Contrastive Intra- and Inter-Modality Generation for Enhancing Incomplete Multimedia Recommendation

Zhenghong Lin
Fuzhou University
Fuzhou, Fujian, China
hongzhenglin970323@gmail.com

Yanchao Tan*
Fuzhou University
Fuzhou, Fujian, China
yctan@fzu.edu.cn

Yunfei Zhan
Fuzhou University
Fuzhou, Fujian, China
z2149699627@gmail.com

Weiming Liu
Zhejiang University
Hangzhou, Zhejiang, China
21831010@zju.edu.cn

Fan Wang
Zhejiang University
Hangzhou, Zhejiang, China
fanwang97@zju.edu.cn

Chaochao Chen
Zhejiang University
Hangzhou, Zhejiang, China
zjuccc@zju.edu.cn

Shiping Wang
Fuzhou University
Fuzhou, Fujian, China
shipingwangphd@163.com

Carl Yang
Emory University
Atlanta, Georgia, USA
j.carlyang@emory.edu

ABSTRACT

With the rapid growth of multimedia-sharing platforms (e.g. Twitter and TikTok), multimedia recommender systems have become fundamental for helping users alleviate information overload and discover items of interest. Existing multimedia recommendation methods often incorporate various auxiliary modalities (e.g., visual, textual, and acoustic) to describe item characteristics and improve task performance. However, these methods usually assume that each item is associated with complete modalities, ignoring the prevalence of missing modality issues in real-world scenarios. To deal with the challenge of missing modalities, in this paper, we propose a novel framework of Contrastive Intra- and Inter-Modality Generation (CI^2MG) for enhancing incomplete multimedia recommendation. We first develop a contrastive intra- and inter-modality generation module for the missing modalities, where the intra-modality representation is updated through clustering-based hypergraph convolution and inter-modality representation is obtained by optimal transport between different modalities. To tackle the challenge of insufficient and incomplete supervision labels during intra- and inter-modality generation, a modality-aware contrastive learning paradigm is introduced based on an augmentation between the intra-modality view and inter-modality view. Furthermore, to learn task-related representations from the generative modalities and further improve the performance of recommendation, we design an enhanced multimedia recommendation module to alleviate the influences driven by task-irrelevant noise. Extensive experiments on real-world datasets show the superiority of our proposed CI^2MG framework in offering great potential for personalized multimedia recommendation over the state-of-the-art baselines regarding Recall, NDCG, and Precision metrics.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Multimedia information systems**.

KEYWORDS

Incomplete Multimedia Recommendation, Modality Generation, Multimodal Learning, Missing Modality Completion

ACM Reference Format:
Zhenghong Lin, Yanchao Tan, Yunfei Zhan, Weiming Liu, Fan Wang, Chaochao Chen, Shiping Wang, and Carl Yang. 2023. Contrastive Intra- and Inter-Modality Generation for Enhancing Incomplete Multimedia Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612362

1 INTRODUCTION

In recent years, the numbers of users and multimedia contents are dramatically growing on multimedia-sharing platforms (e.g. Twitter and TikTok). In consequence, how to help users find their interesting items without browsing the numerous different genres of multimedia information becomes more challenging than ever [8]. To alleviate the above information overload, multimedia recommender systems have been designed, where the core is exploiting the historical user-item interactions and rich multimodal item features (e.g., visual, textual, and acoustic) to recommend multimedia items for users according to their preferences [20, 38, 43].

However, most existing multimedia recommendation methods usually assume that each item is associated with complete modalities and all modalities are always available, which can be unrealistic in real-world applications. Missing modality is a common issue on multimedia platforms and the absence of a modality means that

*Yanchao Tan is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612362

all features in this modality are missing [6]. For example, Figure 1 illustrates the scenarios of missing modalities on the online TikTok platform. The textual modality may be absent because the user only uploaded a short video without textual descriptions (for item i_1); The visual modality may be absent because the user is willing to share a song or other acoustic contents (for item i_2); The acoustic modality may be absent because the video contains the enormous ambient noise during recording, leading to the unavailable acoustic information (for item i_3). When partial modalities are missing, most of the multimedia recommendation models may fail or obtain rather inaccurate representations of items [40].

To handle missing modalities in the multimodal area, many previous works simply discarded missing modalities and achieve some performance improvements [23, 28]. However, these approaches cannot be applied in recommendation where the data is scarce and dropping the incomplete features may loose the full intra- and inter-modality information of multi-modalities [41]. As shown in Figure 1, the short videos i_1, i_2 and i_3 have the similar theme about traveling. The **intra-modality** features may be the important factor for the user to choose items, such as "like" of the scene, or just "view" of the scene when traveling in the textual modality reflecting the different preferences of users. Besides, different modalities may contain modality-specific features. The visual modality for i_3 may involve coarse-grained image information with appearance or color, while textual details are fine-grained descriptions for the item. In consequence, it is necessary to take the **inter-modality** features between different modalities into consideration.

To fully exploit both intra- and inter-modality features, we propose to generate missing modalities without additional supervision. In this way, we can alleviate the problem caused by the modalities that are incomplete or not available, and then enhance the multimedia recommendation. However, the way about how to generate missing modality is a non-trivial task and there are challenges from several perspectives:

Challenge I: *What metric(s) should be used to capture the intra-modality and inter-modality similarity for generating incomplete modalities?* A straightforward way to generate incomplete modalities is naturally exploiting the complete features similar to missing features within the same modality (intra-modality) and features from different modalities within the same item (inter-modality), as illustrated in Figure 1. For intra-modality features, existing methods usually infer the similarity between a single item with missing modality feature and a single item with complete modality feature, where such similarity is used to complete the missing modalities. However, the above way is unrealistic for the real-world process because the missing features may be similar to several complete features. In other words, the generation of incomplete modalities is a complex set-to-set problem (from complete set to missing set). How to model the set-to-set correlations and generate the incomplete modalities based on the set similarity remains unknown. Besides, for inter-modality similarity, the features of different modalities may suffer from heterogeneous representations (e.g., the features in textual modality often contain word position or relation, while the line and color of features are in the visual modality). How to obtain the similarity between different modalities and align the distribution of heterogeneous features based on this similarity to generate the incomplete modalities are still unsolved problem.

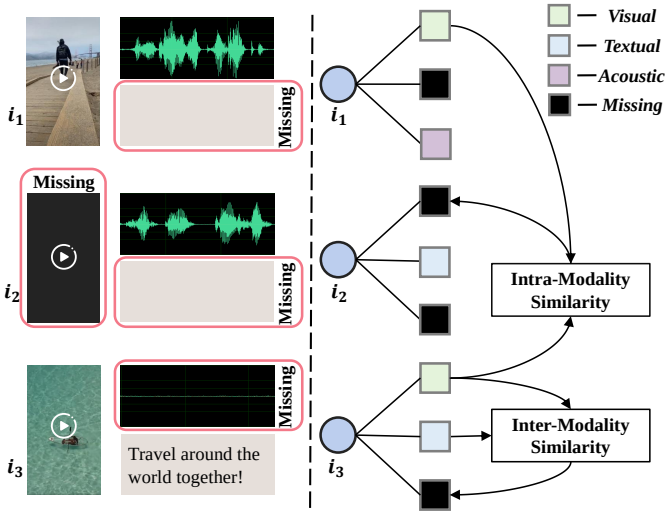

Figure 1: A toy example of missing modalities on the TikTok multimedia-sharing platform. i_1, i_2 and i_3 are similar short-video items about the traveling category. Intra-modality similarity is defined as the similarity between the same modality of different items and inter-modality similarity is the similarity between different modalities of the same item.

