Comprehensive Interest Modeling and Relational Mining for Multi-modal Recommendation

HaoYu Wang[1] (✉) and HongBin Xia[1,2]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi,
Jiangsu 214122, China. 6233111056@stu.jiangnan.edu.cn
[2] Jiangsu Key University Laboratory of Software and Media Technology under
Human-Computer Cooperation (Jiangnan University), Wuxi, China. hbxia@163.com

Abstract. Multimodal recommendation systems attract significant attention due to their ability to integrate user feedback alongside item image and text information, addressing performance limitations caused by data sparsity. While previous studies have made notable progress in improving user and item representations through various fusion and alignment techniques, two critical issues remain. First, comprehensive user preference representation is crucial for enhancing recommendation diversity and accuracy. However, previous research has focused on indepth exploration within each modality, lacking a broader, cross-modal perspective, which limits recommendation diversity and range. Second, the insufficient capture of fine-grained item similarities in multidimensional information fusion results in semantic distances that fail to reflect true associations between items. To address these challenges, we propose a novel Comprehensive Interest Modeling and Relational Mining (CMR) approach. Specifically, we construct two types of item graphs based on interaction data and multimodal information to improve item representations. We then model users' broad preferences by integrating graph convolutional networks with personalized features and identity information from ID embeddings. Finally, we design cross-dimensional contrastive learning tasks to minimize the semantic distance between related items, enhancing multimodal information fusion accuracy. Extensive experiments on three public datasets demonstrate the effectiveness of our model.

Keywords: Multi-modal Fusion · Graph Convolutional Networks · Contrastive Learning · Multimodal Recommendation.

1 Introduction

With the rapid growth of information, recommendation systems are essential for filtering meaningful content. Traditional systems [2, 5] rely on user-item interactions to capture collaborative signals, but data sparsity limits their ability to handle diverse user preferences, leading to repetitive recommendations and cold start issues. As multimodal data (e.g., images, text, videos) grows, multimodal recommendation systems [11] integrate multiple modalities, offering a richer representation of user interests and item features, thus overcoming the limitations of traditional methods.

Various approaches model user modality features. DRAGON [12] uses attention and user co-occurrence graphs to enhance unimodal representations. MGCN [10] captures user preferences by combining user-item interactions and item similarities through graph convolution. GUME [3] improves user feature generalizability with multimodal alignment and denoising strategies. However, these methods overlook interaction patterns and personalized identity features, while simple feature aggregation often introduces noise.

For optimization, MF-BPR [6] uses BPR loss to maximize the score gap between positive and negative samples. MGCN [10] enhances multimodal fusion by increasing mutual information between behavioral and fused features. MENTOR [8] employs a cross-modal alignment loss and graph perturbation to improve representations. Yet, these strategies fail to fully exploit useful information and struggle with distinguishing positive and negative samples, leading to instability during training.

To address these challenges, we propose the Comprehensive Interest Modeling and Relational Mining (CMR) approach for multimodal recommendation. We first construct an item collaboration graph from interaction data to capture relationships and update item ID embeddings. Next, we align multimodal features with ID embeddings in a shared space to enhance interaction information. We then create an item association graph based on modality similarity for independent enhancement across modalities. For user preferences, we combine feedback with enhanced item representations to capture both broad and fine-grained preferences. Finally, we design auxiliary contrastive learning tasks to leverage complementary relationships across user and item dimensions. Extensive experiments on three publicly available datasets demonstrate the advantages of our approach. Our main contributions are summarized as follows:

- We propose a novel strategy for utilizing the item collaboration graph to learn latent relationships between items, thereby enhancing item representations.
- We develop a user modality preference enhancement module that captures a broader range of user modality preferences, leading to a more complete representation of the user.
- We design contrastive learning auxiliary tasks for different dimensional information of users and items, achieving synergistic reinforcement.

