[1]

Unified Heterogeneous Hypergraph Construction for Incomplete Multimedia Recommendation

ZHENGHONG LIN, YANCHAO TAN* and JIAMIN CHEN, Fuzhou University, China
HENGYU ZHANG, Macquarie University, Australia
CHAOCHAO CHEN, Zhejiang University, China
SHIPING WANG, Fuzhou University, China
CARL YANG, Emory University, USA

In the dynamic environment of multimedia-sharing platforms like Twitter and TikTok, multimedia recommendation systems have been widely used to help users discover items of interest. However, traditional approaches often fall short, when the item modalities are incomplete, a common issue in real-world scenarios. To this end, we introduce the unified heterogeneous Hypergraph construction for Incomplete multimedia REcommendation (HIRE), a novel framework designed to jointly learn a heterogeneous hypergraph and perform accurate recommendations under incomplete scenarios. HIRE first initializes the unified heterogeneous hypergraph for modality completion and employs self-supervised learning aligned with the contrastive text-centered view for multimedia recommendation. Such integration effectively handles the challenges posed by incomplete modalities, leading to improved recommendation accuracy. Furthermore, we find that the hypergraph directly learned from the HIRE is a dense structure which can be inaccurate and coarse. Therefore, we devise the HIRE framework with Sparse constraint named HIRES, which uniquely integrates optimal transport and a $\ell_{2,1}$-norm to refine the hypergraph structure. Our extensive experiments across various datasets demonstrate the superiority of HIRES in addressing incomplete modalities, establishing it as a powerful tool for personalized multimedia recommendations.

CCS Concepts: • **Information systems → Recommender systems**.

Additional Key Words and Phrases: Incomplete Multimedia Recommendation, Heterogeneous Hypergraph, Sparse Constraint, Multimodal Representation Learning.

ACM Reference Format:
Zhenghong Lin, Yanchao Tan*, Jiamin Chen, Hengyu Zhang, Chaochao Chen, Shiping Wang, and Carl Yang. 2024. Unified Heterogeneous Hypergraph Construction for Incomplete Multimedia Recommendation. *ACM Trans. Inf. Syst.* 1, 1 (May 2024), 30 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

With the rapid growth of multimedia-sharing platforms such as Twitter and TikTok, multimedia recommender systems have been widely used in many online applications [35, 37, 38]. Benefiting from the available multimodal

[1]* Corresponding author.

Authors' Contact Information: Zhenghong Lin, Yanchao Tan*; Jiamin Chen, Fuzhou University, China, hongzhenglin970323@gmail.com, yctan@fzu.edu.cn, Jiamin020316@163.com; Hengyu Zhang, Macquarie University, Australia, hengyu.zhang3@hdr.mq.edu.au; Chaochao Chen, Zhejiang University, China, zjuccc@zju.edu.cn; Shiping Wang, Fuzhou University, China, shipingwangphd@163.com; Carl Yang, Emory University, USA, j.carlyang@emory.edu.

ACM 1558-2868/2024/5-ART
https://doi.org/XXXXXXX.XXXXXXX

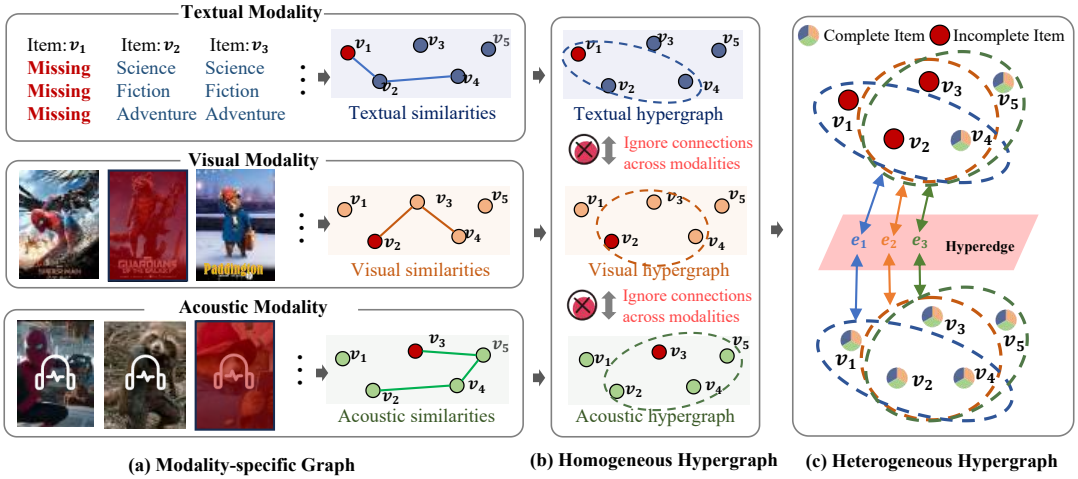


Fig. 1. A toy example of missing modalities on the TikTok multimedia-sharing platform. (a) Introduction of incomplete modalities and high-order similarities. (b) Illustration of unified hypergraph structure. A dotted circle represents a hyperedge in the hypergraph.

information like visual, textual, and acoustic contents, multimedia recommendation systems can obtain more accurate item and user representations compared to the general recommendation models when the interaction data is sparse.

To make good use of rich multimedia information from the online active users, some efforts have already been attempted to align the different modalities into a common latent space [63]. Recently, graph-based collaborative filtering [39] has been demonstrated as an effective mechanism for multimedia recommendations, where the core is to inject multimodal auxiliary information into the latent representations for users or items with user-item interactions. However, these methods often require complete modalities for each item, which are not readily available in every situation [65]. For example, as shown in Fig. 1 (a), the textual modality is absent when the user only want to upload a short video without textual descriptions; The visual modality is missing because users only left text contents in the comment of the multimedia platforms; The acoustic modality may be unavailable while the multimedia video with the enormous ambient noise during recording. In most real-life scenarios, traditional approaches often fall short or may not work well, when the real-incomplete conditions are not considered.

For deploying recommendation models having high performance under complete assumption to real-incomplete scenarios, a natural idea of completing missing modalities was proposed. The existing graph-based completing paradigm is to calculate the modality-specific similarity between nodes with or without missing values and complete the missing modality based on such pair-wise relations. For example, in Fig. 1(a), item v_1 with missing values is completed by the textual similarity with v_2. However, such local pair-wise (one-to-one) connections ignore higher-order global (set-to-set) relations, which is common in real-incomplete scenarios[36, 71]. Under real-world recommendations, the missing features may be similar to several complete features. In other words, the completion of incomplete modalities is a complex set-to-set problem (from complete set to missing set) [57]. To complete the missing values with set-to-set higher-order similarities, the hypergraph-based models are introduced into the collaborative filtering recommender systems. The hypergraph, composed of some hypernodes and hyperedges, is generalizing the concept of edges in graphs to hyperedges [72]. Since the hyperedges can

contain any number of nodes, we can use them to represent the set-to-set higher-order correlations of items in the different modalities, shown in Fig. 1 (b).

Although the hypergraphs have been successful in various domains, the well-designed hypergraph is either unavailable in every situation or artificially constructed, which is costly and time-consuming [15]. Therefore, given the interactions among users and items in incomplete multimedia recommendations and the incomplete multimodal representations, a natural question is: *Can we jointly construct a hypergraph structure and perform incomplete multimedia recommendations?*

To solve the above question, in this work, we introduce the unified heterogeneous Hypergraph construction for the Incomplete multimedia REcommendation, (HIRE), a novel framework designed to jointly learn a heterogeneous hypergraph and perform accurate recommendations under incomplete scenarios. The task is challenging in several perspectives.

Firstly, the existing paradigm of hypergraph generating is based on homogeneous higher-order connections [8], where the hypergraph structure is defined as single-modality similarities and the modality-specific hypergraph convolution is exploited to complete the incomplete missing values, shown in Fig. 1(b). However, such construction mechanism ignores the rich higher-order similarities across modalities under multimedia scenarios [13]. For example, item v_1 and v_2 have similar textual descriptions. Besides, item v_2 and v_3 belong to the same category when considering the visual modality. Across the bridge of the item v_2 across textual and visual modality, v_1 and v_3 may also have a high similarity. Different from the previous homogeneous hypergraph construction methods, we uniformly term the above relations across different modalities as heterogeneous higher-order similarity and seek to model such correlations into a unified heterogeneous hypergraph structure, which can be exploited to complete the missing modalities. Specifically, we proposed the unified heterogeneous hypergraph construction mechanism, where we formulate the hypergraph construction as a process of clustering, and perform the unified hypergraph convolution to complete the missing modalities shown in Fig. 1(c). In this way, we can group the similar items with the heterogeneous higher-order correlations into a unified structure. More details about the construction process are presented in Section 3.2.

Secondly, although the heterogeneous higher-order relations can explore the similarities across modalities to enhance the completion of missing values, most existing hypergraph-based recommendation methods often rely on the homogeneous connections in a single modality [48]. Therefore, how to perform the incomplete multimedia recommendation based on the constructed heterogeneous hypergraph structure is a non-trivial problem, where we need to maintain the existing well-explored homogeneous connections and inject the heterogeneous relationships into the recommendation framework simultaneously. Inspired by the recent success of language models in heterogeneous representation learning, our solution is to design a novel contrastive multimodel recommendation module, which contains a self-supervised contrastive mechanism aligned with the textual modality and an enhanced multimedia recommendation. Specifically, we perform contrastive learning to align the different homogeneous and heterogeneous relations with the textual view. Then, we inject the aligned multimodal relations into the id-based collaborative filtering recommendation framework. More details about the incomplete multimedia recommendation is presented in Section 3.3.

Finally, we find that the hypergraph directly constructed from the clustering mechanism in HIRE is a dense matrix, which can be inaccurate and coarse for incomplete scenarios, since the assumption that items with missing values may introduce more noise than complete items. However, the standard optimization objectives of such hypergraph structure are not designed towards the learning under incomplete scenarios, where such dense hypergraph structure may also assign the weights to incomplete nodes and it may enhance the incomplete noises with the increase of convolution layers of hypergraph. In light of this, we propose to leverage the HIRE framework with Sparse constrain named HIRES to reduce the unreliable interactions of items with missing modalities. Specifically, we devise a novel sparse optimal transport framework, which uniquely integrates optimal transport and a $\ell_{2,1}$-norm constraint to refine the hypergraph structure. Then, to obtain the optimal solution for

the proposed sparse optimal transport, we design a differentiable optimization strategy, calibrating the gradient by the Frank-Wolfe algorithm. More details about the optimization are presented in Section 4.

We evaluate both HIRE and HIRES with extensive experiments on four real-world benchmark datasets for incomplete recommendations. We compare them with 20 comprehensive methods focusing on state-of-the-art collaborative filtering and incomplete multimedia recommendation methods. Extensive experimental results show that HIRES is able to significantly improve the recommendation overall baselines (e.g., with up to 9.09% improvements in P@20 on the sports dataset over the best baseline).

In summary, we mainly make the following contributions:

- *Formulation of unified heterogeneous hypergraph construction*: HIRES and HIRE are the first incomplete multimedia recommendation framework with a unified hypergraph structure, which can capture the higher-order similarities across modalities to complete the missing modalities.
- *Effective model designs*: In HIRE, we exploit the clustering mechanism and textual-aligned self-supervised mechanism to jointly construct the unified heterogeneous hypergraph and perform the enhanced incomplete multimedia recommendations. In HIRES, we devise a novel sparse transport mechanism to constrain the hypergraph structure for items with missing modalities.
- *Extensive experiments on four real-world datasets*: We conduct extensive experiments on four real-world datasets, which demonstrate significant improvements of the proposed HIRES framework on incomplete multimedia recommendation with highly accurate and interpretable results of unified heterogeneous hypergraph structure.

2 Related work

2.1 Multimedia Recommendation

Collaborative filtering (CF), which achieves relatively high performance with graph neural networks, has been widely used in the recommender system. For example, NCL [32] explicitly captured the potential node relatedness into contrastive learning for the graph collaborative filtering. MILK [1]designed a cross-modality alignment module to keep semantic consistency from pretrained multimedia item features. MICRO [70] designed a novel modality-aware structure learning module to learn item-item relationships for each modality. GCCF [7] modeled the user preference by the residual preference prediction and the linear embedding propagation. GDSRec [6] treated the biases as vectors and fused them into the process of learning user and item representations. Although the above methods achieve promising performance, the learned representations that rely on user-item interactions only are limited by data sparsity. Consequently, many studies have incorporated multimodal information, which can alleviate the data sparsity problem. LATTICE [69] leveraged graph structure learning to discover latent item relationships underlying multimodal features. MKGAT [46] introduced the multi-modal knowledge graph to the recommendation system innovatively. Recently, some works have attempted to construct item graph based on similarity within every modality to inject the multi-modal information into the multimedia recommendation. MICRO [70] is proposed to model item-item relationships and conduct fine-grained multimodal fusion with a modality-specific graph to inject the multimodal high-order relations into the item representations. FREEDOM [73] frozen the item-item graph during the training of multimodal process and provide a tighter upper bound on the graph spectrum.

However, all the above methods based on item-item graphs assume that each modality is complete and all the modalities are always available, while in the real-world applications, the missing modality scenarios are more common. When deploying models to more realistic incomplete recommendation scenarios, such item-graph-based mechanisms may fail or fall short because the graph cannot be constructed among missing modalities. Besides, previous graph methods only capture the multimodal high-order relations within a specific modality, while ignoring the relations across modalities, which may hinder the full potential of multimedia recommendations. Therefore, different from the existing methods, our proposed HIRE and HIRES framework are the first incomplete

multimedia recommendation framework with a unified hypergraph structure, which can capture the relations across different modality jointly.

2.2 Incomplete Multi-modal Models

In order to solve the problem of missing modalities in the multimodal area, many previous work directly excluded missing modalities and achieved some performance improvements. For example, Wang et el. [51] proposed a framework based on knowledge distillation, utilizing the supplementary information from all modalities, and avoiding imputation and noise associated with it. ModDrop++ [34] has been applied to MS lesion segmentation to achieve the state-of-the-art performance with missing modality. However, these methods may lose valuable multimodal information by discarding missing modalities. Consequently, many studies that aim to generate missing modalities have been proposed. LIDO [21] investigated the problem of how to reconstruct the topology of a diffusion network with incomplete observations of the node infection statuses. Lin et al. [31] proposed a contrastive intra- and inter-modality generation for enhancing incomplete multimedia recommendation, which alleviated the challenge with missing modalities. Zhou et al. [75] designed anovel self-supervised learning framework BM3 for multi-modal recommendation, removed the requirement of randomly sampled negative examples in modeling the interactions between users and items. Sun et al. [45] described a balance-guided approach for incomplete multi-view spectral clustering, which aids in handling the discrepancies across different views. Lin et al. [30] introduced a dual contrastive prediction model for incomplete multi-view representation learning, which provides a robust framework for dealing with partial data. Similarly, Huang et al. [19] developed an incomplete multi-view clustering network using nonlinear manifold embedding and a probability-induced loss, which improves clustering performance by effectively managing incomplete views.