Challenge II: *How to jointly perform modality generation and multimedia recommendation?* A straightforward way to inject the generative features into the multimedia recommendation is exploiting the generative modalities to learn the latent representations of items directly. However, the generative modalities may involve various task-relevant and task-irrelevant information. For example, users may be attracted by the movies due to high-order semantic features in the visual modality (like special effects or actions), which need to be considered by recommendation tasks, while ignoring the contour/outline features more beneficial for the computer vision task of object detection [9]. Performing multimedia recommendation with the combination of both task-relevant and task-irrelevant information may lead to inferior performance[41].

To address these challenges, we propose a novel framework named Contrastive Intra- and Inter-Modality Generation for enhancing incomplete multimedia recommendation (CI^{2}MG), which consists of two pivotal technical modules: (1) Contrastive Intra- and Inter-Modality Generation for **Challenge I**; (2) Enhanced Multimedia Recommendation for **Challenge II**. Specifically, in (1) contrastive intra- and inter-modality generation, we first design a clustering-based mechanism to select items with similar modality and regard such clustering similarity as the hypergraph structure. By leveraging the intra-modality similarity between missing modalities and complete modalities in hypergraph structure, we perform the hypergraph convolution with 'node-hyperedge-node' feature transformation to update missing modality features through hypergraph convolutional neural networks (HGNN). Then, to obtain the aligned representations from heterogeneous features, we adopt the optimal transport (OT) to align the distribution between different heterogeneous multimedia contents and extract the

shared part information between different modalities to generate inter-modality features. Finally, to tackle the challenge of insufficient and incomplete supervision labels in recommendation, we design modality-aware contrastive learning based on an augmentation between the intra-modality view and inter-modality view. In (2) enhanced multimedia recommendation, we devise task-guided modality generation module to filter the task-irrelevant features and incorporate the generative modalities into the recommendation framework for enhancing multimedia recommendation.

Our overall contributions in this work are summarized as follows:

- *Formulation of intra- and inter-modality generation for incomplete multimedia recommendation.* CI^{2}MG is the first incomplete multimedia recommendation framework based on contrastive intra- and inter-modality generation, which can enhance multimedia recommendation with task-related generative modality features. (Section 3.1).
- *Effective model designs.* In the contrastive intra- and inter-modality generation, we design missing modalities generation to complete the missing features with both intra- and inter-modalities and propose modality-aware contrastive learning to enhance the generative modality representations (Section 3.2). In the enhanced multimedia recommendation, we devise task-guided modality generation to capture the task-relevant features of the recommendation task and perform complete multi-modality recommendation based on generative modalities (Section 3.3).
- *Extensive experiments on real-world multimedia datasets.* We conduct comprehensive experimental evaluations on multimedia recommendation tasks against state-of-the-art approaches over public benchmark datasets. Extensive experimental results demonstrate the superiority of our proposed CI^{2}MG framework (Section 4).

2 RELATED WORK

2.1 Missing Attribution Generation

Most existing multimodal analyses often assume that all modalities are available or complete [1, 2], where such a strong assumption is not always available in practice. To tackle the missing attribute problem, AGCN [36] was designed to iteratively perform two steps with graph embedding learning with previously learned attribute values and attribute update procedure to update the input of graph embedding learning. Zhu et al. proposed a novel data/model jointly driven framework to generate high-quality cases for power system DSA applications [44]. The above methods focused on the issue of missing at random within a feature vector. However, missing modality means the whole values in this modality are missing. In consequence, Ma et al. systematically studied the problem by proposing multimodal learning with severely missing modality and leveraged Bayesian meta-learning in uniformly achieving the objective [19]. Wu et al. proposed a multimodal news recommendation method [34] that could incorporate both textual and visual information.

Different from the above methods, either only utilizing complementary inter-modality information or consistent inter-modality features, our proposed CI^{2}MG uses both inter-modality and intra-modality features and injects the recommend-driven information into the representation learning for multimedia recommendation.

2.2 Multimedia Recommendation

Among all recommendation algorithms, collaborative filtering has become one of the most popular methods in both industry and research communities. NCL [16] incorporated the potential neighbors into contrastive pairs for the preference of users over items. GCCF [4] designed a residual structure to combine the multi-order information. To make use of the various multimedia contents, multimedia recommender systems incorporated various auxiliary modalities into the latent representation learning. Graph convolutional networks used in MMGCN [33] and GRCN [32] and graph attention mechanism applied in MKGAT [26] highly proved that graph neural networks captured high-order dependent structures among users and items playing a crucial role in modality-enhanced collaborative filtering model. Multimodal recommender systems [42] leveraging multimodal information was a recent trend in capturing the preferences of users over items.

Different from the above methods assuming that each item is associated with complete modalities and all modalities are always available, our CI^{2}MG framework is under the missing modality scenarios, which is common in real-world applications.

3 METHODOLOGY

In this section, we mainly show the details of the proposed CI^{2}MG framework. Firstly, we formally formulate the problem statement and perform an overview of CI^{2}MG architecture. Secondly, a contrastive intra- and inter-modality generation is proposed to complete the missing modalities based on the features from the shared part of different modalities and modality-specific correlations. Finally, we provide more accurate modality representations with task-relevant information and combine the multimedia contents into the latent representation space of items to acquire a high-quality multimedia recommendation.

3.1 Problem Statement and CI^{2}MG Overview

Our goal of CI^{2}MG aims to generate the missing modalities based on the intra-modality information of items belonging to the same cluster and the inter-modality features of the shared part among different modalities for recommendation. Given a set $\boldsymbol{U}$ of N_U users and a set $\boldsymbol{V}$ of N_V items, we use $\boldsymbol{R} \in \mathbb{R}^{N_U \times N_V}$ to represent their interaction matrix. The value of $\boldsymbol{R_{ij}}$ is set to 1 if the $i-$th user $\boldsymbol{u}_i \in \boldsymbol{U}$ has interacted with the $j-$th item $\boldsymbol{v}_j \in \boldsymbol{V}$. Otherwise, $\boldsymbol{R}_{ij} = 0$. In addition, the modality information of each item is provided, involving the visual, acoustic, and textual modalities ($M = 3$). We define $\overline{\boldsymbol{f}}_i^{V,m} \in \mathbb{R}^d$ as a d-dimensional vector to represent the each raw feature of i-th item under the m-th modality and we combine the whole item features as $\overline{\boldsymbol{F}}^{V,m} = \{\overline{\boldsymbol{f}}_1^m; \cdots; \overline{\boldsymbol{f}}_{N_V}^m\}$, where we preprocess the values of missing modalities in $\overline{\boldsymbol{F}}^{V,m} \in \mathbb{R}^{N_V \times d}$ to the average values of other item modalities following [36] and we use $\boldsymbol{F}^{U,m}$ to represent the users' preference on items. To exploit the intra-modality information for generating the missing modalities, we initialize prototype $\boldsymbol{E}^m = \{\boldsymbol{e}_1^m, \boldsymbol{e}_2^m, ..., \boldsymbol{e}_K^m\}$ to group the items with

