2025 11th International Conference on Big Data and Information Analytics (BigDIA) | 979-8-3315-5695-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/BigDIA68682.2025.11382555

Denoising Attentive Hypergraph Learning for Multi-Modal Recommendation

1st Mengfan Kong
National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
kongmf23@nudt.edu.cn

2nd Fei Cai*
National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
caifei08@nudt.edu.cn

3rd Aimin Luo*
National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
amluo@nudt.edu.cn

4th Chonghao Chen
National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
chenchonghao@nudt.edu.cn

5th Zhiqiang Pan
National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
panzhiqiang@nudt.edu.cn

***Abstract*—Multi-modal recommendation has received significant attention in recent years. It typically leverages the item modality features to enhance the item semantics and then incorporates these features into the final u ser/item representations. Existing studies primarily focus on graph-based multi-modal recommendation. However, they often neglect the impact of noisy interaction, and this impact tends to be amplified as neighborhood messages are aggregated. Further, we argue that existing methods fail to distinguish the relative importance of modality-specific g raph e mbeddings i n t he f usion p rocess. To address these issues, we propose a novel Denoising Attentive Hypergraph Learning method for Multi-Modal Recommendation (DAHM). Specifically, w e fi rst de vise a Di richlet degree-sensitive method to prune the potential noisy edges. Then, we introduce a local user interest graph encoder and a global item attribute hypergraph encoder to capture the local user preferences and the global similarities among the item modality attributes, respectively. Next, a projection-enhancement hypergraph attention mechanism is applied to adaptively allocate the fusion weights to modality-specific g raph e mbeddings. E xtensive experiments on three real-world datasets verify the effectiveness of DAHM, presenting an average improvement of 9.55% in terms of Recall@20 against the best baseline. We release our code at https://github.com/fanko79/DAHM2025.**

***Index Terms*—Multi-modal Recommendation, Hypergraph Learning, Projection Enhancement**

I. Introduction

With the explosive growth of multi-modal contents on online platforms, leveraging the item modality features for recommendation has become increasingly prevalent, as item modality can reflect a u ser's fi ne-grained pr eferences fr om a semantic perspective. Multi-modal recommendation systems integrate information from different modalities to construct more comprehensive item representations and better understand user interests, thereby enhancing the effectiveness of recommendation.

Early efforts in multi-modal recommendation extend matrix factorization framework by incorporating visual modality (1). Subsequently, some studies (2; 3) integrated the attention mechanism with collaborative filtering to model the interaction between users and modality-specific features. Existing approaches for multi-modal recommendation mainly resort to graph convolution networks (GCNs), since user-item interactions naturally form a bipartite graph. Such GCN-based multi-modal models typically integrate item modality features into the user interaction graph so as to capture the collaborative and modality semantic signals by propagating the message among the neighboring nodes. For instance, some methods (4; 5) directly construct the modality-specific user-item graph to enrich node representations. Other methods (6; 7) mine the latent relationship between item modalities using a modality-aware item-item graph to supplement the collaborative signal for user-item graph.

Despite the notable success, we argue that existing GCN-based multi-modal recommendation methods still suffer from the following limitations: (1) **Natural noise.** The original user-item interactions inevitably bring the noise (e.g., misclicks and popularity bias). These false-positive interactions contaminate node representations and mislead user preferences modeling during graph convolution. Even worse, in GCN-based multi-modal recommenders, the impact of noisy interactions is further amplified as the multi-modal methods construct more modality-specific interaction graphs than models that only utilize the historical user-item interactions. (2) **Equally treating**

979-8-3315-5695-2/25/$31.00 © 2025 IEEE

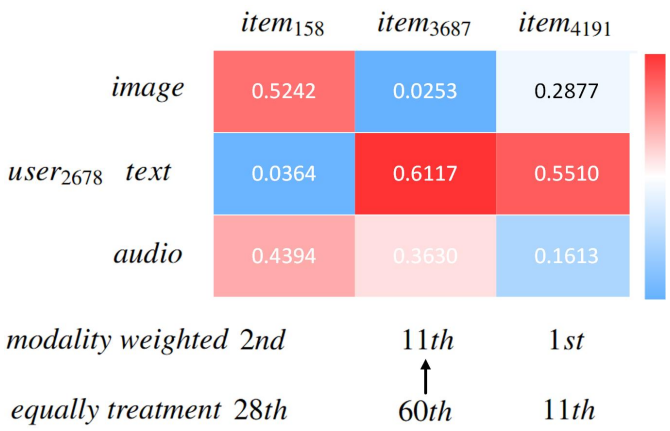

Fig. 1: Ranking of positive items under different fusion strategies in the TikTok dataset.