Different from the above methods, we model the high-order similarities through a unified hypergraph structure, which can be jointly refined by the multimedia recommendation task for completing the missing modalities.

2.3 Hypergraph Neural Networks

Hypergraph has attracted tremendous attention due to its effectiveness in modeling higher-order interactions. Hypergraph Neural Network (HGNN) [10], presented for data representation learning, which can encode high-order data correlation in a hypergraph structure. DHLCF [29] constructed the unprovided hypergraph structures based on the collaborative filtering user-item interactions and proposed an adaptive lightweight neural network to inject the high-order relations of hypergraphs. However, the hypergraph structure is the basis of the hypergraph convolution operation, which is not always available. Consequently, some previous studies solved the problem of the lack of hypergraph incidence matrix by generating hypergraphs. For example, MHCN [64] worked on multiple motif-induced hypergraphs to enhance the social recommendation by leveraging high-order user relations. HSL [3] learned an informative and concised hypergraph structure that is optimized for downstream tasks. QHGN [17] constructed hypergraphs based on the visual objects detected in the video. Gong et al. [16] considered the problem of embedding a hypergraph into the low-dimensional Euclidean space so that most interactions are short-range.

Different from the above method of learning the hypergraph structure, we construct the hypergraph structure with the multimodal information from user preference and we introduce the priori knowledge tailored for incomplete scenarios. Besides, we also design a sparsity optimization to remove the noise information in the hypergraph.

2.4 Sparse Constraint

Sparse constraint mechanisms have emerged as crucial components in optimization. Carreira-Perpinan and Idelbayev [4] proposed an alternative formulation for the constrained optimization problem using “auxiliary

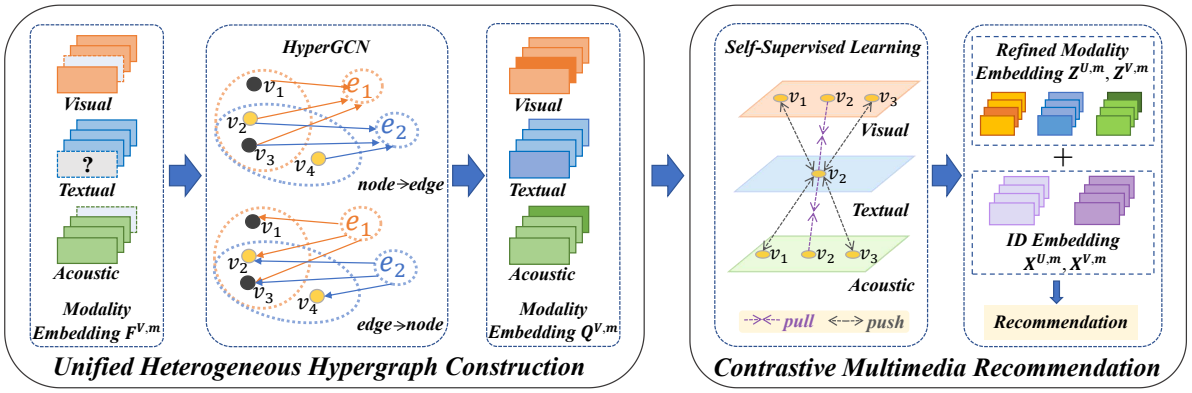


Fig. 2. The overall architecture about HIRES. In the left part, a heterogeneous hypergraph is employed to capture the high-order relations among multiple modalities. In the right part, the Enhanced Multimedia Recommendation Module is designed to jointly perform multimedia recommendations and optimize heterogeneous hypergraph construction.

variables". However, their empirical evaluation was limited to small-scale models and datasets. In order to avoid fine-tuning pruning rates separately for each layer in the network, ProbMask [74] automatically determined the amount of weight redundancy through sparse constraint. Lemaire et al. [26] adopted budgeted regularization for pruning to handle the constrained issue. Gallego-Posada et al. [12] put emphasis on training models with limited levels of sparsity, enabling the training of sparse neural networks using constrained regularization.

Furthermore, the imposition of sparsity constraints has significantly contributed to optimizing the efficiency of the acquired graph structures. For example, Subbareddy et al. [44] considered a new sparsity based graph learning model to calculate the eigenvectors of the graph matrix and determine the eigenvalues based on the eigenvectors obtained. Kuroda et al. [25] proposed sparse regularization as a convex method for sparse reconstruction of graphstructured data. Similarly, we optimize hypergraph structure with sparse constraint to perform more accurate multimedia recommendation under incomplete scenarios.

Different from the existing sparse constraint to denoise the process of optimization, our proposed sparse optimization is a modality-aware process, which is tailored for incomplete multimedia recommendation scenarios.

3 The HIRE Framework

In this section, we will present the main designs of our proposed HIRE framework and discuss the details of each component. Firstly, we give the problem statement and show the overall architecture of our proposed HIRE. Secondly, a unified heterogeneous hypergraph construction mechanism is designed to inject the heterogeneous information into node representations under scenarios with missing modalities. Finally, we introduce the contrastive multimodel recommendation module to align the representations under each modality with the text-centered self-supervised contrastive learning mechanism to jointly enhance the hypergraph structure learning and multimedia recommendation.

3.1 Problem Statement and HIRE Overview

The goal of our proposed HIRE framework is to jointly construct a unified heterogeneous hypergraph and perform the accurate multimodal recommendation based on the unified hypergraph structure under incomplete scenarios. Assume we have initially historical user-item interaction graph, containing user node set $\boldsymbol{U}$ with N_U users and

Table 1. Notations and Definitions.

Notations	Descriptions
N_V	The numbers of items
N_U	The numbers of users
K	The number of hyperedges group
$\boldsymbol{c}_i$	The mean (i.e., centroid) of points $\boldsymbol{j}$ belonging $\boldsymbol{H}_i$
$\boldsymbol{Q}^{m,(l)}$	The m-th modal complete embedding with hypergraph convolution of l layers
$\boldsymbol{W}$	The trainable weight of hypergraph neural networks
$\boldsymbol{Q}^{m,(0)}$	The zero-order representation obtained by item features $\boldsymbol{F}^{V,m}$
$\boldsymbol{q}_i^T, \boldsymbol{q}_i^V, \boldsymbol{q}_i^A$	The i-th item embedding of hypergraph representation $\boldsymbol{Q}^m$
τ	The temperature hyperparameter
$\boldsymbol{X}^{U,(l)} \in \mathbb{R}^{N_U \times d}$, $\boldsymbol{X}^{V,(l)} \in \mathbb{R}^{N_V \times d}$	The ID-corresponding embedding of users and items in the l-th layer
$\sigma(\cdot)$	The activation function to introduce the nonlinear factors
$\boldsymbol{\Theta}^{U,(l)}, \boldsymbol{\Theta}^{V,(l)}$	The trainable user and item weights of the l-layer for GNNs
$\hat{\boldsymbol{X}}^U \in \mathbb{R}^{N_U \times Md}$,$\hat{\boldsymbol{X}}^V \in \mathbb{R}^{N_V \times Md}$	The final representations of users and items
$\boldsymbol{Z}^{\boldsymbol{V},\boldsymbol{m}}$	The m-th modal complete embedding in equation 4
$\|\|\Theta\|\|^2$	The weight-decay regularization against over-fitting
$\boldsymbol{H}$	The probability of joint distribution between item embeddings and hyperedges
$\ell_{2,1}$	The ℓ_2-norm for the column and the ℓ_1-norm for the row
$\mathbf{1}_K$ or $\mathbf{1}_{N_V}$	The vector of ones to calculate the sum of row or column in hypergraph structure
$< \cdot, \cdot >$	The Frobenius dot-product
$\boldsymbol{M}^U$	The matrix stands for the cost of transport.
$\boldsymbol{M}^U$	The formulation of cost matrix obtained by hypernode embedding
$\boldsymbol{z}_i^{deg^-}$	The indegree of i-th item
$\boldsymbol{z}_i^{deg^+}$	The outdegree of i-th item
$\boldsymbol{z}_i^{int}$	The number of interacted modalities (complete modalities) by item i
$\text{vec}(\cdot)$	The process of vectorizing a matrix
$\boldsymbol{s} \in \mathbb{R}^{NK}$	The optimization variable
$\boldsymbol{As} = \boldsymbol{b}$	The equality constraint
$\boldsymbol{Fs} \leq 0$	The inequality constraint
$\boldsymbol{\mu}$	The equality constraint
$\boldsymbol{\lambda} \geq 0$	The dual variables on the inequality constraint
θ	The problem parameter relates to the earlier layers
$J_\theta \tilde{\mathbf{s}}$	The partial Jacobian of $\tilde{\mathbf{s}}$ with the respect to θ
$\boldsymbol{h}^{(k)}$	The embedding at the k-th iteratio
$\mathcal{L}_{FW}$	The loss of sparse optimization for users and items hypergraphs
$\mathcal{L}_s$	The loss of contrastive learning formulated in Equation 4
$\mathcal{L}_{BPR}$	The loss of recommendation tasks formulated in Equation 7
λ	The hyperparameter to control the weights of the loss

item node set $\boldsymbol{V}$ with N_V items. Then, we can use ranking matrix $\boldsymbol{R} \in \mathbb{R}^{N_U \times N_V}$ to define the user-item graph interactions. 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, each item contains M modalities and we use the visual, acoustic, and textual modalities in our experimental setting. ($M = 3$). We define $\boldsymbol{f}_i^{V,m} \in \mathbb{R}^d$ 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 $\boldsymbol{F}^{V,m} = \{\boldsymbol{f}_1^m; \cdots; \boldsymbol{f}_{N_V}^m\}$, where we assign the values of missing modalities to zero. To construct the heterogeneous hypergraph for capturing the heterogeneous information under incomplete modalities, we first initialize the hypergraph structure $\boldsymbol{H}$, where $\boldsymbol{H} \in \mathbb{R}^{N_V \times K}$and K is the number of hyperedge group. After hypergraph convolution with unified hypergraph

structure $\boldsymbol{H}$, we use the $\boldsymbol{Q}^m \in \mathbb{R}^{N_V \times d}$ to represent the hypernode embedding in m-th modalities with self-supervised learning. Finally, we can obtain a prediction probability matrix $\hat{\boldsymbol{R}}$, where the value of $\hat{\boldsymbol{R}}_{ij}$ represents the probability that the j-th item is recommended to the i-th user. In summary, the input and output are defined as follows:

- *Input:* user-item interaction matrix $\boldsymbol{R}$ and incomplete modality features $\boldsymbol{F}^{V,m}$.
- *Output:* a prediction probability matrix $\hat{\boldsymbol{R}}$ for incomplete multimedia recommendation.

We summarize the main components of the HIRE framework in Fig. 2 and provide an overview about the architecture. Our proposed model has two main parts: (1) In the Heterogeneous Hypergraph Construction Module, we initialize the hypergraph structure by K-means algorithm. Then, we exploit a unified heterogeneous hypergraph convolution mechanism in order to complete the missing multimodal features by high-order relations across different modalities. (2) In the Contrastive Multimedia Recommendation Module, we propose a contrastive multimedia recommendation mechanism containing a novel self-supervised mechanism aligned by the textual-modality view, and perform the multimedia recommendation based on the enhanced multimodal representations.

3.2 Heterogeneous Hypergraph Construction

Inspired by the recent success of hypergraphs in various domains, our approach seeks to construct a hypergraph structure containing multiple modalities. The key limitation of constructing the well-designed structure is how to inject the heterogeneous high-order relations [5] across modalities into the unified hypergraph. Benefiting from clustering methods in heterogeneous scenarios [66], we innovatively propose the use of category (clustering centroid) representation to define higher-order correlations across modalities. Our motivation comes from the assumption that items from different modalities may share the same category, e.g. items of the textual football description and visual basketball image all belong to the "Sports" category. Therefore, to construct a unified hypergraph that captures higher-order information in order to complete missing modalities, we design the heterogeneous hypergraph construction module. Specifically, we first initialize the hypergraph structure $\boldsymbol{H} \in \mathbb{R}^{N_V \times K}$ by K-means algorithm, where N_V is the numbers of items and K is the number of hyperedges group.

$$\min_{\boldsymbol{H}} \frac{1}{2} \sum_{k=1}^{K} \sum_{j \in \boldsymbol{H}_i} ||\boldsymbol{F}_j - \boldsymbol{c}_i||_2^2. \tag{1}$$

Here, K-means algorithm to cluster the items with similar characteristics into K groups by clustering mechanism, which can be seen as K hyperedges. $\boldsymbol{F}$ is the joint embedding and can be defined as $\boldsymbol{F} = Concat(\boldsymbol{F}^1, \boldsymbol{F}^2, ..., \boldsymbol{F}^M)$, where $\boldsymbol{F} \in \mathbb{R}^{N_V \times Md}$ and $Concat(\cdot)$ is the concatenation function. Set $\boldsymbol{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_K\}$ represents the unified heterogeneous hypergraph structure with K item-item hyperedges. Besides, the clustering centroid $\boldsymbol{c}_i$ means the embedding of i-th hyperedge, where the value of $\boldsymbol{H}_{ji}$ represents the hypernode $\boldsymbol{j}$ belonging $\boldsymbol{c}_i$.

Then, we exploit the general hypergraph convolutional networks to capture the heterogeneous relations for completing representations of missing modalities:

$$\boldsymbol{Q}^{m,(l+1)} = \sigma(\boldsymbol{D}^{-1}\boldsymbol{H}\boldsymbol{W}\boldsymbol{B}^{-1}\boldsymbol{H}^T\boldsymbol{Q}^{m,(l)}), \tag{2}$$

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

3.3 Contrastive Multimedia Recommendation

Inspired by the recent success of language model, we first perform contrastive learning between different modalities to refine the multimodal representations, aligned with the textual view. The above process can be written as follows:

$$
\begin{aligned}
l(\boldsymbol{q}_i^V, \boldsymbol{q}_i^T, \boldsymbol{q}_i^A) \;=\; & \underbrace{e^{\theta(q_i^T, q_i^V)/\tau}}_{\text{T-V pos}} + \underbrace{e^{\theta(q_i^T, q_i^A)/\tau}}_{\text{T-A pos}} \\
+ & \underbrace{\sum_{k\neq i} e^{\theta(q_i^T, q_k^T)/\tau}}_{T-T\ neg} + \underbrace{\sum_{k\neq i} e^{\theta(q_i^T, q_k^V)/\tau}}_{T-V\ neg} + \underbrace{\sum_{k\neq i} e^{\theta(q_i^T, q_k^A)/\tau}}_{T-A\ neg},
\end{aligned} \tag{3}
$$

$$
\min \mathcal{L}_s = \min \log \frac{e^{\theta(q_i^T, q_i^V)/\tau} + e^{\theta(q_i^T, q_i^A)/\tau}}{l(\boldsymbol{q}_i^V, \boldsymbol{q}_i^T, \boldsymbol{q}_i^A)}, \tag{4}
$$

where $\boldsymbol{q}_i^T$, $\boldsymbol{q}_i^V$ and $\boldsymbol{q}_i^A$ are the i-th item embedding of hypergraph representation $\boldsymbol{Q}^m$, τ is the temperature hyperparameter. Where pos means positive and neg means negative, T means Textual, V means Visual and A means Acoustic.