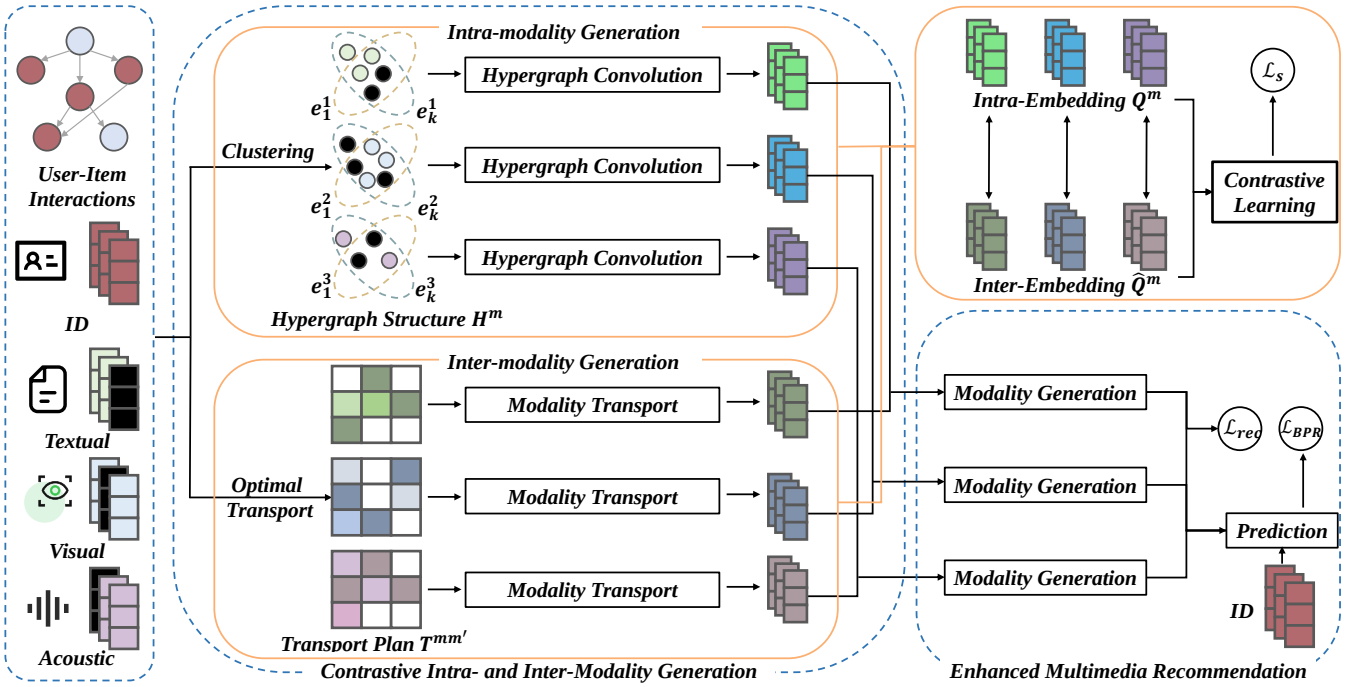

Figure 2: The overall architecture about CI^{2}MG. In the given case, the missing modalities are textual, visual and acoustic. The ID embeddings of users and items are initialized from a trainable lookup table and we use the black color to represent missing. e is the clustering centroid (seen as hyperedge in hypergraph neural networks). In Contrastive Intra- and Inter-Modality Generation, the modalities with missing values are grouped by the clustering mechanism, and perform the hypergraph convolution and optimal transport to generate missing modality features. Then, the modality-aware contrastive learning is conducted on the two different views and the enhanced features are fed to Enhanced Multimedia Recommendation.

the same cluster and K is the number of prototypes. Based on the learnable prototype $\boldsymbol{E}^m \in \mathbb{R}^{d\times K}$, the modality-specific hypergraph structure $\boldsymbol{H}^m \in \mathbb{R}^{N_V\times K}$ can be obtained. We denote $\boldsymbol{Q}^m \in \mathbb{R}^{N_V\times d}$ as the intra-modality representation updated by hypergraph-based message transformation with the clustering hypergraph structure $\boldsymbol{H}^m$. To deal with the heterogeneity, we use $\hat{\boldsymbol{Q}}^m \in \mathbb{R}^{N_V\times d}$ as the aligned representation. With the intra- and inter-modalities features, a probability matrix $\hat{\boldsymbol{R}}$ is computed, where the value of $\hat{\boldsymbol{R}}_{ij}$ represents the probability that the j-th item is recommended to the i-th user. The input and output are defined as follows:

- *Input:* user-item interaction matrix $\boldsymbol{R}$ and missing modalities features $\{\overline{\boldsymbol{F}}^{V,1}, \overline{\boldsymbol{F}}^{V,2}, ...\overline{\boldsymbol{F}}^{V,m}, ...\overline{\boldsymbol{F}}^{V,M}\}$.
- *Output:* a probability matrix $\hat{\boldsymbol{R}}$ for recommendation.

We summarize the main modules of the CI^{2}MG framework in Figure 2 and provide an overview. Our proposed model has two main parts: (1) Contrastive Intra- and Inter-Modality Generation and (2) Enhanced Multimedia Recommendation. In contrastive intra- and inter-modality generation module, we initialize the user and item embeddings $\boldsymbol{F}^{U,m}$ and $\boldsymbol{F}^{V,m}$ by the graph structure $\boldsymbol{R}$ to incorporate the collaborative effects into embedding learning. To complete the missing modalities, we first exploit modality-specific item features from the same cluster obtained by the clustering mechanism of prototype $\boldsymbol{E}^m$ and the generative intra-modality representation $\boldsymbol{Q}^m$ is obtained through the hypergraph neural networks. Then, we adopt the optimal transport to update the inter-modality representation $\boldsymbol{P}^m$ with the shared information across different modalities. To inject both intra- and inter-modality features into the generative representation, the contrastive learning mechanism is designed and the fusion representation $\boldsymbol{\Psi}^m$ of $\boldsymbol{Q}^m$ and $\boldsymbol{P}^m$ is obtained. In enhanced multimedia recommendation, we enhance the generative representation with task-relevant information and score the user preference $\hat{\boldsymbol{R}}$ via enhancing multimodal and id-corresponding representation. Finally, we perform prediction for multimedia recommendation based on the preference probability matrix $\hat{\boldsymbol{R}}$.