all modality features during fusion. Existing methods often neglect the relative importance among different modalities during the fusion process. However, users exhibit diverse preferences for different modalities when interacting with items. As shown in Fig. 1, for $user_{2678}$, DAHM adaptively assigns weights to different modalities (image, text, audio) for each user-item interaction. In comparison to equally-weighted fusion, such modality-aware fusion leads to significantly better rankings for positive items. Therefore, recommendation performance can be improved by assigning appropriate weights to each modality based on user-specific preferences during the fusion process.

To address the aforementioned limitations, we propose a novel **D**enoising **A**ttentive **H**ypergraph Learning method for **M**ultimodal recommendation (**DAHM**). Specifically, we first introduce a Dirichlet degree-sensitive rule to prune the potential noisy interactions, enabling the model to aggregate more reliable neighbor information. Built upon the denoised interaction graph, we apply a local user interest graph encoder to capture collaborative and semantic-related local user preferences. Then, we construct modality-specific item attribute hypergraphs to connect items with the same attribute via hyperedges. After obtaining the modality-specific hypergraph embeddings, we design a projection-enhancement hypergraph attention mechanism to compute the fusion weight of the modality features. Finally, we introduce a self-supervised task to fuse the obtained embeddings. The results of comprehensive experiments on three real-world datasets demonstrate that DAHM outperforms various baselines. DAHM achieves an average improvement of 9.55% in Recall@20 and 10.40% in NDCG@20.

II. Related Works

A. Hypergraph Learning for Recommendation

Recently, many works have explored the integration of hypergraph learning with recommendation to capture complex user–item relationships. For example, SDHID (8) introduces hypergraph and dual hypergraph to capture intra and inter session patterns in session-based recommendation, respectively. LGMRec (4) first generates hypergraphs with item multi-modal features, and treats each type of global user interest as a hyperedge to capture both local and global user interests. Different from existing hypergraph-based recommenders, our model considers the relative importance of different modalities during fusion and distinguishes between modality-common and modality-specific features.

B. Multi-modal Recommendation

Many studies are conducted to improve recommendation performance by incorporating massive multi-modal contents into user historical interactions. As user historical behaviors can be naturally represented as a bipartite graph, graph neural networks have been applied to capture high-order dependencies among users and items in multi-modal recommendation. For example, MMGCN (9) constructs modality-specific graph convolution network, and concatenates the obtained embeddings as the final item representations. LATTICE (10) learn potential semantic structures by constructing item-item modality graph as supplements to promote recommendation. However, the original interaction graph inevitably carries natural noise, which can contaminate node representations and mislead the modeling of user preferences. In this work, we introduce a Dirichlet degree-sensitive method to prune the noisy edges, which is more effective in filtering noisy interactions compared to other edge pruning methods.

III. Preliminaries

We denote a user as u in a user set $\mathcal{U}$ and an item as i in an item set $\mathcal{I}$, respectively. Naturally, the user historical interaction data can be represented as a bipartite graph $\mathcal{G} = (\mathcal{U}, \mathcal{I}, \mathcal{E})$. The edges $\mathcal{E}$ in $\mathcal{G}$ are defined by the adjacency matrix $R \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$. The transition matrix A is constructed from R.

$$A = \begin{bmatrix} 0 & R \\ R^T & 0 \end{bmatrix}, \tag{1}$$

Next, we denote the ID embeddings of user u and item i as $e_u^{id} \in \mathbb{R}^{|\mathcal{U}|\times d}$ and $e_i^{id} \in \mathbb{R}^{|\mathcal{I}|\times d}$. Additionally, each item i is associated with the multi-modal features $e_i^m \in \mathbb{R}^{|\mathcal{I}|\times d_m}$, where d_m is the modality feature dimension and $d \ll d_m$. We define the modality set $M \in \{v, t, a\}$, where v, t, and a represent visual, textual, and audio modalities, respectively. The aim of multi-modal recommendation is to model user preferences $\hat{y}_{u,i}$ accurately according to ID embeddings and item modality features and to predict whether a new item will be adopted by a user.

