Multi-modal Recommendation Algorithm based on Graph Structure Learning

Jianqiao Liu, Peng Wang, Gang Sun*

Abstract—**Current multi-modal recommendation techniques often integrate various modal attributes into item ID embeddings to improve item representation, yet it is difficult to identify possible semantic structural relationships among items. Therefore, this paper proposes a multi-modal recommendation algorithm based on graph structure learning. Initially, the algorithm independently discerns item interconnections for each modality via a modal perception structure learning layer, then consolidates various modalities to form a latent item graph. Following this, the graph structure optimizer is employed to remove noise from the user-item interaction graph, and the relationship among higher-order items is integrated into the item representation via graph convolution. Ultimately, the detailed representations of items are integrated into the cooperative filtering framework and merged with Content ID contrastive learning activities to realize synergistic benefits between content and ID. Experimental data from three real datasets indicate a notable enhancement in the recommendation performance of this algorithm.**

Keywords: recommendation algorithm, multi-modal, graph structure, contrastive learning

I. INTRODUCTION

In recent years, graph-based recommendation systems [1,2] have integrated advanced connectivity into the embedding procedure to enhance representation learning, leading to significant success. Lately, there have been numerous efforts to amalgamate diverse modal content within graph-oriented recommendation frameworks. MMGCN [3] establishes user-item interaction graphs for different modalities, obtains user preferences for different modalities, and models user and item representations more accurately. Following MMGCN, GRCN [4] employs a variety of features to enhance the graph of user-item interactions, pinpointing incorrect positive responses and reducing related noise edges. LATTICE [5] employs the advanced interaction semantics inherent in the user-item graph along with the potential semantics of item content derived from the item's structure. HUIGN [6] developed an interactive graph of items, with edges representing item pairs engaged by a single user. HUIGN is capable of extracting user intentions at various stages by segmenting the item graph.

Despite the success of earlier efforts, current research has not been able to thoroughly simulate the relationships between items, a crucial aspect in recommendation systems [7]. Particularly, the co-occurrence of high-order items, users, and items is the sole factor taking into account cooperative relationships [8]. However, the semantic relationships that reflect item content information have not been explicitly modeled. In addition, most recommendation systems typically use ID to represent users and items, which poses a cold start problem. Considering the association between items and rich multi-modal content features in multimedia recommendation, there are rich semantic relationships in multi-modal content, which helps recommendation models to comprehensively discover candidate items. Therefore, by learning the underlying semantic item-item structure and content ID (C-ID) contrastive learning task behind multi-modal content, more content embeddings and broader ID embeddings can be obtained, further improving recommendation performance.

Compared with other multi-modal recommendation algorithms, the main contributions of the multi-modal recommendation algorithm based on graph structure learning proposed in this paper include:

1) The modal perception structure learning layer enables independent learning of item relationships for each modality, followed by aggregating various modalities to form a latent item graph that captures the hidden semantic structural connections among items.

2) By performing graph convolution operations, the affinity between high-order items is explicitly injected into the item representations. These rich item representations are embedded into existing collaborative filtering models and combined with Content ID contrastive learning tasks to achieve complementary enhancement.

3) The tests was performed on three publicly accessible datasets, revealing that the algorithm introduced in this study exhibits commendable recommendation efficacy.

II. RELATED WORK

A Contrastive learning

Self-supervised learning, a nascent technology, acquires representations via self-defined supervised signals from unprocessed data, independent of annotated labels [9]. Contrastive Learning (CL) has evolved into a significant segment of self-supervised learning, focusing on creating strong and distinct representations by clustering positive samples nearer and negative ones more distantly [10]. In order to generate negative samples of visual data, a hierarchical enhancement process was employed, which

Manuscript received April 16, 2025; revised August 7, 2025.

The work was supported in part by Open Project of Anhui Engineering Research Center for Intelligent Computing and Information Innovation (FYKFKT24049, ICII202508).

Jianqiao Liu is a laboratory technician of School of Foreign Languages, Fuyang Normal University, Fuyang 236037, China (e-mail: liujianqiaowgy@163.com).

Peng Wang is a postgraduate student of School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China (e-mail: fengchui024@139.com).

Gang Sun is a professor of School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China (corresponding author, e-mail: ahfysungang@163.com).

includes operations such as color jitter, random flipping, cropping, rotation, and resizing [11]. Recent advancements have broadened the scope of self-supervised learning to include graph representation learning. Velickovic and others [12] presented a goal-oriented function to assess the mutual information (MI) linking global graph embeddings with local node embeddings. GraphCL [13] and GRACE [14] suggested a contrastive goal at the node level to streamline earlier studies. Furthermore, Zhu and colleagues [15] suggested a method for adaptive enhancement contrast, integrating different prior assumptions in the graph's topology and semantics. Broadly, the majority of CL studies are distinct from one another regarding the creation of negative samples and contrasting targets.

B Graph structure learning

GNN have shown remarkable proficiency in the analysis of graph-structured data and are extensively utilized across diverse graph analysis areas, such as classifying nodes, predicting links, retrieving information, among others. Nonetheless, the majority of GNN techniques are acutely attuned to graph structure quality, often necessitating an ideal graph structure, a challenge to develop in real-world scenarios [16]. The iterative nature of GNN, which accumulates data recursively from a node's vicinity to determine its embedding, leads to a cascading impact. Minor disturbances within the graph will spread to adjacent nodes, influencing the embeddings of numerous other nodes. Furthermore, numerous practical uses exist where the original graph configuration is absent. Lately, there's been a surge in literature focusing on the core concept of Graph Structure Learning (GSL), targeting the collective learning of refined graph frameworks and their respective representations. Three varieties of GSL techniques exist: metric learning, probabilistic modeling, and direct optimization approaches. Within some suggestions, despite the inherent ability of user-item interactions to create two-part graphs, the exploration of item-item connections remains infrequent. This study utilizes a metric learning approach to depict edge weights as indicators of distance between two endpoint nodes, effectively modeling the item relationship. This method is ideal for multimedia recommendation due to its capacity to hold detailed content information for assessing the semantic connection between two items.

