fuse different modalities. The fused item representations are combined with user–item interactions to complement user information. Finally, we perform graph contrastive learning on the completed global graph to maximize mutual information between users and items and learn more accurate embedding representations. Extensive experiments on four benchmark datasets demonstrate the effective performance of our proposed model versus several state-of-the-art methods in scenarios with missing modality.

1. Introduction

In multimodal recommendation, multimodal data such as acoustic, visual, and textual features of items are involved, which can represent user preferences at a fine-grained modality level. Traditional recommendation methods [1–4] primarily rely on historical user–item interactions to model user preferences and generate recommendation lists. Exploration for multimodal information of items remains insufficient, which can adversely affect the recommendation performance. Aiming at enhancing the recommendation accuracy, recent studies on multimodal recommendation have integrated multimodal information of items into the traditional user–item recommendation paradigm to comprehensively represent item attributes and user interests. In particular, methods such as He et al. [5] concatenate the potential representation of items with multimodal features, while Liu et al. [6] and Chen et al. [7] use attention mechanisms to capture item representations by exploring user's preferences for multimodal features. With the development of research on graph-structured data, Wei et al. [8], Wang et al. [9], Li et al. [10], et al. employ graph neural networks (GNNs) to capture multimodal information of items. Wang et al. [9] performs graph convolution operations on single-modality user–item graphs to learn user preferences for each modality. Fang et al. [11] simultaneously processes multimodal data and semantic information under a unified framework while introducing a projection mechanism to effectively handle multiple modalities and additional semantic information. Furthermore, some methods have been proposed to enhance user and item representations via auxiliary graphs (such as user–user relation graphs or item–item relation graphs) and have delivered good recommendation performances. For example, Zhang et al. [12] extracts item representations by mining latent item–item semantic relation graphs from multimodal features. Khelloufi et al. [13] learns item–item structures and aggregates various modalities to obtain potential item graphs.

However, the abovementioned methods rely on the fundamental assumption that the multimodal data are sufficiently ideal to provide sufficient information (Fig. 1(a)). Unfortunately, this situation is often invalid in real-world scenarios. The modality information in the actual user–item graph may be missing because of privacy concerns and human errors in the data collection process (Fig. 1(b)). In extreme cases, not only the modality information of items is incomplete but also the user information may be unavailable. Researchers have proposed

* Corresponding author.
E-mail addresses: zhaowq0119@163.com (W. Zhao), yangk@fudan.edu.cn (K. Yang), isdingpeijin@163.com (P. Ding), nace1127@163.com (C. Na), evakattylee@gmail.com (W. Li).

https://doi.org/10.1016/j.knosys.2025.113035
Received 23 July 2024; Received in revised form 9 December 2024; Accepted 16 January 2025
Available online 31 January 2025

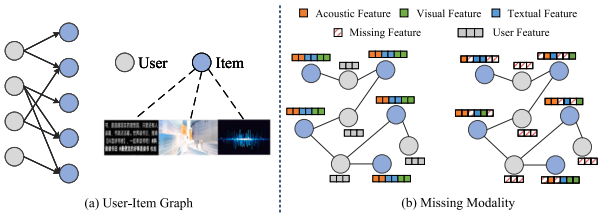


Fig. 1. Illustration of user–item graph and missing modality.

various methods to solve the missing modality problem. Ding et al. [14] integrates modalities on the basis of sensitivity to various tumor regions and develops a segmentation-based regularizer to address the problem of imbalanced training. Xu et al. [15] employed contrastive learning and knowledge distillation to capture semantic consistency between text–image pairs, enhancing the model's generalization capabilities for missing modalities within individual items. Wang et al. [16] exploited a score-based diffusion model to recover missing modalities for conforming to the original distribution and embedded available modalities to guide and refine the diffusion-based recovery process. These methods tackled the issue of missing modalities from various perspectives. However, very few existing approaches involved handling of missing modalities in multimodal recommendation tasks. For instance, Wang et al. [17] introduced modalities dropout and a multimodal sequential autoencoder to learn multimodal representations for complementation and inputting of missing modalities. Lin et al. [18] developed a contrastive intramodality and intermodality generation mechanism for missing modalities. However, these approaches did not consider the cases wherein item modalities and user information were simultaneously missing.

In this paper, we propose a modality-completion model that combines contrastive learning and attention mechanisms, referred to as MMGACL. We first construct modality-aware user–item graphs for each modality. We extract item subgraphs by considering the associations between items belonging to the same user and perform item complementation in the context of a single modality within subgraphs. Furthermore, different modalities can provide complementary information (e.g., the text modality (e.g., dialogue) can aid the acoustic modality in a video). Therefore, we employ bimodal attention aggregation to capture the intermodality associations and extract crucial information to obtain final item embeddings. These item embeddings are used to complete user information via user–item interactions. To prevent the complemented user and item representations from being excessive similar, we introduce contrastive learning to maximize the mutual information between users and items. Finally, the representations obtained are used for the recommendation task. The contributions of this study can be summarized as follows:

- We propose a novel graph attention contrastive learning with missing modality (MMGACL) model to enhance multimodal recommendations. In particular, we suppose an extreme scenario wherein item modalities and user information are simultaneously missing in multimodal data.
- MMGACL performs diffusion-based modality complementation and bimodal attention aggregation for complementation of intramodality and intermodality information, respectively. The missing information is effectively reconstructed via the exploitation of neighbor information and cross-modality associations.
- To prevent the complemented user and item features from becoming excessively similar, MMGACL introduces contrastive learning to maximize the mutual information between them so as to obtain more accurate representations.
- We conduct extensive experiments demonstrating that MMGACL outperforms baseline methods in multimodal recommendation with missing modalities.

2. Related works

In this section, we briefly review various works related to multimodal recommendation and further examine some studies on missing modalities in multimodal learning.

2.1. Multimodal recommendation

Collaborative filtering (CF) is a classical recommendation algorithm in many fields both in industry and academia. Many multimodal recommendation models use deep learning techniques to explore user preferences on the basis of the CF paradigm. He et al. [5] incorporates visual features extracted by pre-trained deep networks into user opinion prediction, and it connects latent visual features with item ID embeddings as the final representation of items. Aiming at further capturing of user preferences toward items, Liu et al. [6] employs attention neural networks to evaluate users' special attention from different aspects of items, following which it integrates the obtained attention into metric-based learning methods. Lin et al. [19] introduces a user's/item's neighbor from both the graph structure and semantic space, merging such latent neighbors to perform contrastive learning. Chen et al. [20] designs two types of contrastive learning that integrate long- and short-term user preferences via implementation of personalized travel recommendation and keywords generation tasks for better learning of user and item representations.

The development of GNNs had provided another research avenue for multimodal recommendation. GNNs learn additional information from the structure of user–item interaction graphs and auxiliary graphs to enhance user and item representations. Wei et al. [8] utilizes user–item interactions in each modality to guide representation learning. Wei et al. [21] introduces a graph refining layer to identify and prune noisy edges, adapting the structure of interaction graphs on the basis of the model's training status. Li et al. [10] trains a specialized GNN on each single-modality user–item bipartite graph to obtain single-modality embeddings, which are used as mutual supervision signals to reveal and synchronize the latent semantic relations across various modalities. In addition, Zhang et al. [22] and Khelloufi et al. [13] leverage item–item graphs to capture latent item features, further enhancing recommendation performance.

Item information used by these methods is generally complete and available, while our MMGACL is applied in scenarios with missing modalities, thus being more consistent with real-world data.