3.2 Contrastive Intra- and Inter-Modality Generation

In multimedia contents, both intra- and inter-modality features are significant for modeling user preferences over items. Ignoring the correlations of modalities may lead to less accurate multimedia recommendation [31]. For example, the intra-modality features can reflect the modality-specific information of items, such as the amazing visual content of videos, acoustic melody of music, and detailed textual descriptions of products. Besides, the inter-modality features contain the shared semantic information of different modalities, such as the characteristics and categories of items.

However, the modality missing is a common issue in real-world scenarios [19] Discarding the incomplete modalities may fail to capture the consistency and complementarity among different modalities, which can not be applied in areas where data is scarce [22]. To inject the modality-aware information into the representation learning of items, we propose our contrastive intra- and inter-modality generation with two steps: (1) missing modalities generation and (2) modality-aware contrastive learning.

3.2.1 Missing Modalities Generation. To capture sufficient information from sparse user-item interactions and incorporate the collaborative modality effects into representation learning of items, we first adopt the graph neural networks (GNNs) to initialize the embeddings of inputs. Suppose $\boldsymbol{F}^{V,m} \in \mathbb{R}^{N_V \times d}$ and $\boldsymbol{F}^{U,m} \in \mathbb{R}^{N_U \times d}$ be the item embedding with missing values under the m-modality and the user embedding with latent dimension d, respectively. $\boldsymbol{F}^{U,m}$ can be seen as the preference of users over these items. Here, N_V and N_U are the numbers of items and users. Then, we can obtain the input modality representation of users and items by the graph message aggregation:

$$\mathbf{F}^{U,m} = \boldsymbol{D}^{U^{-1}} \boldsymbol{R} \overline{\boldsymbol{F}}^{V,m}, \qquad \mathbf{F}^{V,m} = \boldsymbol{D}^{V^{-1}} \boldsymbol{R}^T \boldsymbol{F}^{U,m}. \tag{1}$$

Here, $\boldsymbol{D}^U \in \mathbb{R}^{N_U \times N_U}$ and $\boldsymbol{D}^V \in \mathbb{R}^{N_V \times N_V}$ are the diagonal degree matrices of user-item and item-user interaction matrices $\boldsymbol{R} \in \mathbb{R}^{N_U \times N_V}$ and $\boldsymbol{R}^T \in \mathbb{R}^{N_V \times N_U}$. $\overline{\boldsymbol{F}}^{V,m} = [\overline{\boldsymbol{f}}_1^m; \cdots; \overline{\boldsymbol{f}}_{N_V}^m]$ is the raw item features of the m-th modality with $\overline{\boldsymbol{f}}_i^m \in \mathbb{R}^d$. Furthermore, we also perform the id-corresponding graph information aggregation over the neighbors of users and items:

$$\begin{aligned} \mathbf{X}^{U,(l+1)} &= \sigma(\boldsymbol{D}^{U^{-1}} \boldsymbol{R} \mathbf{X}^{V,(l)} \mathbf{W}^{U,(l)}), \\ \mathbf{X}^{V,(l+1)} &= \sigma(\boldsymbol{D}^{V^{-1}} \boldsymbol{R}^T \mathbf{X}^{U,(l)} \mathbf{W}^{V,(l)}). \end{aligned} \tag{2}$$

We define $\mathbf{X}^{U,(l)} \in \mathbb{R}^{N_U \times d}, \mathbf{X}^{V,(l)} \in \mathbb{R}^{N_V \times d}$ as the id-corresponding embedding of users and items in the l-th layer of graph neural networks, where the zero-layer embeddings $\mathbf{X}^{U,(0)}$ and $\mathbf{X}^{V,(0)}$ are initialized from a trainable lookup table. $\sigma(\cdot)$ is the activation function to introduce the nonlinear factors. $\boldsymbol{W}^{U,(l)}$ and $\boldsymbol{W}^{V,(l)}$ are trainable user and item weights of the l-layer for GNNs.

Intra-modality Generation. To generate the missing values with intra-modality features, we design intra-modality missing generation module updating the missing features with modality-specific items from semantic similar neighbors, e.g. modality-specific items with the same category. Specifically, we first devise a prototype-based clustering mechanism to group items under each modality with similar characteristics. Inspired by the advantages of prototype-based clustering in dealing with insufficient, incomplete, or distorted data [24], we initialize the matrix $\boldsymbol{E}^m = [\boldsymbol{e}_1^m, \ldots, \boldsymbol{e}_i^m, \ldots, \boldsymbol{e}_K^m]$ of the m-th modality, where $\boldsymbol{e}_i^m \in \mathbb{R}^d$ is the i-th cluster prototype and K is the number of clustering centroid. In our clustering mechanism, we aim to find which nodes may be classified into the same cluster. In other words, the problem can be formulated as calculating the likelihood of items with m-th modality belonging to the cluster k:

$$pro(\boldsymbol{H}^m | \boldsymbol{F}^{V,m}, \boldsymbol{E}^m) = pro(\boldsymbol{H}^m[i,k] = 1 | \boldsymbol{f}_i^{V,m}, \boldsymbol{e}_k^m), \tag{3}$$

where $\boldsymbol{H}^m \in \mathbb{R}^{N_V \times K}$ is the probability matrix and $\boldsymbol{f}_i^{V,m} \in \boldsymbol{F}^{V,m}$ is the i-th item representation. Then, we can obtain $\boldsymbol{H}^m[i,k]$ through the cosine similarity:

$$\boldsymbol{H}^m[i,k] = \boldsymbol{f}_i^{V,m} \cdot \boldsymbol{e}_k^m / (||\boldsymbol{f}_i^{V,m}||_2 \cdot ||\boldsymbol{e}_k^m||_2), \tag{4}$$

where $\boldsymbol{H}^m[i,k]$ is the probability assigning item i to cluster centroid k. Then, we can update the representation of the missing modality by its semantic neighbors within the same cluster.

However, the problem of the above process can be seen as using the set of existing modalities to complete the set of missing modalities, where the two sets belong to the same cluster. To model such set-to-set correlations, the concept of hypergraph is involved naturally [11]. The hypergraph is composed of some hypernodes and hyperedges, where the hyperedge is generalizing the concept of edges and a hyperedge can contain any number of nodes [10]. In consequence, we can use the cluster centroid $\boldsymbol{E}^m$ to represent the hyperedge embeddings and exploit clustering probability matrix $\boldsymbol{H}^m$ to be the hypergraph incidence matrix indicating which hyperedge the hypernodes may belong to. Then, we adopt the hypergraph convolutional networks to perform the set-to-set compensation for updating representations of missing modalities:

$$\boldsymbol{Q}^{m,(l+1)} = \sigma(\boldsymbol{D}^{-1} \boldsymbol{H}^m \boldsymbol{W} \boldsymbol{B}^{-1} \boldsymbol{H}^{mT} \boldsymbol{Q}^{m,(l)}), \tag{5}$$

where $\boldsymbol{Q}^{m,(l)}$ is the complete embedding of the m-th modality after hypergraph convolution of l layers. $\boldsymbol{D}$ and $\boldsymbol{B}$ are diagonal matrices. $\boldsymbol{W}$ is the trainable weight of hypergraph neural networks and the zero-order representation $\boldsymbol{Q}^{m,(0)}$ is obtained by item features $\boldsymbol{F}^{V,m}$.