III. Symbol Definition

In this paper, U,$I(|I|=N)$ symbolizes sets of users and items, respectively. Every user $u \in U$ is linked to a group of items I^u, each receiving positive responses, signifying a preference rating $y_{ui} = 1$ for $i \in I^u$. x_u, $x_i \in \mathbb{R}^d$ represents the input ID embedding for u and i, with d being the dimension of embedding. Beyond the interaction between users and items, multi-modal attributes also function as informational content for items. Depict the modal attributes of item i as $e_i^m \in \mathbb{R}^{d_m}$, with d_m symbolizing the feature dimensions, $m \in M$ is the modality, and M is the collection of modalities. Multi-modal recommendation aims to prioritize user choices based on the anticipated preference score $\hat{y}_{ui}$, thereby precisely forecasting user preferences. This paper focuses solely on the visual and textual formats depicted by $M = \{v, t\}$.

IV. Methodology

The recommendation algorithm based on latent item graph structure learning (LIGSL) proposed in this paper is divided into five parts: latent structure graph construction module, graph structure optimizer module, graph convolution module, feature information fusion module, and downstream collaborative filtering method. The structure of the algorithm is shown in Figure 1.

A Construction of potential structural graph

Features with multiple modes offer detailed and significant information about the content of items. Initially, for every modality m, the initial kNN modality perception map S^m is created using the original characteristics. Assuming that items with similarities tend to interact more than those with dissimilarities, the semantic link between two items is measured by their similarity. Widely used techniques for assessing node resemblance encompass cosine similarity, kernel-based functions, and mechanisms of attention. In this paper, a simple cosine similarity without parameters was chosen. The similarity matrix $S^m \in \mathbb{R}^{N \times N}$ is shown in equation (1):

$$S_{ij}^m = \frac{(e_i^m)^\top e_j^m}{\|e_i^m\| \|e_j^m\|} \tag{1}$$

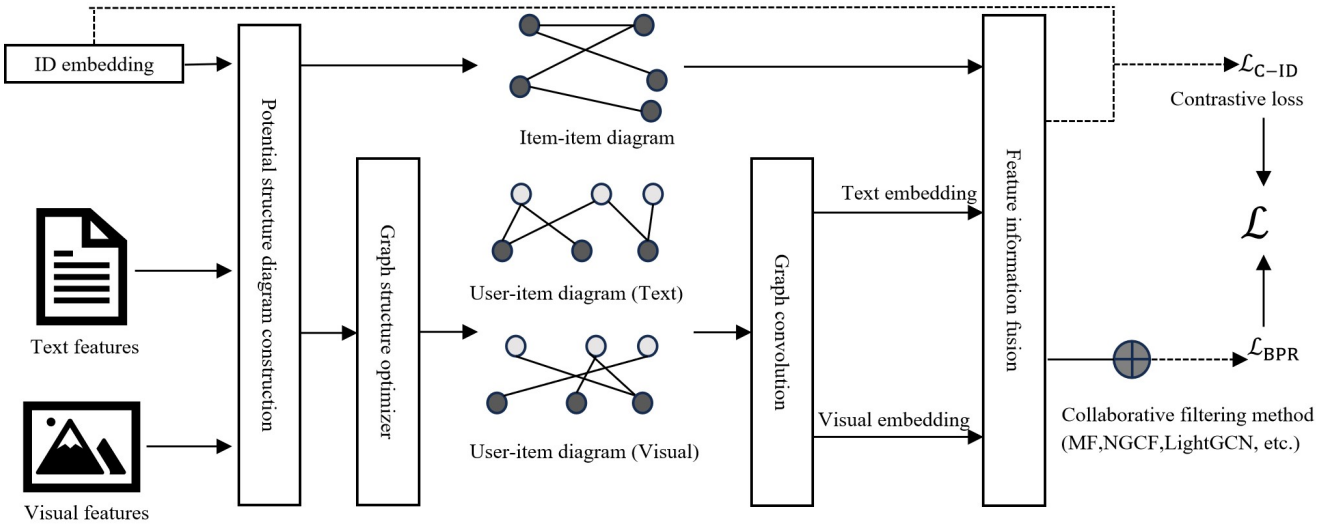

Fig. 1. LIGSL Framework Diagram

where S_{ij}^m represents the *i-th* row and *j-th* column of matrix $S^m \in \mathbb{R}^{N\times N}$. Additionally, apply kNN sparsification to convert the weighted S^m into an unweighted matrix $\hat{S}_{ij}^m$, as depicted in the equation(2). In other words, for every item *i*, solely the linkage of its most similar edges is preserved.

$$\hat{S}_{ij}^m = \begin{cases} 1, & S_{ij}^m \in top\text{-}k(S_i^m) \\ 0, & otherwise \end{cases} \quad (2)$$