IV. Approach

A. Dirichlet Degree-sensitive Edge Drop

Recent researches have explored pruning superfluous edges in the interaction graph, demonstrating its effectiveness in graph denoising. Inspired by this, we further sparsify the graph $\mathcal{G}$ by pruning superfluous edges by introducing a Dirichlet degree-sensitive rule. Specifically, we calculate the degree of

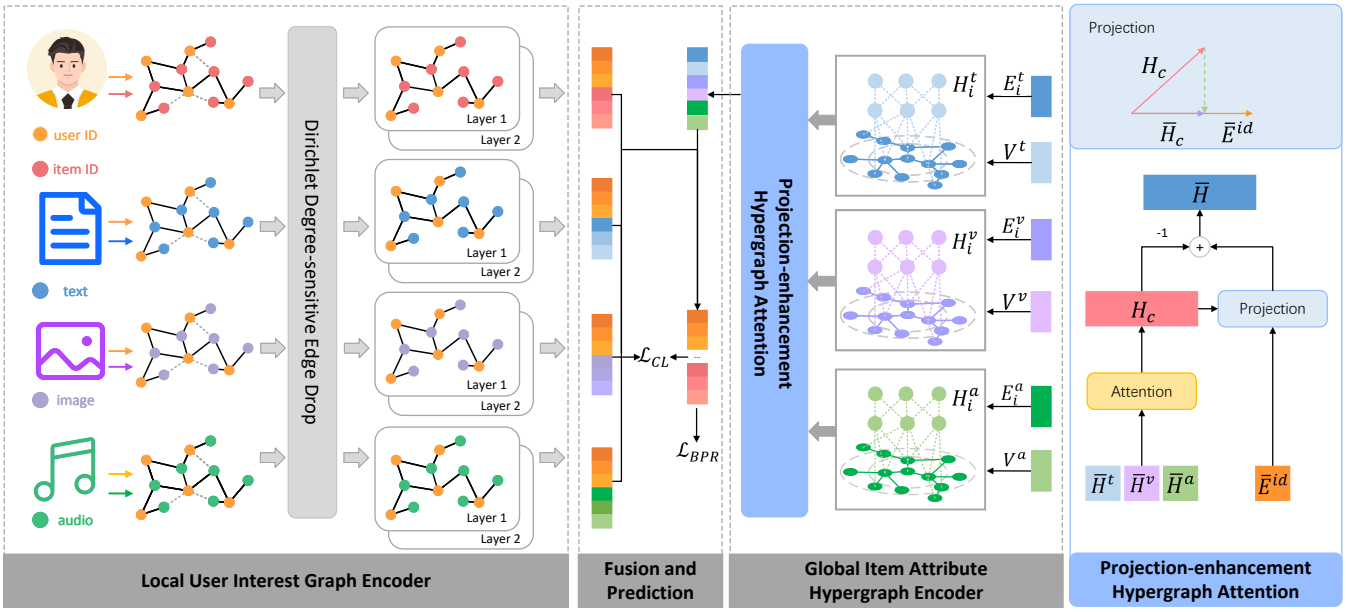

Fig. 2: The framework of DAHM.

each node in $\mathcal{G}$, used as parameters for Dirichlet distribution to obtain the reserved probability of each edge before each training epoch. Subsequently, we sample $k = (1 - \lambda_1)|E|$ edges according to the resulting multinomial probability distribution. Here, λ_1 is the edge drop ratio, a hyperparameter. Formally, this process can be formulated as:

$$\begin{aligned} d_{e_k} &= \frac{1}{\sqrt{d_i}\sqrt{d_j}}, \\ p_{e_k} &\sim Dirichlet(d_{e_k}), \\ A_d &\sim Multinomial(k; p_{e_1}, p_{e_2}, \ldots, p_{|E|}), \end{aligned} \tag{2}$$

where d_i and d_j denote the degrees of node i and j, respectively, and p_{e_k} represents the retention probability of each edge. Consequently, we obtain a denoised adjacency matrix A_d, which is used during the training stage for information propagation. Moreover, we adopt a Dirichlet distribution to derive the edge retention probabilities p_{e_k} rather than directly using d_{e_k}. This is because the Dirichlet distribution considers the relative importance among edges instead of processing each edge independently, thereby smoothing the values of d_{e_k} and mitigating the negative effects of extreme values.