2.2. Missing modality in multimodal learning

Researchers have developed various methods to address the issue of missing modalities in multimodal learning. Havaei et al. [23] learns embeddings of multimodal information by computing the mean and variance of features from any number of available modalities. Dorent et al. [24] extends the [23] model via generating pixel-level classification based on mean and variance features. By leveraging their powerful feature representation capabilities, deep learning-based methods can better estimate missing modalities. Tran et al. [25] employs a cascaded residual autoencoder to recover missing modalities. Wang et al. [26], Xu et al. [15], and Wei et al. [27] combine knowledge distillation for reconstruction of missing modality information. In particular, Wang et al. [26] employs knowledge distillation to explore the importance of various modalities and extract knowledge from them to address the missing modality problem in a cross-modality manner. Lian et al. [28] uses GNNs to reconstruct missing modalities. Duan et al. [29] employs transformer-based progressive hierarchical fusion to thoroughly explore multimodal interaction information and align various modalities in both the embedding space and probability space.

However, these methods focus on the missing modalities of items, without consideration of completeness of user information. Meanwhile,

few studies have addressed the issue of missing modalities in multimodal recommendations. Our MMGACL not only takes into account missing item modalities but also considers missing user information. In addition, we evaluate its performance in the multimodal recommendation task.

3. Preliminaries

Definition 1 (*Multimodal user–item Graphs*)**.** Multimodal user–item graphs combine multiple modalities to describe users and items. They are mainly used to depict the relationship between users and items for helping comprehend users' behaviors, preferences, and needs. We represent the interaction data as a user–item bipartite graph $G = \{(u, i) \mid u \in U, i \in I\}$, where U and I denote the set of users and items, respectively. The edge $e_{ui} = 1$ represents an interaction between user u and item i; otherwise, $e_{ui} = 0$.

In addition, the user–item interaction data include various modality information — visual, acoustic, and textual features. We use $m \in M = \{A, V, T\}$ as the modality indicator, where A, V, and T denote acoustic, visual, and textual modalities, respectively.

Definition 2 (*Missing Modality*)**.**

In multimodal user–item graphs, missing modality refers to a situation wherein a particular modality (e.g., acoustic, visual, or textual) is missing from the representation of the user or item. This situation can result in missing information regarding the user or item, thus negatively affecting the data analysis and recommendation tasks. We denote item features with different missing modalities by $X_i^{m \in M = \{A,V,T\}}$ and user features with missing information by X_u.

Missing modalities affect the performance of multimodal recommendations and limit the recommendation system's comprehensive understanding of user interests and preferences [30]. Fortunately, this issue of missing modalities can be addressed via data complementation, cross-modality learning, or multimodal information integration.

Definition 3 (*Graph Contrastive Learning (GCL)*)**.**

GCL is a self-supervised learning method for learning the graph or node embedding representations via construction of contrastive tasks. The core idea is to enhance model representations without labeled data by maximizing the similarity between positive samples and minimizing that between negative samples.

GCL typically comprises two key components: (a) positive/negative samples and (b) contrastive loss. One way of setting positive samples is the same node from different views while the negative samples are the other nodes [31]. Selection of positive and negative samples affects the quality of the graph or node embedding learned by the model [32]. Contrastive loss is employed to measure the similarity between the positive samples with respect to the similarity between negative samples. A common contrastive loss function, InfoNCE loss [33], is defined as follows:

$$\mathcal{L}_{InfoNCE} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j))}{\exp(\mathrm{sim}(z_i, z_j)) + \sum_{k \in N_i} \exp(\mathrm{sim}(z_i, z_k))}, \tag{1}$$

where z_i and z_j denote the embeddings of positive sample for nodes i and j, respectively. N_i represents the set of negative samples for node i, and $\mathrm{sim}(\cdot, \cdot)$ denotes the similarity function.

4. Method

In this section, we elaborate on the proposed MMGACL model. As shown in Fig. 2, our model comprises four components: (a) Modality-aware user–item graphs, (b) diffusion-based modal complementation, (c) attention aggregation layer, and (d) contrastive learning via user–item graph.

4.1. Construction of user–item graphs

In multimodal recommendation, missing modality implies that specific-modality data for an item are incomplete, which can be comprehended as absence of certain feature vector entries within the content of modality.

We individually consider each modality information to model the user–item graph from the perspective of each modality. This process involves construction of three graphs, the acoustic user–item graph G_a, visual user–item graph G_v, and textual user–item graph G_t.

Considering the unimodal approach ensures that each graph completely represents the information of a specific modality (e.g., G_a focuses on acoustic information, ensuring that all audio-related information is preserved), this approach maximizes the use of information from each modality for complementation of item information. In addition, by separately processing each modality, it can prevent potential introduction of noise and bias that may result from mixing of data across different modalities.

4.2. Diffusion-based modality complementation (intramodality)

In multimodal recommendation, we assume that when two items i_m and i_n in a user–item graph are connected by a same user u_k, the features between items i_m and i_n may be more similar than the features of unconnected items. Items i_m and i_n may belong to a same category or have similar features that attract user u_k. Via the assumption, we can speculate that the set of first-order neighbor items $\{i_1, i_2, \ldots\}$ of user u_k have higher feature correlations and the correlation decreases with distance for higher-order neighbors. Therefore, for user–item graphs with missing modality information, neighbor information can be used for complementation of missing parts.

In particular, diffusion-based modality complementation is based on approximate topic-sensitive PageRank diffusion [34], which comprises two steps: item information complementation and user information complementation. One can refer to the details shown in Fig. 3 to better understand the process of modality complementation.

Item information complementation. We extract item subgraphs from user–item graphs in the context of three modalities. If two items are connected to a same user, an edge is added between these two items. On the item subgraph, diffusion complementation is denoted as:

$$\begin{aligned} \mathbf{X}_i^{(0)} &= \mathbf{X}_i^m, \\ \mathbf{X}_i^{(k+1)} &= (1-\alpha)\tilde{\mathbf{A}}_i \mathbf{X}_i^{(k)} + \alpha \mathbf{X}_i^m, \\ \mathbf{X}_i' &= \mathbf{X}_i^{(K)}, \end{aligned} \tag{2}$$

where $\alpha \in (0, 1]$ denotes restart probability, $\tilde{\mathbf{A}}_i$ is normalized adjacency matrix of item, $\mathbf{X}_i'$ represents complemented item feature matrix, K is iteration number of complementation, and $k \in [0, K-1]$.

After item information complementation, as shown in Fig. 4, PCA-based transformation maps $\mathbf{X}_i'$ in different modalities are projected into the attention fusion space, referred to as feature projection:

$$\mathbf{X}_i'' = Proj.(\mathbf{X}_i'), \tag{3}$$

where $\mathbf{X}_i''$ denotes the feature matrix after projection.

User information complementation. In the user–item graph, upon noting that user information is missing, meaning that we cannot directly access the characteristics or attributes of the user. We utilize the item information to complement the user information:

$$\begin{aligned} \mathbf{X}_u^{(0)} &= \mathbf{X}_u, \\ \mathbf{X}_u^{(k+1)} &= (1-\alpha)\tilde{\mathbf{A}}_u \mathbf{X}_u^{(k)} + \alpha \mathbf{X}_u, \\ \overline{\mathbf{X}_u} &= \mathbf{X}_u^{(K)}, \end{aligned} \tag{4}$$

where $\overline{\mathbf{X}_u}$ denotes the complemented user feature matrix.

