LGMRec: Local and Global Graph Learning for Multimodal Recommendation

Zhiqiang Guo\textsuperscript{1}, Jianjun Li\textsuperscript{*}, Guohui Li\textsuperscript{2}, Chaoyang Wang\textsuperscript{3}, Si Shi\textsuperscript{4}, Bin Ruan\textsuperscript{1}

\textsuperscript{1} School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
\textsuperscript{2} School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China
\textsuperscript{3} Wuhan Digital Engineering Institute, Wuhan, China
\textsuperscript{4} Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
\{zhiqiangguo, jianjunli, guohuili\}@hust.edu.cn, sunwardtree@outlook.com, shisi@gml.ac.cn, binruan0227@gmail.com

Abstract

The multimodal recommendation has gradually become the infrastructure of online media platforms, enabling them to provide personalized service to users through a joint modeling of user historical behaviors (e.g., purchases, clicks) and item various modalities (e.g., visual and textual). The majority of existing studies typically focus on utilizing modal features or modal-related graph structure to learn user local interests. Nevertheless, these approaches encounter two limitations: (1) Shared updates of user ID embeddings result in the consequential coupling between collaboration and multimodal signals; (2) Lack of exploration into robust global user interests to alleviate the sparse interaction problems faced by local interest modeling. To address these issues, we propose a novel Local and Global Graph Learning-guided Multimodal Recommender (LGMRec), which jointly models local and global user interests. Specifically, we present a local graph embedding module to independently learn collaborative-related and modality-related embeddings of users and items with local topological relations. Moreover, a global hypergraph embedding module is designed to capture global user interests from the user-item graph that integrates multimodal contents. The global embeddings acquired within the hypergraph embedding space can then be combined with two decoupled local embeddings to improve the accuracy and robustness of recommendations. Extensive experiments conducted on three benchmark datasets demonstrate the superiority of our LGMRec over various state-of-the-art recommendation baselines, showcasing its effectiveness in modeling both local and global user interests.

Introduction

With the explosive growth of massive multimedia information (e.g., images, texts, and videos) on online media platforms, such as YouTube and TikTok, a lot of efforts have been devoted to multimodal recommender systems (MRSs) to assist these platforms in providing personalized services to users. Nowadays, the primary task of MRSs is to design an effective way to integrate item multimodal information into traditional user-item interaction modeling frameworks to capture comprehensive user interests.

Some early studies on MRSs adopt either the linear fusion between item modal features and their ID embeddings (He et al. 2016; Liu et al. 2017; Wei et al. 2019) or the attention mechanism on item modalities (Chen et al. 2017, 2019; Liu et al. 2019) to model representations of users and items. However, the efficacy of these models is somewhat constrained as they only model low-order user-item interactions. The surge of research on graph-based recommendations (Wang et al. 2019; He et al. 2020; Mao et al. 2021; Wu et al. 2021) has sparked a wave of explorations in using graph neural networks (GNN) to enhance multimodal recommendations. These works typically capture higher-order user interests from the user-item graph that integrates multimodal contents (Wei et al. 2019, 2020; Wang et al. 2021; Yi et al. 2022; Tao et al. 2022; Wei et al. 2023), or construct modality-aware auxiliary graph structures to transfer multimodal knowledge into item and user embeddings (Zhang et al. 2021, 2022a; Zhou et al. 2023).

Though achieving remarkable progress, existing studies on MRSs still suffer from the following two limitations in modeling user interests. (1) **Coupling.** Firstly, collabora-
tion and multimodal information provide different avenues for exploring user interests. In general, collaborative signals emphasize similar user behavior patterns, while modal knowledge is reflected through content similarity. However, prior works (Wei et al. 2019; Yi et al. 2022) often overlook this matter and share user ID embeddings in both collaborative and multimodal modeling modules (red line in Figure 1 (a)) to learn user interests that couple collaborative and multimodal signals. Experimentally, we randomly select two users from the Baby dataset and exhibit the gradient comparison of their ID embeddings (with 64 dimensions) from the collaborative and multimodal modeling modules in Figure 1 (b). In the early stages of training, the ratio of gradients with opposite directions (orange bar) from the two modules in all dimensions exceeds 50% for each user, which demonstrates that collaborative and multimodal signals generally have different guidance for user embedding learning. Though this ratio slightly decreases as the training continues, the coupling design still restricts stable updates of user embeddings. (2) Locality. Secondly, most existing methods (Tao et al. 2022; Zhou et al. 2023) only learn local user interests from the interaction graph (Figure 1 (c)), lacking the exploration of user global interests. Sparse user-item interactions limit their modeling of robust user interests. As shown in Figure 1 (d), user global (general) interests are usually related to item attribute labels that do not rely on the local interactions. Specifically, items usually have multiple common attributions from visual space, such as color, style, shape. Users have different interests in various attributes. For example, u1 may like clothes with bright colors, while u2 prefers a simple style. A method that modeling only local interests may recommend the shirt i1 to u2 based on similar behaviors, i.e., same purchases (i2, i3, i4) between u1 and u2. But, the global interests of u2 can provide additional guidance, making it more likely to recommend the outerwear i5 with simple style that match u2’s true interests.