Then, inspired by the recent success of id-based collaborative filtering paradigm in recommendation, we perform the ID-corresponding aggregation through the user-item interactions over and combine the representations among user and item neighbors:

$$
\begin{aligned}
\boldsymbol{X}^{U,(l+1)} &= \sigma(\boldsymbol{D}^{U^{-1}} \boldsymbol{R} \boldsymbol{X}^{V,(l)} \boldsymbol{\Theta}^{U,(l)}), \\
\boldsymbol{X}^{V,(l+1)} &= \sigma(\boldsymbol{D}^{V^{-1}} \boldsymbol{R}^T \boldsymbol{X}^{U,(l)} \boldsymbol{\Theta}^{V,(l)}).
\end{aligned} \tag{5}
$$

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

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:

$$
\begin{aligned}
\hat{\boldsymbol{X}}^{\boldsymbol{U}} &= Concat(\boldsymbol{X}^{\boldsymbol{U}}, \boldsymbol{Z}^{\boldsymbol{U},\boldsymbol{1}}, \boldsymbol{Z}^{\boldsymbol{U},\boldsymbol{2}}, ..., \boldsymbol{Z}^{\boldsymbol{U},\boldsymbol{M}}), \\
\hat{\boldsymbol{X}}^{\boldsymbol{V}} &= Concat(\boldsymbol{X}^{\boldsymbol{V}}, \boldsymbol{Z}^{\boldsymbol{V},\boldsymbol{1}}, \boldsymbol{Z}^{\boldsymbol{V},\boldsymbol{2}}, ..., \boldsymbol{Z}^{\boldsymbol{V},\boldsymbol{M}}).
\end{aligned} \tag{6}
$$

We denote $\hat{\boldsymbol{X}}^U \in \mathbb{R}^{N_U \times (M+1)d}$ and $\hat{\boldsymbol{X}}^V \in \mathbb{R}^{N_V \times (M+1)d}$ as the final representations of users and items, where $Concat(\cdot)$ is a concatenation function. $\boldsymbol{Z}^{\boldsymbol{V},\boldsymbol{m}}$ is the complete embedding of the m-th modality after self-supervised learning in equation 4. We denote $\boldsymbol{Z}^{\boldsymbol{U},\boldsymbol{m}} = \boldsymbol{D}^{U^{-1}} \boldsymbol{R} \boldsymbol{Z}^{\boldsymbol{V},\boldsymbol{m}}$, here, $\boldsymbol{D}^U \in \mathbb{R}^{N_U \times N_U}$ and is the diagonal degree matrix of user-item and item-user interaction matrices $\boldsymbol{R} \in \mathbb{R}^{N_U \times N_V}$ and $\boldsymbol{R}^T \in \mathbb{R}^{N_V \times N_U}$. Then, we exploit the Multilayer Perceptron (MLP) to project the fusion representations after concatenation onto a common latent subspace. The formulation of the above process can be written as $\tilde{\boldsymbol{X}}^U = \sigma(\hat{\boldsymbol{X}}^U \hat{\boldsymbol{W}}^U + \hat{\boldsymbol{B}}^U)$ and $\tilde{\boldsymbol{X}}^V = \sigma(\hat{\boldsymbol{X}}^V \hat{\boldsymbol{W}}^V + \hat{\boldsymbol{B}}^V)$, where $\hat{\boldsymbol{W}}^U \in \mathbb{R}^{Md \times d}$, $\hat{\boldsymbol{W}}^V \in \mathbb{R}^{Md \times d}$ are the projection weights and $\hat{\boldsymbol{B}}^U \in \mathbb{R}^{N_U \times d}$, $\hat{\boldsymbol{B}}^V \in \mathbb{R}^{N_V \times d}$ are the bias in MLP. The preference score $\hat{\boldsymbol{R}}$ can be predicted by $\hat{\boldsymbol{R}} = \tilde{\boldsymbol{X}}^U (\tilde{\boldsymbol{X}}^V)^T$, where $\tilde{\boldsymbol{X}}^U \in \mathbb{R}^{N_U \times d}$ and $\tilde{\boldsymbol{X}}^V \in \mathbb{R}^{N_V \times d}$. The value $\hat{\boldsymbol{r}}_{ij}$ in $\hat{\boldsymbol{R}}$ means the probability of item j recommended to user i. For enhanced multimodal recommendation, we adopt the BPR loss, wich is a common loss function in recommendation tasks.

$$
\mathcal{L}_{BPR} = \sum_{(i,j_p,j_n)}^{|\mathcal{E}|} -log(sigm(\hat{\boldsymbol{r}}_{ij_p} - \hat{\boldsymbol{r}}_{ij_n})), \tag{7}
$$

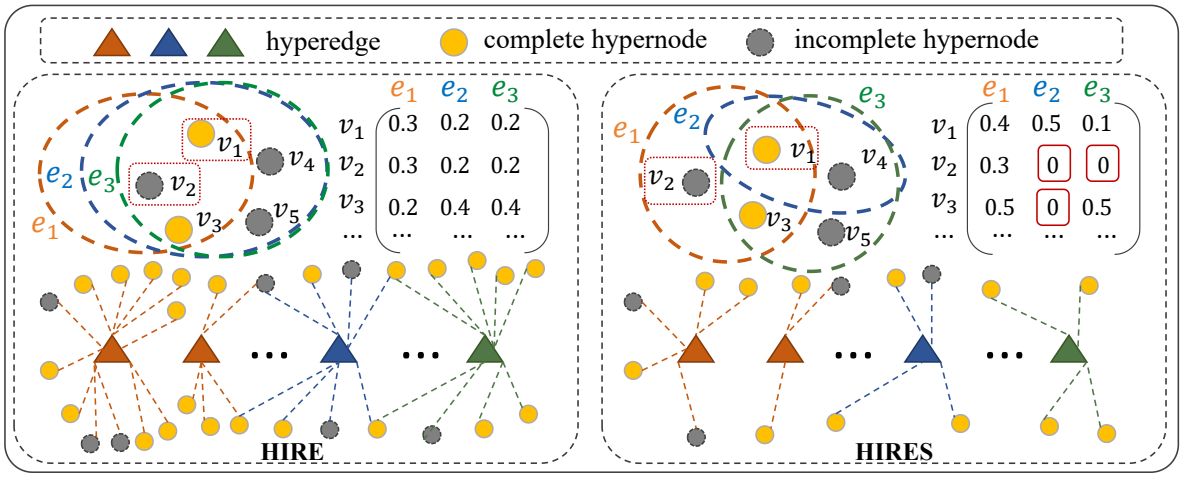


Fig. 3. An illustration of sparse optimization. Dashed circles are hyperedges, and solid circles are hypernodes. The red dashed rectangles represent the two hypernodes belonging to different categories.

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 HIRE:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda \mathcal{L}_s \tag{8}$$

Where λ is the hyperparameter.

4 HIRE with Sparse Constraint (HIRES)

Benefiting from the heterogeneous clustering mechanism in Equation 1, HIRE can essentially capture the cross-modality high-order relations to complete missing modalities for multimedia recommendation. However, the above unified hypergraph directly clustered from the original item representations cannot take prior knowledge under incomplete scenarios into account, which may hinder the full potential of incomplete multimedia recommendations. For example, shown in HIRE of Fig. 3, the item v_1 contains the complete modalities and abundant interactions while v_2 has sparser interactions, meanwhile, with the absence of acoustic and textual modalities. Therefore, v_1 may have more information to enhance the incomplete multimedia recommendation but v_2 may have more incomplete information with noise, deepened with the aggregation of hypergraph-based message mechanisms. Therefore, the dense clustering structure can be inaccurate and coarse for incomplete scenarios [33]. In this way, we hope to obtain a sparse unified heterogeneous hypergraph structure to reduce the unreliable interactions of items with missing modalities.

In order to deal with the above problem, we design the novel framework of **HIRE** with **S**parse constraint (**HIRES**) shown in Fig. 3, inspired by the sparse regularization. Our goal of HIRES is to remove the irrelevant relations within the hypergraph by assigning the zero value to noisy hypernodes in the hypergraph structure matrix, where the more noise relationships (the more zero values) are removed, the sparser the hypergraph structure becomes. Specifically, to design the details of HIRES, we replace the clustering-based heterogeneous hypergraph construction in Section 3.2 with a sparse optimal transport mechanism. Furthermore, to obtain the optimal solution for the proposed sparse optimal transport, we design a differentiable optimization strategy, calibrating the gradient by the Frank-Wolfe algorithm [11].

4.1 Sparse Optimal Transport Framework

Different from unconstrained mechanisms in clustering, the inspiration of HIRES comes from two straightforward insights. **The first insight is modality-driven constrain:** items (row in the hypergraph structure) with complete modalities are more trustworthy than those with missing values. **The second insight is interaction-driven constrain:** items with abundant interactions should have a large impact on incomplete multimedia recommendations, since such items may have richer semantic information for recommender systems to complete the missing modalities.

For **modality-driven constrain:** to reduce the weight of low-reliability items with missing modalities, we exploit sparse regularization as the constraint to properly rearrange the hypergraph structure. Compared with the existing unsupervised sparse constraints, e.g., ℓ_1-norm and ℓ_2-norm, we exploit the $\ell_{2,1}$-norm, which constrains row sparsity in the hypergraph structure, since we hope to make the items with missing modalities (row in the hypergraph structure) have fewer interactions with hyperedges. The definition of $\ell_{2,1}$-norm is shown:

$$\|\boldsymbol{H}\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{K} (\boldsymbol{H}_{ij})^2}, \tag{9}$$

where K is the number of hyperedges setting to control the column constraint and N is the number of items setting to control the row sparsity. The hypergraph structure $\boldsymbol{H}$ represents a learnable matrix under the column and row constraints, and we can use $\boldsymbol{H}_{ij}$ to measure the similarity of i-th and j-th probability distribution, the hypernode distribution i and hyperedge distribution j. In other words, $\ell_{2,1}$ is equivalent to calculating the column sparsity of ℓ_2-norm first and then finding the row constraint of ℓ_1-norm. To balance the impact on both item and hyperedge in the structure, we propose to leverage the optimal transport technique with row sparse regularization to ensure a reliable hypergraph structure. The formulation of above process can be written:

$$\begin{aligned} &\min_{\boldsymbol{H}\in\Delta} J = < \boldsymbol{H}, \boldsymbol{M} > + \eta \|\boldsymbol{H}\|_{2,1} \\ &s.t. \Delta = \{\boldsymbol{H} \in \boldsymbol{R}_{+}^{N_V \times K} | \boldsymbol{H}\mathbf{1}_K = \frac{\mathbf{1}_N}{N_V}, (\boldsymbol{H})^T \mathbf{1}_{N_V} = \frac{\mathbf{1}_K}{K}\}. \end{aligned} \tag{10}$$

Here, η is a hyperparameter to control the strength of the sparsity setting to 1. Δ is the constraint condition in optimal transport. The symbols of $\mathbf{1}_K$ or $\mathbf{1}_{N_V}$ are the vector of all ones with the K-dimensional and N_V-dimensional number, which can be seen as the operator to calculate the sum of row and column in hypergraph structure $\boldsymbol{H}$ and $< \cdot, \cdot >$ is the Frobenius dot-product to calculate the total transport costs. The cost matrix $\boldsymbol{M}$ represents represents the cost per unit distance. To obtain the cost matrix $\boldsymbol{M}$, we can use metric (e.g. the cosine similarity or mean square error) to calculate the distance between hypernodes and hyperedges for $\boldsymbol{M}$. Here, we use the $L2$ distance to obtain cost matrix $\boldsymbol{M}$ and the formulation can be calculated by measuring the distance between hypernode embedding $\boldsymbol{F}$ and hyperedge embedding $\boldsymbol{E}$ by

$$\boldsymbol{M}_{ij} = ||\boldsymbol{F}_i - \boldsymbol{E}_j||_2^2, \tag{11}$$

where $\boldsymbol{M}_{ij}$ can be also seen as the similarity between hypernode i and hyperedge j and a small value of $\boldsymbol{M}_{ij}$ means that the i-th item has higher similarity belonging to the j-th corresponding hyperedge. i and j represents the row of $\boldsymbol{F}$ and column of $\boldsymbol{E}$, respectively. $\boldsymbol{E}$ is initialized from the Xavier distribution [59].

For **interaction-driven constrain**, to assign large weights to the items with rich interactions, we use the degree of the node as the measurement for interactions, which can be formulated as $\boldsymbol{z}_i^{deg} = \boldsymbol{z}_i^{deg^-} + \boldsymbol{z}_i^{deg^+}$. Here, $\boldsymbol{z}_i^{deg^-}$ and $\boldsymbol{z}_i^{deg^+}$ are indegree and outdegree of i-th item, respectively. Besides, we also consider the item-modality interactions and use the $\boldsymbol{z}_i^{int}$ to represent the number of interacted modalities (complete modalities) by item i, the value in $\boldsymbol{z}_i^{int}$ from the range $\{0, 1, 2, 3\}$. Then, we adopt $\boldsymbol{z} = \boldsymbol{z}_i^{deg} \times \boldsymbol{z}_i^{int}$ as the final number for interaction-driven

measurement. To inject the interaction-driven constrain into optimal transport theory [47], we modify the distribution constraint of items in Δ from uniform assigning equal weight to each node (i.e., $\boldsymbol{H}\mathbf{1}_K = \frac{\mathbf{1}_N}{N_V}$) to interaction-based (i.e., $\boldsymbol{H}\mathbf{1}_K = \boldsymbol{p}$), where $\boldsymbol{p}_i = \boldsymbol{z}_i^{deg} / \sum_{j=1}^{N_V} \boldsymbol{z}_j^{deg}$. Therefore, the final formulation of the sparse optimal transport mechanism can be formulated as

$$\begin{aligned} &\min_{\boldsymbol{H}\in\Delta} J = < \boldsymbol{H}, \boldsymbol{M} > + \eta \|\boldsymbol{H}\|_{2,1} \\ &s.t. \Delta = \{\boldsymbol{H} \in \boldsymbol{R}_+^{N_V \times K} | \boldsymbol{H}\mathbf{1}_K = \boldsymbol{p}, (\boldsymbol{H})^T \mathbf{1}_{N_V} = \frac{\mathbf{1}_K}{K}\}. \end{aligned} \tag{12}$$

4.2 Differentiable Optimization Strategy

It seems that solving the problem in Equation 12 is difficult as the term of $\ell_{2,1}$-norm is non-smooth. Therefore, we design a differentiable sparse optimization strategy, calibrating the gradient by the Frank-Wolfe algorithm.