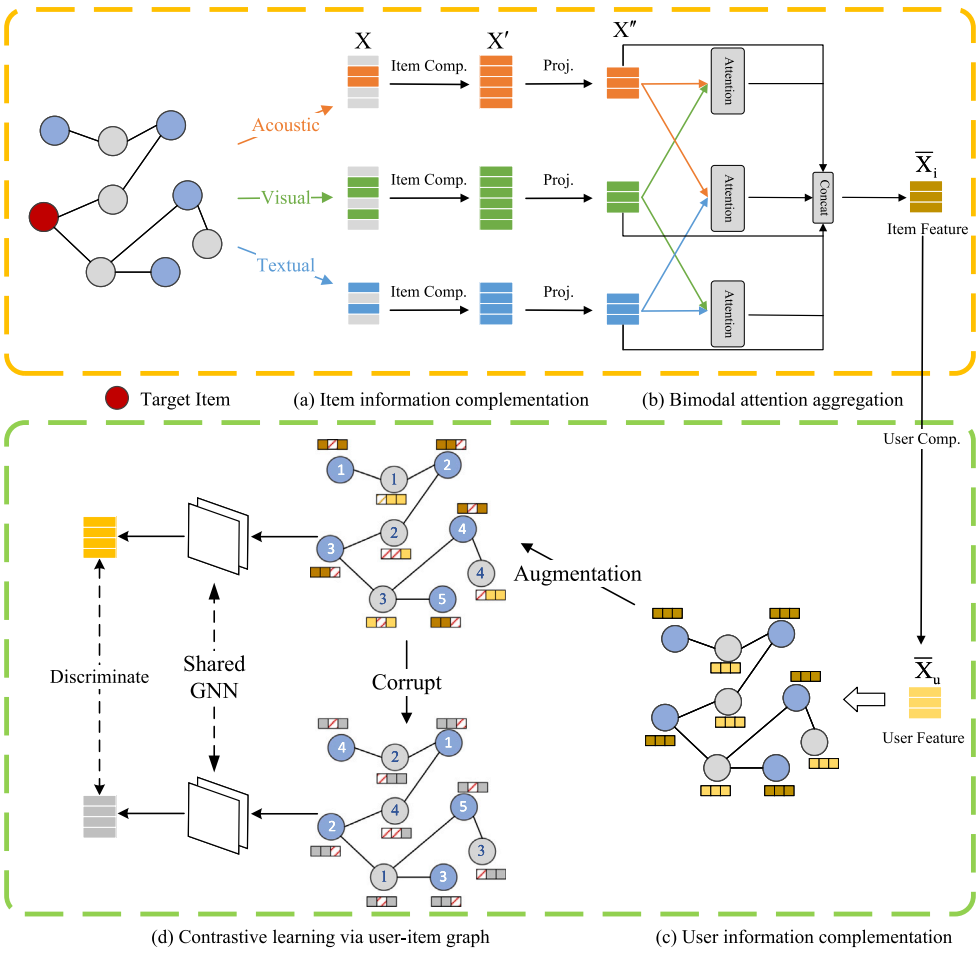


Fig. 2. The overall framework of MMGCAL.

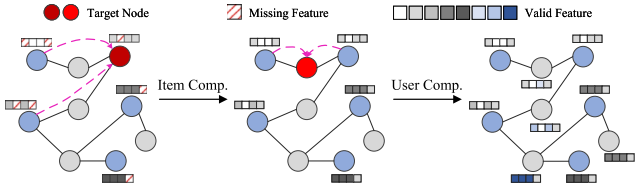


Fig. 3. Diffusion-based modality complementation.

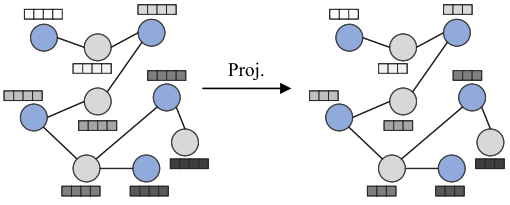


Fig. 4. Feature Projection.

4.3. Bimodal attention aggregation (intermodality)

Different modalities (e.g., acoustic, visual, and text) contain rich and diverse information that is often complementary and interrelated. We need to utilize the useful information in an intermodal manner and reduce the effect of redundant information. We adopt bimodal attention aggregation by applying the attention function to the representations of pairwise modalities, i.e., acoustic-visual (A-V), visual-text (V-T), and acoustic-text (A-T). Finally, the outputs of pairwise attentions are concatenated with the original representations.

We employ the pair of A-V modality attention computation as an example, denoted as BA_{AV}. Modality representation matrices $\mathbf{A}$ and $\mathbf{V}$ are obtained from the item information complementation in Section 4.2. We compute a pair of matching matrices $\mathbf{M}_1$, $\mathbf{M}_2$ on the two representations to represent the cross-modality information:

$$\mathbf{M}_1 = \mathbf{A} \cdot \mathbf{V}^T \quad \& \quad \mathbf{M}_2 = \mathbf{V} \cdot \mathbf{A}^T. \tag{5}$$

Considering that information between different modalities exhibits complementary or correlated characteristics, we can leverage the contextual information of each item under different modalities for computation. We use the softmax function to obtain the probability distribution scores $\mathbf{S}_1$, $\mathbf{S}_2$ of the bimodal attention matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ over each item. This is essentially for calculating the attention weights of item contexts. Thereafter, soft attention is applied for multimodal attention matrices to obtain the modality attention representations:

$$\begin{aligned} \mathbf{S}_1(i,j) &= \frac{e^{\mathbf{M}_1(i,j)}}{\sum_{k=1}^{N_i} e^{\mathbf{M}_1(i,k)}} \quad \text{for} \quad i,j = 1,..,N_i, \\ \mathbf{S}_2(i,j) &= \frac{e^{\mathbf{M}_2(i,j)}}{\sum_{k=1}^{N_i} e^{\mathbf{M}_2(i,k)}} \quad \text{for} \quad i,j = 1,..,N_i, \\ \mathbf{R}_1 &= \mathbf{S}_1 \cdot \mathbf{V} \quad \& \quad \mathbf{R}_2 = \mathbf{S}_2 \cdot \mathbf{A}, \end{aligned} \tag{6}$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ denote the representation matrices of modality attention. N_i is the number of items

Finally, we combine the attention representations of each modality with another modality via a multiplicative gating function [35]:

$$\mathbf{A}_1 = \mathbf{R}_1 \odot \mathbf{A} \quad \& \quad \mathbf{A}_2 = \mathbf{R}_2 \odot \mathbf{V}, \tag{7}$$

where $\mathbf{A}_1$, $\mathbf{A}_2$ is the attention metrics, $\odot$ denotes element-wise matrix multiplication.

Via concatenation of $\mathbf{A}_1$ and $\mathbf{A}_2$, we obtain the BA_{AV} between modalities $\mathbf{A}$ and $\mathbf{V}$:

$$BA_{AV} = concat[\mathbf{A}_1, \mathbf{A}_2]. \tag{8}$$

We follow the same procedure to compute BA_{VT} and BA_{AT} as that used for BA_{AV}. Thereafter, we concatenate the bimodal attention pairs with the individual modalities $\mathbf{A}$, $\mathbf{V}$, and $\mathbf{T}$:

$$\overline{\mathbf{X}_i} = concat[BA_{AV}, BA_{VT}, BA_{AT}, \mathbf{A}, \mathbf{V}, \mathbf{T}], \tag{9}$$

where $\overline{\mathbf{X}_i}$ represents the final item feature matrix used for user information complementation.