2 Methodology

In this section, we first define the relevant concepts and then provide a comprehensive overview of the architecture and optimization process of the proposed CMR model.

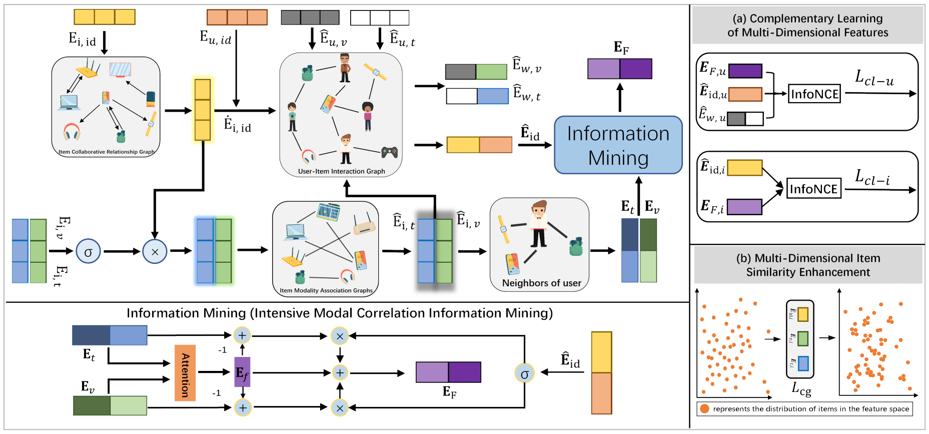


Fig. 1: The overview architecture of our proposed CMR model. Information Mining is a key component within the framework, while (a) and (b) represent two distinct optimization strategies for contrastive learning.

2.1 Preliminaries

Let $\mathcal{U} = \{u\}$ and $\mathcal{I} = \{i\}$ denote the user and item sets, respectively. The ID embeddings are initialized separately as $E_{u,id} \in \mathbb{R}^{d\times|\mathcal{U}|}$ for users and $E_{i,id} \in \mathbb{R}^{d\times|\mathcal{I}|}$ for items, where d is the embedding dimension. The modality embeddings for items are represented as $E_{i,m} \in \mathbb{R}^{d_m\times|\mathcal{I}|}$, with $m \in \mathcal{M} = \{v, t\}$ denoting visual and textual modalities, respectively, and d_m indicating the dimension of modality m.

We define the user-item interaction matrix $\mathcal{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $\mathcal{R}_{u,i} = 1$ indicates an interaction between user u and item i, and $\mathcal{R}_{u,i} = 0$ indicates no interaction. Based on this matrix, we create a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$ as the set of nodes and $\mathcal{E} = \{(u,i) \mid u \in \mathcal{U}, i \in \mathcal{I}, \mathcal{R}_{u,i} = 1\}$ as the set of edges.

2.2 Interweaving and Deepening of Item Information

Accurate dependency relationships and enriched item representations are crucial for precise recommendations. We propose an advanced item representation learning framework that integrates item collaboration relationships and modality-specific associations to achieve this goal.

Item Collaborative Relationship Graph. The complementary relationship between items is a key concept in economics, reflecting the interdependence between different products. For example, while electric toys and batteries may not be related semantically, they are often purchased together. Drawing inspiration from this idea and [9], we propose applying it to recommendation systems to better capture interaction data.

We construct an item-item collaboration matrix C based on interaction data, where each pair (i, j) is represented by $C_{i,j}$, indicating the frequency of user interactions with that pair. Since item complementarity is observed empirically rather than through a strict formula, we introduce a threshold T to reduce noise, refining the top-k selection for more accurate identification of collaborative relationships. To maintain inherent properties when $i = j$, matrix elements are set to 1, forming self-loops.

Item Modality Association Graphs. We compute a similarity score matrix S_m for each modality using cosine similarity to quantify item similarity based on raw features, avoiding parameter dependency and reducing computational complexity. The results are then refined using KNN to mitigate noise from weakly related edges, retrieving the top-k most similar items and constructing a sparse graph based on these neighbors' indices and similarity values.