Every component in $\tilde{S}^m$ can be either 0 or 1, with 1 symbolizing a possible link between two elements. Normalize the discretized adjacency matrix $\tilde{S}^m$ to $\tilde{S}^m = (D^m)^{-\frac{1}{2}}\hat{S}^m(D^m)^{-\frac{1}{2}}$, where $D^m \in \mathbb{R}^{N\times N}$ is the diagonal matrix of $\tilde{S}^m$ and $D_{ii}^m = \sum_j \hat{S}_{ij}^m$. By employing the derived modality-aware adjacency matrix, a prospective item-item graph is formed through the aggregation of each modality's structure, as depicted in the equation (3):

$$S = \sum_{m\in M} \alpha_m \tilde{S}^m \quad (3)$$

where $S \in \mathbb{R}^{N\times N}$, α_m represents the crucial score for the modality, while M denotes the collection of modalities. One can ascertain the significance score using parameter functions. In this instance, we simplify the model's parameters by incorporating a hyperparameter (symbolizing the significance of visual modality in the creation of *S*), as depicted in the equation:

$$\alpha_t = 1 - \alpha_v \quad (4)$$

If user *u* interacts with item *i*, set the value of each entry A_{ui} in A to 1, otherwise set the value of A_{ui} to 0.

B Diagram Structure Optimization

Drawing inspiration from reference [17], optimizing graph structure involves reducing the graph's size by eliminating surplus edges based on probabilities sensitive to degree. Depict the user-item diagram using G = (V, E), with V representing the set of nodes and E the set of edges. In the user-item graph, the count of users and items stands at M and N, respectively, M + N = |V|, with $|\cdot|$ indicating the set's cardinality. Formulate a balanced adjacency matrix $R \in \mathbb{R}^{M\times N}$ using the user-item interaction matrix E, as depicted in the equation (5):

$$A = \begin{pmatrix} 0 & R \\ R^\top & 0 \end{pmatrix} \quad (5)$$

Given a specific edge $e_k \in E$, calculate its probability $p_k = \frac{1}{\sqrt{\omega_i}\sqrt{\omega_j}}$, wherein ω_i and ω_j represent the degrees of nodes *i* and *j* in graph G. Typically, cut a specific percentage of edges within the graph. Consequently, select edges from a polynomial distribution characterized by index *n* and parameter vector $p = \prec p_0, p_1, \cdots, p_{|E|-1} \succ$. Thereby significantly reducing the likelihood of sampling highly connected nodes' edges in the graph. In other words, there's a higher probability of trimming these edges in G. Subsequently, sample these edges to form a symmetric adjacency matrix A_ρ. In line with the earlier proposed item-item graph, a normalization process was applied to A_ρ, yielding result $\hat{A}_\rho$. Echoing DropEdge [18], LIGSL adjusts the user-item graph and progressively standardizes the sampled adjacency matrix throughout each training phase. Yet, for deducing models, the initial normalized adjacency matrix $\hat{A} = D^{-1/2}AD^{-1/2}$ is employed.

C Graph Convolution

Once the potential structure is established, graph convolution is executed by integrating item affinity into the embedding space, enhancing the learning of item representations. The process of graph convolution may be viewed as the dissemination and compilation of messages. Through the dissemination of item representations from adjacent items, an item is capable of consolidating data in its immediate vicinity. Furthermore, the arrangement of several graph convolutional layers enables the capture of complex item-item relationships.

Utilizing straightforward message dissemination and grouping techniques, bypassing feature alteration and non-linear activation, proves to be both efficient and computationally sound [19]. Within the *l* ayer, the equation illustrates how messages are transmitted and aggregated(6):

$$h_i^{(l)} = \sum_{j\in N(i)} A_{ij} h_j^{(l-1)} \quad (6)$$

In this context, N(i) represents the item next to, and $h_i^{(l)}$ denotes the *l-th* layer's depiction of item *i*. Envision input item $h_i^{(0)}$ as its respective ID embedding vector x_i. Using item ID embedding as input representation instead of multimodal features, given that graph convolution directly measures item-item affinity. Post layering *L* layers, the encoding of complex item-item relationships derived from diverse modal data proves advantageous for subsequent collaborative filtering techniques.

D Feature information fusion

Employing an attention mechanism to merge various item embedding modalities $H_{(L)}^m$, excluding the index *(L)*, and utilizing h_i^m to depict the *i-th* row of $H_{(L)}^m$, representing the graph convolution's output embedding for item *i*. The significance of each modality linked to item *i* is depicted in the equation(7):

$$w_i^m = q^\top \tanh(Wh_i^m + b) \quad (7)$$

where in $q \in \mathbb{R}^d$ symbolizes the attention vector, $W \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$ denote the weight matrix and the bias vector. All modalities are subject to these specified parameters. Once the significance of various modalities is determined, standardize them to derive weight coefficients, as illustrated in the equation(8):

$$\alpha_i^m = \frac{\exp(w_i^m)}{\sum_{m=1}^{|M|} \exp(w_i^m)} \quad (8)$$

Then, the multi-modal fusion embedding of item *i* is shown in equation (9):

$$h_i = \sum_{m=1}^{|M|} \alpha_i^m h_i^m \quad (9)$$

E Combining collaborative filtering

LIGSL obtains representations of items via various modal features and amalgamates them with ensuing cooperative filtering methods to mimic user-item interactions. The flexibility of this model enables it to serve as a ready-to-use element in diverse collaborative filtering methods.