To address the aforementioned issues, we propose a novel Local and Global Graph Learning-guided Multimodal Recommender (LGMRec), which explores capturing and exploiting both local and global representations of users and items to facilitate multimodal recommendation. Specifically, to address the first limitation, we present the local graph embedding module to independently capture collaborative-related and modality-related local user interests by performing message propagation on user-item interaction graphs with ID embeddings and modal features, respectively. In view of the many-to-many dependency relationship between attributes and items is similar to that between hyperedges and nodes in hypergraphs, we further consider each implicit attribute as a hyperedge, and present a global hypergraph embedding module to model hypergraph structure dependencies, so as to address the second limitation. Extensive experimental results on three real-world datasets demonstrate that LGMRec surpasses various recommendation baselines significantly, and verify its effectiveness and robustness in modeling local and global user interests.

**Related Work**

**Graph-based Recommendation** The powerful ability of graph neural networks (Kipf and Welling 2016; Hu et al. 2019) in modeling high-order connectivity has greatly promoted the development of recommender systems. Specifically, graph-based recommendation methods model user and item representations by naturally converting the user history interactions into a user-item bipartite graph. Early studies directly inherit the message propagation mechanism of vanilla graph neural network to aggregate high-order neighbor information to represent users and items (Berg, Kipf, and Welling 2017; Ying et al. 2018; Wang et al. 2019). Later, by simplifying the message propagation process, some graph-based recommendation methods further improve recommendation performance (Chen et al. 2020; He et al. 2020; Mao et al. 2021). Additionally, some other methods explore more node dependencies to enhance the representations of users and items (Ma et al. 2019; Sun et al. 2019, 2020a; Li et al. 2022). Later, contrastive learning is also adopted to enhance graph-based recommendations (Lee et al. 2021; Wu et al. 2021; Yu et al. 2022; Lin et al. 2022; Yang et al. 2021; Cai et al. 2023) to construct contrastive views. However, since no modality features are considered, their modeling abilities are limited by sparse interactions.

**Hypergraph learning for Recommendation** By constructing the hyperedge structure containing more than two nodes, hypergraph learning (Feng et al. 2019; Gao et al. 2020) can enhance the generalization ability of the model via capturing complex node dependencies. Some recommendation methods (Ji et al. 2020; Wang et al. 2020; He et al. 2021; Yu et al. 2021; Xia et al. 2021; Zhang et al. 2022b) try to build hypergraph structures and node-hyperedge connections to capture high-order interaction patterns and achieve substantial performance improvements. To further improve performance, several recently developed methods (Xia et al. 2022; Xia, Huang, and Zhang 2022) combine self-supervised learning and hypergraph learning to model robust user and item representations. For example, HCCF (Xia et al. 2022) enhances collaborative filtering with the hypergraph-guided self-supervised learning paradigm. Different from these works that generate hypergraph dependencies via only collaborative embeddings, our work achieves hypergraph structure learning with the modeling of modality-aware global relations.

**Multi-modal Recommendation** The multi-modal recommendation has become the basic application on online media platforms to provide personalized services to users by analyzing the massive multi-modal information (e.g., images and textual descriptions) and user historical behaviors (e.g., reviews, clicks). Early studies on MRSs usually incorporate multi-modal contents as side information to extend the vanilla CF framework (He and McAuley 2016; Chen, He, and Kan 2016; Gao, Zhang, and Xu 2017; Du et al. 2020) or utilize deep autoencoder to model modal features (Guo et al. 2022; Liu et al. 2022). Inspired by the great success of graph-based recommendation methods (He et al. 2020;
Ma et al. 2019; Mao et al. 2021), some studies directly model user high-order interests on modality-specific interaction graphs (Wei et al. 2019, 2020; Sun et al. 2020b; Du et al. 2022; Kim et al. 2022). For instance, MMGCN (Wei et al. 2019) incorporates modality information into the graph message passing to infer modality-related user preferences. Another line utilizes auxiliary semantic graph structures learned from multimodal features to enhance user or item representations (Wang et al. 2021; Zhang et al. 2021). For example, LATTICE (Zhang et al. 2021) is a representative method that exploits modal content similarity to generate representations (Wang et al. 2021; Zhang et al. 2021). For example, LATTICE (Zhang et al. 2021) is a representative method that exploits modal content similarity to generate auxiliary latent item semantic relations to promote recommendation. Recently, Some works (Wei et al. 2021; Yi et al. 2022; Tao et al. 2022; Zhang et al. 2022a; Zhong et al. 2023; Wei et al. 2023) introduce contrastive learning into MRSs to model robust user and item representations. However, these methods usually perform message passing along the edges of user-item interactions to obtain local user interests, failing to explore modality-aware comprehensive user interests.