B. Local User Interest Graph Encoder

The initial ID embeddings (e.g., $e_u^{id,0}, e_i^{id,0}$) are vectors in $\mathbb{R}^d$. Then, DAHM aggregates the neighborhood information using a simplified graph convolutional network:

$$E^{id,k+1} = LightGCN(E^{id,k}) = (D^{-1/2} A_d D^{-1/2}) E^{id,k}, \tag{3}$$

where D is the diagonal matrix of A_d and the initial embedding matrix is set as $E^{id,0} = \left[e_{u1}^{id,0}, \ldots, e_{u|\mathcal{U}|}^{id,0}, e_{i1}^{id,0}, \ldots, e_{i|\mathcal{I}|}^{id,0}\right]$. In particular, we adopt mean aggregator to integrate all embedding layers:

$$\bar{E}^{id} = Mean(E^{id,0}, E^{id,1}, \ldots E^{id,k+1}), \tag{4}$$

where $\bar{E}^{id}$ is the user/item ID embedding with local topology structure information.

ID embeddings capture user-item co-occurrence patterns from a statistical perspective. In contrast, item modality features reflect users' fine-grained preferences from a semantic perspective. Then, DAHM transforms pre-trained modality features $e_i^m \in \mathbb{R}^{d_m}$ into initial item modality embedding $e_i^{m,0} \in \mathbb{R}^d$ with a unified embedding dimension d:

$$e_i^{m,0} = W_m e_i^m + b_m. \tag{5}$$

Then, DAHM initializes the user embeddings $e_u^{m,0}$ independently. Similarly, DAHM implements modality-specific information propagation via $LightGCN(\cdot)$:

$$E^{m,k+1} = LightGCN(E^{m,k}) = (D^{-1/2}) A_d (D^{-1/2}) E^{m,k}, \tag{6}$$

where $E^{m,k}$ represents the k-th layer modality embeddings and $E^{m,0} = [e_{u1}^{m,0}, \ldots, e_{u|\mathcal{U}|}^{m,0}, e_{i1}^{m,0}, \ldots, e_{i|\mathcal{I}|}^{m,0}]$. Then, we choose the last layer $E^{m,k}$ as the local modality-specific embeddings $\bar{E}^m$.

C. Global Item Attribute Hypergraph Encoder

The global item attribute hypergraph encoder is designed to capture the global similarities among modalities, which might be uncovered from the local neighborhood perspective.

1) Modality-specific Hypergraph Construction: Hypergraph is capable of unifying nodes beyond pairwise relations. Inspired by this, we construct modality-specific item-item hypergraphs to adaptively learn the global dependencies among different item modality attributes. Specifically, we

denote modality-specific learnable vectors $V^m \in \mathbb{R}^{d_m \times h}$ as hyperedge embeddings, where h is the number of hyperedges, a hyperparameter. Therefore, the modality-specific item hypergraph matrices H_i^m can be formulated as:

$$H_i^m = E_i^m V^{m^T}, \tag{7}$$

where $H_i^m \in \mathbb{R}^{|\mathcal{I}| \times h}$, $E_i^m = [e_{i1}^m, \ldots, e_{i|\mathcal{I}|}^m]$ is the pre-trained modality features, and $V^m = [V_1^m, \ldots, V_h^m]$ is the hyperedge vector matrices. Hyperedges act as intermediate hubs that enable message passing between vertices, even if they are far apart in the original graph. Naturally, items with similar modality attributes are more likely to be connected to the same hyperedge. Considering that if a node is connected with too many hyperedges may introduce noise and redundancy, we apply the Gumbel-Softmax (11) on H_i^m:

$$\bar{H}_i^{m,k+1} = \varrho\left(\bar{H}_i^m\right) \cdot \varrho\left(\bar{H}_i^{m^T}\right) \bar{H}_i^{m,k}. \tag{8}$$

Specifically, $H_i^{m,0} = \bar{E}_i^{id}$ mentioned in the local user interest graph encoder, and $\varrho(\cdot)$ denotes the dropout operation. Then, we take the last layer as the final item modality hypergraph embeddings $\bar{H}_i^m$. Further, we define user modality hypergraph embeddings as:

$$\bar{H}_u^m = R \cdot \bar{H}_i^m, \tag{9}$$

where R is the adjacency matrix mentioned in Section III.