4.4. Contrastive learning via user–item graph

We obtain the user and item features after modality complementation. However, user feature complementation is performed on the basis of item features and user–item interactions, which may result in high similarity between user–item features. When the user and item features are excessively similar, the model may be overfitted, making distinguishing between them difficult [36] and affecting the recommendation performance. We introduce contrastive learning to help user–item features be distinguishable. As shown in Fig. 2(d), this section mainly consists of augmentation, corruption, GNN encoder, and discrimination.

Augmentation. For a given graph G and feature matrix $\mathbf{X}$, we use feature dropout as an augmentation technique to create $\ddot{G}$ and $\ddot{\mathbf{X}}$:

$$(\ddot{G}, \ddot{\mathbf{X}}) = Dropout(G, \mathbf{X}). \tag{10}$$

We follow the augmentation strategy mentioned in You et al. [37]. In particular, feature dropout randomly zeroes some of the elements of the input with a probability during training. $\ddot{G}$ and $\ddot{\mathbf{X}}$ change with augmentation in each training iteration, making the model less dependent on the graph for a fixed feature distribution.

Corruption. Thereafter, $\ddot{G}$ and $\ddot{\mathbf{X}}$ are corrupted to obtain $\widetilde{G}$ and $\widetilde{\mathbf{X}}$, respectively, which are used to generate user and item embeddings in the negative group. The process corrupts the topology of $\ddot{G}$ by randomly altering the order of the users and items in $\ddot{\mathbf{X}}$. The corrupted $\widetilde{G}$ and $\widetilde{\mathbf{X}}$ are utilized to generate the representations of nodes in a network connection that differs from the original one:

$$(\widetilde{G}, \widetilde{\mathbf{X}}) = Corrupt(\ddot{G}, \ddot{\mathbf{X}}). \tag{11}$$

GNN encoder and projector. We select a GNN encoder as well as a projector to output node representations for the user–item graph. In our work, we use graph convolutional network (GCN) [38] as an encoder and a multi-layer perception network as a projector with adjustable layers. A shared encoder and projector are used for the generation of node embeddings for positive and negative groups. We regard node embeddings from these two groups as a dataset that contains $2N$ data samples for discrimination. Prior to discrimination, all data samples are aggregated via the sum-based aggregation technique.

Discrimination. In the discrimination process, we adopt binary cross entropy (BCE) loss to discriminate between the two groups of node samples:

$$\mathcal{L}_{BCE} = -\frac{1}{2N}\left(\sum_{k=1}^{2N} z_k \log \hat{z}_k + (1 - z_k) \log (1 - \hat{z}_k)\right), \tag{12}$$

where z_k denotes the indicator of node k (if node k is corrupted, z_k is 0; otherwise it is 1), and $\hat{z}_k$ represents the prediction for a node sample k.

Table 1
Statistics of evaluation datasets.

Dataset	TikTok			Amazon					
				Office		Baby		Sports	
Modality	A	V	T	V	T	V	T	V	T
Embed Dim	128	128	768	4096	512	4096	1024	4096	1024
User	9308			4874		19445		35598	
Item	6710			2406		7050		18357	
Interactions	68722			52957		160792		296337	
	99.90%		99.96%						

A, V, and T denote the item modality Acoustic, Visual, and Textual, respectively.

4.5. Model optimization

We use Bayesian personalized ranking (BPR) [39] loss as the loss function to predict the interaction between users and items. We model a triplet of one user and two items, one of which is an observed item and the other one is an unobserved item:

$$\mathcal{R} = \{(u, i, i') \mid (u, i) \in G, (u, i') \notin G\}, \tag{13}$$

where $\mathcal{R}$ denotes a set of triples for training, i denotes the observed item while i' represents the unobserved one. Furthermore, we assume that the user prefers the observed item to the unobserved one. The BPR loss function can be defined as:

$$\mathcal{L}_{BPR} = \sum_{(u,i,i') \in \mathcal{R}} -\ln \mu\left(\hat{y}(u, i) - \hat{y}(u, i')\right) + \lambda \parallel \Theta \parallel_2^2, \tag{14}$$

where $\mu(\cdot)$ is the sigmoid function, $\hat{y}$ denotes the vector inner product operation between user u and item i, λ and Θ represent the regularization weight and the parameters of the model, respectively.

We obtain the following overall objective function upon combining the user–item recommendation task and the contrastive constraint:

$$\mathcal{L} = \omega \mathcal{L}_{BPR} + (1 - \omega)\mathcal{L}_{BCE}, \tag{15}$$

where ω represents the trade-off parameter governing the importance of the contrastive constraint terms. We simultaneously train all components of MMGACL.

5. Experiments

In this section, we perform an extensive empirical evaluation of MMGACL in scenarios with missing modality. Our experiments answer the following Research Questions (RQs):

- **RQ1**: How effective is our method in the missing modality scenarios?
- **RQ2**: Regarding the key designs of modality complementation in our MMGACL, what are their effects in boosting the recommendation performance?
- **RQ3**: How do the hyperparameters influence the recommendation performance of our method?

5.1. General settings

5.1.1. Datasets

The experiments were conducted on four publicly available multimodal recommendation datasets, i.e., TikTok, Amazon-Office, Amazon-Baby, and Amazon-Sports (with modality information preprocessed referring to Wei et al. [40] and Liu et al. [6]). Details and statistics of the datasets are presented in Table 1.

Table 2
Performance comparison of baselines on different datasets in terms of R@20, P@20 and N@20. OOM represents Out-Of-Memory on a 16 GB GPU.

Dataset	Metrics	GraphSAGE NIPS'17	SGC ICML'19	GRACE ICML'20	GGD NIPS'22	MMGCN MM'19	BM3 WWW'23	GCN_{MF} FGCS'21	FP LOG'22	MMGACL
TikTok	R@20	0.0813	0.0886	0.1332	0.0583	0.0697	0.1156	0.1346	0.1797	**0.2301**
	P@20	0.0043	0.0040	0.0065	0.0025	0.0038	0.0055	0.0067	0.0085	**0.0115**
	N@20	0.0355	0.0229	0.0765	0.0185	0.0243	0.0505	0.0736	0.0868	**0.1076**
Office	R@20	0.0299	0.0416	0.0449	0.0156	0.0474	0.0564	0.0545	0.0668	**0.0819**
	P@20	0.0045	0.0065	0.0101	0.0025	0.0072	0.0110	0.0107	0.0110	**0.0135**
	N@20	0.0176	0.0165	0.0249	0.0065	0.0243	0.0299	0.0303	0.0365	**0.0488**
Baby	R@20	0.0662	0.0583	0.0679	0.0723	0.0498	0.0658	0.0704	0.0733	**0.1650**
	P@20	0.0030	0.0035	0.0036	0.0035	0.0029	0.0040	0.0035	0.0040	**0.0085**
	N@20	0.0264	0.0339	0.0295	0.0284	0.0153	0.0263	0.0416	0.0488	**0.1004**
Sports	R@20	0.0533	0.0602		0.0698	0.0513	0.1214	0.1250	0.1301	**0.1804**
	P@20	0.0035	0.0033	OOM	0.0041	0.0032	0.0061	0.0066	0.0080	**0.0091**
	N@20	0.0258	0.0322		0.0343	0.0245	0.0671	0.0598	0.0717	**0.1053**

The best and runner-up results are highlighted with **bold** and underline, respectively.