**Methodology**

In this section, we first formulate the problem of multimodal recommendation and present the overall framework of our LGMRec, and then introduce each component in detail.

**Problem Statement and Overview**

We set the user set as \( \mathcal{U} = \{u\} \) and item set as \( \mathcal{I} = \{i\} \). The ID embeddings of each user \( u \in \mathcal{U} \) and item \( i \in \mathcal{I} \) are denoted as \( e_u \in \mathbb{R}^d \) and \( e_i \in \mathbb{R}^d \), respectively, where \( d \) is the embedding dimension. The user-item interactions can be represented as a matrix \( R \in \mathbb{R}^{d|\mathcal{U}| \times |\mathcal{I}|} \), in which the element \( r_{u,i} = 1 \) if user \( u \) interacts with item \( i \), and \( r_{u,i} = 0 \) otherwise. Based on interaction matrix \( R \), we can construct the user-item interaction graph \( \mathcal{G} = \{\mathcal{U} \cup \mathcal{I}, \mathcal{E}\} \), where \( \mathcal{E} \) is an edge set build on observed interactions, i.e., a nonzero \( r_{u,i} \) corresponds to an edge between user \( u \) and item \( i \) on the graph \( \mathcal{G} \). Further, we incorporate item multimodal contents and denote the original modality feature of item \( i \) generated from pre-trained models as \( e_{m}^{i} \in \mathbb{R}^{d_m} \) under modality \( m \in \mathcal{M} \), where \( \mathcal{M} \) is the set of modalities and \( d_m \) denotes the dimension of modal features. In this work, we consider two mainstream modalities, vision \( v \) and text \( t \), i.e., \( \mathcal{M} = \{v, t\} \).

Given the above settings, the multimodal recommendation algorithm aims to learn a prediction function to forecast the score \( \hat{r}_{u,i} \) of an item \( i \) adopted by a user \( u \) via joint modeling user behaviors and multimodal contents. Formally,

\[
\hat{r}_{u,i} = \text{PREDICTION}(R, E^{id}, \{E^{m}_{i}\}_{m \in \mathcal{M}})
\]

where \( \text{PREDICTION}(\cdot) \) is the prediction function, \( E^{id} = [e_{u}, e_{v;i}, e_{v;u}, e_{t;i}, e_{t;u}] \in \mathbb{R}^{(d|\mathcal{U}|+|\mathcal{I}|) \times d} \) denotes the ID embedding matrix by stacking all the ID embeddings of users and items, \( E^{m}_{i} = [e_{v;i}^{m}, e_{t;i}^{m}] \in \mathbb{R}^{d \times d_m} \) is the item modality feature matrix under modality \( m \).

**Overview.** As illustrated in Figure 2, the framework of LGMRec consists of three major components: (i) Local graph embedding (LGE) module, which adopts GNN to capture collaborative-related and modality-related user local interests on user-item interaction graph with ID embeddings and modal features, respectively; (ii) Global hypergraph embedding (GHE) module, which learns the global user and item representations by capturing the global hypergraph structure dependencies from different item modal feature spaces; and (iii) Fusion and prediction module, which fuses both local and global embeddings to predict final user preference scores for items.

**Local Graph Embedding (LGE) Module**

The LGE module is designed to independently learn the collaborative-related and modality-related user and item rep-
resentations with local topology structure for avoiding unstable updates of user embeddings and promoting decoupled user interest learning.