Inter-modality Generation. To generate the missing values with inter-modality features, we propose an inter-modality missing generation module completing the missing with shared parts of different modalities, e.g. a girl may be attracted by both the amazing image and textual description about the pattern of a dress. Different from the intra-modality representation learning, the cross inter-modality features may suffer from the heterogeneous issue with different modalities [12]. The features in textual modality often contain word position or relation [5], while the line or color of the feature is in the visual modality [17]. Inspired by the optimal transport in dealing with heterogeneous data, we design the optimal transport for inter-modality features with distributional alignment to bridge the heterogeneity gap. Firstly, we calculate the distribution distance between different modalities:

$$\boldsymbol{M}_{ij}^{mm'} = ||\boldsymbol{Q}_i^m - \boldsymbol{Q}_j^{m'}||_2^2, \tag{6}$$

where $\boldsymbol{M}_{ij}^{mm'}$ represent the transport cost from modality m to modality m' and $\boldsymbol{Q}_i^m, \boldsymbol{Q}_j^{m'}$ are the row vectors of $\boldsymbol{Q}^m, \boldsymbol{Q}^{m'}$. The whole distance formulation of optimal transport can be formulated:

$$d_M(\mathbf{r}, \mathbf{c}) := \min_{\mathbf{T}^{mm'} \in U(\mathbf{r},\mathbf{c})} < \mathbf{T}^{mm'}, \boldsymbol{M}_{ij}^{mm'} > . \tag{7}$$

We set $\mathbf{r}$ and $\mathbf{c}$ as all one vectors for empirical feature distributions of modalities m and m'. $\mathbf{T}^{mm'}$ means the transport plan and needs to be under the constraints of optimal transport settings defined as

$$U(\mathbf{r}, \mathbf{c}) := \{\mathbf{T}^{mm'} \in \mathbb{R}_+^{N_V \times N_V} | \mathbf{T}^{mm'} \mathbf{1}_m = \mathbf{r}, \mathbf{T}^{mm'^T} \mathbf{1}_{m'} = \mathbf{c}\}, \tag{8}$$

where we define $\mathbf{1}_m$ or $\mathbf{1}_{m'}$ as the m-dimensional or m'-dimensional vector of ones to calculate the sum of row or column in $\mathbf{T}^{mm'}$. For

solving such optimal transport problem, the Sinkhorn [7] optimization is adopted.

In the second step, we can use the mapping function $\mathbf{T}^{mm'}$ to align the embeddings of different modalities and update the missing modality representation with aligned distributions in the same item:

$$\hat{\boldsymbol{Q}}^m = \frac{1}{M-1} \sum_{m \neq m'} diag(\frac{1}{\boldsymbol{\mu}})(\boldsymbol{T}^{m'} + \Delta T^{m'})\boldsymbol{Q}^{m'}. \tag{9}$$

We define the generative embedding with complete modality as $\hat{\boldsymbol{Q}}^m$. m is the missing modality for compensating and m' is the different modality of the same item, except for m. M is the number of all modalities ($m \leq M$ and $m' \leq M$), where all modalities contain visual, textual, and acoustic in this paper ($M = 3$). Inspired by [3], we introduce an adjustable transport parameter $\Delta T^{m'}$ and $\boldsymbol{\mu}$ is the probabilistic simplex of embedding $\hat{\boldsymbol{Q}}^m$.

3.2.2 Modality-aware Contrastive Learning. Contrastive learning is an effective self-supervised framework to capture the consistency of features under different views [30]. To tackle the challenge of insufficient and incomplete supervision labels in recommendation [37], we design a modality-aware contrastive learning to perform an augment between the intra-modality view and inter-modality view:

$$\mathcal{L}_s = \sum_{m=1}^{M} \sum_{k=1}^{N_V} -log \frac{exp(s(\mathbf{q}_k^m, \hat{\mathbf{q}}_k^m)/\tau)}{\sum_{k'=1}^{N_v} exp(s(\mathbf{q}_k^m, \hat{\mathbf{q}}_{k'}^m)/\tau)}. \tag{10}$$

Here, the contrastive loss of InfoNCE [21] is adopted to supervise the generative item representations from intra-view and inter-view, where $\mathbf{q}_k^m$ and $\hat{\mathbf{q}}_k^m$ are the k-th item embedding of intra-modality representation $\mathbf{Q}^m$ and inter-modality representation $\hat{\mathbf{Q}}^m$, respectively. The temperature parameter τ is used to control the strength of the gradient for better balance and $s(\cdot)$ is the similarity measurement function. Through the loss of modality-aware contrastive learning, the auxiliary supervision signals can help the representation $\mathbf{Q}^m$ and $\hat{\mathbf{Q}}^m$ learn from each other, mutually.

3.3 Enhanced Multimedia Recommendation

The above section describes how to generate the missing modalities in a self-supervised way. However, the generative modalities may involve task-relevant and task-irrelevant information. For example, users may be attracted by the movies due to high-order semantic features in the visual modality (like special effects or actions), which need to be considered by recommendation tasks, while ignoring the contour/outline features more beneficial for object detection (a task in computer vision [9]). Hence, we propose enhanced multimedia recommendation with two steps: (1) task-guided modality inference and (2) complete multi-modality recommendation.

3.3.1 Task-guided Modality Generation. To take both the intra-modality features and inter-modality features into consideration, we combine two representations and denote $\boldsymbol{\Psi}^m \in \mathbb{R}^{N_V \times (d+d)} = [\boldsymbol{Q}^m, \boldsymbol{V}^m]$ as the fusion embeddings. The operation of modality inference can be written as

$$\boldsymbol{Z}^m = \sigma(BN(\boldsymbol{\Psi}^m \boldsymbol{W} + \boldsymbol{B})), \tag{11}$$

where $\boldsymbol{Z}^m \in \mathbb{R}^{N_V \times d}$ is the inferred task-related item embedding of the m-th modality and $BN(\cdot)$ is the batch-normalization layer. $\boldsymbol{W} \in \mathbb{R}^{2d \times d}$ is the trainable weights and $\boldsymbol{B} \in \mathbb{R}^{N_V \times d}$ is the bias matrix. The activation function $\sigma(\cdot)$ is exploited to introduce the nonlinear factors. To ensure the effectiveness and robustness of the inference process, we reconstruct our inferred embedding $\boldsymbol{Z}^m$ and the raw item features $\overline{\boldsymbol{F}}^{V,m}$ among features of non-missing modalities in the original inputs. The loss of reconstruction is formulated:

$$\mathcal{L}_{rec} = \sum_{m=1}^{M} \left\| \boldsymbol{Z}_i^m - \overline{\boldsymbol{F}}_i^{V,m} \right\|_F^2, \tag{12}$$

where i is the index of non-missing modality features of the original inputs. The goal of reconstruction loss aims to ensure that the inferred values of the parts without missing modalities remain consistent with the original values.