Depict the results of user and item embeddings using joint filtering techniques as $\tilde{x}_u, \tilde{x}_i \in \mathbb{R}^d$, and augment these embeddings by incorporating normalized multi-modal fusion item embeddings h_i, as illustrated in the equation (10):

$$\hat{\boldsymbol{x}}_i = \tilde{\boldsymbol{x}}_i + \frac{\boldsymbol{h}_i}{\|\boldsymbol{h}_i\|_2} \quad (10)$$

Determine the preference score between user and item by computing the inner product of user and enhanced item embedding, as depicted in the equation(11):

$$\hat{y} = \tilde{x}\ \hat{x} \quad (11)$$

F Joint optimization

Employing Bayesian Personalized Ranking (BPR) loss for determining pairwise rankings promotes the prediction of observed items above those not observed. The BPR loss is shown in equation (12):

$$\mathcal{L}_{\text{BPR}} = -\sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{I}_u}\sum_{j\notin\mathcal{I}_u}\ln\sigma\left(\hat{y}_{ui}-\hat{y}_{uj}\right) \quad (12)$$

where I^u represents observation items related to user u, (u, i, j) represents paired training triplets, where $i \in I^u$ represents a positive element and $j \notin I^u$ a negative element derived from unseen interactions. The sigmoid function is denoted as $\sigma(\cdot)$.

To compensate for the cold start deficiency in ID embedding by contrastive learning tasks with content ID, given item i, first select several important ID embeddings and connect them as id_i. Next, linearly transform the multi-modal content embedding h_i and the ID embedding id_i into the same vector space, as shown in equations (13) and (14):

$$C_i = f(h_i) \quad (13)$$

$$I_i = f'(\text{id}_i) \quad (14)$$

where f and f' are fully connected layers, while C_i and I_i are output embeddings with the same dimension. In the following contrastive learning, they are positive samples of each other. Randomly select H negative samples for each C_i and I_i from the training batch, defined as $C_i^- = \{C_{i1}^-, C_{i1}^-, \ldots, C_{iH}^-\}$ and $I_i^- = \{I_{i1}^-, I_{i1}^-, \ldots, I_{iH}^-\}$. Subsequently, apply the negative logarithmic likelihood function to enhance the resemblance of each positive pairs while reducing the likeness of negative pairs, as demonstrated in equations (15) and (16):

$$L_{\text{C2I}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(s(C_i,I_i)/\tau\right)}{\exp\left(s(C_i,I_i)/\tau\right)+\sum_{j=1}^{H}\exp\left(s(C_i,I_{ij}^-)/\tau\right)} \quad (15)$$

$$L_{\text{I2C}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(s(I_i,C_i)/\tau\right)}{\exp\left(s(I_i,C_i)/\tau\right)+\sum_{j=1}^{H}\exp\left(s(I_i,C_{ij}^-)/\tau\right)} \quad (16)$$

where $s(\cdot)$ represents cosine similarity and τ is temperature parameter. Finally, add their average to the ranking loss, and the overall loss function is shown in equation (17):

$$L = L_{\text{BPR}} + 0.5\alpha\cdot(L_{C2I} + L_{I2C}) \quad (17)$$

V. Experiments

A Experimental datasets

Experiments using the Baby, Sports, and Clothing categories from Amazon's publicly accessible dataset were carried out to validate the algorithm's efficacy as suggested in this study. Previous studies have extensively utilized the Amazon review dataset [20, 21]. Each dataset's unprocessed data undergoes preprocessing through a 5-core configuration for both items and users. Table I displays the statistical details of these datasets.

B Evaluation indicators

The experiment employs Recall and NDCG techniques to assess the algorithm's performance in recommending.

1) Recall

Formally, Recall@K is defined as shown in equation (18):

$$\text{Recall}@K = \frac{1}{|U^T|}\sum_{u\in\mathcal{U}^T}\frac{\sum_{i=1}^{K} I\left[I_u^r(i)\in I_u^t\right]}{|I_u^t|} \quad (18)$$

where U^T represents the user set included in the test data, and $I_u^r(i)$ symbolizes the *i-th* item recommended suggested by the user u. The instruction function $I[\cdot]$ serves to determine the quantity of suggested items within the set I_u^t, which represents the items in the test data that interact with user u.

2) NDCG

NDCG@K is defined as shown in equation (19):

$$\text{NDCG}@K = \frac{\text{DCG}@K}{\text{IDCG}@K} = \frac{\sum_{i=1}^{K}\frac{2^{rel_i}-1}{\log_2(i+1)}}{\sum_{i=1}^{|REL_K|}\frac{2^{rel_i}-1}{\log_2(i+1)}} \quad (19)$$

where NDCG@K represents the ideal ranking scenario where the item interacting with the user is located at the top.

C Comparison Models

The effectiveness of LIGSL is demonstrated through a comparison with various recommendation models, such as the universal CF recommendation model and the multi-modal recommendation model.

1) BPR: A model for matrix factorization refined using Bayesian pairwise ranking loss.

2) VBPR: Enhanced the traditional MF structure by integrating visual elements into the BPR loss process. Following earlier studies, merge the item's multi-modal attributes to create its visual characteristics, utilizing these elements for learning user preferences.

3) LightGCN: Presented here is a streamlined graph convolutional network, exclusively executing linear propagation and aggregation among adjacent nodes. Calculate the mean of the hidden layer embeddings to ascertain the ultimate anticipated user and item embeddings.

4) GRCN: Educating enhanced graphs through the lens of user and item depictions. Following the enhancement of the graph, graph convolution is executed to derive depictions of users and items.

5) DualGNN: Develop a supplementary user-user correlation chart to improve user depiction in GCN.

6) LATTICE: Employing the structure of graphs to uncover hidden semantic connections among items, explicitly understood through their multi-modal attributes.

7) SLMRec: Incorporating autonomous learning into suggestions for multimedia. The proposal includes three varieties of data enhancement to uncover diverse patterns in data, aiding in contrastive learning.