Collaborative Graph Embedding (CGE) We first capture the high-order connectivity via the message propagation on the user-item interaction graph with ID embeddings. In particular, the collaborative graph propagation function CGPROG(·) in the \((l+1)\)-th layer can be formatted as,

\[
E^{l+1} = \text{CGPROG}(E^l) = \left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right)E^l,
\]

where CGPROG(·) function inherits the lightweight form of the simplified graph convolutional network (Chen et al. 2020; He et al. 2020), \(A \in \mathbb{R}^{(|I|+|Z|) \times (|I|+|Z|)}\) is the adjacency matrix constructed from interaction matrix \(R\) and \(D\) is the diagonal matrix of \(A\). Each diagonal element \(D_{i,i}\) in \(D\) denotes the number of nonzero entries in the \(j\)-th row vector of matrix \(A\). The initial embeddings matrix is set as \(E^0 = E^{id}\). Then, we adopt the layer combination (He et al. 2020) to integrate all embeddings from hidden layers,

\[
E^{id}_{lge} = \text{LAYERCOMB}(E^0, E^1, E^2, \ldots, E^{l}),
\]

where \(E_{lge}^{id} \in \mathbb{R}^{(|I|+|Z|) \times d}\) is collaborative-related embeddings of users and items with local neighborhood information. We use the mean function to achieve LAYERCOMB(·) for embedding integration.

Modality Graph Embedding (MGE) Considering the semantic differences between modalities, we further independently infer the modality-related embeddings of users and items on the interaction graphs with modal features. The original modal features of items are usually generated from semantic differences between modalities, we further independently inferring modal feature embeddings of users and items with local neighborhood information. We use the mean function to achieve LAYERCOMB(·) for embedding integration.

Global Hypergraph Embedding (GHE) Module The GHE module is designed to capture the modality-aware global representations of users and items against sparse and noisy user behaviors.

Hypergraph Dependency Constructing Explicit attribution information of item modalities is often unavailable, especially for visual modalities. Hence, we define learnable implicit attribute vectors \(\{v^m_a\}_{a=1}^{|A|} \in \mathbb{R}^{d_m}\) as hyperedge embeddings under modality \(m\) to adaptively learn the dependencies between implicit attributes and items/users, where \(A\) is the number of hyperedges. Specifically, we obtain hypergraph dependency matrices in low-dimensional embedding space by,

\[
H^m_i = E^m_i \cdot V^m_T, \quad H^m_u = A_u \cdot H^m_{i}^T, \quad (7)
\]

where \(H^m_i \in \mathbb{R}^{|I| \times |A|}\) and \(H^m_u \in \mathbb{R}^{|I| \times |A|}\) are the item-hyperedge and user-hyperedge dependency matrices, respectively. \(E^m_i\) is the raw item modal feature matrix, \(V^m = [v^m_1, \ldots, v^m_A] \in \mathbb{R}^{A \times d_m}\) is the hyperedge vector matrix, and \(A_u \in \mathbb{R}^{|I| \times |A|}\) is the user-related adjacency matrix extracted from \(A\). Intuitively, items with similar modal features are more likely to be connected to the same hyperedge. The user-hyperedge dependencies are indirectly derived from the user-item interactions, which implies the user behavior intention, i.e., the more frequently users interact with items under a certain attribute, the more they may prefer the attribute.

To further avoid the negative impact of meaningless relationships, we employ the Gumbel-Softmax reparameterization (Jang, Gu, and Poole 2017) to ensure that an item is attached to only one hyperedge as much as possible,

\[
\text{h}^m_{i,*} = \text{SOFTMAX} \left( \log \delta - \log(1-\delta) + h^m_{i,*} \right), \quad (8)
\]

where \(h^m_{i,*} \in \mathbb{R}^A\) is the \(i\)-th row vector of \(H^m_i\) that reflects the relations between item \(i\) and all hyperedges. \(\delta \in \mathbb{R}^A\) is a noise vector, where each value \(\delta_j \sim \text{Uniform}(0, 1)\), and \(\tau\) is a temperature hyperparameter. Afterwards, we can get the augmented item-attribute hypergraph dependency matrix \(\tilde{H}^m_i\). By performing similar operations on \(\tilde{H}^m_u\), we can obtain the augmented user-attribute relation matrix \(\tilde{H}^m_u\).

Hypergraph Message Passing By taking the attribute hyperedge as an intermediate hub, we achieve hypergraph message passing to deliver global information to users and items without being limited by hop distances. Formally,

\[
E^{m,h+1}_{i} = \text{DROP}(\tilde{H}^m_i) \cdot \text{DROP}(\tilde{H}^m_{i}^T) \cdot E^{m,h}_i, \quad (9)
\]

where \(E^{m,h}_i\) is the global embedding matrix of items in the \(h\)-th hypergraph layer, and \(\text{DROP}(\cdot)\) denotes a dropout function. We take collaborative embedding \(E_{lge}^{id}\) of items
as the initial global embedding matrix when \( h = 0 \). Further, we can calculate the global user embedding matrix as,
\[
E_{u}^{m,h+1} = \text{DROP}(H_{u}^{m}) \cdot \text{DROP}(H_{i}^{m}) \cdot E_{i}^{m,h}.
\]