Specifically, we will use the Frank-Wolfe direction as the approximation to formulate the non-differentiable objective function as the form with KKT conditions. First, we hope to obtain the solution of hypergraph structure $\boldsymbol{H}$ by optimizing the objective function. We take the derivative of Equation 12 for $\boldsymbol{H}$ and a diagonal $\boldsymbol{D}$ is defined to simplify expression:

$$\begin{aligned} \nabla J(\boldsymbol{H}) &= \boldsymbol{M} - \boldsymbol{H}\boldsymbol{D} \\ &= \boldsymbol{M} - \boldsymbol{H} \begin{pmatrix} -\frac{\eta}{\|\boldsymbol{H}_1\|_2} & & \\ & \ddots & \\ & & -\frac{\eta}{\|\boldsymbol{H}_j\|_2} \end{pmatrix}. \end{aligned} \tag{13}$$

If we introduce $\boldsymbol{D}$ and define the values of $\boldsymbol{D}$ as columns of ℓ_2-norm in $\boldsymbol{H}$, written by $\|\boldsymbol{H}_j\|_2$. Then, we can simplify the Equation 13 and adopt the Frank-Wolfe algorithm to calculate the calibrated gradient to optimize the current parameters in unified hypergraph structure $\boldsymbol{H}^U$ as the following objectives:

$$\min_{\boldsymbol{H}^U \in \Delta} J_H = < \boldsymbol{H}, \nabla J(\boldsymbol{H}) > . \tag{14}$$

Here, we define the calibrated Frank-Wolfe direction of gradient as the opposite direction most inconsistent with the current optimization direction. If we find the iteration point $\boldsymbol{s}$ with calibrated Frank-Wolfe direction, we can reformulate the objective function by Frank-Wolfe algorithm:

$$\boldsymbol{s} = \underset{\boldsymbol{s}\in\Delta}{argmin}\, \boldsymbol{G}\boldsymbol{s}, \quad \boldsymbol{G} = \mathrm{vec}(\nabla J(\boldsymbol{H})), \quad \boldsymbol{s} = \mathrm{vec}(\boldsymbol{H}), \tag{15}$$

where function $\mathrm{vec}(\cdot)$ represents the process of converting a matrix into a vector. To optimize the current iteration points $\boldsymbol{s}$ and obtain the approximate solution in a differentiable way, we introduce the optimization strategy of DeepEMD algorithm[67]. Hence, we can rewrite the Equation 15 into the linear programming problem following the KKT conditions:

$$\min_{\boldsymbol{s}} \boldsymbol{G}\boldsymbol{s} \quad \text{s.t. } \boldsymbol{A}\boldsymbol{s} = \boldsymbol{b}, \quad \boldsymbol{F}\boldsymbol{s} \leq 0. \tag{16}$$

Here, $\boldsymbol{s} \in \mathbb{R}^{NK}$ is the vector to be solved under Frank-Wolfe algorithm. The equality constraint $\boldsymbol{A}\boldsymbol{s} = \boldsymbol{b}$ represents all boundary conditions constructed with equality conditions. Besides, the inequality constraint $\boldsymbol{F}\boldsymbol{s} \leq 0$ denotes the all feasible solution ranges in the variable domain under KKT conditions. Then, following the Lagrangian principle which has been well-studied in the LP problem, it can be reformulated as follows:

$$\mathcal{L}_{FW}(\theta, \boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \boldsymbol{G}\boldsymbol{s} + \boldsymbol{\lambda}^T \boldsymbol{F}\boldsymbol{s} + \boldsymbol{\mu}^T (\boldsymbol{A}\boldsymbol{s} - \boldsymbol{b}), \tag{17}$$

where $\boldsymbol{\mu}$ is the set containing all equality constraint in Equation 17. Similar to the above definition, $\boldsymbol{\lambda} \geq 0$ is the set containing all the inequality constraint. Here, θ is the learnable parameter in our designed neural network.

According to the EMD principle under KKT conditions, we can calculate the optimum $(\tilde{\mathbf{s}}, \tilde{\mu}, \tilde{\lambda})$ of loss function through $g(\theta, \boldsymbol{s}, \mu, \lambda) = 0$ and the formulation is given by

$$g(\theta, \boldsymbol{s}, \mu, \lambda) = \begin{bmatrix} \nabla_\theta L_{FW}(\theta, \boldsymbol{s}, \mu, \lambda) \\ \boldsymbol{diag}(\lambda)\boldsymbol{F}(\theta)\boldsymbol{s} \\ \boldsymbol{A}(\theta)\boldsymbol{s} - \boldsymbol{b}(\theta) \end{bmatrix}. \tag{18}$$

Due to the convex optimization theory which has been proposed in [2], the variables $\tilde{\mathbf{s}}$ and θ have the implicit function and can be solved by

$$J_\theta \tilde{\mathbf{s}} = -J_s g(\theta, \tilde{\lambda}, \tilde{\mu}, \tilde{\mathbf{s}})^{-1} J_\theta g(\theta, \tilde{\mathbf{s}}, \tilde{\mu}, \tilde{\lambda}). \tag{19}$$

Here, we analyze the the implicit function of Equation 19. To calculate the optimum variables $\tilde{\mathbf{s}}$ and θ, we need to obtain the Jacobian matrix of the solution and we can use $J_\theta \tilde{\mathbf{s}}$ to represent the partial Jacobian of $\tilde{\mathbf{s}}$ with the respect to θ. Based on the implicit function theory about the Jacobian[24], we can calculate the formula of Jacobian and obtain it by a differentiable way. Then, we can obtain a gradient concluding $\tilde{\mathbf{s}}$ with closed-form solution targeting for parameter θ. In other words, we can use the automaticly differentiable framework (e.g. Pytorch or TensorFlow) for optimization of parameters in our neural network. Specifically, we flatten the optimized hypergraph matrix as a vectorized formulation. For example, we split $\boldsymbol{H} = [\boldsymbol{h}_1; \boldsymbol{h}_2; \cdots; \boldsymbol{h}_N] \in \mathbb{R}^{N \times K}$ with $\boldsymbol{h}_i = [\boldsymbol{H}_{i1}, \boldsymbol{H}_{i2}, \cdots, \boldsymbol{H}_{iK}] \in \mathbb{R}^K$, into $\boldsymbol{h} = [\boldsymbol{h}_1^T; \boldsymbol{h}_2^T; \cdots; \boldsymbol{h}_N^T] \in \mathbb{R}^{NK}$ as a vector. Then, we use $\boldsymbol{h}^{(k)}$ represent the embedding at the k-th iteration, which is a fixed point, and we can obtain the final iterative formula with

$$\boldsymbol{h}^{(k+1)} = (1 - \gamma)\boldsymbol{h}^{(k)} + \gamma \boldsymbol{s}, \tag{20}$$

where γ is the hyperparameter to maintain the current iteration point. The optimization formulation of hyperdege $\boldsymbol{E}$ in Equation 11 can be obtain by setting the derivation to 0:

$$\frac{\partial J}{\partial \boldsymbol{E}_j} = 0, \quad \boldsymbol{E}_j = \frac{\sum_{i=1}^{N_V} \boldsymbol{H}_{ij} \boldsymbol{F}_i}{\sum_{i=1}^{N_V} \boldsymbol{H}_{ij}}. \tag{21}$$

Then, we can obtain the closed-form solution of $\boldsymbol{E}$, which also can be solved by a differentiable way.

Finally, we combine the BPR recommendation loss in Equation 7, the loss of differentiable sparse optimization for hypergraphs in Equation 17 and the loss for contrastive learning in Equation 4 to obtain the final proposed objective as follows:

$$\mathcal{L} = \mathcal{L}_{BPR} + \mathcal{L}_{FW} + \lambda \mathcal{L}_s. \tag{22}$$

Here, λ is the hyperparameter of the loss function in contrastive learning to balance the text-aligned strength of the training process. By integrating the three losses, we can learn a hypergraph structure tailored for incomplete scenarios and multimedia recommendations jointly. The whole optimization is presented in Algorithm 1.

4.3 Complexity Analysis

Through our proposed HIRE and HIRES, we can achieve promising results under incomplete multimedia recommendation scenarios. However, the trade-off between model performance and running complexity is essential during the deploying process. Therefore, in this section, we will provide our discussion about complexity analysis of our HIRE and HIRES.

For our HIRE framework, the running time is limited by two steps: hypergraph construction and multimedia recommendation. In the hypergraph construction, we exploit the deep K-means clustering technique in Equation 1 with $O(N_V K d)$ and the complexity of our hypergraph convolution is $O(2N_V K d)$. Here, N_V is the number of items, K is the number of hyperedges and d is the dimension of latent embeddings. In the multimedia recommendation step, our contrastive learning is implemented by mini-batch sampling, which can be ignored compared with the global complexity. The multimedia recommendation contains the multimodal fusion module with $O(3N_V d^2)$ and

the recommendation module with $O(N_U N_V d)$, where N_U is the number of users. Therefore, the total complexity of our HIRE is $O(3N_V Kd + 3N_V d^2 + N_U N_V d)$

For our HIRES framework, compared with HIRE framework, we replace the hypergraph construction with sparse optimal transport. In the sparse optimal transport step, the running time is limited by calculating the cost matrix $\boldsymbol{M}$ with $O(N_V Kd)$ and the derivative $\nabla J(\boldsymbol{H})$ in Equation 13. The complexity of $\nabla J(\boldsymbol{H})$ is $O(N_V K^2 + N_V K)$, where we need to calculate the values of $\boldsymbol{D}$ with $O(N_V K)$ and the derivative with $O(N_V K^2)$. Note that, since we use a differentiable technique to approximate the sparse optimal transport, the other steps can be calculated by the gradient framework automatically, e.g. Pytorch or TensorFlow. Therefore, the total complexity of our HIRES is $O(N_V Kd + N_V K^2 + N_V K + 3N_V d^2 + N_U N_V d)$.

Algorithm 1 Sparse Optimal Transport Optimization

Input: Multimedia item representation $\boldsymbol{F}$, number of hyperedge K, hyperparameter γ, number of dimension d.
Output: Final differentiable loss function $\mathcal{L}_{FW}$.

1: Initial hyperedge embedding $\boldsymbol{E}$ and unified heterogeneous hypergraph structure $\boldsymbol{H}$;
2: **while** not convergent **do**
3: Calculate the cost matrix $\boldsymbol{M}$ by $\boldsymbol{M}_{ij} = ||\boldsymbol{F}_i - \boldsymbol{E}_j||_2^2$ in Equation 11;
4: Update the hypergraph structure by $\min_{\boldsymbol{H}\in\Delta} J =< \boldsymbol{H}, \boldsymbol{M} > +\eta\|\boldsymbol{H}\|_{2,1}$ in Equation 12;
5: Take the derivative of J by Equation 13;
6: exploit the Frank-Wolfe algorithm to optimize the current hypergraph matrix $\boldsymbol{H}$ with Equation 14;
7: Find the iteration point $\boldsymbol{s}$ with calibrated Frank-Wolfe direction by Equation 16.
8: **end while**
9: Calculate $\mathcal{L}_{FW}(\theta, \boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \boldsymbol{G}\boldsymbol{s} + \boldsymbol{\lambda}^T \boldsymbol{F}\boldsymbol{s} + \boldsymbol{\mu}^T(\boldsymbol{A}\boldsymbol{s} - \boldsymbol{b})$ in Equation 17.

5 Experiments

In this section, we evaluate the effectiveness of our proposed HIRES on four public multimedia datasets. First, we provide a brief description of datasets and experimental settings. Then, we evaluate our proposed HIRES framework from the following four research questions:

- **RQ1:** How do HIRE framework and HIRES framework perform compared to the state-of-the-art single space recommendation methods?
- **RQ2:** How does the HIRE and HIRES propose each component contributes to performance improvement?
- **RQ3:** How do the hyperparameters impact the performance of recommendation and how can optimal values be chosen?
- **RQ4:** How does HIRES improve the modeling of multiple facets of users and items?

5.1 Experimental Setup

5.1.1 Datasets. To comprehensively demonstrate the effectiveness of the methods we are comparing, we use four real-world datasets from different application domains,which have different sizes and interaction densities. The statistics are summarized in Table 2.

- **Amazon.** We use three datasets from Amazon, which are Amazon-Baby, Amazon-Sports, and Amazon-Elec. The images and textual details of products are used to generate 4096-dimensional visual feature embeddings and textual feature embeddings. The textual features are encoded with Sentence-Bert [41].
- **TikTok.** This dataset comes consists of short videos on the Tiktok platform. The short videos contain visual, acoustic, and textual features, which are considered as multi-modal features. The textual features are also

Table 2. Statistics of experimented datasets with multimodal item Visual (V), Acoustic (A), and Textual (T) contents.

Dataset	Amazon-Baby		Amazon-Sports		Amazon-Elec		Tiktok		
Modality	V	T	V	T	V	T	V	A	T
Embedding Dim	4096	1024	4096	1024	4096	384	128	128	768
User	19445		35598		192403		9319		
Item	7050		18357		63001		6710		
Interactions	160792		296337		1689188		59541		
Sparsity	99.883%		99.955%		99.990%		99.904%		

encoded with Sentence-Bert. Besides, the acoustic features are extracted by the Transformer [14] tailored for acoustic learning.

In these datasets, each item review rating is treated as a record of positive user-item interaction and we use a graph structure to store the matrix. Following previous works [9, 40], we adopt their selected multimedia pre-processing settings, like [28, 31], to build our incomplete multimedia scenarios. Specifically, we filter out the items that have fewer than 5 interactions and the users that have fewer than 5 interactions in their domains as previous works [75]. Note that we maintain the user having ratings with items with score greater than 3 as the positive samples to model the real scenarios. The modality missing means the whole features under the modality are missed and we use the zero value to mask the features as the missing process. Under each modality, we randomly select a certain rate of nodes for missing operations, which is defined as the missing rate. Besides, we conduct 10 tests on each round of experimental results and report the average of the results as the final performance. In order to guarantee a fair comparison, we follow the latest multimedia models [1] to split the dataset into training, test and validation sets with the ratio of 7:2:1.