8) BM3: Introduced a new contrastive learning method to guide multi-modal recommendation of user and item representations, eliminating the need for negative samples.

9) MMGCN: Merges the various representations created by GCN across different item modalities for the purpose of recommendation.

D Experimental setup

Based on PyTorch to implement the algorithm in this paper, in order to ensure fair comparison, Xavier was used to initialize parameters and Adam was used to optimize the model [21]. Set the regularization coefficient to $\lambda = 10^{-3}$, Set the C-ID contrastive learning temperature parameter to $\tau = 0.5$ and look for the edge trimming ratio ρ sensitive to degree, ranging from {0.8, 0.9}. Regarding convergence, the initial halt and the overall epoch are set at 20 and 1000.

TABLE I
STATISTICAL INFORMATION OF THE DATASETS

Dataset	User	Item	Interaction Quantity	Data Density
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	278,677	99.97%

E Experimental results and analysis

Comparisons were made between the LIGSL and other models using the Baby, Sports, and Clothing datasets, with the findings presented in Table II. The data in Table II reveals that LIGSL markedly outperforms alternative techniques. Specifically, LIGSL has improved the evaluation metrics Recall and NDCG in the Clothing, Sports, and Baby datasets. This method first learns the relationships between items independently for each modality through a modal perception structure learning layer, and aggregates multiple modalities to construct a latent item graph, which can learn rich semantic relationships in multi-modal content. Subsequently, the graph structure optimizer is used to denoise the user-item interaction graph, and the affinity between high-order items is injected into the item representation through graph convolution, which can obtain high-order relationships between items. Finally, the rich item representations are embedded into the collaborative filtering model and combined with Content ID contrastive learning tasks to achieve complementary advantages between content and ID. The experimental results also demonstrate the effectiveness of the proposed method for joint loss optimization through learning potential item-item relationships and contrastive learning tasks.

F Ablation experiment

To investigate how various elements of LIGSL influence performance, these versions of LIGSL were created to analyze their impact via ablation experiments, using LightGCN as the standard collaborative filtering technique in this study:

1) LIGSL-cl: Remove C-ID and only use BPR loss for contrastive learning task.

2) LIGSL-g: Remove the graph structure optimizer module and use random edge dropout.

3) LIGSL-m: The downstream collaborative filtering method is MF.

4) LIGSL-n: The downstream collaborative filtering method is NGCF.

Figure 2 and Figure 3 displays the ablation experiment. The illustration reveals that the content-aware approach typically outperforms the collaborative filtering technique in performance. This suggests that features involving multiple modes offer detailed content insights about the item, enhancing the precision of recommendations. On three datasets, GRCN surpasses other baseline models due to its ability to identify and reduce incorrect positive connections in graphs depicting user-item interactions. Even with its intricate design process, GRCN remains less effective than LATTICE, underscoring the need for precise recording of item-item connections. Current recommendation models, cognizant of content, depend greatly on the multi-modal features' representativeness to attain varying results across diverse datasets.

The Clothing dataset, along with VBPR, MMGCN, and GRCN, essential for uncovering item characteristics via visual attributes, surpass all CF techniques, LightGCN included. In the case of the remaining two datasets, where multi-modal attributes might not directly disclose item characteristics, these content aware methods have yielded only minor enhancements. The efficacy of VBPR and MMGCN falls short when compared to the cooperative filtering technique of LightGCN. Unlike current content-aware techniques, LIGSL distinguishes itself by identifying possible item relationships in multi-modal features instead of directly utilizing them as supplementary data. The likelihood of item relationships is reduced based on the representativeness of multi-modal attributes, thereby enhancing performance. Compared with other variant models, LIGSL consistently maintains the best recommendation performance.

TABLE II
PERFORMANCE COMPARISON

Algorithm	Baby				Sports				Clothing			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0206	0.0303	0.0114	0.0138
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0361	0.0544	0.0197	0.0243
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
GRCN	0.0532	0.0824	0.0282	0.0358	0.0599	0.0919	0.0330	0.0413	0.0421	0.0657	0.0224	0.0284
DualGNN	0.0513	0.0803	0.0278	0.0352	0.0588	0.0899	0.0324	0.0404	0.0452	0.0675	0.0242	0.0298
LATTICE	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
SLMRec	0.0521	0.0772	0.0289	0.0354	0.0663	0.9900	0.0365	0.0450	0.0442	0.0659	0.0241	0.0296
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
MMGCN	0.0421	0.0660	0.0220	0.0282	0.0401	0.0636	0.0209	0.0270	0.0227	0.0361	0.0120	0.0154
LIGSL	**0.0642**	**0.1020**	**0.0339**	**0.0436**	**0.0721**	**0.1107**	**0.0390**	**0.0489**	**0.0640**	**0.0946**	**0.0344**	**0.0426**