(10)

Apparently, the hypergraph passing explicitly enables global information transfer by taking the item collaborative embedding and modality-aware hypergraph dependencies as input. Then, we can obtain the global embeddings matrix \( E_{ghe} \) by aggregating global embeddings from all modalities,
\[
E_{ghe} = \sum_{m \in M} E_{u}^{m,H} \cdot E_{i}^{m,H} = [E_{u}^{m,H}, E_{i}^{m,H}],
\]

(11)

where \( E_{u}^{m,H} \in \mathbb{R}^{|U| \times d} \) and \( E_{i}^{m,H} \in \mathbb{R}^{|I| \times d} \) are global embedding matrices of user \( u \) and item \( i \) obtained in the \( H \)-th hypergraph layer under modality \( m \), respectively.

To further achieve the robust fusion of global embeddings among different modalities, we develop cross-modal hypergraph contrastive learning to distill the self-supervision signals for global interest consistency. Specifically, we take the global embeddings of users acquired in different modalities as positive pairs and different users as negative pairs, and then employ the InfoNCE (Gutmann and Hyvärinen 2010) function to formally define user-side hypergraph contrastive loss as,
\[
L_{HCL}^{u} = \sum_{u \in U} - \log \frac{\exp(s(E_{u}^{m,H}, E_{u}^{m,H})/\tau)}{\sum_{u' \in U} \exp(s(E_{u}^{m,H}, E_{u'}^{m,H})/\tau)},
\]

(12)

where \( s(\cdot) \) is the cosine function, and \( \tau \) is the temperature factor, generally set to 0.2. Note here we only consider visual and textual modalities, i.e., \( m \in \{ v, t \} \).

Similarly, we can define item-side cross-modal contrastive loss \( L_{HCL}^{i} \).

Fusion and Prediction

We acquire the final representations \( E^{*} \) of users and items by fusing their two types of local embeddings \( E_{lge}, E_{lge}^{m} \) and global embeddings \( E_{ghe} \),
\[
E^{*} = E_{lge}^{id} + \sum_{m \in M} \text{NORM}(E_{lge}^{m}) + \alpha \cdot \text{NORM}(E_{ghe}),
\]

(13)

where \( \text{NORM}(\cdot) \) is a normalization function to alleviate the value scale difference among embeddings, \( \alpha \) is an adjustable factor to control the integration of global embeddings.

We then use inner product to calculate the preference score \( \hat{r}_{u,i} \) of user \( u \) towards item \( i \), i.e., \( \hat{r}_{u,i} = e_{u}^{*} \cdot e_{i}^{*\top} \).

The Bayesian personalized ranking (BPR) loss (Rendle et al. 2012) is employed to optimize model parameters,
\[
L_{BPR} = - \sum_{(u,i^{+},i^{-}) \in \mathcal{R}} \ln \sigma(\tilde{r}_{u,i^{+}} - \tilde{r}_{u,i^{-}}) + \lambda_{1}||\Theta||_2^2,
\]

(14)

where \( \mathcal{R} = \{(u,i^{+},i^{-})|(u,i^{+}) \in G_{u}, (u,i^{-}) \notin G_{u}\} \) is a set of triples for training, \( \sigma(\cdot) \) is the sigmoid function, and \( \lambda_{1} \) and \( \Theta \) represent the regularization coefficient and model parameters, respectively.

Finally, we integrate hypergraph contrastive loss with the BPR (Rendle et al. 2012) loss into a unified objective as,
\[
L = L_{BPR} + \lambda_{2} \cdot (L_{HCL}^{u} + L_{HCL}^{i})
\]

(15)

where \( \lambda_{2} \) is a hyperparameter for loss term weighting. We minimize the joint objective \( L \) by using Adam optimizer (Kingma and Ba 2014). The weight-decay regularization term is applied over model parameters \( \Theta \).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#User</th>
<th>#Item</th>
<th>#Interaction</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baby</td>
<td>19,445</td>
<td>7,050</td>
<td>160,792</td>
<td>99.983%</td>
</tr>
<tr>
<td>Sports</td>
<td>35,598</td>
<td>18,357</td>
<td>296,337</td>
<td>99.955%</td>
</tr>
<tr>
<td>Clothing</td>
<td>39,387</td>
<td>23,033</td>
<td>278,677</td>
<td>99.969%</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the three evaluation datasets

Experimental Settings

Datasets

To evaluate our proposed model, we conduct comprehensive experiments on three widely used Amazon datasets (McAuley et al. 2015): Baby, Sports, and Outdoors, Clothing Shoes and Jewelry. We refer to them as Baby, Sports, Clothing for brief. We adopt the 5-core setting to filter users and items for each dataset. The three datasets include both visual and textual modal features. In this work, we use the 4096-dimensional original visual features and 384-dimensional original textual features that have been extracted and published in prior work (Zhou et al. 2023). The statistics of the three datasets are summarized in Table 1.