After graph construction, we enhance item representations by updating $E_{i,id}$ with structured item associations, applying linear and nonlinear transformations to raw modality features, and element-wise multiplying these with ID embeddings. The fused features are then propagated through a GCN module to generate refined item embeddings:

$$\hat{E}_{i,m} = \dot{\mathcal{S}}_m((\dot{\mathcal{C}} \cdot E_{i,id}) \odot \sigma(\mathbf{W}_1 E_{i,m} + \mathbf{b}_1)). \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_m}$ is the trainable transformation matrix, and $\mathbf{b}_1 \in \mathbb{R}^d$ is the bias vector, both modality-specific.

2.3 Comprehensive Representation of User Characteristics

To comprehensively capture user characteristics, we combine detailed profiling of user preferences with broad contextual relationships, integrating both depth and breadth in user representation.

Deep Preference Profiling. We model the user's modality features indirectly through user feedback to capture shallow preferences. Specifically, we define the user's modality features as $E_{u,m} \in \mathbb{R}^{d_m \times |\mathcal{U}|}$, generated through weighted aggregation of the modality features of interacted items. The user's shallow modality features are then concatenated with the item's modality features to form a unified modality-based representation $\mathbf{E}_m$.

Next, we calculate an importance score for each modality m using a self-attention mechanism and apply softmax to normalize the weights. Subsequently, we perform a weighted aggregation of each modality to obtain the shallow fused representation $\mathbf{E}_f$:

$$E_f = \sum_{m=1}^{|\mathcal{M}|} \frac{\exp(\mathbf{q}^\top \tanh(\mathbf{W}_3 \mathbf{E}_m + \mathbf{b}_3))}{\sum_{m=1}^{|\mathcal{M}|} \exp(\mathbf{q}^\top \tanh(\mathbf{W}_3 \mathbf{E}_m + \mathbf{b}_3))} \mathbf{E}_m, \tag{2}$$

where $\mathbf{q} \in \mathbb{R}^d$ denotes the attention vector, $\mathbf{W}_3 \in \mathbb{R}^{d \times d}$ is the weight matrix, and $\mathbf{b}_3 \in \mathbb{R}^d$ is the bias vector. These parameters are shared across all modalities.

However, deep-level information associated with specific modalities is often underutilized due to over-reliance on a single modality, leading to incomplete representation of key features. To address this, followed by [10], we introduce interaction-level collaborative signals to guide user preference representation. We construct a GCN module to propagate user and item ID embeddings across the interaction graph:

$$\mathbf{E}_{id}^{(l)} = \mathbf{A}\mathbf{E}_{id}^{(l-1)} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} 0 & \mathcal{R} \\ \mathcal{R}^\top & 0 \end{pmatrix}, \tag{3}$$

where $\mathbf{E}_{id}^{(l)}$ represents enhanced user and item representations at layer l, and $\mathbf{E}_{id}^{(0)} = \{E_{u,id} \parallel \dot{E}_{i,id}\}$ are the initial ID embeddings. The final deep modality representation $\mathbf{E}_F$ is defined as:

$$\mathbf{E}_F = \frac{1}{1+|\mathcal{M}|}(\sum_{m=1}^{|\mathcal{M}|} \Big(\sigma(\mathbf{W}_4\hat{\mathbf{E}}_{id} + \mathbf{b}_4) \odot (\mathbf{E}_m - \mathbf{E}_f)\Big) + \mathbf{E}_f), \tag{4}$$

where $\mathbf{W}_4 \in \mathbb{R}^{d \times d}$ is the weight matrix, and $\mathbf{b}_4 \in \mathbb{R}^d$ is the bias vector. These parameters are modality-specific. $\hat{\mathbf{E}}_{id}$ is the average of embeddings across all layers.