5.1.2 Evaluation protocols. We evaluate the recommendation performance using three metrics: Recall@K , Precision@K and Normalized Discounted Cumulative Gain (NDCG@K) to test the accuracy of our proposed HIRE and HIRES. Besides, to illustrate the diversity of our proposed model, we also introduce the Intra-List Category Similarity (ILCS) to calculate the micro diversity for our tailored hypergraph methods for incomplete multimedia recommendations. The specific formulae are as follows:

$$\text{Recall@k} = \frac{\text{TP@k}}{\text{TP@k} + \text{FN@k}} \tag{23}$$

Recall@k is defined as the ratio of the number of true positive cases identified within the top k recommendations to the total number of positive cases. It measures the ability of the system to retrieve relevant items from the total set of relevant items.*TP@k* means the number of positive cases that are correctly identified within the top k recommendations or predictions. These are the relevant items that the system successfully retrieves. *FN@k* means the number of positive cases that are not identified within the top k recommendations or predictions. These are the relevant items that the system fails to retrieve.

$$\text{Precision@k} = \frac{\text{TP@k}}{\text{TP@k} + \text{FP@k}} \tag{24}$$

Precision@K is the accuracy of evaluating the recommendation system in the Top-K recommendation results. *FP@k* means the number of samples that were incorrectly predicted as positive in the first k predictions. That is, these samples are actually negative classes, but the model incorrectly predicts them as positive.

$$\text{DCG@K} = \sum_{i=1}^{K} \frac{2^{F(v_i T(u))} - 1}{\log_2(i+1)} \tag{25}$$

$$\text{IDCG}\,@K = \sum_{i=1}^{K} \frac{1}{\log_2(i+1)} \tag{26}$$

$$\text{NDCG@ K} = \frac{\text{DCG@ K}}{\text{IDCG@ K}} \tag{27}$$

$NDCG@K$ is an indicator that is sensitive to the order of recommended items. The function $F(v_i T(u))$ is an indicator function that determines whether item v is in the set $T(u)$. If it is, the value is 1, otherwise, it is 0. DCG represents Discounted Cumulative Gain, and $IDCG$ represents the Ideal Discounted Cumulative Gain. Following the settings in [53], we all-rank item evaluation strategy is used to measure the accuracy. To measure the accuracy, the average scores over all users are reported in the test set.

$$ILCS = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|\hat{\boldsymbol{R}}_u|(|\hat{\boldsymbol{R}}_u| - 1)} \sum_{(i,j) \in \hat{\boldsymbol{R}}_u} \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{28}$$

Intra-List Category Similarity (ILCS) is the metric to measure the micro diversity of the recommendation sets for specific users, which has been widely exploited in previous works [56]. ILCS can measure the diversity between all pairs of items in $\hat{\boldsymbol{R}}_u$ for each user u, and then takes the average. The fewer categories C of items i and j that overlap in the recommended candidate set means the smaller ILCS values, which is is supposed to indicate greater diversity of the proposed framework.

5.1.3 Baselines. To illustrate the influences of the inclusion of missing modalities on the models and verify the effectiveness of our proposed framework in completing missing modalities, we compare with the following representative and recent baselines, which can be divided into two categories. (1) collaborative filtering without multimedia contents (MF-BPR, NGCF, LightGCN, SGL, NCL, HCCF, MMGCN, LATTICE, and CLCRec), and (2) multimedia recommendation with missing modalities (LightGCN-*M*, MMGCL, SLMRec, MMSSL, AGCN, DualGNN, M^{3}care, GCIMH, T2-GNN, MoMKE and CI^{2}MG):

- **BPR [42]:** The full name of BPR is Bayesian Personalized Ranking, which is a sorting algorithm for Top-N recommendation and is suitable for implicit feedback data.
- **NGCF [52]:** NGCF exploits the user-item graph structure by propagating embeddings on it, effectively injecting the collaborative signal into the embedding process in an explicit manner.
- **LightGCN [18]:** LightGCN learns embeddings by linearly propagating them on the user-item interaction graph, and uses weighted sum of embeddings learned at all layers as the final embedding.
- **SGL [20]:** SGL generates multiple views of a node, maximizing the agreement between different views of the same node compared to that of other nodes.
- **NCL [27]:** NCL consists of two core components, which focus on the individual supervised learning for each single expert and the knowledge transferring among multiple experts, respectively.
- **HCCF [60]:** HCCF is proposed to jointly capture local and global collaborative relations with a hypergraph-enhanced cross-view contrastive learning architecture.
- **MMGCN [55]:** MMGCN, built upon the message-passing idea of graph neural networks, can yield modal-specific representations of users and micro-videos to better capture user preferences.
- **LATTICE [69]:** LATTICE, which is proposed for multimodal recommendation, leverages graph structure learning to discover latent item relationships underlying multimodal features.

- **CLCRec [54]:** CLCRec reformulates the representation learning for cold-start item from an information-theoretic standpoint.
- **LightGCN-M:** This is generated on the basis of model LightGCN combined with multimodal features.
- **MMGCL [62]:** MMGCL aims to explicitly enhance multi-modal representation learning in a self-supervised learning manner.
- **SLMRec [49]:** In order to capture multi-modal patterns in the data itself, SLMRec goes beyond the supervised learning paradigm.
- **MMSSL [53]:** MMSSL is a modality-aware structure learning paradigm via adversarial perturbations for data augmentation to characterize the inter-dependency.
- **AGCN [58]:** AGCN is proposed for joint item recommendation and attribute inference in an attributed user-item bipartite graph with missing attribute values.
- **DualGNN [50]:** leverages the correlation between users to mine the particular fusion pattern for each user.
- **M^{3}care [68]:** M^{3}Care imputes the task-related information of the missing modalities in the latent space by the auxiliary information from each patient's similar neighborss.
- **GCIMH [43]:** GCIMH employs graph convolutional autoencoding and multi-modal hashing to generate hash codes from incomplete multi-modal data for efficient retrieval.
- **T2-GNN [22]:** T2-GNN enhances GNN performance on incomplete graphs using teacher-student distillation with specialized feature and structure guidance.
- **MoMKE [61]:** MoMKE leverages unimodal and joint representations learned from all modality experts through a two-stage training process to robustly handle incomplete multimodal data.
- **CI^{2}MG [31]:** CI^{2}MG is proposed for enhancing incomplete multimedia recommendation, in order to address the challenge of missing modalities.

5.1.4 Experimental Setup Details. To ensure fair comparison, we have carefully tuned the hyperparameter of dimension for all baselines through cross-validation as suggested in their original papers to achieve their best performance. Besides, for multi-modality methods, the dimension given is the total dimension after concatenation which is the same for all baselines to achieve a fair comparison, following the most existing settings [23, 53]. We implement HIRE and HIRES using PyTorch, which will be made publicly available upon the acceptance of this work. AdamW and Adam are adopted 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 number of graph layer in $\{1, 2, 3, 4\}$. In our experiments, except the LightGCN-*M* implemented by ourselves, we use official implementations proposed by the original paper of other baselines. For all the baselines, we exploit the same value with the common hyperparameters. For example, the embedding dimension d is set to 64, the batch size to 1024. For the specific hyperparameters in the baselines, we use the values reported in their original literature. Additionally, for the experimental environment, we implement the proposed method and other baseline models on a standard Ubuntu-16.04 operating system. Furthermore, most experiments reported in this paper are trained using four Nvidia Tesla P100 and two GeForce RTX 3090 GPUs with 128G memory.

5.2 Overall Performance Comparison (RQ1)

In order to verify the effectiveness of our proposed model, We compare the recommendation performance of HIRE and HIRES against the selected benchmarks using Recall@20, Precision@20, and NDCG@20 metrics. From Table 3, we can make the following observations.

Overall, HIRE and HIRES surpass all 20 baselines across the evaluation metrics on the four multimedia datasets. This answers **RQ1** and illustrates the effectiveness of jointly training heterogeneous hypergraph and multimedia recommendation under the incomplete scenarios. Compared with the HIRE framework, the performance gains of HIRES on Amazon-baby, Amazon-sports, Amazon-elec, and Tiktok range from reasonably large (2.17% achieved

Table 3. Performance (%) comparison of baselines with the 90% missing rate in terms of Recall@20, Precision@20 and NDCG@20 on Amazon-Baby, Amazon-Sports, Amazon-Elec and Tiktok multimedia datasets, where ∗ denotes a significant improvement according to the wilcoxon signed-rank test.

Baseline	Amazon-Baby			Amazon-Sports		
	Recall@20	Precision@20	NDCG@20	Recall@20	Precision@20	NDCG@20
MF-BPR	4.51±0.11	0.24±0.002	2.06±0.16	4.03±0.15	0.19±0.009	1.97±0.18
NGCF	6.11±0.12	0.34±0.003	2.64±0.10	6.80±0.13	0.34±0.006	3.08±0.10
LightGCN	7.24±0.20	0.38±0.002	3.32±0.16	6.31±0.13	0.40±0.007	3.55±0.15
SGL	6.85±0.14	0.36±0.007	2.91±0.14	7.33±0.15	0.37±0.006	3.51±0.14
NCL	7.37±0.19	0.36±0.007	3.06±0.19	7.56±0.17	0.37±0.005	3.50±0.20
HCCF	7.20±0.14	0.38±0.007	3.05±0.17	6.95±0.14	0.39±0.005	3.21±0.15
MMGCN	5.44±0.19	0.27±0.001	2.39±0.17	4.37±0.14	0.22±0.004	2.17±0.15
LATTICE	6.77±0.19	0.39±0.001	3.02±0.12	6.68±0.20	0.28±0.002	2.74±0.11
CLCRec	6.31±0.15	0.36±0.002	2.78±0.19	5.97±0.18	0.28±0.002	2.57±0.15
LightGCN-*M*	5.38±0.27	0.28±0.002	2.18±0.10	3.65±0.16	0.23±0.004	2.35±0.15
MMGCL	5.70±0.25	0.31±0.002	2.57±0.17	6.90±0.13	0.37±0.001	3.28±0.19
SLMRec	6.91±0.18	0.41±0.008	3.12±0.15	7.61±0.18	0.37±0.003	3.43±0.17
MMSSL	8.00±0.20	0.41±0.001	3.47±0.10	8.16±0.10	0.35±0.002	3.58±0.05
AGCN	7.05±0.15	0.33±0.001	3.06±0.09	5.25±0.14	0.32±0.002	2.90±0.09
DualGNN	6.15±0.21	0.31±0.002	2.82±0.13	5.38±0.24	0.29±0.001	2.31±0.14
M^3care	7.31±0.10	0.37±0.001	3.01±0.15	6.44±0.15	0.29±0.002	2.31±0.13
GCIMH	6.89±0.19	0.28±0.002	2.97±0.12	5.01±0.13	0.29±0.003	2.71±0.10
T2-GNN	7.01±0.21	0.30±0.002	3.02±0.11	5.16±0.14	0.32±0.001	2.93±0.13
MoMKE	8.21±0.17	0.42±0.001	3.33±0.09	8.16±0.15	0.40±0.002	3.40±0.09
CI^2MG	8.26±0.13	0.44±0.004	3.55±0.10	8.36±0.12	0.44±0.006	3.31±0.12
HIRE	8.54±0.18*	0.46±0.004*	3.70±0.10*	8.63±0.12*	0.46±0.003*	3.45±0.10*
HIRES	**8.81±0.16***	**0.47±0.004***	**3.78±0.09***	**8.98±0.12***	**0.48±0.003***	**3.59±0.08***
Imp of HIRE	3.39%	4.55%	4.23%	3.23%	4.55%	4.23%
Imp of HIRES	6.66%	6.82%	6.48%	7.42%	9.09%	8.46%
Baseline	Amazon-Elec			Tiktok		
	Recall@20	Precision@20	NDCG@20	Recall@20	Precision@20	NDCG@20
MF-BPR	3.03±0.13	0.16±0.004	1.43±0.13	3.23±0.19	0.19±0.003	1.25±0.15
NGCF	4.11±0.18	0.22±0.002	1.75±0.15	5.92±0.14	0.31±0.004	2.20±0.21
LightGCN	4.71±0.13	0.27±0.005	2.29±0.14	5.85±0.14	0.32±0.004	3.21±0.16
SGL	4.94±0.13	0.24±0.001	2.04±0.17	6.58±0.14	0.28±0.006	1.28±0.19
NCL	4.94±0.15	0.24±0.002	2.04±0.15	6.58±0.18	0.28±0.001	1.28±0.14
HCCF	4.70±0.13	0.27±0.005	2.24±0.16	6.49±0.18	0.26±0.009	2.71±0.11
MMGCN	3.90±0.15	0.19±0.003	1.61±0.11	6.99±0.17	0.33±0.002	2.46±0.12
LATTICE	4.96±0.17	0.27±0.003	2.08±0.07	6.26±0.18	0.33±0.002	3.19±0.08
CLCRec	4.71±0.19	0.24±0.001	2.08±0.13	5.86±0.17	0.34±0.003	3.20±0.10
LightGCN-*M*	3.59±0.17	0.21±0.006	1.59±0.10	6.47±0.11	0.38±0.003	2.86±0.18
MMGCL	3.82±0.18	0.19±0.002	1.70±0.09	5.84±0.13	0.29±0.006	2.59±0.13
SLMRec	5.12±0.16	0.29±0.001	2.19±0.19	8.03±0.13	0.36±0.001	4.36±0.10
MMSSL	5.45±0.15	0.28±0.007	2.31±0.06	8.56±0.13	0.39±0.005	4.39±0.04
AGCN	5.08±0.12	0.25±0.001	2.04±0.10	7.63±0.13	0.32±0.003	2.83±0.13
DualGNN	4.38±0.17	0.23±0.005	2.11±0.09	5.56±0.17	0.31±0.002	2.73±0.03
M^3care	4.38±0.16	0.23±0.004	2.11±0.12	5.56±0.17	0.31±0.001	3.66±0.09
GCIMH	4.83±0.15	0.23±0.002	1.93±0.11	7.47±0.14	0.30±0.003	3.62±0.11
T2-GNN	5.01±0.18	0.24±0.003	2.01±0.08	7.54±0.12	0.31±0.002	3.77±0.09
MoMKE	5.26±0.16	0.29±0.002	2.51±0.12	8.42±0.11	0.38±0.003	4.55±0.10
CI^2MG	5.54±0.15	0.30±0.001	2.68±0.13	8.57±0.17	0.39±0.004	4.86±0.08
HIRE	5.76±0.13*	0.31±0.004*	2.79±0.13*	8.85±0.13*	0.40±0.005*	5.02±0.04*
HIRES	**5.92±0.11***	**0.32±0.003***	**2.87±0.09***	**9.08±0.10***	**0.42±0.001***	**5.21±0.02***
Imp of HIRE	3.97%	3.33%	4.10%	3.27%	2.56%	3.29%
Imp of HIRES	6.86%	6.67%	7.09%	5.95%	7.69%	7.20%

with Precision@20 on Amazon-Baby) to significantly large (5.00% achieved with Precision@20 on Tiktok). This experimental results show that our proposed modality-driven sparse constrain in HIRES is more effective in capturing the higher-order relations with missing modalities where we can assign smaller weights to the missing modalities, as we will further demonstrate in the optimization section. Compared with the second-best performance, HIRES has significant performance improvements in Recall, Precision, and NDCG range from (5.95% achieved with Recall@20 on Tiktok) to (9.09% achieved with Rrecision@20 on Amazon-Sports). It's worth noting that the improvements of HIRES are particularly significant in scenarios where user-item interactions are sparse, like with Tiktok, which supports the appropriate design of unified heterogeneous hypergraph to make full use of high-order similarity in the incomplete multimedia recommendation.