Evaluation Protocols

For each dataset, we randomly split historical interactions into training, validation, and testing sets with \( 8 : 1 : 1 \) ratio. Two widely used protocols are used to evaluate the performance of top-\( n \) recommendation: Recall (R@\( n \)) and Normalized Discounted Cumulative Gain (He et al. 2015) (N@\( n \)). We tune \( n \) in \( \{10, 20\} \) and report the average results for all users in the testing set.

Parameter Settings

For a fair comparison, we optimize all models with the default batch size 2048, learning rate 0.001, and embedding size \( d = 64 \). For all graph-based methods, the number \( L \) of collaborative graph prorogation layers is set to 2. In addition, we initialize the model parameters with the Xavier method (Glorot and Bengio 2010). For our model, the optimal hyper-parameters are determined via grid search on the validation set. Specifically, the number of modal graph embedding layers and hypergraph embedding layers \( (K \) and \( H \) ) are tuned in \( \{1, 2, 3, 4\} \). The number \( A \) of hyperedge is searched in \( \{1, 2, 4, 8, 16, 32, 64, 128, 256\} \). The dropout ratio \( \rho \) and the adjust factor \( \alpha \) are tuned in \( \{0.1, 0.2, \ldots, 1.0\} \).

We search both the adjst weight \( \lambda_{3} \) of contrastive loss and the regularization coefficient \( \lambda_{1} \) in \( \{e^{-6}, e^{-5}, \ldots, 0.1\} \). The early stop mechanism is adopted, i.e., the training will stop when R@20 on the verification set does not increase for 20 successive epochs. We implement LGMRec\(^2\) with MMRRec (Zhou 2023).

Baselines

We compare our proposed LGMRec with the following four groups of recommendation baselines, including (1) General CF Models: BPR (Rendle et al. 2012); (2) Graph-based Recommenders: LightGCN (He et al. 2020), SGL (Wu et al. 2021), NCL (Lin et al. 2022); (3) Hypergraph-based Recommenders: HCCF (Xia et al. 2022), SHT (Xia, Huang, and Zhang 2022); and (4) Multi-Modal Recommenders: VBPR (He and McAuley 2016), MMGCN (Wei et al. 2019), GRCN (Wei et al. 2020).

\(^{2}\)https://github.com/georgeguo-cn/LGMRec
The performance comparison for all methods on the three datasets is summarized in Table 2, from which we have the following key observations: (1) **The superiority of LGMRec**. LGMRec substantially outperforms all other baselines and achieves promising performance across different datasets. We attribute such significant improvements to: i) The modeling of separated local embeddings that excavates user decoupled interests; ii) The hypergraph learning injects the modality-related global dependencies to local graph embeddings to mitigate interactive sparsity. (2) **The effectiveness of modal features**. Introducing knowledge-rich modality information is beneficial for boosting performance. Experimentally, though only linearly fusing the ID embeddings and modal features of items, the performance of VBPR still outperforms its counterpart (i.e., BPR). By effectively modeling the modal information, the multimodal recommenders (e.g., MMGCN, LATTICE, SLMRec, BM3) with LightGCN as the backbone network basically achieve better results than LightGCN. (3) **The effectiveness of hypergraph learning**. Hypergraph-based recommenders (i.e., HCCF and SHT) outperform the graph-based CF model LightGCN, suggesting the effectiveness of modeling global dependencies under hypergraph architecture. Besides, the significant improvement of LGMRec over competitive baselines further demonstrates the potential of hypergraph networks in modeling modality-aware global dependencies.

### Ablation Study

We conduct ablation studies to explore the compositional effects of LGMRec. From the results reported in Table 3, we can find: (1) The variant w/o MM without multimodal contents degenerates into LightGCN and achieves the worst performance, indicating that introducing modality features can greatly improve accuracy. (2) Removing either LGE or GHE can cause performance drops of LGMRec, demonstrating the benefits of modeling both local and global user interests. Notably, the variant w/o LGE performs worse than w/o GHE, which indicates that local interests directly related to user behavior are more important, and global interests can serve as a supplement. (3) In local graph embeddings, the variant w/o CGE (with MGE only) achieves better performance than w/o MGE (with CGE only) on all datasets, which reveals the importance of integrating multimodal features into user-item interaction modeling. (4) The variant w/o HCL removes hypergraph contrastive learning and only linearly adds all global embeddings. Its performances indicate that contrastive fusion of global embeddings of different modalities can improve performance by modeling the inter-modal global semantic consistency. (5) The variant w/ SUID that still shares user ID embeddings in both MGE and CGE modules performs worse than LGMRec, verifying the benefits of

### Table 2: Overall performances of LGMRec and other baselines on three datasets. The best result is in boldface and the second best is underlined. The t-tests validate the significance of performance improvements with \( p \leq 0.05 \).