Global Relationship Capture. We then introduce a GCN to capture complex contextual relationships between users and items, inspired by the work of [3]. Additionally, we integrate ID embeddings to preserve each node's unique interaction features:

$$\mathbf{E}_{w,m}^{(l)} = \mathbf{A}\mathbf{E}_{w,m}^{(l-1)} + \gamma\hat{\mathbf{E}}_{id}, \tag{5}$$

where $\mathbf{E}_{w,m}^{(0)} = \{\hat{E}_{u,m} \parallel \hat{E}_{i,m}\}$ are the initial ID embeddings. We use a newly initialized user representation $\hat{E}_{u,m}$ to address the new task and obtain a broad user modality representation, where γ is a hyper-parameter controlling subtle information. By averaging the embeddings across all layers, we obtain $\hat{\mathbf{E}}_{w,m}$, while the final representation $\mathbf{E}_W$ is computed as follows:

$$\mathbf{E}_W = \sum_{m=1}^{|\mathcal{M}|} \hat{\mathbf{E}}_{w,m}, \quad \text{and} \quad \hat{\mathbf{E}}_{w,m} = \frac{1}{L+1}\sum_{i=0}^{L} \mathbf{E}_{w,m}^{(l)}. \tag{6}$$

2.4 Optimization

In recommendation systems, user and item features span multiple dimensions with inherent complementarity and similarity across these dimensions. To effectively capture these complex relationships, we design specific loss functions and apply targeted contrastive learning strategies to optimize user and item representations, enhancing the model's ability to capture interdependencies among multi-dimensional features.

Multi-Dimensional Item Similarity Enhancement. We represent $\mathbf{E}_d \in \mathcal{D} = \{\hat{E}_{i,t}, \hat{E}_{i,v}, \hat{\mathbf{E}}_{i,id}\}$ as single-dimensional information corresponding to textual, visual, and ID-related features. Using limited samples and multi-dimensional information, we compute item similarity within each dimension. We then apply a dual strategy to identify relevant items: First, we aggregate similarity metrics across all dimensions and use the top-k method to retrieve the most relevant items, denoted as $\mathcal{I}_{m,d}$. Second, we apply the top-k method within individual dimensions to retrieve the most relevant items in each dimension, denoted as $\mathcal{I}_{s,d}$.

This dual approach validates information across sources and incorporates insights from individual dimensions, enhancing the completeness of retrieved data. Finally, we compute multi-dimensional similarity scores $\mathbf{Sco}_{m,d}(i)$, single-dimensional similarity scores $\mathbf{Sco}_{s,d}(i)$, and overall similarity scores $\mathbf{Sco}_{all,d}(i)$:

$$\begin{aligned} \mathbf{Sco}_{m,d}(i) &= \exp(\mathbf{E}_d(i) \cdot \mathbf{E}_d(\mathcal{I}_{m,d})/\phi), \\ \mathbf{Sco}_{s,d}(i) &= \exp(\mathbf{E}_d(i) \cdot \mathbf{E}_d(\mathcal{I}_{s,d})/\phi), \\ \mathbf{Sco}_{all,d}(i) &= \textstyle\sum_{k=1}^{|\mathcal{I}|} \exp\left(\mathbf{E}_d(i) \cdot \mathbf{E}_d(k)^\top/\phi\right), \end{aligned} \tag{7}$$

where ϕ is the temperature hyper-parameter of the softmax function. Finally, we define a contrastive loss function $\mathcal{L}_{cg}$, which utilizes $\mathbf{Sco}_{m,d}(i)$ and $\mathbf{Sco}_{s,d}(i)$ to maximize the relative similarity of relevant items:

$$\mathcal{L}_{cg} = \sum_{d=1}^{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{I}|} \left(-\log\left(\frac{\mathbf{Sco}_{m,d}(i)}{\mathbf{Sco}_{all,d}(i)}\right) - \log\left(\frac{\mathbf{Sco}_{s,d}(i)}{\mathbf{Sco}_{all,d}(i) - \mathbf{Sco}_{m,d}(i)}\right)\right). \tag{8}$$