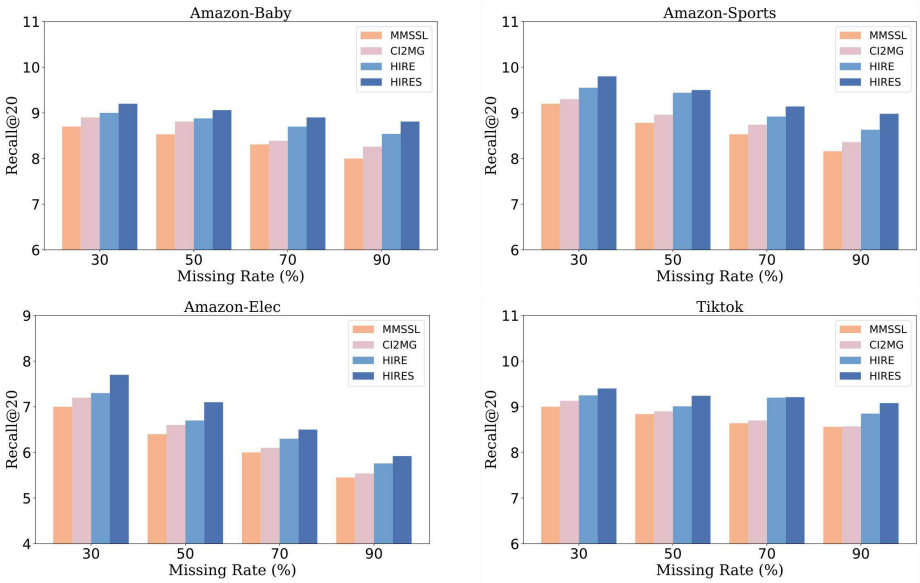


Fig. 4. Performance about the comparison with different missing rates for multimedia recommendation regarding Recall@20 of the HIRES on the Amazon-Baby, Amazon-Sports, Amazon-Elec and Tiktok datasets.

In particular, by considering both implicit feedback and multi-modal features for recommendation, HIRES performs better than single-modality baselines, which supports the appropriate using of multimodal information that can enhance the recommender systems. The graph-based paradigm is an efficient way to exploit the user-item interactions to involve the multimedia information. Compared with the existing method (e.g. LightGCN) with only graph encoders, our proposed HIRE performs better than graph-based methods in many cases. For example, HIRES framework outperforms LightGCN from 5.01% (achieved in Precision@20 on Tiktok) to 43.03% (achieved in NDCG@20 on Tiktok). Moreover, the performance gains of proposed HIRES over graph-based method ranges from 11.26% (achieved in NDCG@20 on Amazon-Sports) to 62.30% (achieved in NDCG@20 on Tiktok). This observation illustrates that introducing the hypergraph structure can capture more effective set-to-set higher-order relations in mining and completing the heterogeneous multimodal representation under the scenarios with missing modalities. The main differences of hypergraph structures between HIRE and HIRES reside in properly designing the missing-aware sparse constraint among multimodal node representations, where the experimental results show that sparse hypergraph structures can enhance the incomplete multimedia recommendations. Besides, by considering the hypergraph structure in multimedia scenarios, we can not only capture the local information but also the global collaborative filtering signals. However, exploiting the hypergraph directly may ignore the multimodal relations across the modalities, which is helpful to the missing modalities. For example, although HCCF can jointly capture local and global user-item relations with a hypergraph-enhanced cross-view contrastive learning architecture, it fails to capture the complex relations across modalities. In this way, HIRE outperforms HCCF up to 21.31% in NDCG@20 on Baby dataset, up to 24.17% in Recall@20 on Sports dataset, up to 24.55% in NDCG@20 on Electronics dataset, and up to 53.84% in Precision@20 on Tiktok dataset. HIRES can outperform HCCF by up to 22.36% in Recall@20 on Baby dataset, up to 29.21% in Recall@20 on Sports dataset, up to 25.96% in Recall@20 on Electronics dataset, and up to 39.91% in Recall@20 on Tiktok dataset.

Table 4. The ablation study on the HIRE and HIRES performance across Baby, Sports, Elec and Tiktok.

+Homo-Hypergraph	+Hete-Hypergraph	+Contrastive	+Constraint	Amazon-Baby		
				Recall@20	Precision@20	NDCG@20
×	×	×	×	5.38	0.28	2.18
✓	×	×	×	6.01	0.33	2.86
✓	✓	×	×	7.65	0.40	3.24
✓	✓	✓	×	8.54	0.46	3.00
✓	✓	✓	✓	8.81	0.47	3.78
+Homo-Hypergraph	+Hete-Hypergraph	+Contrastive	+Constraint	Amazon-Sports		
				Recall@20	Precision@20	NDCG@20
×	×	×	×	3.65	0.23	2.35
✓	×	×	×	4.86	0.36	2.97
✓	✓	×	×	7.52	0.41	3.12
✓	✓	✓	×	8.63	0.46	3.45
✓	✓	✓	✓	8.98	0.48	3.59
+Homo-Hypergraph	+Hete-Hypergraph	+Contrastive	+Constraint	Amazon-Elec		
				Recall@20	Precision@20	NDCG@20
×	×	×	×	3.59	0.21	1.59
✓	×	×	×	4.68	0.25	1.98
✓	✓	×	×	5.03	0.28	2.45
✓	✓	✓	×	5.57	0.31	2.79
✓	✓	✓	✓	5.92	0.32	2.87
+Homo-Hypergraph	+Hete-Hypergraph	+Contrastive	+Constraint	Tiktok		
				Recall@20	Precision@20	NDCG@20
×	×	×	×	6.47	0.38	2.86
✓	×	×	×	7.46	0.38	3.56
✓	✓	×	×	8.08	0.39	4.77
✓	✓	✓	×	8.85	0.40	5.02
✓	✓	✓	✓	9.08	0.42	5.21

Moreover, the multimodal baselines (i.e., MMGCL, MMSSL, and CI^{2}MG) can obviously outperform the without multimedia competitors (e.g., MF-BPR, LightGCN, HCCF), which is consistent with the results in previous work [69]. When the incomplete conditions (e.g., LATTICE and MMSSL) are not all considerable, those multimedia-based methods cannot sustain competitive results. For example, the performance gains of HIRE over the multimodal method without completing achieve an improvement up to 12.19% on Precision@20 on Baby dataset, up to 31.42% on Recall@20 on Sports dataset, up to 20.77% on NDCG@20 on Elec datasets, and 14.35% on Precision@20 on Tiktok dataset, respectively. Besides, the performance gains of HIRES over the multimodal method without completing achieve an improvement up to 14.63% on Precision@20 on Baby dataset, up to 37.14% on Recall@20 on Sports dataset, up to 24.24% on NDCG@20 on Elec datasets, and 18.67% on Precision@20 on Tiktok dataset, respectively. These results also show that most existing multimedia methods will suffer significant performance degradation while simply discarding incomplete modalities in real missing scenarios, which also support the significant design with modality completion.

Compared with the multi-modal methods targeting the incomplete scenarios (e.g. AGCN, DualGnn M^{3}care, GCIMH, T2-GNN, MoMKE and CI^{2}MG), our proposed HIRE and HIRES can capture the multi-modal higher-order user-item interactions tailored for the multimedia recommendation, which enhance the process of completing multi-modal representation with missing values. For example, our proposed HIRE and HIRES methods gain an improvement up to 5.76% and 10.04% with Rcall@20 in the Sports dataset over the MoMKE framework respectively, which illustrates our proposed methods have more efficient capabilities of feature completing. Although CI^{2}MG with modality completing can achieve the second-best performance via reconstructing user-item interactions and aligning modality features from both inter- and intra-modality perspectives in many cases, our proposed HIRES can still gain improvements by capturing the complex high-order relations within multimodal data. The

performance gains of HIRE and HIRES over CI^{2}MG achieve an average improvement up to 4.63% and 7.53%, respectively. Such observations strongly indicate that the optimal transport with sparse constraints is more useful in the more difficult incomplete scenarios, which consists of modality-driven constrain and interaction-driven constrain. Note that LightGcn ranks the second in NDCG@20 on Sports dataset. This is because missing values may introduce additional noise, especially for the scenarios with high missing rates and sparser interactions. However, existing incomplete multimedia recommendation methods assign the equal weights to both missing and complete modalities, which can lead to negative transferring.

Furthermore, 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 4, following the setting [31]. In general, from the observations with the increase of missing modality, the overall performance shows a downward trend, which shows that the missing modality will affect the model performance. The proposed HIRES method can achieve best Recall@20 scores with all missing settings (30%, 50%, 70% and 90%) on all datasets, for example the value of Recall@20 is up to 9.8 in Sports baseline with missing setting is 30%, which illustrate HIRE and HIRES can construct heterogeneous hypergraph and execute recommender system, for completing missing modality. Note that, our proposed HIRES framework may achieve the similar results under both 70% and 90% missing rates on the Baby and Tiktok datasets. This is because we have the tailored design of the modality-driven constrain and interaction-driven constrain, where we look for the node with the highest confidence as the completion basis. The experimental results are also particularly evident for the robustness of our proposed HIRES towards incomplete recommendation.

5.3 Ablation Experiment (RQ2)

To better understand our proposed techniques, we conducted ablation experiments on our framework by incrementally adding components to answer **RQ2**. We begin with a widely-adopted general recommendation model, LightGCN, as our base model. Subsequently, we introduce the homogeneous hypergraph structure (+Homo-Hypergraph) by constructing modality-specific hypergraph, e.g., texual, visual, acoustic, and combine the local graph with the global hypergraph representations to address incomplete multimedia recommendations. Then, following the same architecture with +Homo-Hypergraph, we replace the homogeneous hypergraph structure with a unified heterogeneous hypergraph structure (+Hete-Hypergraph) by the clustering-based mechanism, where we perform a unified hypergraph convolution to update the multimodal representations under each modality. Moreover, to make full use of the different modality information to supervise the learning of the unified hypergraph structure, we add the self-supervised contrastive learning aligned with the textual view (+Contrastive) to verify the effectiveness of the contrastive module. Finally, we analyze the sparse constraint with a differentiable optimization strategy (+Constraint). Insights gleaned from Table 4 lead to the following observations:

Compared with the base models, the performance gains of Base+Homo-Hypergraph on four datasets fluctuate, e.g., ranging from 26.38% to 33.15% (achieved in NDCG@20 and Recall@20 on Amazon-Sports). Such observations strongly indicate that the hypergraph structure is capable of capturing higher-order relationships between sets to enhance the multimedia recommendations. Similarly, we replace the homogeneous hypergraph structure with a unified heterogeneous hypergraph structure (Base+Homo-Hypergraph+Hete-Hypergraph), by considering the higher-order multimodal relationships across modalities, our tailored design of unified heterogeneous hypergraph structure can achieve an improvement up to 11.71% in Recall@20, up to 17.85% in Precision@20, and up to 13.28 in NDCG@20 on Amazon-Baby. Then, to verify the ability of our proposed HIRES to handle multimedia information, we ablate our design of the self-supervised contrastive learning aligned with the textual view. The performance gains of Base+Homo-Hypergraph+Hete-Hypergraph+Contrastive over Base+Homo-Hypergraph+Hete-Hypergraph on Amazon-Elec datasets range from 10.73% on Recall@20 to 13.87%

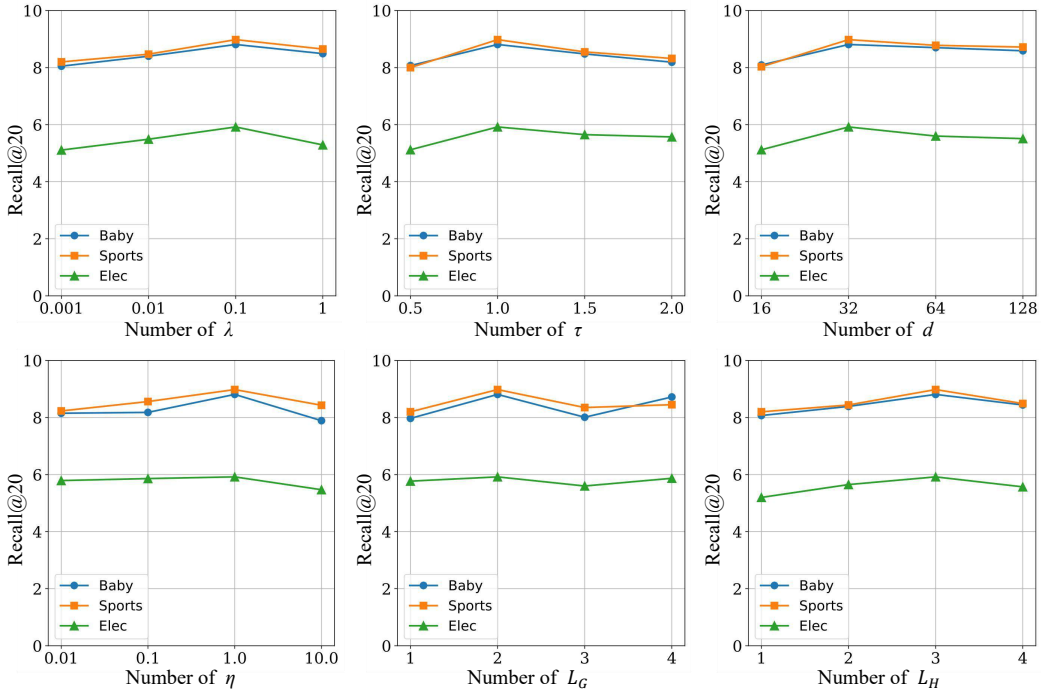


Fig. 5. Performance of hyperparameter study regarding Recall@20 of the HIRES framework with varying hyperparameters on Baby, Sports and Electronics datasets.

on NDCG@20, showing introducing text invariance into contrastive learning can enhance the extraction of multimodal features. Finally, we analyze our sparse optimal transport framework with the modality-driven constrain, modality-driven constrain and interaction-driven constrain. From the observations, our proposed Base+Homo-Hypergraph+Hete-Hypergraph+Contrastive+Constraint achieves the highest performance. Compared with the base model, Base+Homo-Hypergraph+Hete-Hypergraph+Contrastive+Constraint achieve a significant improvement from reasonably large 5.01% (achieved in Precision@20 on Tiktok) to significant large 43.03% (achieved in NDCG@20 on Tiktok). These results align with those presented in Table 3, demonstrating the effectiveness of our proposed techniques.

5.4 Effect of Hyperparameters (RQ3)

Our proposed HIRES framework primarily introduces the hyperparameters, i.e., λ, τ, η, L_G, L_H and d. In order to clarify the optimal number of hyperedges in the proposed hypergraph, we especially analyze the selecting of hyperedges in our proposed HIRE and HIRES, where K_m and K_s means the number of hyperedges in HIRE and HIRES, respectively. Here we show the impact of these hyperparameters on performance and explain their optimal settings.