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Metrics</th>
<th>Baby R@10</th>
<th>Baby R@20</th>
<th>Baby N@10</th>
<th>Baby N@20</th>
<th>Sports R@10</th>
<th>Sports R@20</th>
<th>Sports N@10</th>
<th>Sports N@20</th>
<th>Clothing R@10</th>
<th>Clothing R@20</th>
<th>Clothing N@10</th>
<th>Clothing N@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>LightGCN</td>
<td>0.0464</td>
<td>0.0732</td>
<td>0.0251</td>
<td>0.0320</td>
<td>0.0553</td>
<td>0.0829</td>
<td>0.0307</td>
<td>0.0379</td>
<td>0.0331</td>
<td>0.0514</td>
<td>0.0181</td>
<td>0.0227</td>
<td></td>
</tr>
<tr>
<td>SGL</td>
<td>0.0532</td>
<td>0.0820</td>
<td>0.0289</td>
<td>0.0363</td>
<td>0.0620</td>
<td>0.0945</td>
<td>0.0339</td>
<td>0.0423</td>
<td>0.0392</td>
<td>0.0586</td>
<td>0.0216</td>
<td>0.0266</td>
<td></td>
</tr>
<tr>
<td>NCL</td>
<td>0.0538</td>
<td>0.0836</td>
<td>0.0292</td>
<td>0.0369</td>
<td>0.0616</td>
<td>0.0940</td>
<td>0.0339</td>
<td>0.0421</td>
<td>0.0410</td>
<td>0.0607</td>
<td>0.0228</td>
<td>0.0275</td>
<td></td>
</tr>
<tr>
<td>HCCF</td>
<td>0.0480</td>
<td>0.0756</td>
<td>0.0259</td>
<td>0.0332</td>
<td>0.0573</td>
<td>0.0857</td>
<td>0.0317</td>
<td>0.0394</td>
<td>0.0342</td>
<td>0.0533</td>
<td>0.0187</td>
<td>0.0235</td>
<td></td>
</tr>
<tr>
<td>SHT</td>
<td>0.0470</td>
<td>0.0748</td>
<td>0.0256</td>
<td>0.0329</td>
<td>0.0564</td>
<td>0.0838</td>
<td>0.0306</td>
<td>0.0384</td>
<td>0.0345</td>
<td>0.0541</td>
<td>0.0192</td>
<td>0.0243</td>
<td></td>
</tr>
<tr>
<td>VBPR</td>
<td>0.0424</td>
<td>0.0663</td>
<td>0.0223</td>
<td>0.0284</td>
<td>0.0556</td>
<td>0.0854</td>
<td>0.0301</td>
<td>0.0378</td>
<td>0.0281</td>
<td>0.0412</td>
<td>0.0158</td>
<td>0.0191</td>
<td></td>
</tr>
<tr>
<td>MMGCN</td>
<td>0.0498</td>
<td>0.0749</td>
<td>0.0261</td>
<td>0.0315</td>
<td>0.0582</td>
<td>0.0825</td>
<td>0.0305</td>
<td>0.0382</td>
<td>0.0329</td>
<td>0.0564</td>
<td>0.0219</td>
<td>0.0253</td>
<td></td>
</tr>
<tr>
<td>GRCN</td>
<td>0.0531</td>
<td>0.0835</td>
<td>0.0291</td>
<td>0.0370</td>
<td>0.0600</td>
<td>0.0921</td>
<td>0.0324</td>
<td>0.0407</td>
<td>0.0431</td>
<td>0.0664</td>
<td>0.0230</td>
<td>0.0289</td>
<td></td>
</tr>
<tr>
<td>LATTICE</td>
<td>0.0536</td>
<td>0.0858</td>
<td>0.0287</td>
<td>0.0370</td>
<td>0.0618</td>
<td>0.0950</td>
<td>0.0337</td>
<td>0.0423</td>
<td>0.0459</td>
<td>0.0702</td>
<td>0.0253</td>
<td>0.0306</td>
<td></td>
</tr>
<tr>
<td>MMGCL</td>
<td>0.0522</td>
<td>0.0778</td>
<td>0.0289</td>
<td>0.0355</td>
<td>0.0660</td>
<td>0.0994</td>
<td>0.0362</td>
<td>0.0448</td>
<td>0.0438</td>
<td>0.0669</td>
<td>0.0239</td>
<td>0.0297</td>
<td></td>
</tr>
<tr>
<td>MICRO</td>
<td>0.0570</td>
<td>0.0905</td>
<td>0.0310</td>
<td>0.0406</td>
<td>0.0675</td>
<td>0.1026</td>
<td>0.0365</td>
<td>0.0463</td>
<td>0.0496</td>
<td>0.0743</td>
<td>0.0264</td>
<td>0.0332</td>
<td></td>
</tr>
<tr>
<td>SLMRec</td>
<td>0.0540</td>
<td>0.0810</td>
<td>0.0296</td>
<td>0.0361</td>
<td>0.0676</td>
<td>0.1007</td>
<td>0.0374</td>
<td>0.0462</td>
<td>0.0452</td>
<td>0.0675</td>
<td>0.0247</td>
<td>0.0303</td>
<td></td>
</tr>
<tr>
<td>BM3</td>
<td>0.0538</td>
<td>0.0857</td>
<td>0.0301</td>
<td>0.0378</td>
<td>0.0659</td>
<td>0.0979</td>
<td>0.0354</td>
<td>0.0437</td>
<td>0.0450</td>
<td>0.0669</td>
<td>0.0243</td>
<td>0.0295</td>
<td></td>
</tr>
<tr>
<td>LGMRec</td>
<td>0.0644</td>
<td>0.1002</td>
<td>0.0349</td>
<td>0.0440</td>
<td>0.0720</td>
<td>0.1068</td>
<td>0.0390</td>
<td>0.0480</td>
<td>0.0555</td>
<td>0.0828</td>
<td>0.0302</td>
<td>0.0371</td>
<td></td>
</tr>
</tbody>
</table>