3.3.2 Complete Multi-modality Recommendation. To incorporate the learned features of items into the recommendation framework, we combine the id embedding and item contents as the complete multi-modality embeddings:

$$\begin{aligned} \hat{\mathbf{X}}^U &= Concat(X^U, F^{U,1}, F^{U,2}, ..., F^{U,M}), \\ \hat{\mathbf{X}}^V &= Concat(X^V, Z^{V,1}, Z^{V,2}, ..., Z^{U,M}). \end{aligned} \tag{13}$$

We denote $\hat{\mathbf{X}}^U \in \mathbb{R}^{N_U \times Md}$ and $\hat{\mathbf{X}}^V \in \mathbb{R}^{N_V \times Md}$ as the final representations of users and items, where $Concat(\cdot)$ is a concatenation function. Then, the preference score $\hat{\mathbf{R}}$ can be predicted by $\hat{\mathbf{R}} = \hat{\mathbf{X}}^U (\hat{\mathbf{X}}^V)^T$ and the value $\hat{\mathbf{r}}_{ij}$ in $\hat{\mathbf{R}}$ means the probability of item j recommended to user i. For enhanced multimedia recommendation, we adopt the BPR loss, wich is a common loss function in recommendation tasks.

$$\mathcal{L}_{BPR} = \sum_{(i,j_p,j_n)}^{|\mathcal{E}|} -log(sigm(\hat{\mathbf{r}}_{ij_p} - \hat{\mathbf{r}}_{ij_n})), \tag{14}$$

where j_p and j_n denotes the positive and negative samples for user i. Finally, We train our recommender systems with the combination loss to jointly optimize CI^{2}MG:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_{rec} + \lambda_3 ||\Theta||^2. \tag{15}$$

The last term $||\Theta||^2$ is the weight-decay regularization against overfitting and $\lambda_1, \lambda_2, \lambda_3$ are all hyperparameters.

4 EXPERIMENTS AND ANALYSES

In this section, the effectiveness of our proposed CI^{2}MG is evaluated on four public multimedia datasets. First, we start with a brief description of conducted datasets and experimental settings. Then, we evaluate our proposed CI^{2}MG framework focusing on the following research questions:

- **RQ1:** How does CI^{2}MG perform in comparison with other state-of-the-art models for multimedia recommendation?
- **RQ2:** How does each component devised in the CI^{2}MG contribute to performance improvement?
- **RQ3:** How do the hyperparameters affect the prediction performance and how to choose optimal values?

4.1 Experimental Settings

4.1.1 Datasets. We use four real-world multi-modal recommendation datasets summarized in Table 1. **TikTok** comes from the short

Table 1: Statistics of experimented datasets with multimodal item Visual (V), Acoustic (A), and Textual (T) contents.

Dataset	Amazon				Tiktok			Allrecipes	
	Sports		Baby						
Modality	V	T	V	T	V	A	T	V	T
Embed Dim	4096	1024	4096	1024	128	128	768	2048	20
User	35598		19445		9319			19805	
Item	18357		7050		6710			10067	
Interactions	256308		139110		59541			58922	
Sparsity	99.961%		99.899%		99.904%			99.970%	

videos on the Tiktok platform. The videos' visual, acoustic, and textual features are considered as multi-modal features. The textual features are encoded with Sentence-Bert[25]. Two datasets Amazon-Baby and Amazon-Sports come from **Amazon**. The images and textual details of products are used to generate 4096-dimensional visual feature embeddings and textual feature embeddings. The textual features are also encoded with Sentence-Bert. **Allrecipes** comes from one of the largest recipe social networking sites which has 52, 821 recipes in 27 different categories. Images of each recipe are considered as visual features, and 20 ingredients are sampled as textual features.

4.1.2 Baselines. To illustrate the influences of missing modalities on the models and verify the effectiveness of our proposed framework in completing the missing modalities, the following representative state-of-the-art baselines for comparison can be divided into (1) collaborative filtering based methods without multimedia contents (MF-BPR, NGCF [29], LightGCN [13], SGL [35], NCL [15], and HCCF [37]), and (2) multimedia recommendation with missing modalities (LightGCN-*M*, MMGCL [39], SLMRec [27] and MMSSL [31]).

4.1.3 Hyperparameter Settings. We implement the proposed framework with Pytorch. We adopt AdamW[18] and Adam[14] as the optimizer for the generator. In particular, we set learning rate in {4.5e-4, 5e-4, 5.4e-3, 5.6e-3} and {2.5e-4, 3e-4, 3.5e-3}, the decay of L_2 regularization term in {1.2e-2, 1.4e-2, 1.6e-2}, the number of graph layer in {1, 2, 3, 4}. We set the embedding dimension to 64 for our model and other compared methods. The hyperparameters λ_1, λ_2 and λ_3 are in {1e-3, 1e-2, 1e-1, 1} and τ in contrastive learning ranges from [0,1].

4.2 Overall Performance Comparison (RQ1)

To verify the effectiveness of our proposed model, we consider the case of extreme modality missing, where the missing rate is set to 90%. We compare recommendation results of our proposed CI^{2}MG in Table 2. In general, our CI^{2}MG outperforms all baselines across all evaluation metrics.

Specifically, we compare the conventional methods of collaborative filtering without multimedia contents with our CI^{2}MG. However, we observe that in some cases, discarding the missing modalities may achieve an improvement. The performance gain of LightGCN only with historical interactions, on the Amazon-Baby dataset, is 28.16% regarding *R*@20, compared with LightGCN-*M* with multi-modalities. The results validate the missing modalities may affect the modeling of user-item correlations. Moreover, the

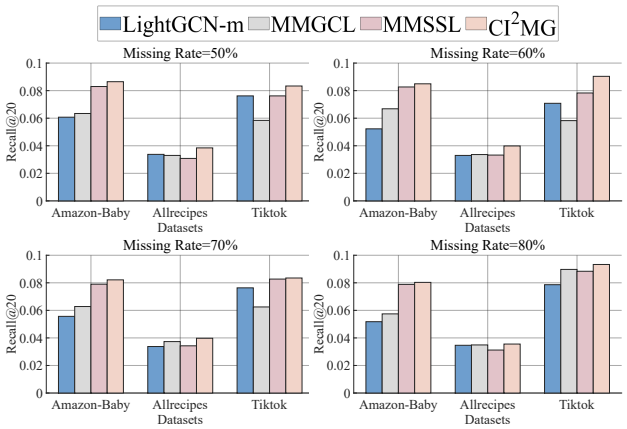

Figure 3: Performance about the comparison with SOTA methods with different missing rates for multimedia recommendation regarding Recall@20 of the CI^{2}MG on the Amazon-Baby, Allrecipes and Tiktok datasets.

ranking of many baselines is fluctuating across datasets as we see the second best performance scattered among different models like MMSSL and MMGCL. Compared with the second best performance, the performance gains of CI^{2}MG on the Amazon-Baby, Amazon-Sports, Allrecipes and Tiktok datasets range from reasonably large (3.71% achieved with *N*@20 on the Tiktok dataset) to significantly large (9.40% achieved with *R*@20 on the Amazon-Baby dataset).