5.1.2. Evaluation protocols

In the top-K item recommendation task, we evaluated our approach across three common metrics: Recall (R@K), Precision (P@K), and Normalized Discounted Cumulative Gain (N@K). We employed an all-ranking strategy [41] to avoid potential biases in test sampling. Each experiment was repeated for five rounds, following which the average value was taken as the result of the experiment.

5.1.3. Baselines

- **GraphSAGE** [42]: This method is a GNN model based on domain aggregation. For each node, it aggregates information regarding the neighboring nodes and uses this information to update the embedding representation of the concerned node.
- **SGC** [43]: This method reduces unnecessary complexity in GCN by simplifying the model. Model simplification is achieved by removing the nonlinearities and collapsing the weight matrices between consecutive layers.
- **GRACE** [31]: This method generates two graph views by corrupting the original graph at the structure and attribute levels. Thereafter, it learns the node representations by maximizing the consistency of the node representations in both these views.
- **GGD** [44]: This method proposes a new paradigm of self-supervised learning by observing technical deficiencies, namely group discrimination. Instead of calculating the similarity between two nodes, it directly discriminates between the two groups of node samples.
- **MMGCN** [8]: This method leverages GCN to learn user preference for various modalities via the construction of specific-modality graphs. It combines representations learned from each modality to generate final user and item embeddings.
- **BM3** [45]: This method jointly optimizes the three modality objectives to learn user and item representations simply via dropout to create contrastive views while adjusting modality features both in intermodal and intramodal perspectives.
- **GCN_{MF}** [46]: This method adapts GCN for graphs with missing features. It calculates the expected activations of neurons in the first hidden layer of the GCN while keeping the neurons in the other layers unchanged.
- **FP** [30]: This method addresses the problem of missing features in graphs via implementation of feature propagation based on Dirichlet energy minimization and diffusion-type differential equations on graphs.

5.1.4. Implementation details

We implemented our model with PyTorch. To construct the missing modality, we randomly replace 50% of entries in each modality feature matrix with 0. The user features used in the baseline were obtained by random generation. Hyper-parameters of MMGACL were selected via grid search during training as follows: user and item output dimension [16, 32, 64, 128], encoder layer depth [1, 2], projector layer depth [1, 2], batch size [512, 1024, 2048], α of modality complementation [0.01, 0.05, 0.1, 0.15], learning rate of the Adam optimizer [47] [0.0001, 0.001, 0.005], and dropout rate of augmentation in contrastive learning [0.1, 0.2, 0.3]. The best hyper-parameter values were selected based on the result obtained from a test set: encoder layer depth = 1, projector layer depth = 1, batch size = 2048, α of modality complementation = 0.01, learning rate = 0.001, and dropout rate of augmentation = 0.1. The user and item output dimension was 16 in the TikTok and Office datasets, while in the Baby and Sports datasets was 64. We conducted the experiments on a PC server equipped with a Tesla V100-SXM2 GPU.

5.2. Performance analysis (*RQ1*)

The results of our model and the baselines on the four datasets are presented in Table 2. Clearly, our proposed method outperformed all baselines in terms of R@20, P@20, and N@20, indicating that our method can effectively deal with the missing modality problem. In particular, MMGACL improves the highest baseline of R@20 by 28.05%, 22.60%, and 38.66% on the TikTok, Office, and Sports, respectively. Especially on the Baby dataset, the R@20 was more than twice the result of FP. Moreover, the following observations were made by categorizing all the methods into four classes:

- In the first and second blocks, we compared the performance with two GNN-based models (GraphSAGE and SGC) and two self-supervised learning models (GRACE and GGD). In general, GRACE and GGD outperformed the GNN models, except for GGD on the TikTok and Office datasets. This suggested the effectiveness of incorporating contrastive learning into GNNs for handling multimodal data. Contrastive learning helps better capture the structural information of the user–item bipartite graph, making the user–item features more discriminative while containing both modality and structural information. In addition, the results of GRACE and GGD were notably inferior to that of MMGACL. This phenomenon may be because in missing modality scenarios, more features or edges are dropped during generation of the contrastive view. This results in insufficient information contained in the users and items, thereby negatively affecting the model's performance.
- The third block comprises two models for multimodal recommendation, MMGCN and BM3. Their results in handling missing modality data highlighted the importance of modality information for recommendation. Richer contextual information between users and items helps the recommendation systems more comprehensively understand user behavior. Moreover, MMGCN and BM3 may have redundant information when processing modality information for recommendation. However, MMGACL employs bimodal attention aggregation, which considers correlations

Table 3
Ablation study of intramodality complementation in MMGACL. "w/o" means removing the corresponding module from the complete model.

Datasets	Variants	R@20	P@20	N@20
TikTok	$\text{MMGACL}_{w/o\ u\&i}$	0.0161	0.0005	0.0063
	$\text{MMGACL}_{w/o\ i}$	0.2103	0.0105	0.0924
	$\text{MMGACL}_{w/o\ u}$	0.0110	0.0004	0.0015
	MMGACL	**0.2301**	**0.0115**	**0.1076**
Office	$\text{MMGACL}_{w/o\ u\&i}$	0.0029	0.0010	0.0024
	$\text{MMGACL}_{w/o\ i}$	0.0460	0.0070	0.0265
	$\text{MMGACL}_{w/o\ u}$	0.0037	0.0015	0.0021
	MMGACL	**0.0819**	**0.0135**	**0.0488**
Baby	$\text{MMGACL}_{w/o\ u\&i}$	0.0124	0.0002	0.0018
	$\text{MMGACL}_{w/o\ i}$	0.1105	0.0055	0.0560
	$\text{MMGACL}_{w/o\ u}$	0.0210	0.0011	0.0052
	MMGACL	**0.1650**	**0.0085**	**0.1004**
Sports	$\text{MMGACL}_{w/o\ u\&i}$	0.0182	0.0005	0.0029
	$\text{MMGACL}_{w/o\ i}$	0.0950	0.0048	0.0324
	$\text{MMGACL}_{w/o\ u}$	0.0110	0.0002	0.0016
	MMGACL	**0.1804**	**0.0091**	**0.1053**

and complementarity between pairs of modalities and extracts important information. This approach effectively enhanced the recommendation performance.

- GCL_{MF} and FP in the fourth block are designed to handle the problem of missing features. The results of MMGACL demonstrated the advantages offered by modality-aware user–item graphs and modality information complementation. Complementation of modality information within a single modality-specific user–item graph is more effective than integrating all modalities before complementation. Information correlations within the same modality are closer and can provide more critical information for the missing parts. MMGACL separates the complementation processes on users and items, hence allowing for more precise capture of user preferences of items. This is achieved by user information complementation on the basis of user–item interactions while ensuring usability of item information. In addition, bimodal attention aggregation extracts complementary information between different modalities, providing more data to compensate for missing modalities.

5.3. Ablation study (*RQ2*)

5.3.1. Intramodality complementation

To evaluate the effect of intramodality complementation, we conducted ablation experiments by separating removing different complementations. In particular, we designed the following variants of MMGACL.

- $\text{MMGACL}_{w/o\ u\&i}$: This variant neither complements the user nor the item, and user information is obtained via random generation.
- $\text{MMGACL}_{w/o\ i}$: This variant does not complement the item and user information is obtained based on the item information.
- $\text{MMGACL}_{w/o\ u}$: This variant complements the item, and user information is obtained via random generation.

Table 3 presents the recommendation performances of MMGACL and variants thereof, i.e., $\text{MMGACL}_{w/o\ u\&i}$, $\text{MMGACL}_{w/o\ i}$, and $\text{MMGACL}_{w/o\ u}$ on the four datasets.