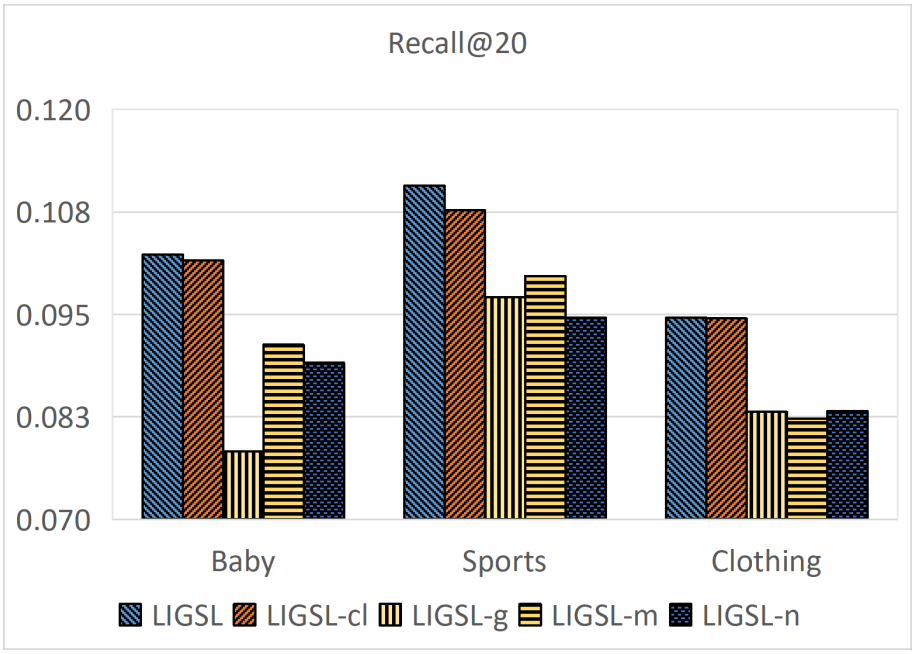

Fig. 2. Recall@20 in ablation experiments

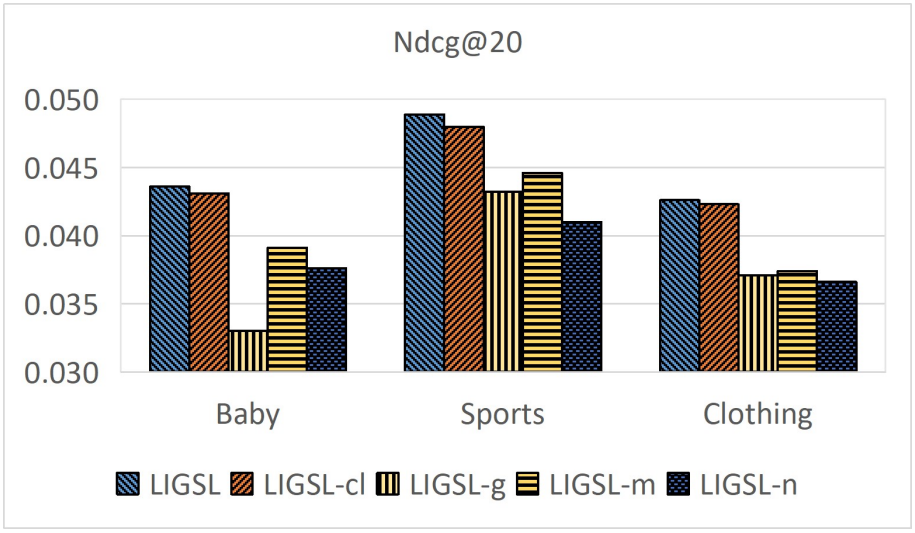

Fig. 3. NDCG@20 in ablation experiments

G *Effects of parameters*

This part delves into the impact of hyperparameters within LIGSL on its recommendation performance.

Adjacency number k

In order to prevent the dispersion of messages from disparate items, a diagram depicting the relationship between items was created using just k items that were most alike. Research indicates that $k=10$ typically represents the optimal figure for the count of neighboring items, as depicted in Figure 4, Figure 5 and Figure 6. While the ideal k value might differ based on the situation, opting for lower values can diminish interference from non-related neighbors.

Regularization coefficient λ

LIGSL utilizes unprocessed text and visual features to create possible item-item graphs. Initially, a experiment study aims to clarify how multi-modal data affects LIGSL's efficiency by methodically altering the input of visual elements from 0 to 1.0 throughout the creation of the graph. A ratio of 0 signifies that item-item graph creation is solely dependent on textual elements, while a ratio of 1.0 indicates that the graph's construction depends exclusively on visual aspects. On three datasets Baby, Sports, and Clothing, the experimental results of Recall@20 and NDCG@20 are shown in Figure 7 and Figure 8. The results indicate that when constructing effective item-item graphs, textual features contain more information than visual features.

VI. Conclusions

In this paper, a recommendation algorithm LIGSL based on graph structure learning is proposed. The multi-modal structure learning framework initiates with a modality-aware graph learning layer that autonomously extracts intra-modal relational patterns among items, subsequently synthesizing a latent heterogeneous graph through cross-modal aggregation. Subsequently, a graph topology refinement module is applied to the user-item interaction network, implementing noise suppression mechanisms to enhance the model's resilience against sparse and noisy observational data. On this basis, the model explicitly injects the affinity between high-order items into the item representation through graph convolution

operation. Finally, this method embeds these rich item representations into existing collaborative filtering models and combines them with content ID contrastive learning tasks to leverage the representation advantages of content and ID, achieving complementary enhancement. This algorithm has certain theoretical significance and practical value. The recommendation algorithm proposed in this paper does not utilize the implicit information in knowledge graphs and social networks. Future research will consider mining relationships between items through knowledge graphs, analyzing relationships between users through social networks, and using knowledge graphs and social networks to model users and items more accurately, in order to improve the performance of recommendation algorithms.