Complementary Learning of Multi-Dimensional Features. Complementary relationships exist among features from different dimensions. By incorporating contrastive learning, we aim to align user and item representations across these dimensions, thereby establishing interdependencies among multi-dimensional features. To streamline presentation and reduce equation complexity, we define a general contrastive loss function:

$$\mathbf{L}(S, T, t) = \sum_{u \in \mathcal{U}} -\log\left(\frac{\exp\left(S_u \cdot T_u/\tau\right)}{\sum_{v \in \mathcal{U}} \exp\left(S_v \cdot T_v/\tau\right)}\right), \tag{9}$$

where S and T represent feature representations from different dimensions, and τ is the temperature hyperparameter for the softmax function.

To capture the complexity and diversity of user behavior and interests, we apply pairwise contrastive learning across a rich set of user features, including deep features, wide features, and ID embeddings. The resulting loss is denoted as $\mathcal{L}_{cl-u}$. Additionally, to improve the stability of feature representations, we introduce noise perturbations to the multi-dimensional feature space and perform contrastive learning on these perturbed representations, denoted as $\mathcal{L}_{ap}$. These paired representations are then processed using the general contrastive loss function defined in Eq.(9), ensuring consistent feature alignment.

Table 1: Statistics of the experimental datasets

Dataset	**#User**	**#Item**	**#Behavior**	**#Sparsity**
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Electronics	192,403	63,001	1,689,188	99.99%

For item representations, our focus is on establishing contrastive learning between deep modality and interaction information to capture their complementary aspects. This is formulated as $\mathcal{L}_{cl-i}$:

$$\mathcal{L}_{cl-i} = \sum_{i \in \mathcal{I}} -\log \frac{\exp\left(\hat{\mathbf{E}}_{i,id} \cdot \mathbf{E}_{i,F} / \tau\right)}{\sum_{j \in \mathcal{I}} \exp\left(\hat{\mathbf{E}}_{j,id} \cdot \mathbf{E}_{j,F} / \tau\right)}, \tag{10}$$

Finally, the total loss is:

$$\mathcal{L} = \mathcal{L}_{bpr} + \lambda_1 (\mathcal{L}_{cl-u} + \mathcal{L}_{cl-i}) + \lambda_2 \mathcal{L}_{ap} + \lambda_{cg} \mathcal{L}_{cg}, \tag{11}$$

where λ_1, λ_2, and λ_{cg} are hyper-parameters for different comparison targets.

3 EXPERIMENTS

In this section, we conduct comprehensive experiments to evaluate the effectiveness of our CMR model on three widely used real-world datasets. These experimental results are designed to address the following four research questions:

- **RQ1**: How does the effectiveness of the CMR model compare with state-of-the-art traditional and multimedia recommendation methods?
- **RQ2**: Why multi-dimensional item similarity enhancement can achieve better recommendation performance?

3.1 Experimental Settings

Dataset. Following previous studies [15, 4], we perform experiments using three subsets of the popular Amazon dataset: (a) Baby, (b) Sports and Outdoors (denoted by Sports), and (c) Electronics (denoted by Elec). Each dataset's raw data is pre-processed with a 5-core configuration for both users and items, and the filtered 5-core results are summarized in Table 1.

Compared Methods. To validate the effectiveness of our proposed model CMR, we compare it with state-of-the-art recommendation methods, which we categorize into two groups: traditional recommendation models and multi-modal

Table 2: Performance comparison of various models across datasets in terms of Recall@K (R@K) and NDCG@K (N@K). The improvements are significant under paired-t test (p < 0.05).