Table 3 shows that the results of the $\text{MMGACL}_{w/o\ u\&i}$ and $\text{MMGACL}_{w/o\ u}$ variants were unsatisfactory, while the $\text{MMGACL}_{w/o\ i}$ variant performed better but unsatisfactorily. This indicated the effectiveness of user information complementation on the basis of item data. In the user–item graph, the interaction between users and items exhibits a certain degree of similarity. Moreover, the $\text{MMGACL}_{w/o\ i}$ variant performed worse than our MMGACL demonstrating the validity of item complementation. Missing information can be more accurately filled by utilizing correlation and consistency of features within the same modality.

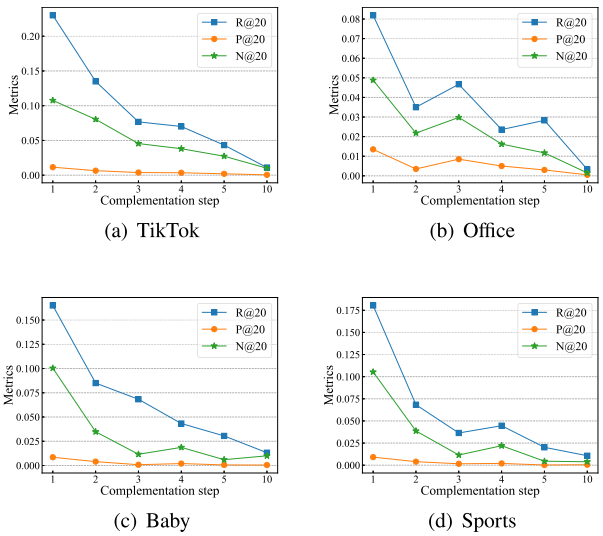


Fig. 5. Effect of complementation step.

Table 4
Ablation study of intermodality complementation in MMGACL.

Datasets	Variants	R@20	P@20	N@20
TikTok	MMGACL_{concat}	0.0796	0.0035	0.0509
	MMGACL_{add}	0.0952	0.0045	0.0585
	$\text{MMGACL}_{multiple}$	0.0566	0.0023	0.0298
	MMGACL	**0.2301**	**0.0115**	**0.1076**
Baby	MMGACL_{concat}	0.1533	0.0080	0.0963
	MMGACL_{add}	0.1574	0.0082	0.0993
	$\text{MMGACL}_{multiple}$	0.0896	0.0043	0.0412
	MMGACL	**0.1650**	**0.0085**	**0.1004**

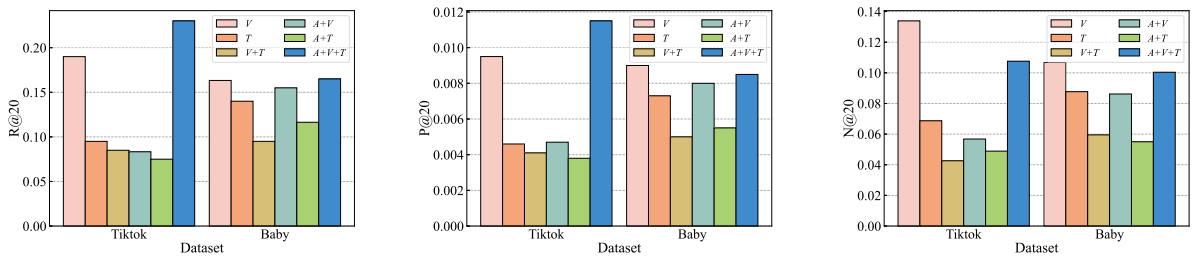
5.3.2. Intermodality complementation

Bimodal attention aggregation serves as intermodality complementation in MMGACL. We further studied the behavior of MMGACL by varying the ways of aggregating item information.

- MMGACL_{concat}: This variant concatenates the three modality item features into a larger tensor.
- MMGACL_{add}: This variant involves use of element-wise addition, which adds elements at corresponding positions in the item features of the three modalities.
- $\text{MMGACL}_{multiple}$: This variant uses element-wise multiplication, which multiplies elements at corresponding positions in the item features of the three modalities.

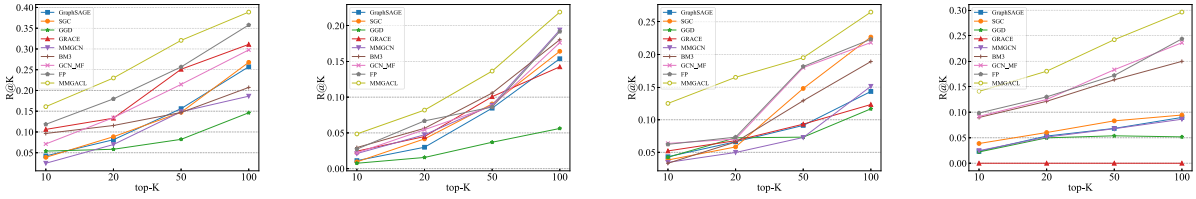
Table 4 presents the performance delivered by MMGACL, MMGACL_{concat}, MMGACL_{add}, and $\text{MMGACL}_{multiple}$ on the four datasets.

Table 4 shows that MMGACL outperformed its three variants, demonstrating that bimodal attention aggregation is effective in complementation of intermodality information. In particular, the MMGACL_{concat} variant concatenates information from the three modalities, retaining all information of the item complementation under each modality. However, it is unable to effectively select or emphasize certain important information and performs worse than MMGACL in terms of information extraction. The MMGACL_{add} and $\text{MMGACL}_{multiple}$ variants aggregate item features via element-wise addition/multiplication, which cannot be adjusted on the basis of different item input data. These variants fail to capture the correlation of information between different modalities, resulting in less effective information integration compared with that achieved by MMGACL.



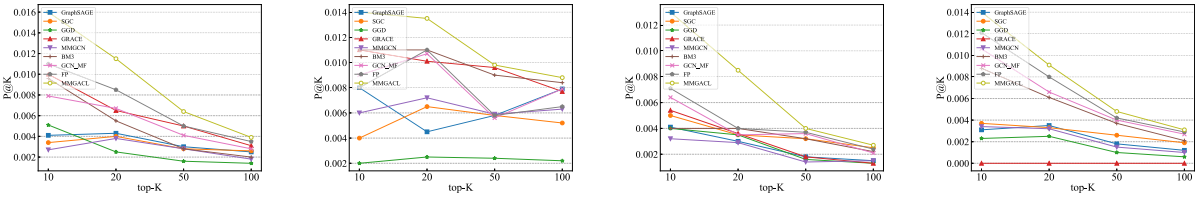
(a) R@20 (b) P@20 (c) N@20

Fig. 6. Effect of different modalities on TikTok and Baby.



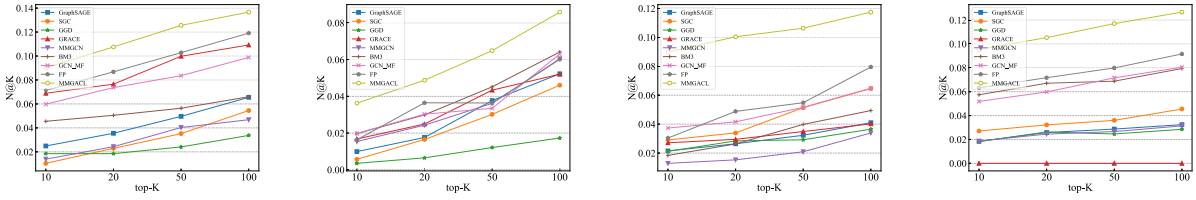
(a) TikTok (b) Office (c) Baby (d) Sports

Fig. 7. Effect of top-K recommendation in terms of R@K.