From Fig. 5, we have the following observations: (1) λ is the weight which controls the strength of the $\mathcal{L}_s$. Too small λ will cause the weakening of strength for the positive modality pairs, while too large λ will cause the overfitting problem. The optimal values are approximately 0.01 and 0.1. This experimental result demonstrates that HIRES is sensitive to λ. The setting $\lambda = 0.1$ appears to be the rule-of-thumb. (2) τ is the temperature parameter

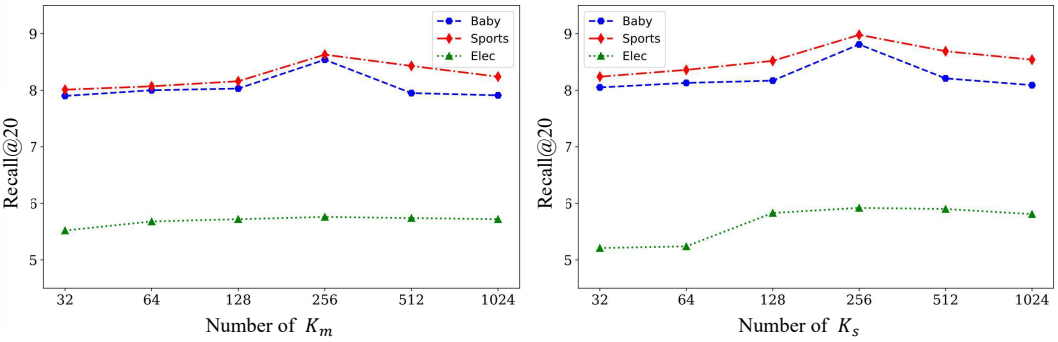
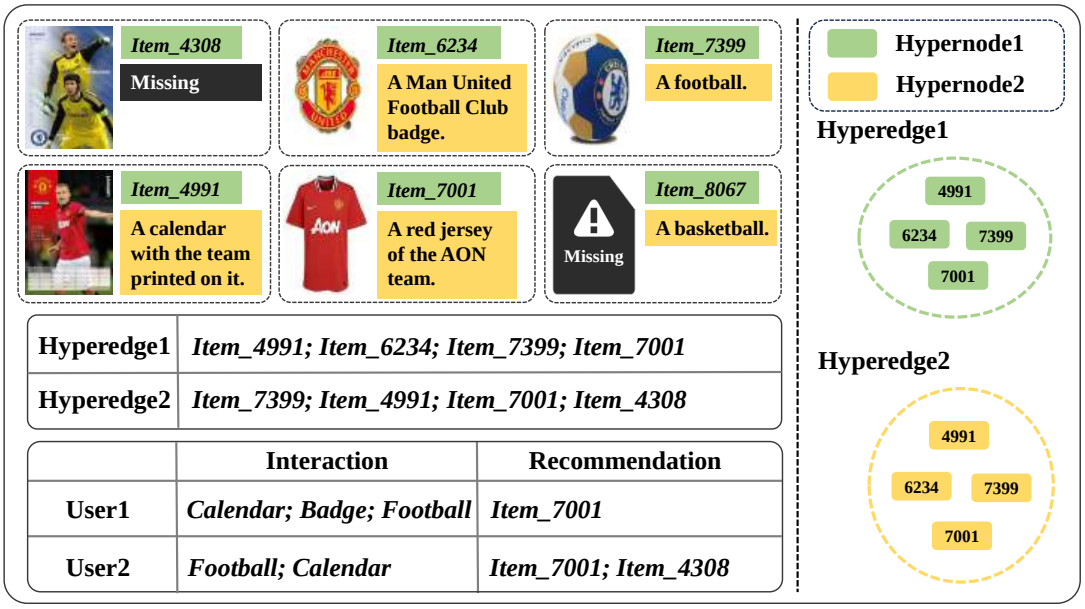


Fig. 6. Hyperparameter analysis for the number of hyperedges within the unified hypergraph in HIRE and HIRES on Baby, Sports and Electronics datasets.

of contrastive learning. For hyperparameter τ, we found the optimal τ for different datasets is consistently 1.0. Specifically, we observe that the effectiveness of τ increases with its value, leading to better performance when the values are small. However, increasing τ beyond an optimal value tends to degrade performance. In practical terms, τ = 1.0 seems to be the rule-of-thumb. (3) d means the dimension of latent representations, where the optimal value is 64 on all datasets. In particular, we observe the effectiveness of dimension d always leads to the performance gains when d is less than 64. However, increasing it beyond the optimal value tends to deteriorate the performance. In practice, setting d as 64 appears to be the rule-of-thumb. (4) η is the hyperparameter to control the strength of sparse regularization in sparse optimal transport. From the experimental results, we can observe that the optimal number is selected with η set to 1, which illustrates the sparse constraint of appropriate strength can effectively enhance the learning of hypergraph structures. From the observation, The model has not achieved convergence when η is set too small. However, the strong strength of sparse constraint can cause the hypergraph structure to discard some side information, which may lead to the overfitting problem. Therefore, it seems to be the rule-of-thumb to set the hyperparameter η with optimal number 1. (5) L_G means the graph layers of HIRES, where the optimal values are about 2 layers. We can observe the performance improves when the values are small. Besides, when layer greater than 2, as the number of layers increases, the performance actually decreases, which is called overfitting problem. Given that additional message passing and aggregation can exacerbate data sparsity issues, we set $L = 2$ to alleviate the problem of over-smoothing. (6) L_H means the hypergraph convolution layers of our proposed HIRES. If L_H is too small, the higher-order multimodal relations between each modality may become weaker. However, too large L_H will likely cause the performance decrease from the experimental results. This may be due to the introduction of noise as the number of convolution layers increases, which is a common problem in graph convolution.

Moreover, to illustrate how the varying K impacts model performance, we analyze the number of hyperedges within the hypergraph in our proposed HIRE and HIRES, as shown in Fig. 6. From the observations, we can find that our proposed HIRE framework can achieve a significant performance improvement as the number of hyperedges K_m increases. However, the improvement of model performance is not positively correlated with the number of hyperedges, where too large number of hyperedges K_m can lead to redundant parameters in our HIRE, resulting in overfitting problem. Besides, hypergraph with larger number of hyperedges may bring additional computational cost. Therefore, to achieve the trade-offs between complexity and perfomance, the number of K_m is set to 256. Besides, the number of hyperedges K_s in our proposed HIRES has the same trend as K_m, where the optimal value is 256 on all datasets. Compared with the hypergraph proposed in HIRE, our proposed hypergraph



Hyperedge1	***Item_4991; Item_6234; Item_7399; Item_7001***
Hyperedge2	***Item_7399; Item_4991; Item_7001; Item_4308***

	Interaction	**Recommendation**
User1	***Calendar; Badge; Football***	***Item_7001***
User2	***Football; Calendar***	***Item_7001; Item_4308***

Fig. 7. Examples of incomplete multimedia recommendation by the proposed HIRES with sparse hypergraph structure and contrastive multimedia recommendation. The corresponding items are also recommended by proposed HIRES.

of HIRES is more sensitive to hyperparameter K_s. This is because our sparse optimal transport mechanism in HIRES introduces prior knowledge of missing-aware and interacting-aware information. Then, we can also observe K_s is not sensitive on the Amazon-Elec dataset with sparser user-item interactions, which illustrates the effectiveness of our sparse optimization mechanism under sparse scenarios.

5.5 Case Study (RQ4)

In order to demonstrate the advantages of our proposed HIRES, we first provide interpretable recommendations for more insights. Specifically, we demonstrate two example users on sports dataset. All of the recommended corresponding items are retrieved based on the user-item relations and sparse hypergraph structure, which are learned by HIRES. Furthermore, we visualize the clustering results of HIRE and HIRES to demonstrate the effectiveness of structural learning. Then, we show the hypergraph structure with our sparse optimization and the diversity analysis to give the more clear clarification.

Capturing high-order similarities with user-item interactions. As shown in Fig. 7, we can capture the high-order simialrities modeled by the unified hypergraph structure via the user-item interactions. For example, the preference of user1 is consistent with the category $football$ from the historical interactions. Consequently, $item\_4991$ and $item\_7399$ may have the high-order similarities across the different modalities, due to the same interactions with user1, which is modeled by the hyperedge1 in HIRES. In this way, we can jointly construct our unified hypergraph structure to capture the high-order correlations with the help of user-item interactions.

Enhancing the incomplete multimedia recommendation with the sparse hypergraph structure. As shown in Fig. 7, with the well-designed sparse hypergraph structure (e.g., hyperedge2), we can further leverage the high-order similarties to distinguish user preferences, thereby achieving fine-grained recommendations. Specifically, $item\_4308$ and $item\_4991$ belong to the same hyperedge modeled by the unified hypergraph structure

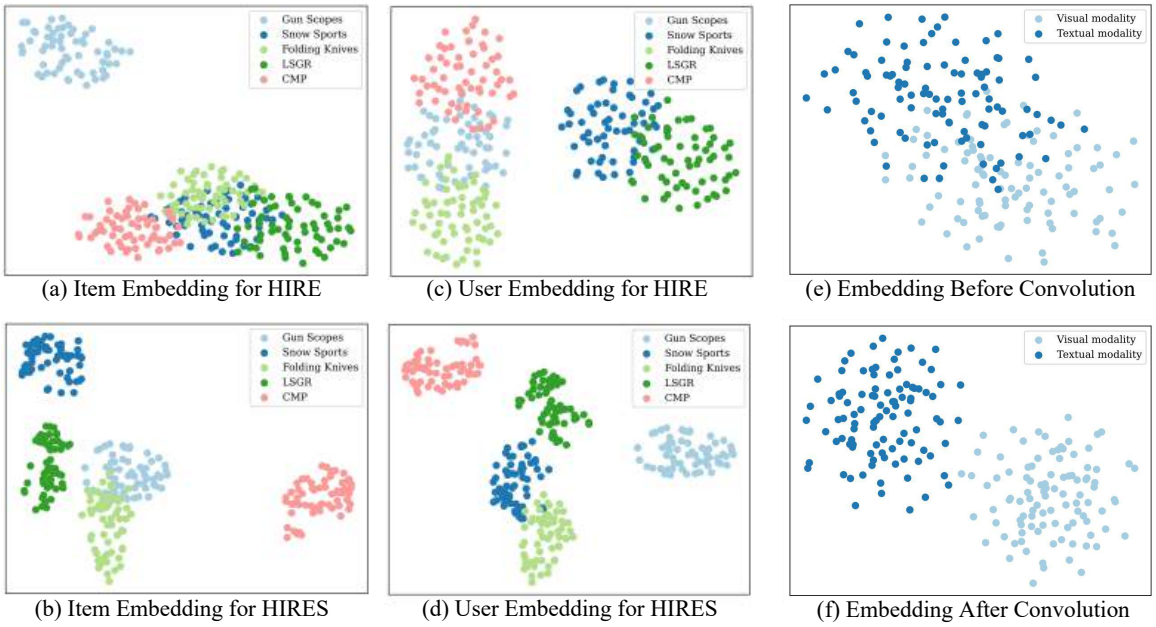


Fig. 8. Visualizations of item embeddings learned by the HIRE and HIRES with the unified hypergraph convolution on the sports dataset.

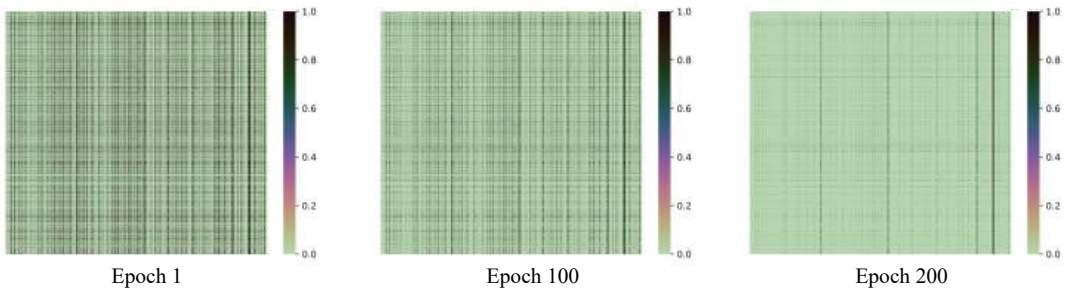


Fig. 9. Visualizations for the unified hypergraph structure under sparse optimal transport mechanism of HIRES

since the visual modalities are relatively close. Therefore, we exploit such hypergraph structure to complete missing modalities to enhance the incomplete multimedia recommendation, where the $item\_4308$ and $item\_4991$ are recommended to user2 with the calendar category. Besides, the sparse hypergraph structure can also enhance the representations with complete modalities, e.g., the recommended $item\_7001$ for user2. In this way, we can jointly conduct hypergraph structure construction and incomplete multimedia recommendation, allowing them supervise each other.

Moreover, to discuss how different modalities interact within the hypergraph and provide more insights into our proposed HIRE and HIRES framework, we use T-SNE to visualize the representations of our proposed HIRE

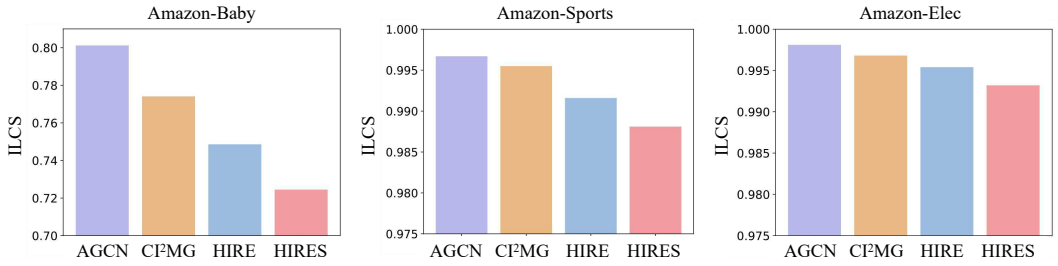


Fig. 10. Diversity analysis with ILCS metric for incomplete multimedia recommendation on Baby, Sports and Elec datasets.

and HIRES after our sparse optimization strategy. We first use the nodes with same color clustering to represent the real different item categories in the sports dataset in Fig. 8 (a)-(d). Then, we provide the visualization of item embedding with our hypergraph convolution under the sparse optimization strategy in in Fig. 8 (e)-(f). As shown in Fig. 8 (a)-(d), Gun Scopes, Snow Sports, Folding Knives, LSGR and CMP represent labeled sports. Besides, we select 200 nodes and construct five categories to illustrate our performance of HIRE and HIRES. From the experimental T-SNE results, we can see that both HIRE and HIRES are effective in accurately grouping all items into five categories only based on user-item interactions within the dataset. The visualization of the experimental results indicates that our proposed HIRE and HIRES can learn distinctive representations with the help of interactions across different modalities. Besides, benefiting from the complementary information between different modalities, clear decision boundaries can be characterized when incomplete recommendation decisions are executed. As shown in Fig. 8 (e)-(f), we illustrate more details on how different modalities interact within the hypergraph. We first visualize the initialized embeddings under visual and textual modalities, where the experimental results show that the initialized modality representations cannot be distinguished, especially for the nodes near the decision boundaries. When we perform the unified hypergraph convolution to capture the cross-modality relations within the hypergraph structure, the hard nodes near the decision boundaries can use the higher-order relations from different modalities within a hyperedge to enhance the representation learning. From the observation, we can find that interacting within the hypergraph can make the nodes near the modality boundaries more discriminative to enhance the incomplete multimedia recommendation. From Fig. 8, we can also observe that both HIRE and HIRES can perform well on dataset, where HIRES can achieve better performance than HIRE.