Model	**Baby**				**Sports**				**Elec**			
	R@10	**R@20**	**N@10**	**N@20**	**R@10**	**R@20**	**N@10**	**N@20**	**R@10**	**R@20**	**N@10**	**N@20**
MF-BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0235	0.0367	0.0127	0.0161
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0313	0.0387	0.0363	0.0540	0.0204	0.0250
DualGNN	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0363	0.0541	0.0202	0.0248
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0437	0.0648	0.0247	0.0302
MGCN	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0442	0.0650	0.0246	0.0302
FREEDOM	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0382	0.0588	0.0209	0.0262
DRAGON	0.0662	0.1021	0.0345	0.0435	0.0749	0.1124	0.0403	0.0500	0.0453	0.0673	0.0249	0.0306
MENTOR	0.0678	0.1048	0.0362	0.0450	0.0763	0.1139	0.0409	0.0511	0.0439	0.0655	0.0244	0.0300
GUME	0.0673	0.1042	0.0365	0.0460	0.0778	0.1165	0.0427	0.0527	0.0458	0.0680	0.0253	0.0310
CMR*	**0.0700**	**0.1080**	**0.0381**	**0.0480**	**0.0813**	**0.1209**	**0.0446**	**0.0548**	**0.0480**	**0.0721**	**0.0268**	**0.0328**
Improved	**3.245%**	**3.053%**	**4.384%**	**4.348%**	**4.499%**	**3.777%**	**4.450%**	**3.985%**	**4.803%**	**6.029%**	**5.703%**	**5.806%**

*: The CMR model is run on an NVIDIA 4090D GPU with 24 GB.

recommendation models. For traditional methods, we include MF-BPR [6] and LightGCN [1]. In the multi-modal category, we evaluate BM3 [15], DualGNN [7], DRAGON [12], FREEDOM [14], MGCN [10], GUME [3], and MENTOR [8]. These models serve as baselines to assess CMR's performance across both collaborative filtering and multi-modal feature integration scenarios.

Evaluation Protocols. To ensure fair performance evaluation, we use two commonly adopted metrics: Recall@K (R@K) and NDCG@K (N@K). The recommendation performance of each method is then assessed by computing the average metrics across all users in the test set, with K values of 10 and 20 for top-K evaluation.

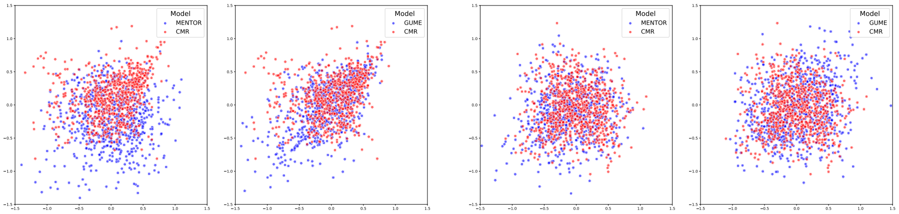
Implementation Details. We implement our proposed CMR model using the MMRec framework [13]. Consistent with previous studies [15], we set the embedding size of both users and items to 64 across all models. The optimal hyperparameters are determined via grid search on the validation set. Specifically, the hyperparameters λ_1, λ_2 and λ_{cg} in $\left\{1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}\right\}$. The hyperparameters τ, γ and ϕ in {0.1, 0.2, 0.3, 0.4}. We set k = 10 for the k-Nearest Neighbors method and set the threshold $T = 2$ to build the collaborative relationship graph of items. For convergence consideration, the early stopping and total epochs are fixed at 20 and 1,000, respectively.

3.2 Overall Performance (RQ1)

Table 2 presents a performance comparison between the proposed CMR model and various baseline methods across three datasets. Several key observations can be drawn from the results shown in the table.

Our CMR model significantly outperforms both conventional recommendation techniques and existing multimodal approaches. Specifically, in terms of Recall@20 for Baby, Sports, and Electronics, CMR achieves improvements of 3.05%, 3.78%, and 6.03%, respectively. For NDCG@20, the performance gains are 4.35%, 3.98%, and 5.81%. These results highlight CMR's ability to handle diverse dataset sizes, varying sparsity levels, multi-modal content distribution patterns, and different application contexts.