(a) TikTok (b) Office (c) Baby (d) Sports

Fig. 8. Effect of top-K recommendation in terms of P@K.



(a) TikTok (b) Office (c) Baby (d) Sports

Fig. 9. Effect of top-K recommendation in terms of N@K.

5.4. Parameter sensitivity (**RQ3**)

In this section, we examine the sensitivity of several important parameters of our MMGACL on different datasets.

5.4.1. Complementation step

We investigated the effect of the number of complementation steps involved in the modality complementation module. The results are shown in Fig. 5. We can observe that R@20, P@20, and N@20 values decreased as number of complementation steps increased on all datasets. This validated our speculation discussed in Section 4.2: during the complementation of information for a target node by using neighbor nodes, information from closer nodes is more relevant because of higher feature correlation. However, the correlation of higher-order nodes diminishes with distance between them. This decline is attributed to the presence of noise from higher-order nodes, negatively affecting the recommendation performance.

5.4.2. Modality analysis

We evaluated the recommendation accuracy of the MMGACL model by feeding into it single-modality and bimodal features. Experiments were conducted on the TikTok and Baby datasets, with the results shown in Fig. 6. The importance of different modalities was observed to vary with datasets. In particular, on the TikTok dataset, models with single-modality features outperformed the ones with multimodal features. Models incorporating acoustic features performed worse than those containing visual or textual features. This indicates that visual features were more important than other two modalities. Similar to that for the TikTok dataset, models with single-modality features performed better on the Baby dataset too, particularly those with visual features. However, unlike the TikTok dataset, the Baby dataset saw models with acoustic features outperforming those with visual or textual features. Overall, on both datasets, fusion of the three modalities generally yielded better results than those achieved by single and bimodal combinations. Notably, on the TikTok dataset, the N@20 metric value achieved via fusion of the three modalities was inferior to that obtained

upon using visual features; however, on the Baby dataset, models with visual features performed comparable to MMGACL in terms of the P@20 and N@20.

5.4.3. Top-K analysis

In top-K recommendations, we selected K = 20 as our main measure of recommendation performance. We also conducted a study for K = 10, 50, and 100, with the results shown in Figs. 7–9. Our model MMGACL achieves optimally performed on all datasets. With increase in K, the R@K and N@K values improved but P@K decreased. This was because a larger recommendation list can include more items and thus increase the proportion of relevant items retrieved from R@K. Similarly for N@K, a higher K value allows for more relevant items to be recommended, thereby improving the score. For P@K, a longer recommendation list is more likely to include irrelevant items, leading to decreased precision. The results for different K values demonstrated that MMGACL was suitable for missing modality scenarios.

5.5. Complexity analysis

Time complexity of MMGACL depended on three fundamental components: modality complementation, bimodal attention aggregation, and contrastive learning. Modality complementation included item complementation and user complementation, each having a complexity of $O(TNd)$, where T is iteration number of complementation, N represents number of nodes, and d denotes node feature dimension. Complexity of the bimodal attention aggregation component was determined by modality feature dimensions and number of nodes, with the component repeatedly called thrice, each time with a complexity of $O(N^2d)$. The part of contrastive learning took into account the encoder, projector, aggregation, and BCE loss with a complexity of $O(2B(LNd + LNd^2 + MNd^2 + Nd + N))$, where B is training batch size, L denotes number of encoder layers and M represents number of projector layers. In conclusion, the overall complexity of the model was $O(2TNd + 3N^2d + 2B(LNd + LNd^2 + MNd^2 + Nd + N)) \rightarrow O(Nd(T + N + BL + BLd + BMd))$. We can see that time complexity was mainly governed by the bimodal attention aggregation, encoder, and projector.

6. Conclusion and future work

We propose the MMGACL model to tackle the problem of missing modality information. MMGACL performed item complementation via modeling of user–item graphs individually under different modalities. Thereafter, our model aggregated the three modalities through a bimodal attention module, considering intermodality correlations to obtain final item representation. In extreme missing modality scenarios, we might not be able to obtain user information. Therefore, we used item features and user–item interactions for user information complementation. To eliminate the effect of excessively similar user and item features after complementation, we employed contrastive learning to maximize their mutual information and distinguish users from items. We evaluated the recommendation performance of the proposed MMGACL model on four multimodal datasets to demonstrate its applicability in scenarios with missing modality. The experimental results showed that MMGACL notably improved accuracy in comparison with the existing multi-category models.

However, results of experimental modality analysis showed that MMGACL performed less effectively with multiple modalities than with a single modality when handling missing information on different datasets. This revealed the model's scalability limitations on the datasets. Future research will focus on developing adaptive mechanisms to dynamically address variations in missing modalities. The high computational cost of the MMGACL algorithm on large datasets will also be an important direction for optimization. In addition, our model will be extended to multimodal knowledge graphs and knowledge hypergraphs [48] via integrating triplet information to enhance the representation quality. In the future, MMGACL will be applied to broader domains such as social media, healthcare systems, and cyber security [49,50], tackling challenges associated with missing modalities.

CRediT authorship contribution statement

Wenqian Zhao: Writing – review & editing, Writing – original draft, Validation, Methodology. **Kai Yang:** Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Peijin Ding:** Writing – review & editing, Data curation. **Ce Na:** Writing – review & editing. **Wen Li:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 22KJD120002).

Data availability

Data will be made available on request.

References

[1] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, Lightgcn: Simplifying and powering graph convolution network for recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 639–648, http://dx.doi.org/10.1145/3397271.3401063.

[2] Y. Himeur, A. Alsalemi, A. Al-Kababji, F. Bensaali, A. Amira, C. Sardianos, G. Dimitrakopoulos, I. Varlamis, A survey of recommender systems for energy efficiency in buildings: Principles, challenges and prospects, Inf. Fusion 72 (2021) 1–21, http://dx.doi.org/10.1016/j.inffus.2021.02.002.

[3] X. Zhou, D. Lin, Y. Liu, C. Miao, Layer-refined graph convolutional networks for recommendation, in: 2023 IEEE 39th International Conference on Data Engineering, ICDE, IEEE, 2023, pp. 1247–1259, http://dx.doi.org/10.1109/ICDE55515.2023.00100.

[4] U. Fang, M. Li, J. Li, L. Gao, T. Jia, Y. Zhang, A comprehensive survey on multi-view clustering, IEEE Trans. Knowl. Data Eng. 35 (12) (2023) 12350–12368.

[5] R. He, J. McAuley, VBPR: visual bayesian personalized ranking from implicit feedback, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, (1) 2016, http://dx.doi.org/10.1609/aaai.v30i1.9973.

[6] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, M. Kankanhalli, User diverse preference modeling by multimodal attentive metric learning, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1526–1534, http://dx.doi.org/10.1145/3343031.3350953.

[7] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha, Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 765–774, http://dx.doi.org/10.1145/3331184.3331254.

[8] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T.-S. Chua, MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1437–1445, http://dx.doi.org/10.1145/3343031.3351034.

[9] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, L. Nie, Dualgnn: Dual graph neural network for multimedia recommendation, IEEE Trans. Multimed. 25 (2021) 1074–1084, http://dx.doi.org/10.1109/TMM.2021.3138298.