2) Projection-enhancement Hypergraph Attention: For item i, we refer to the item features observable across all modalities as modality-common features, and those that can only be observed through a specific modality as modality-specific features. Then, we use the attention mechanism (12) to compute the relative importance of modality-common features across different modalities, where the attention weight of each modality hypergraph embedding is calculated as:

$$\alpha_m = \frac{exp\left(a^T \sigma\left(W_a \bar{H}^m + b_a\right)\right)}{\sum_{m' \in M}^{|M|} exp\left(a^T \sigma\left(W_a \bar{H}^{m'} + b_a\right)\right)}, \tag{10}$$

where $a \in \mathbb{R}^d$ denotes attention vector. $\bar{H}^m$ is obtained by concatenating $\bar{H}_u^m$ and $\bar{H}_i^m$. $\sigma(\cdot)$ denotes the $Tanh$ nonlinearity. Therefore, the modality-common hypergraph features can be calculated as:

$$H_c = \sum_{m \in M}^{|M|} \alpha_m \bar{H}^m. \tag{11}$$

Projection is a method that decouples implicit feedback from explicit feedback to enhance its representational effectiveness. Inspired by (13), we leverage a projection-enhancement method to modality-common features and collaborative signals (as shown in the upper-right part of Fig.2). The projection-enhancement modality-common features can be formulated as:

$$\bar{H}_c = \frac{H_c \cdot \bar{E}^{id}}{\left|\bar{E}^{id}\right|} \frac{\bar{E}^{id}}{\left|\bar{E}^{id}\right|}, \tag{12}$$

where $\bar{E}^{id}$ is collaborative signal mentioned in Eq. (4). Then, modality-specific hypergraph features $\bar{H}_s^m$ are calculated by modality-specific hypergraph embedding matrix $\bar{H}^m$ minus modality-common hypergraph features $\bar{H}_c$.

$$\bar{H}_s^m = \bar{H}^m - \bar{H}_c. \tag{13}$$

Finally, we obtain the modality hypergraph embeddings $\bar{H}$:

$$\bar{H} = \bar{H}_c + \frac{1}{|M|} \sum_{m \in M}^{|M|} \bar{H}_s^m. \tag{14}$$

D. Fusion and Prediction

After obtaining two types of local embeddings $\bar{E}^{id}$, $\bar{E}^m$ and global hypergraph embeddings $\bar{H}$, we acquire the final user/item representations by fusing them together:

$$E^f = \bar{E}^{id} + \sum_{m \in M}^{|M|} \bar{E}^m + \alpha \bar{H}, \tag{15}$$

where α is a hyperparameter to adjust the integration of hypergraph embeddings.

Furthermore, to promote the exploration of collaborative and modality-related information, a self-supervised auxiliary task has been introduced. The task aims to distill self-supervision signals for dependency modeling between final user preferences E^f and modality-specific user preferences $\bar{E}^m$. The mathematical expression of this task is as follows:

$$\mathcal{L}_{CL} = -\sum_{m \in M}^{|M|} \sum_{u \in \mathcal{U}}^{|\mathcal{U}|} \log \frac{exp\ s\left(e_u^f, \bar{e}_u^m\right)}{\sum_{u' \in \mathcal{U}} \left(exp\ s\left(e_{u'}^f, \bar{e}_u^m\right) + exp\ s\left(e_{u'}^m, \bar{e}_u^m\right)\right)}, \tag{16}$$

$$s\left(e_u^f, \bar{e}_u^m\right) = \frac{e_u^{f^T}, \bar{e}_u^m}{\tau' \cdot ||e_u^f|| \cdot ||\bar{e}_u^m||}, \tag{17}$$

where $s(\cdot)$ denotes the similarity function, and τ' is the temperature factor of softmax. e_u^f and $\bar{e}_m^f$ are the user u in E^f and $\bar{E}^m$, respectively. We use the inner product to quantify the likelihood of interaction (e.g. $\hat{y}_{u,i} = e_u^f * e_i^{f^T}$). The Bayesian Personalized Ranking loss (14) is employed as follows:

$$\mathcal{L}_{BPR} = \sum_{(u,i^+i^-)} -log\left(sigmoid\left(\hat{y}_{u,i^+} - \hat{y}_{u,i^-}\right)\right). \tag{18}$$

Then, we integrate self-supervised loss to jointly update the representations:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda_2 \cdot \mathcal{L}_{CL} + \lambda_3 \cdot \|\theta\|_2^2, \tag{19}$$

where λ_2 and λ_3 are hyperparameters for loss term weighting.