Finally, we also show the unified hypergraph structure during the training process of our proposed HIRES in Fig. 9 to help to understand how sparse optimal transport mechanism refines the learned structure. Specifically, we first show the initialized dense hypergraph structure containing many irrelevant noisy connections. With the execution of training, benefiting from the sparse optimal transport mechanism, our tailored cross-modality hypergraph can remove the modality-driven and interaction-driven noisy relations out of the hyperedge, which leads to a clear sparse structure. Besides, due to the exploiting of $\ell_{2,1}$-norm, our sparse unified hypergraph structure have the column sparsity, which can enhance the representation learning of hyperedges in our proposed unified hypergraph convolutions.

Diversity analysis for the incomplete multimedia recommendation. To illustrate the diversity of our recommendation results, we show the diversity comparison with respect to the intra-List Category Similarity (ILCS) metric in Fig. 10, where a small value of ILCS means great diversity of the recommendation results. From the observation, we can learn that our proposed HIRE and HIRES surpass all baselines for the diversity evaluation, which illustrates our methods can exploit side relations across modalities to enhance the recommendations.

6 Conclusion

In this paper, we propose a novel framework designed to jointly learn a heterogeneous hypergraph and perform accurate recommendations under incomplete scenarios named HIRE. HIRE first initializes the hypergraph structure by K-means algorithm and exploits a unified heterogeneous hypergraph convolution mechanism to complete the missing multimodal features by high-order relations. Then, the contrastive multimodel recommendation is designed with a textual-aligned self-supervised mechanism to enhance the incomplete multimedia recommendation. Besides, we also devise the HIRE framework with Sparse optimization named HIRES. To refine the hypergraph structure, we uniquely integrate optimal transport and a $\ell_{2,1}$-norm constraint and propose a novel optimization strategy. Extensive experiments demonstrate the clear improvements of HIRES over the state-of-the-art baselines and insightful case studies show the accuracy and interpretability of our proposed methods.

7 Acknowledgment

This work is in part supported by the National Natural Science Foundation of China under Grants 62302098, 6230071268, U21A20472 and 62276065; the National Key Research and Development Plan of China under Grant 2021YFB3600503; the Fujian Provincial Youth Education and Scientific Research Project under Grant JAT220811; the Fujian Provincial Natural Science Foundation of China under Grant 2024J01510026; Fuzhou University Fund for Overseas Academic Visits of Outstanding Students. Carl Yang was not supported by any grants from China.

References

[1] H. Bai, L. Wu, M. Hou, M. Cai, Z. He, Y. Zhou, R. Hong, and M. Wang. Multimodality invariant learning for multimedia-based new item recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 677–686, 2024.

[2] S. T. Barratt. On the differentiability of the solution to convex optimization problems, 2018.

[3] D. Cai, M. Song, C. Sun, B. Zhang, linda Qiao, and H. Li. Hypergraph structure learning for hypergraph neural networks. In *International Joint Conference on Artificial Intelligence*, 2022.

[4] M. A. Carreira-Perpinán and Y. Idelbayev. "learning-compression" algorithms for neural net pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8532–8541, 2018.

[5] C. Chen, M. Zhang, Y. Zhang, W. Ma, Y. Liu, and S. Ma. Efficient heterogeneous collaborative filtering without negative sampling for recommendation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 19–26, 2020.

[6] J. Chen, X. Xin, X. Liang, X. He, and J. Liu. Gdsrec: Graph-based decentralized collaborative filtering for social recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4813–4824, 2022.

[7] L. Chen, L. Wu, R. Hong, K. Zhang, and M. Wang. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 27–34, 2020.

[8] X. Chen, K. Xiong, Y. Zhang, L. Xia, D. Yin, and J. X. Huang. Neural feature-aware recommendation with signed hypergraph convolutional network. *ACM Trans. Inf. Syst.*, 39(1):8:1–8:22, 2020.

[9] W. Du, S. Haoyang, C. Nguyen, and J. Sun. Enhancing product representation with multi-form interactions for multimodal conversational recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023-3 November 2023*, pages 6491–6500, 2023.

[10] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao. Hypergraph neural networks. In *AAAI*, pages 3558–3565, 2019.

[11] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[12] J. Gallego-Posada, J. Ramirez, A. Erraqabi, Y. Bengio, and S. Lacoste-Julien. Controlled sparsity via constrained optimization or: How i learned to stop tuning penalties and love constraints. *Advances in Neural Information Processing Systems*, 35:1253–1266, 2022.

[13] J. Gao, X. Zhao, M. Li, M. Zhao, R. Wu, R. Guo, Y. Liu, and D. Yin. Smlp4rec: An efficient all-mlp architecture for sequential recommendations. *ACM Trans. Inf. Syst.*, 42(3):86:1–86:23, 2024.

[14] P. Gao, H. Tian, and J. Qin. Video frame interpolation with flow transformer. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 1933–1942, 2023.

[15] Y. Gao, Y. Feng, S. Ji, and R. Ji. Hgnn$^+$: General hypergraph neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3181–3199, 2023.

[16] X. Gong, D. J. Higham, and K. Zygalakis. Generative hypergraph models and spectral embedding. *Scientific Reports*, 13(1):540, 2023.
[17] Z. Guo, J. Zhao, L. Jiao, X. Liu, and F. Liu. A universal quaternion hypergraph network for multimodal video question answering. *IEEE Trans. Multim.*, 25:38–49, 2023.
[18] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, 2020.
[19] C. Huang, J. Cui, Y. Fu, D. Huang, M. Zhao, and L. Li. Incomplete multi-view clustering network via nonlinear manifold embedding and probability-induced loss. *Neural Networks*, 163:233–243, 2023.
[20] C. Huang, L. Xia, X. Wang, X. He, and D. Yin. Self-supervised learning for recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 5136–5139, New York, NY, USA, 2022. Association for Computing Machinery.
[21] H. Huang, K. Han, B. Xu, and T. Gan. Reconstructing diffusion networks from incomplete data. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 3085–3091. ijcai.org, 2022.
[22] C. Huo, D. Jin, Y. Li, D. He, Y.-B. Yang, and L. Wu. T2-gnn: Graph neural networks for graphs with incomplete features and structure via teacher-student distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4339–4346, 2023.
[23] Y. Kim, T. Kim, W.-Y. Shin, and S.-W. Kim. Monet: Modality-embracing graph convolutional network and target-aware attention for multimedia recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 332–340, 2024.
[24] S. G. Krantz and H. R. Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.
[25] H. Kuroda and D. Kitahara. Graph-structured sparse regularization via convex optimization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5538–5542. IEEE, 2022.
[26] C. Lemaire, A. Achkar, and P.-M. Jodoin. Structured pruning of neural networks with budget-aware regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9108–9116, 2019.
[27] J. Li, Z. Tan, J. Wan, Z. Lei, and G. Guo. Nested collaborative learning for long-tailed visual recognition, 2022.
[28] X. Li, Q. Sun, Z. Ren, and Y. Sun. Dynamic incomplete multi-view imputing and clustering. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3412–3420, 2022.
[29] X. Li, M. Zhang, S. Wu, Z. Liu, L. Wang, and P. S. Yu. Dynamic graph collaborative filtering, 2021.
[30] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and S. Member. Dual Contrastive Prediction for Incomplete Multi-View Representation Learning. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2023.
[31] Z. Lin, Y. Tan, Y. Zhan, W. Liu, F. Wang, C. Chen, S. Wang, and C. Yang. Contrastive intra- and inter-modality generation for enhancing incomplete multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 6234–6242. ACM, 2023.
[32] Z. Lin, C. Tian, Y. Hou, and W. X. Zhao. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2320–2329, 2022.
[33] Z. Lin, Q. Yan, W. Liu, S. Wang, M. Wang, Y. Tan, and C. Yang. Automatic hypergraph generation for enhancing recommendation with sparse optimization. *IEEE Transactions on Multimedia*, 2023.
[34] H. Liu, Y. Fan, H. Li, J. Wang, D. Hu, C. Cui, H. H. Lee, H. Zhang, and I. Oguz. Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities, 2022.
[35] K. Liu, F. Xue, D. Guo, L. Wu, S. Li, and R. Hong. MEGCF: multimodal entity graph collaborative filtering for personalized recommendation. *ACM Trans. Inf. Syst.*, 41(2):30:1–30:27, 2023.
[36] H. Luo, H. E, G. Chen, Y. Zheng, X. Wu, Y. Guo, Q. Lin, Y. Feng, Z. Kuang, M. Song, Y. Zhu, and L. A. Tuan. Hypergraphrag: Retrieval-augmented generation with hypergraph-structured knowledge representation, 2025.
[37] H. Luo, H. E, Y. Guo, Q. Lin, X. Wu, X. Mu, W. Liu, M. Song, Y. Zhu, and L. A. Tuan. Kbqa-o1: Agentic knowledge base question answering with monte carlo tree search, 2025.
[38] H. Luo, H. E, Z. Tang, S. Peng, Y. Guo, W. Zhang, C. Ma, G. Dong, M. Song, W. Lin, Y. Zhu, and A. T. Luu. ChatKBQA: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, 2024.
[39] H. Luo, H. E, Y. Yang, T. Yao, Y. Guo, Z. Tang, W. Zhang, S. Peng, K. Wan, M. Song, W. Lin, Y. Zhu, and A. T. Luu. Text2nkg: Fine-grained n-ary relation extraction for n-ary relational knowledge graph construction. In *Advances in Neural Information Processing Systems*, volume 37, pages 27417–27439, 2024.
[40] H. Ma, Y. Yang, L. Meng, R. Xie, and X. Meng. Multimodal conditioned diffusion model for recommendation. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, pages 1733–1740, 2024.
[41] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.
[42] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback, 2012.

[43] X. Shen, Y. Chen, S. Pan, W. Liu, and Y. Zheng. Graph convolutional incomplete multi-modal hashing. In *Proceedings of the 31st ACM international conference on multimedia*, pages 7029–7037, 2023.
[44] B. Subbareddy, A. Siripuram, and J. Zhang. Graph learning under spectral sparsity constraints. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5405–5409. IEEE, 2021.
[45] L. Sun, J. Wen, C. Liu, L. Fei, and L. Li. Balance guided incomplete multi-view spectral clustering. *Neural Networks*, 166:260–272, 2023.
[46] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1405–1414, New York, NY, USA, 2020. Association for Computing Machinery.
[47] Y. Tan, C. J. Yang, X. Wei, Z. Wu, W. Liu, and X. Zheng. Partial relaxed optimal transport for denoised recommendation. In *Proceedings of the Workshop on Deep Learning for Search and Recommendation (DL4SR 2022) co-located with the 31st ACM International Conference on Information and Knowledge Management (CIKM 2022), Atlanta, Georgia, USA, October 17-21, 2022*, volume 3317, 2022.
[48] X. Tang, L. Chen, H. Shi, and D. Lyu. Dhyper: A recurrent dual hypergraph neural network for event prediction in temporal knowledge graphs. *ACM Trans. Inf. Syst.*, 42(5):129:1–129:23, 2024.
[49] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, and T.-S. Chua. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:5107–5116, 2023.
[50] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 2021.
[51] Q. Wang, L. Zhan, P. Thompson, and J. Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1828–1838, New York, NY, USA, 2020. Association for Computing Machinery.
[52] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 165–174, New York, NY, USA, 2019. Association for Computing Machinery.
[53] W. Wei, C. Huang, L. Xia, and C. Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 790–800, New York, NY, USA, 2023. Association for Computing Machinery.
[54] Y. Wei, X. Wang, Q. Li, L. Nie, Y. Li, X. Li, and T.-S. Chua. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5382–5390, 2021.
[55] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.
[56] C. Wu, S. Shi, C. Wang, Z. Liu, W. Peng, W. Wu, D. Kong, H. Li, and K. Gai. Enhancing recommendation accuracy and diversity with box embedding: A universal framework. In *Proceedings of the ACM on Web Conference 2024*, pages 3756–3766, 2024.
[57] H. Wu, J. Long, N. Li, D. Yu, and M. K. Ng. Adversarial auto-encoder domain adaptation for cold-start recommendation with positive and negative hypergraphs. *ACM Trans. Inf. Syst.*, 41(2):33:1–33:25, 2023.
[58] L. Wu, Y. Yang, K. Zhang, R. Hong, Y. Fu, and M. Wang. 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*, SIGIR '20, page 679–688, New York, NY, USA, 2020. Association for Computing Machinery.
[59] L. Xia, C. Huang, Y. Xu, P. Dai, and L. Bo. Multi-behavior graph neural networks for recommender system. *IEEE Trans. Neural Networks Learn. Syst.*, 35(4):5473–5487, 2024.
[60] L. Xia, C. Huang, Y. Xu, J. Zhao, D. Yin, and J. Huang. Hypergraph contrastive collaborative filtering. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, pages 70–79, 2022.
[61] W. Xu, H. Jiang, and X. Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446, 2024.
[62] Z. Yi, X. Wang, I. Ounis, and C. Macdonald. 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*, SIGIR '22, page 1807–1811, New York, NY, USA, 2022. Association for Computing Machinery.
[63] Z. Yin, K. Han, P. Wang, and X. Zhu. H3GNN: hybrid hierarchical hypergraph neural network for personalized session-based recommendation. *ACM Trans. Inf. Syst.*, 42(3):63:1–63:30, 2024.
[64] J. Yu, H. Yin, J. Li, Q. Wang, N. Q. V. Hung, and X. Zhang. Self-supervised multi-channel hypergraph convolutional network for social recommendation. In *WWW*, pages 413–424, 2021.
[65] J. Zeng, T. Liu, and J. Zhou. 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*, pages 1545–1554, 2022.
[66] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, and S. Ma. Learning discrete representations via constrained clustering for effective and efficient dense retrieval. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6504–6508, 2023.

[67] C. Zhang, Y. Cai, G. Lin, and C. Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*, pages 12200–12210, 2020.
[68] C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, and J. Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2418–2428, 2022.
[69] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3872–3880, 2021.
[70] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, and L. Wang. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9154–9167, 2022.
[71] X. Zhang, B. Xu, F. Ma, C. Li, Y. Lin, and H. Lin. Bi-preference learning heterogeneous hypergraph networks for session-based recommendation. *ACM Trans. Inf. Syst.*, 42(3):68:1–68:28, 2024.
[72] S. Zhao, W. Wei, X. Mao, S. Zhu, M. Yang, Z. Wen, D. Chen, and F. Zhu. Multi-view hypergraph contrastive policy learning for conversational recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval,SIGIR 2023,Taipei, Taiwan, July 23-27,2023*, pages 654–664, 2023.
[73] X. Zhou and Z. Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 935–943, 2023.
[74] X. Zhou, W. Zhang, H. Xu, and T. Zhang. Effective sparsification of neural networks with global sparsity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3599–3608, 2021.
[75] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, and F. Jiang. Bootstrap latent representations for multi-modal recommendation. In Y. Ding, J. Tang, J. F. Sequeda, L. Aroyo, C. Castillo, and G. Houben, editors, *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 845–854. ACM, 2023.