3.3 Visualization Analysis (RQ2)



(a) (b) (c) (d)

Fig. 2: Comparative visualization of item distributions generated by different models.

To validate the effectiveness of the multi-dimensional item similarity enhancement module, we apply t-SNE to project high-dimensional item representations onto a two-dimensional plane. Specifically, we randomly select 800 items from both the baby and sports datasets to provide a balanced view of each model's similarity performance. We then compare the results across three models: CMR, GUME[3], and MENTOR[8]. As shown in Figure 2, the results demonstrate that CMR exhibits a higher concentration of item distributions, presenting a more compact and consistent aggregation of items. This highly concentrated distribution indicates that CMR effectively aggregates multi-dimensional features, thereby enhancing the relational similarity between items.

4 Conclusion

In this study, we propose a novel Comprehensive Interest Modeling and Relational Mining (CMR) approach for multimodal recommendation. We construct item relation graphs and align multimodal features with ID embeddings to enhance interactive information. For user preference modeling, we incorporate user feedback with item representations to capture diverse modality preferences. We also design auxiliary contrastive learning tasks to exploit complementary relationships across dimensions, improving recommendation accuracy. Experimental results show that CMR outperforms state-of-the-art multimodal recommendation methods.

References

1. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: Lightgcn: Simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. pp. 639–648 (2020)
2. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S.: Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. pp. 173–182 (2017)
3. Lin, G., Zhen, M., Wang, D., Long, Q., Zhou, Y., Xiao, M.: Gume: Graphs and user modalities enhancement for long-tail multimodal recommendation. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 1400–1409 (2024)
4. Liu, Y., Zhang, K., Ren, X., Huang, Y., Jin, J., Qin, Y., Su, R., Xu, R., Yu, Y., Zhang, W.: Alignrec: Aligning and training in multimodal recommendations. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 1503–1512 (2024)
5. Nilashi, M., Ibrahim, O., Bagherifard, K.: A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. Expert Systems with Applications **92**, 507–520 (2018)
6. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. arXiv preprint arXiv:1205.2618 (2012)
7. Wang, Q., Wei, Y., Yin, J., Wu, J., Song, X., Nie, L.: Dualgnn: Dual graph neural network for multimedia recommendation. IEEE Transactions on Multimedia **25**, 1074–1084 (2021)
8. Xu, J., Chen, Z., Yang, S., Li, J., Wang, H., Ngai, E.C.H.: Mentor: Multi-level self-supervised learning for multimodal recommendation. arXiv preprint arXiv:2402.19407 (2024)
9. Xv, G., Li, X., Xie, R., Lin, C., Liu, C., Xia, F., Kang, Z., Lin, L.: Improving multi-modal recommender systems by denoising and aligning multi-modal content and user feedback. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3645–3656 (2024)
10. Yu, P., Tan, Z., Lu, G., Bao, B.K.: Multi-view graph convolutional network for multimedia recommendation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 6576–6585 (2023)
11. Zhou, H., Zhou, X., Zeng, Z., Zhang, L., Shen, Z.: A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. arXiv preprint arXiv:2302.04473 (2023)
12. Zhou, H., Zhou, X., Zhang, L., Shen, Z.: Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In: ECAI 2023, pp. 3123–3130. IOS Press (2023)
13. Zhou, X.: Mmrec: Simplifying multimodal recommendation. In: Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops. pp. 1–2 (2023)
14. Zhou, X., Shen, Z.: A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 935–943 (2023)
15. Zhou, X., Zhou, H., Liu, Y., Zeng, Z., Miao, C., Wang, P., You, Y., Jiang, F.: Bootstrap latent representations for multi-modal recommendation. In: Proceedings of the ACM Web Conference 2023. pp. 845–854 (2023)