V. Experiments

A. Experimental Settings

1) Datasets: Experiments are conducted on three datasets: (a) Netflix[1], (b) Allrecipes[2], and (c) TikTok[3]. All contain textual and visual modalities, and TikTok additionally contains an audio modality. Data statistics are summarized in Table I.

[1]https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data

[2]https://www.allrecipes.com/

[3]https://www.biendata.xyz/competition/icmechallenge2019/

TABLE I: Statistics of the multi-modal datasets, containing visual (V), textual (T), and acoustic (A) contents.

Dataset	Netflix		Allrecipes		TikTok		
Modality	V	T	V	T	V	T	A
Feat. Dim.	512	768	2048	20	128	128	128
User	43739		19805		9380		
Item	9872		10086		6710		
Interaction	648287		73494		68722		
Sparsity	99.850%		99.963%		99.890%		

2) Evaluation Protocols: We randomly split historical interactions into training, validation, and testing sets with 8:1:1 ratio. Then, we adopt three metrics for top-K prediction task: Recall@K, Precision@K, and NDCG@K, and set K=20.

3) Baselines Methods: We compare our proposed DAHM with the following baselines: LightGCN (15), LayerGCN (16), VBPR (1), MMGCN (9), LATTICE (10), FREEDOM (17), MGCN (6), BM3 (18), LGMRec (4), SMORE (7).

4) Parameter Settings: We set the other hyperparameters as follows: learning rate = 0.001 and $\tau' = 0.5$ for three datasets, $\lambda_1 = 0.2, 0.5, 0.2$ for Netflix, Allrecipes and TikTok, $\lambda_2 = 0.001, 0.001, 0.1$ for Netflix, Allrecipes and TikTok, $\lambda_3 = 0.001, 0.01, 0.1$ for Netflix, Allrecipes and TikTok, $\alpha = 0.3, 3.0, 0.05$ for Netflix, Allrecipes and TikTok, $h = 256, 8, 16$ for Netflix, Allrecipes and TikTok.

B. Overall performance

Table II shows the performance of DAHM and other baselines on three datasets. Key observations are as follows:

Generally, DAHM significantly outperforms all baselines across all metrics on all datasets, which demonstrates the effectiveness and superiority of our proposed method. Especially on the Allrecipes dataset, DAHM achieves improvements of 17.28%, 15.13%, and 15.79% in terms of Recall@20, NDCG@20, and Precision@20, respectively. Specifically, we can observe that the sparser the dataset, the greater the improvement achieved by our model.

In comparison with the hypergraph-based multi-modal recommender (e.g., LGMRec), the obvious performance improvement of DAHM can be observed. We attribute this to the fact that we devise the projection-enhancement hypergraph attention mechanism to adaptively allocate modality fusion weights. Although SMORE and MGCN fuse modality-specific features under the guidance of collaborative signals, their gating mechanism cannot effectively filter out irrelevant collaborative signals. Our proposed projection-enhancement hypergraph attention mechanism calculates global user attention across different modalities, achieving a more comprehensive fusion of modality features.

Further, although LayerGCN models user preferences using only ID embeddings, its performance still surpasses the performance of some multi-modal recommenders because it employs an edge pruning rule to filter out noisy interactions. However, DAHM achieves even better performance than LayerGCN by adopting a Dirichlet degree-sensitive rule, which more effectively prunes noisy edges.

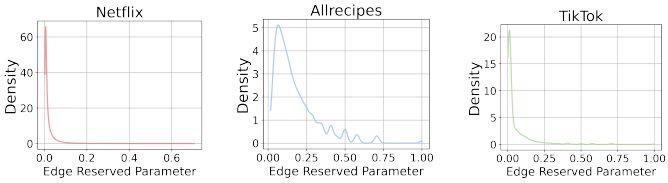

Fig. 3: The edge reserved parameters of each dataset.