[10] J. Li, C. Yang, G. Ye, Q.V.H. Nguyen, Graph neural networks with deep mutual learning for designing multi-modal recommendation systems, Inform. Sci. 654 (2024) 119815, http://dx.doi.org/10.1016/j.ins.2023.119815.

[11] H. Fang, D. Liang, W. Xiang, Single-stage extensive semantic fusion for multi-modal sarcasm detection, Array 22 (2024) 100344.

[12] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, L. Wang, Latent structure mining with contrastive modality fusion for multimedia recommendation, IEEE Trans. Knowl. Data Eng. (2022) http://dx.doi.org/10.1109/TKDE.2022.3221949.

[13] A. Khelloufi, H. Ning, A. Naouri, A.B. Sada, A. Qammar, A. Khalil, L. Mao, S. Dhelim, A multimodal latent-features-based service recommendation system for the social internet of things, IEEE Trans. Comput. Soc. Syst. (2024) http://dx.doi.org/10.1109/TCSS.2024.3360518.

[14] Y. Ding, X. Yu, Y. Yang, Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 3975–3984.

[15] F. Xu, P. Fu, Q. Huang, B. Zou, A. Aw, M. Wang, Leveraging contrastive learning and knowledge distillation for incomplete modality rumor detection, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 13492–13503, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.900.

[16] Y. Wang, Y. Li, Z. Cui, Incomplete multimodality-diffused emotion recognition, Adv. Neural Inf. Process. Syst. 36 (2024).

[17] C. Wang, M. Niepert, H. Li, LRMM: Learning to recommend with missing modalities, 2018, arXiv Preprint, arXiv:1808.06791.

[18] Z. Lin, Y. Tan, Y. Zhan, W. Liu, F. Wang, C. Chen, S. Wang, C. Yang, Contrastive intra-and inter-modality generation for enhancing incomplete multimedia recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 6234–6242, http://dx.doi.org/10.1145/3581783.3612362.

[19] Z. Lin, C. Tian, Y. Hou, W.X. Zhao, Improving graph collaborative filtering with neighborhood-enriched contrastive learning, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2320–2329, http://dx.doi.org/10.1145/3485447.3512104.

[20] L. Chen, G. Zhu, W. Liang, J. Cao, Y. Chen, Keywords-enhanced contrastive learning model for travel recommendation, Inf. Process. Manage. 61 (6) (2024) 103874.

[21] Y. Wei, X. Wang, L. Nie, X. He, T.-S. Chua, Graph-refined convolutional network for multimedia recommendation with implicit feedback, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3541–3549, http://dx.doi.org/10.1145/3394171.3413556.

[22] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, L. Wang, Mining latent structures for multimedia recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3872–3880, http://dx.doi.org/10.1145/3474085.3475259.

[23] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, Hemis: Hetero-modal image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 469–477, http://dx.doi.org/10.1007/978-3-319-46723-8_54.

[24] R. Dorent, S. Joutard, M. Modat, S. Ourselin, T. Vercauteren, Hetero-modal variational encoder-decoder for joint modality completion and segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer, 2019, pp. 74–82, http://dx.doi.org/10.1007/978-3-030-32245-8_9.

[25] L. Tran, X. Liu, J. Zhou, R. Jin, Missing modalities imputation via cascaded residual autoencoder, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 1405–1414.

[26] H. Wang, C. Ma, J. Zhang, Y. Zhang, J. Avery, L. Hull, G. Carneiro, Learnable cross-modal knowledge distillation for multi-modal learning with missing modality, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 216–226, http://dx.doi.org/10.1007/978-3-031-43901-8_21.

[27] S. Wei, C. Luo, Y. Luo, MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 20039–20049.

[28] Z. Lian, L. Chen, L. Sun, B. Liu, J. Tao, Gcnet: Graph completion network for incomplete multimodal learning in conversation, IEEE Trans. Pattern Anal. Mach. Intell. (2023) http://dx.doi.org/10.1109/TPAMI.2023.3234553.

[29] S. Duan, L. Wu, A. Liu, X. Chen, Alignment-enhanced interactive fusion model for complete and incomplete multimodal hand gesture recognition, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 4661–4671, http://dx.doi.org/10.1109/TNSRE.2023.3335101.

[30] E. Rossi, H. Kenlay, M.I. Gorinova, B.P. Chamberlain, X. Dong, M.M. Bronstein, On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features, in: Learning on Graphs Conference, PMLR, 2022, 11–1.

[31] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, L. Wang, Deep graph contrastive representation learning, 2020, arXiv Preprint, arXiv:2006.04131.

[32] K. Yang, Y. Liu, Z. Zhao, P. Ding, W. Zhao, Local structure-aware graph contrastive representation learning, Neural Netw. 172 (2024) 106083, http://dx.doi.org/10.1016/j.neunet.2023.12.037.

[33] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv Preprint, arXiv:1807.03748.

[34] J. Gasteiger, A. Bojchevski, S. Günnemann, Predict then propagate: Graph neural networks meet personalized pagerank, 2018, arXiv Preprint, arXiv:1810.05997.

[35] B. Dhingra, H. Liu, Z. Yang, W.W. Cohen, R. Salakhutdinov, Gated-attention readers for text comprehension, 2016, arXiv Preprint, arXiv:1606.01549.

[36] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, N. Yokoya, Learning joint representations of videos and sentences with web image search, in: Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14, Springer, 2016, pp. 651–667, http://dx.doi.org/10.1007/978-3-319-46604-0_46.

[37] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, Y. Shen, Graph contrastive learning with augmentations, Adv. Neural Inf. Process. Syst. 33 (2020) 5812–5823.

[38] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv Preprint, arXiv:1609.02907.

[39] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, 2012, arXiv Preprint, arXiv:1205.2618.

[40] W. Wei, C. Huang, L. Xia, C. Zhang, Multi-modal self-supervised learning for recommendation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 790–800, http://dx.doi.org/10.1145/3543507.3583206.

[41] Y. Wei, X. Wang, X. He, L. Nie, Y. Rui, T.-S. Chua, Hierarchical user intent graph network for multimedia recommendation, IEEE Trans. Multimed. 24 (2021) 2701–2712, http://dx.doi.org/10.1109/TMM.2021.3088307.

[42] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Adv. Neural Inf. Process. Syst. 30 (2017).

[43] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, K. Weinberger, Simplifying graph convolutional networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6861–6871.

[44] Y. Zheng, S. Pan, V. Lee, Y. Zheng, P.S. Yu, Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination, Adv. Neural Inf. Process. Syst. 35 (2022) 10809–10820.

[45] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, F. Jiang, Bootstrap latent representations for multi-modal recommendation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 845–854, http://dx.doi.org/10.1145/3543507.3583251.

[46] H. Taguchi, X. Liu, T. Murata, Graph convolutional networks for graphs containing missing features, Future Gener. Comput. Syst. 117 (2021) 155–168, http://dx.doi.org/10.1016/j.future.2020.11.016.

[47] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv Preprint, arXiv:1412.6980.

[48] C. Wang, X. Wang, Z. Li, Z. Chen, J. Li, Hyconve: A novel embedding model for knowledge hypergraph link prediction with convolutional neural networks, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 188–198.

[49] Y. Jia, Z. Gu, L. Du, Y. Long, Y. Wang, J. Li, Y. Zhang, Artificial intelligence enabled cyber security defense for smart cities: A novel attack detection framework based on the MDATA model, Knowl.-Based Syst. 276 (2023) 110781.

[50] A. Kiflay, A. Tsokanos, M. Fazlali, R. Kirner, Network intrusion detection leveraging multimodal features, Array 22 (2024) 100349.