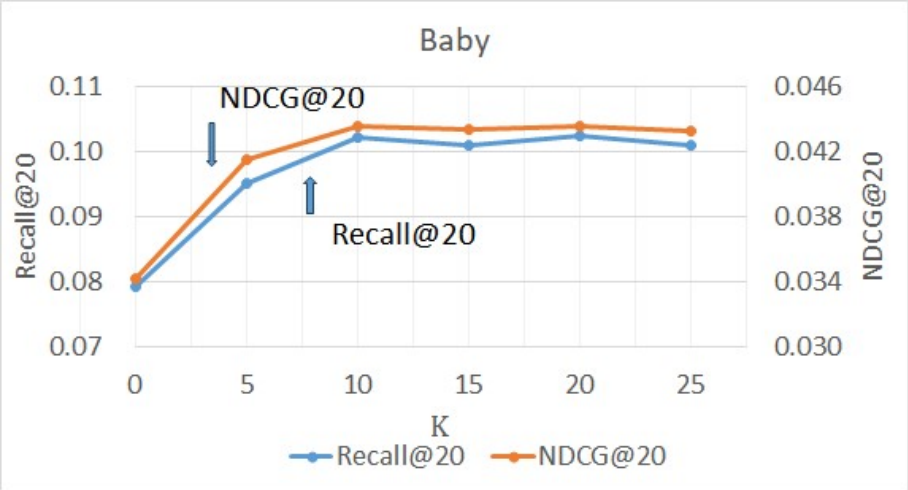

Fig. 4. Performance comparison of different adjacency number *k* in dataset Baby

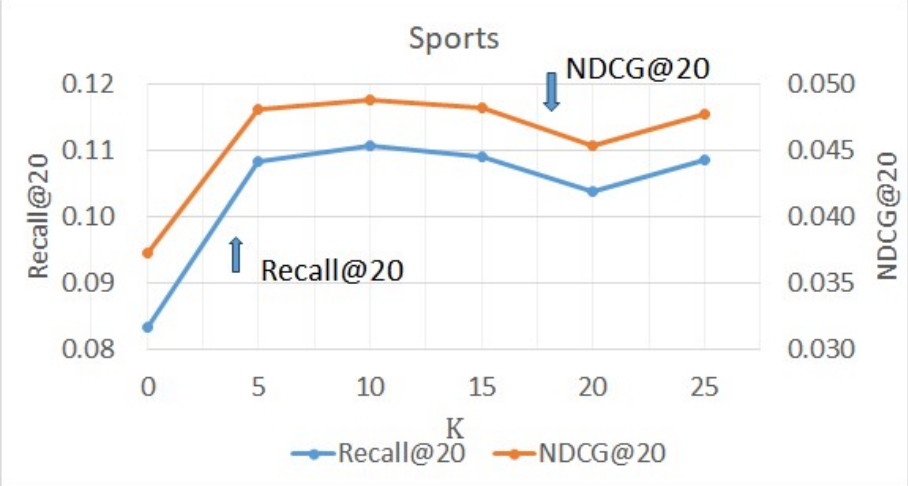

Fig. 5. Performance comparison of different adjacency number *k* in dataset Sports

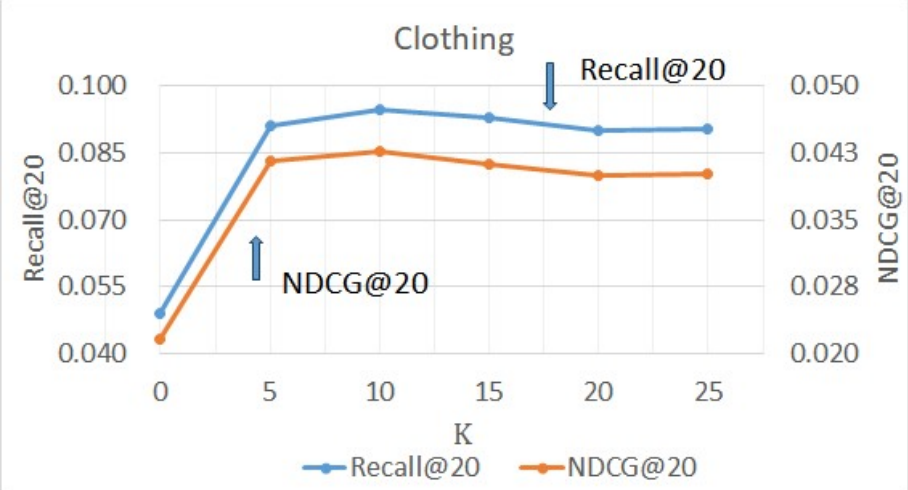

Fig. 6. Performance comparison of different adjacency number *k* in dataset Clothing