C. Study of Edge Pruning Method

In this section, we evaluate the effectiveness of different edge pruning strategies based on the DAHM framework. Specifically, we compare the following variants: 1) $DAHM_r$ (Random Pruning): randomly prunes edges from the original interaction graph. 2) $DAHM_l$ (Degree-Based Pruning): follows the method used in LayerGCN and FREEDOM, pruning based on node degrees. 3) $DAHM_d$ (Dirichlet-Based Pruning): the pruning strategy proposed in DAHM.

The performance comparison is shown in Table III. It can be observed that $DAHM_d$ consistently outperforms the other two methods. To understand this, we analyze the reserved edge parameter d_{e_k} in Eq.(2). As shown in Fig.3, the higher the degree of the nodes connected, the lower the edge reserved parameters. We believe that the effectiveness of different edge pruning methods is influenced by the degree distribution skewness of the dataset. While pruning edges connected to high-degree nodes can alleviate the issue of natural noise, these high-degree nodes often contain important structural information. $DAHM_l$ equipped with degree-sensitive method may excessively prune edges connected to high-degree nodes, disrupting the overall structure of the graph. Therefore, the higher the proportion of high-degree nodes in the dataset, the worse the performance of $DAHM_l$. In contrast, $DAHM_d$ introduces randomness and smoothing via Dirichlet sampling, avoiding over-pruning and reducing reliance on extreme values. This leads to better generalization and robustness, which accounts for its superior performance.

D. Ablation Study

In Table IV, experimental results are presented to demonstrate the importance of key components in DAHM.

The suboptimal performance of the variant w/o GE (i.e., removing the global item attribute hypergraph encoder) demonstrates that modeling global item attribute relationships across different modalities plays a critical role in improving recommendation quality. Interestingly, the performance degradation becomes even more pronounced when we only remove the projection-enhancement hypergraph attention mechanism (w/o HA) from the global item attribute hypergraph encoder. This suggests that assigning modality-specific fusion weights is essential for effective multi-modal representation learning. Our proposed projection-enhancement hypergraph attention mechanism effectively addresses this by learning adaptive

TABLE II: Overall performances of DAHM and other baselines.

Models	Netflix			Allrecipes			TikTok		
	R@20	**N@20**	**P@20**	**R@20**	**N@20**	**P@20**	**R@20**	**N@20**	**P@20**
LightGCN	0.1794	0.0695	0.0090	0.0313	0.0137	0.0016	0.0934	0.0426	0.0047
LayerGCN	0.1851	0.0685	0.0093	0.0335	0.0134	0.0017	0.0940	0.0429	0.0047
VBPR	0.1712	0.0701	0.0086	0.0033	0.0013	0.0002	0.0488	0.0199	0.0024
MMGCN	0.1658	0.0644	0.0083	0.0382	0.0121	0.0019	0.0709	0.0211	0.0035
LATTICE	0.1991	0.0767	0.0100	0.0354	0.0139	0.0018	0.0842	0.0384	0.0042
FREEDOM	0.1548	0.0666	0.0077	0.0321	0.0127	0.0016	0.0898	0.0391	0.0045
MGCN	0.1890	0.0717	0.0094	0.0343	0.0151	0.0017	0.1032	0.0462	0.0052
BM3	0.1303	0.0515	0.0065	0.0362	0.0152	0.0018	0.0989	0.0390	0.0049
LGMRec	0.1883	0.0702	0.0091	0.0335	0.0125	0.0017	0.0787	0.0329	0.0039
SMORE	0.1913	0.0751	0.0096	0.0306	0.0122	0.0015	0.1117	0.0481	0.0056
Ours	**0.2073**	**0.0828**	**0.0104**	**0.0448**	**0.0175**	**0.0022**	**0.1198**	**0.0520**	**0.0060**
Improv.	4.12%	7.95%	4.00%	17.28%	15.13%	15.79%	7.25%	8.11%	7.14%

TABLE III: The performance comparison of different edge pruning methods.