Moreover, to provide a more comprehensive demonstration of our model in handling missing values on recommendation tasks, we present the experimental results with different missing rates, shown in Figure 3. From the observations, the proposed CI^{2}MG method can achieve best *Recall*@20 scores with all missing settings (50%, 60%, 70% and 80%) on all datasets, which is particularly evident for its effectiveness towards incomplete recommendation.

4.3 Ablation Experiment (RQ2)

To better understand the main designs, we closely study our framework by adding the components one by one. From Table 3, we have the following observations:

Firstly, our model only with Intra-modality generation outperforms LightGCN-*m* on the effectiveness metrics on both two modalities (e.g., Amazon-Baby) and more modality scenarios (e.g., Tiktok). Secondly, compared with missing modalities with only intra-modality features, injecting inter-modality information achieves performance gains on the effectiveness metrics by achieving 7.37% improvement in *R*@20, 13.88% improvement in *P*@20 and 6.36% improvement in *N*@20. The results show the effectiveness of our design for contrastive intra- and inter-modality generation. Thirdly, with a modality-aware contrastive learning to perform an augment between the intra-modality view and inter-modality view, the performance gains a slight improvement. Our contrastive learning brings about 4.54% increase in *R*@20, 4.88% increase in *P*@20 and 1.99% increase in *N*@20 on Tiktok datasets. Finally, we jointly perform our modality generation and multimedia recommendation,

Table 2: Performance (%) comparison of baselines with the 90% missing rate in terms of Recall@20, Precision@20 and NDCG@20 on Amazon-Baby, Amazon-Sports, Allrecipes and Tiktok multimedia datasets, where ∗ denotes a significant improvement according to the wilcoxon signed-rank test.

Baseline	Amazon-Baby			Amazon-Sports			Allrecipes			Tiktok		
	R@20	P@20	N@20	R@20	P@20	N@20	R@20	P@20	N@20	R@20	P@20	N@20
MF-BPR	4.40±0.19	0.24±0.007	2.00±0.17	4.30±0.19	0.23±0.006	2.02±0.13	1.37±0.15	0.07±0.004	0.53±0.15	3.46±0.17	0.17±0.003	1.30±0.18
NGCF	5.91±0.13	0.32±0.002	2.61±0.12	6.95±0.12	0.37±0.005	3.18±0.12	1.65±0.19	0.08±0.003	0.59±0.13	6.04±0.13	0.30±0.005	2.38±0.20
LightGCN	6.98±0.19	0.37±0.002	3.19±0.15	7.82±0.15	0.42±0.008	3.69±0.16	2.12±0.13	0.10±0.004	0.76±0.17	6.53±0.16	0.33±0.004	2.82±0.16
SGL	6.78±0.14	0.36±0.004	2.96±0.13	7.79±0.13	0.41±0.007	3.61±0.15	1.91±0.16	0.10±0.002	0.69±0.16	6.03±0.15	0.30±0.006	2.38±0.19
NCL	7.03±0.19	0.38±0.009	3.11±0.18	7.65±0.16	0.40±0.004	3.49±0.17	2.24±0.18	0.10±0.003	0.77±0.18	6.58±0.18	0.34±0.002	2.69±0.17
HCCF	7.05±0.17	0.37±0.008	3.08±0.16	7.79±0.17	0.41±0.008	3.61±0.14	2.25±0.12	0.11±0.004	0.82±0.12	6.62±0.16	0.29±0.006	2.67±0.13
LightGCN-*M*	5.29±0.26	0.28±0.001	2.24±0.09	4.27±0.15	0.23±0.005	2.48±0.12	3.38±0.15	0.17±0.002	1.34±0.11	6.82±0.10	0.34±0.001	2.83±0.20
MMGCL	5.70±0.24	0.31±0.003	2.57±0.19	6.90±0.14	0.37±0.003	3.28±0.13	3.82±0.19	0.19±0.001	1.70±0.07	5.84±0.11	0.29±0.005	2.59±0.11
SLMRec	7.01±0.19	0.39±0.007	3.21±0.12	7.87±0.18	0.40±0.006	3.77±0.16	3.28±0.12	0.15±0.003	1.41±0.11	7.11±0.17	0.32±0.002	4.25±0.09
MMSSL	7.78±0.19	0.41±0.002	3.41±0.09	8.16±0.12	0.43±0.001	3.81±0.07	3.35±0.11	0.17±0.006	1.51±0.08	7.64±0.13	0.38±0.008	4.42±0.07
Our	**8.51±0.17***	**0.45±0.005***	**3.69±0.08***	**8.65±0.10***	**0.46±0.001***	**3.98±0.08***	**4.12±0.10***	**0.21±0.005***	**1.85±0.06***	**8.03±0.15***	**0.41±0.007***	**4.58±0.07***
Improv.	9.40%	9.09%	8.21%	5.99%	5.84%	4.56%	7.69%	7.69%	9.11%	5.10%	7.32%	3.71%

Table 3: Ablation study results (%) on the intra-modality generation, inter-modality generation, modality-aware contrastive learning (Conl) and Task-guided modality generation on Amazon-Baby and Tiktok datasets.

Baseline	Amazon-Baby			Tiktok		
Metric	R@20	P@20	N@20	R@20	P@20	N@20
LighGCN-M	5.29	0.28	2.24	6.82	0.34	2.83
+ Intra	7.46	0.36	3.30	7.65	0.35	3.39
+ Inter	8.01	0.41	3.51	7.72	0.37	3.66
+ Conl	8.14	0.43	3.58	7.91	0.39	3.74
CI²MG	8.51	0.45	3.69	8.03	0.41	4.58

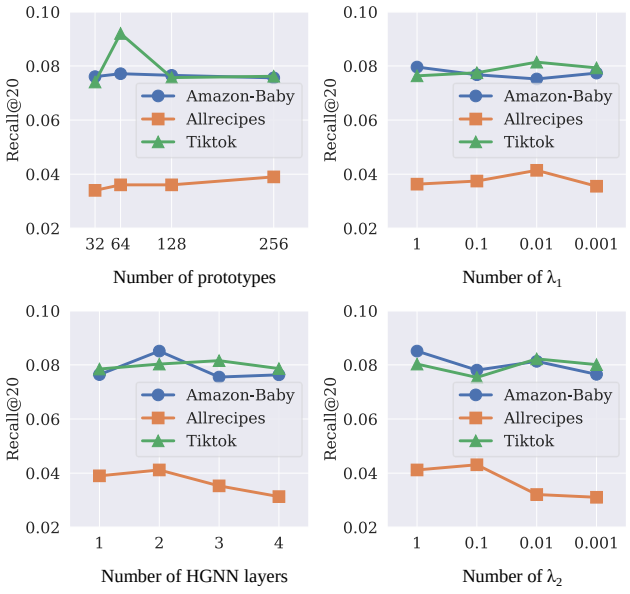

Figure 4: Performance of hyperparameter study regarding Recall@20 of the CI²MG framework with varying hyperparameters on Amazon-Baby, Allrecipes and Tiktok datasets.

to enhance the representations by task-related information. According to experimental results, our CI²MG can achieve the SOTA performance with an average 4.54% improvement over all metrics.