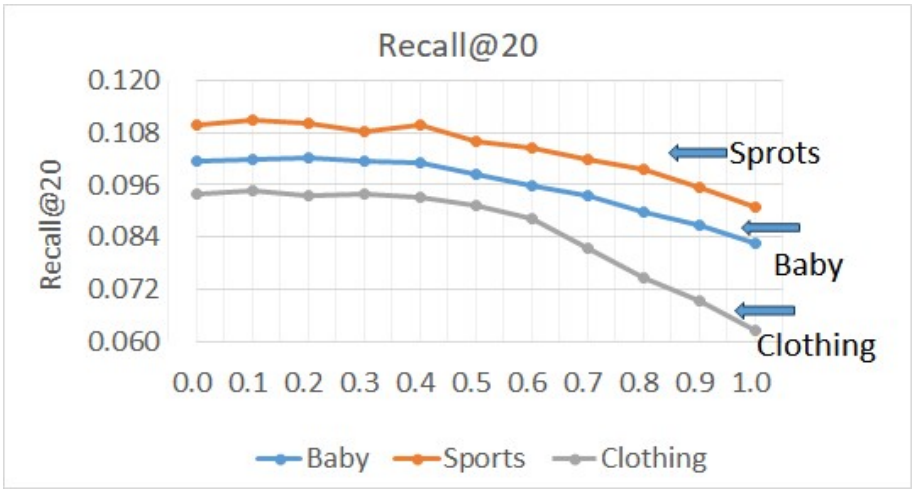

Fig. 7. Recall@20 of different visual feature proportions λ

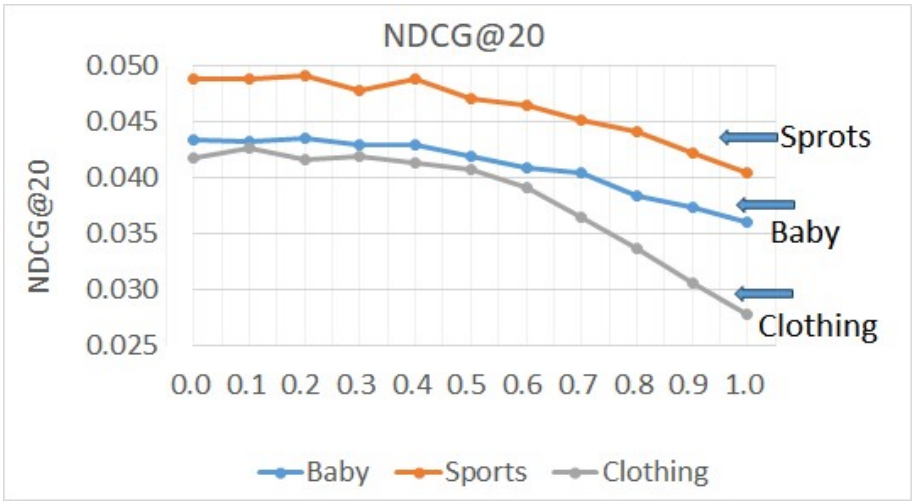

Fig. 8. NDCG@20 of different visual feature proportions λ

REFERENCES

[1] Wang X, He X, Wang M, et al. "Neural Graph Collaborative Filtering", In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp 165-174.

[2] He X, Deng K, Wang X, et al. "LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation", In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp 639-648.

[3] Wei Y, Wang X, Nie L, et al. "MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video", In Proceeding of the ACM Multimedia, 2023, pp 1437-1445.

[4] Wei Y, Wang X, Nie L, et al. "GraphRefined Convolutional Network for Multimedia Recommendation with Implicit Feedbac", In Proceeding of the ACM Multimedia, 2020, pp 3451-3459.

[5] Zhang J, Zhu Y, Liu Q, et al. "Mining Latent Structures for Multimedia Recommendation, In Proceeding of the ACM Multimedia, 2021, pp 3872- 3880.

[6] Zheng HT, Guan R, Wang HW, et al. "Multimodal Graph Fusion for Multimedia Recommendation by Edge-wise Modulation", IEEE Transactions on Multimedia, 2023, pp 378-392.

[7] Sarwar B M, Karypis G, Konstan J A, et al. "Item-based Collaborative Filtering Recommendation Algorithms", In Proceeding of the International Conference of World Wide Web, 2001, pp 285-295.

[8] Wei Y, Wang X, He X, et al. "Hierarchical User Intent Graph Network for Multimedia Recommendation". IEEE Transactions on Multimedia, 2021, pp 384-398.

[9] Liu Q, Wu S, Wang L. Deepstyle: "Learning User Preferences for Visual Recommendation", In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp 841-844.

[10] Rendle S, Feudenthaler C, Gantner Z, et al. BPR: "Bayesian Personalized Ranking from Implicit Feedback", In Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2012, pp 452-461.

[11] Chen X, Chen HX, Xu HT, et al. "Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation", In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp 765-774.

[12] Velickovic P, Fedus W, Hamilton W L, et al. "Deep graph Infomax", In Proceedings of the International Conference on Learning Representations, 2019, pp 446-478.

[13] You Y, Chen T, Sui Y, et al. "Graph Contrastive Learning with Augmentations", NeurIPS, 2020, pp 5812-5823.

[14] Zhu Y, Xu Y, Yu F, et al. "Deep Graph Contrastive Representation Learning", arXiv: 2006.04131, 2020.

[15] Zhu Y, Xu Y, Yu F, et al. "Graph Contrastive Learning with Adaptive Augmentation", In Proceeding of the International Conference of World Wide Web, 2024, pp 2069-2080.

[16] Franceschi L, Nieper M, Pontil M, et al. "Learning Discrete Structures for Graph Neural Networks", In Proceeding of the International Conference on Machine Learning, 2019, pp 1972-1982.

[17] Chen J Y, Zhang HW, He XN, et al. "Attentive Collaborative Filtering: Multimedia Recommendation with Item and Component-Level Attention", In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp 335-344.

[18] Anees K, Luca C, Seyedahamd A et al. "Differentiable Graph Modul for Graph Convolutional Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, pp 1606-1617.

[19] Wu F, Zhang TY, Souza A, et al. "Simplifying Graph Convolutional Networks", In Proceeding of the International Conference on Machine Learning, 2019, pp 6861-6871.

[20] Zhang JH, Zhu YQ, Liu Q, et al. "Mining Latent Structures for Multimedia Recommendation", In Proceedings of the ACM International Conference on Multimedia, 2021, pp 3872-3880.

[21] Hu YF, Koren Y, Chris V. "Collaborative Filtering for Implicit Feedback Datasets", In Proceeding of the IEEE International Conference on Data Mining, 2008, pp 263-272.