| Improv.  | 12.98%  | 10.72%    | 12.58%    | 8.37%       | 6.51%      | 4.09%       | 4.28%       | 3.67%       | 11.90%      | 11.44%      | 14.39%      | 1.75%       |

Table 3: Ablation of different components on LGMRec.
Users
Items
Dependencies
Hyperedge

Users
Interacted
Dependencies
Visual Modality
Textual Modality

Hyperedge
Dependencies
Interacted
Items

Users

<table>
<thead>
<tr>
<th>Hyperedge</th>
<th>Visual Modality</th>
<th>Textual Modality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Case study of learned global dependencies of two users $u_{1344}$ and $u_{4351}$ with four hyperedges on Baby dataset.

Case Study

We qualitatively study the global hypergraph dependencies. Specifically, we randomly select two users $u_{1344}$, $u_{4351}$ with similar global embeddings learned on Baby dataset. Hypergraph dependencies under visual and textual modalities for the two users and the items they interact with are presented in Figure 5. The four hyperedges (squares) are shaded depending on the user-hyperedge dependency score. Moreover, the interacted items (circles) are arranged below the corresponding hyperedges in order, according to the maximum item-hyperedge dependency score. From Figure 5, we can observe that: (1) The user-hyperedge dependencies differ in different modalities. For example, the global interests of user $u_{1344}$ in the visual modality are mainly related to the 4-th attribute hyperedge. Under the textual modality, user $u_{1344}$ has larger dependency scores with the 3-rd hyperedges. Thus, we guess that the four items ($i_{151}$, $i_{906}$, $i_{1167}$, and $i_{2131}$) closely related to head hyperedges can reflect user $u_{1344}$’s true preferences, while item $i_{4663}$ attached to the 1-st hyperedge may be a noise interaction. (2) Although the interacted items are largely non-overlapping, user $u_{4351}$ and user $u_{1344}$ still have similar hyperedge dependencies, demonstrating why their global embeddings are similar. The results further reveal that LGMRec can exploit global hypergraph learning to distill similar knowledge of item modal features for performance improvement.

Conclusion

In this work, we proposed a novel model LGMRec for MRSSs, which captures and utilizes local embeddings with local topological information and global embeddings with hypergraph dependencies. Specifically, we adopted a local graph embedding module to independently learn collaborative-related and modality-related local user interests. A global hypergraph embedding module is further designed to mine global user interests. Extensive experiments on three datasets demonstrated the superiority of our model over various baselines. For future work, we intend to seek better means of modeling the differences and commonalities among modalities for further performance improvement.
Acknowledgements

We would like to thank all anonymous reviewers for their valuable comments. The work was partially supported by the National Key R&D Program of China under Grant No. 2022YFC3802101 and the National Natural Science Foundation of China under Grant No. 62272176.

References