4.4 Hyperparameter Study (RQ3)

Our proposed CI²MG framework mainly introduces four hyperparameters, i.e., the number of prototypes K, HGNN layers L, the weights λ_1 and λ_2, respectively. Here we show how these four hyperparameters impact the performance.

From Figure 4, we can observe the following results: (1) K is the number of prototypes and we found that the optimal values of K are 64 and 256. In consequence, our CI²MG is sensitive to the prototype number K and the optimal parameters can be obtained by slight tuning. (2) L means the layers of HGNN, where the optimal values are about 2 layers. This experimental result illustrates that, like GNN, more message passing and aggregation of HGNN may also aggravate the data sparsity issue, where we set $L = 2$ to alleviate the over-smoothing issue. (3) λ_1 and λ_2 are both weights, which control the strength of the loss $\mathcal{L}_s$ and $\mathcal{L}_{rec}$. Firstly, we show the model performance with varying $\mathcal{L}_s$. λ_1 is the weight of contrastive learning loss $\mathcal{L}_s$, where the optimal values are 1 and 0.01. Because CI²MG is sensitive to λ_1, the setting $\lambda_1 = 0.01$ seems to be the rule-of-thumb. Secondly, for hyperparameter λ_2, the loss $\mathcal{L}_{rec}$ is designed to help the model jointly train modality generation and multimedia recommendation. The optimal values of λ_2 are 0.01, 0.1 and 1 and $\lambda_2 = 1.0$ seems to be the rule-of-thumb.

5 CONCLUSION

In this paper, we propose a novel framework named CI²MG. Specifically, we propose a novel contrastive intra- and inter-modality generation to complete the features of the missing modality. Besides, an enhanced multimedia recommendation is designed to obtain the task-related information. Extensive quantitative experiments demonstrate the clear advantages of our CI²MG over the state-of-the-art baselines of the recommendation.

ACKNOWLEDGMENTS

This work is in part supported by the Fujian Provincial Youth Education and Scientific Research Project under Grant JAT220811; the National Natural Science Foundation of China under Grants U21A20472 and 62276065; the National Key Research and Development Plan of China under Grant 2021YFB3600503.

REFERENCES

[1] Desheng Cai, Shengsheng Qian, Quan Fang, Jun Hu, and Changsheng Xu. 2022. Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network for Personalized Micro-video Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 581–590.

[2] Xianshuai Cao, Yuliang Shi, Jihu Wang, Han Yu, Xinjun Wang, and Zhongmin Yan. 2022. Cross-modal Knowledge Graph Contrastive Learning for Machine Learning Method Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3694–3702.

[3] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport. *Advances in Neural Information Processing Systems* 35 (2022), 39090–39102.

[4] Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. 2020. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 27–34.

[5] Zeming Chen and Qiyue Gao. 2022. Probing Linguistic Information For Logical Inference In Pre-trained Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10509–10517.

[6] Jae Won Cho, Dong-Jin Kim, Jinsoo Choi, Yunjae Jung, and In So Kweon. 2021. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1592–1601.

[7] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013).

[8] Xiaoyu Du, Zike Wu, Fuli Feng, Xiangnan He, and Jinhui Tang. 2022. Invariant Representation Learning for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 619–628.

[9] Hamidreza Fazlali, Yixuan Xu, Yuan Ren, and Bingbing Liu. 2022. A versatile multi-view framework for lidar-based 3d object detection with guidance from panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17192–17201.

[10] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3558–3565.

[11] Jun Fu, Chen Hou, Wei Zhou, Jiahua Xu, and Zhibo Chen. 2022. Adaptive Hypergraph Convolutional Network for No-Reference 360-degree Image Quality Assessment. In *Proceedings of the 30th ACM International Conference on Multimedia*. 961–969.

[12] Hao Geng, Deqing Wang, Fuzhen Zhuang, Xuehua Ming, Chenguang Du, Ting Jiang, Haolong Guo, and Rui Liu. 2022. Modeling Dynamic Heterogeneous Graph and Node Importance for Future Citation Prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 572–581.

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

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. 2022. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6949–6958.

[16] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM Web Conference 2022*. 2320–2329.

[17] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. 2022. Image-adaptive YOLO for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1792–1800.

[18] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[19] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2302–2310.

[20] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning Hybrid Behavior Patterns for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 376–384.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[22] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. 2021. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE transactions on pattern analysis and machine intelligence* 44, 10 (2021), 6839–6853.

[23] Srinivas Parthasarathy and Shiva Sundaram. 2020. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 400–404.

[24] Pengjiang Qian, Yizhang Jiang, Zhaohong Deng, Lingzhi Hu, Shouwei Sun, Shitong Wang, and Raymond F Muzic. 2015. Cluster prototypes and fuzzy memberships jointly leveraged cross-domain maximum entropy clustering. *IEEE transactions on cybernetics* 46, 1 (2015), 181–193.

[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[26] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1405–1414.

[27] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* (2022).

[28] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. 2020. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1828–1838.

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

[30] Wei Wei, Chao Huang, Lianghao Xia, Yong Xu, Jiashu Zhao, and Dawei Yin. 2022. Contrastive meta learning with behavior multiplicity for recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 1120–1128.

[31] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*.

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

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

[34] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2560–2564.

[35] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 726–735.

[36] Le Wu, Yonghui Yang, Kun Zhang, Richang Hong, Yanjie Fu, and Meng Wang. 2020. Joint item recommendation and attribute inference: An adaptive graph convolutional network approach. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 679–688.

[37] Lianghao Xia, Chao Huang, Yong Xu, Jiashu Zhao, Dawei Yin, and Jimmy Huang. 2022. Hypergraph contrastive collaborative filtering. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*. 70–79.

[38] Qi Yang, Sergey Nikolenko, Alfred Huang, and Aleksandr Farseev. 2022. Personality-Driven Social Multimedia Content Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7290–7299.

[39] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1811.

[40] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1545–1554.

[41] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022. M3Care: Learning with Missing Modalities in Multimodal Healthcare Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2418–2428.

[42] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[43] Xiaolin Zheng, Jiajie Su, Weiming Liu, and Chaochao Chen. 2022. DDGHM: Dual Dynamic Graph with Hybrid Metric Training for Cross-Domain Sequential Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 471–481.

[44] Lipeng Zhu and David J Hill. 2021. Data/model jointly driven high-quality case generation for power system dynamic stability assessment. *IEEE Transactions on Industrial Informatics* 18, 8 (2021), 5055–5066.