Datasets	Variants	R@20	R@50	N@20	N@50
Netflix	$DAHM_r$	0.1623	0.2488	0.0626	0.0795
	$DAHM_l$	0.1845	0.2886	0.0737	0.0944
	$DAHM_d$	**0.2073**	**0.3092**	**0.0828**	**0.1030**
Allrecipes	$DAHM_r$	0.0352	0.0665	0.0141	0.0203
	$DAHM_l$	0.0423	0.0710	0.0168	0.0224
	$DAHM_d$	**0.0448**	**0.0712**	**0.0175**	**0.0226**
TikTok	$DAHM_r$	0.1176	0.1806	0.0486	0.0611
	$DAHM_l$	0.1180	0.1811	0.0495	0.0618
	$DAHM_d$	**0.1198**	**0.1842**	**0.0520**	**0.0647**

TABLE IV: Ablation study of DAHM on three datasets.

Datasets	Netflix		Allrecipes		TikTok	
Metrics	**R@20**	**N@20**	**R@20**	**N@20**	**R@20**	**N@20**
w/o GE	0.1594	0.0761	0.0371	0.0164	0.0958	0.0401
w/o HA	0.1853	0.0718	0.0405	0.0168	0.0602	0.0256
w/o D	0.1947	0.0905	0.0424	0.0160	0.1033	0.0436
w/o CL	0.1927	0.0868	0.0416	0.0160	0.0749	0.0392

fusion weights and facilitating the separation of modality-common and modality-specific features.

Additionally, the variant w/o D uses the original interaction matrix. The observed performance drop highlights the necessity of our Dirichlet degree-sensitive edge pruning method, which can effectively filter noisy interactions.

Finally, we ablate the self-supervised task with the variant w/o CL. The suboptimal performance further reveals the importance of exploring the mutual information between the final user representations and multi-modal features.

VI. Conclusion

In this paper, we propose a Denoising Attentive Hypergraph Learning method (DAHM) for multimodal recommendation. In detail, we design a Dirichlet degree-sensitive edge pruning method to filter noisy interactions. Further, we design the projection-enhancement hypergraph attention method to calculate the relative importance of modality-specific hypergraph embeddings. Extensive experiments have been conducted to demonstrate the superiority of our method.

References

[1] R. He and J. McAuley, "Vbpr: Visual bayesian personalized ranking from implicit feedback," in *AAAI*, 2016, pp. 144–150.

[2] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," in *SIGIR*, 2017, pp. 335–344.

[3] Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, and T.-S. Chua, "Mgat: Multimodal graph attention network for recommendation," *Information Processing & Management.*, vol. 57, p. 102277, 2020.

[4] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan, "Lgmrec: Local and global graph learning for multimodal recommendation," in *AAAI*, 2024, pp. 8454–8462.

[5] Y. Wei, X. Wang, L. Nie, X. He, and T.-S. Chua, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in *MM*, 2020, pp. 3541–3549.

[6] P. Yu, Z. Tan, G. Lu, and B.-K. Bao, "Multi-view graph convolutional network for multimedia recommendation," in *MM*, 2023, pp. 6576–6585.

[7] R. K. Ong and A. W. H. Khong, "Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation," in *WSDM '25*, 2025, pp. 773–781.
[8] R. Gao, Y. Tao, Y. Yu, J. Wu, X. Shao, J. Li, and Z. Ye, "Self-supervised dual hypergraph learning with intent disentanglement for session-based recommendation," vol. 270, 2023, p. 110528.
[9] 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 *MM*, 2019, pp. 1437–1445.
[10] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining latent structures for multimedia recommendation," in *MM*, 2021, pp. 3872–3880.
[11] E. Jang, S. S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 6000–6010.
[13] C. Meng, H. Zhang, W. Guo, H. Guo, H. Liu, Y. Zhang, H. Zheng, R. Tang, X. Li, and R. Zhang, "Hierarchical projection enhanced multi-behavior recommendation," in *KDD*, 2023, pp. 4649–4660.
[14] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *UAI*, 2009, pp. 452–461.
[15] 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, pp. 639–648.
[16] X. Zhou, D. Lin, Y. Liu, and C. Miao, "Layer-refined graph convolutional networks for recommendation," in *ICDE*, 2023, pp. 1247–1259.
[17] X. Zhou and Z. Shen, "A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation," in *MM*, 2023, pp. 935–943.
[18] 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 *WWW*, 2023, pp. 845–854.