DMIMRec: Disentangled Multi-Interest Representation Learning for Multimodal Recommendation Systems

Chengyi Zhou[1], Minghua Nuo[1,2,3†], Yuanyuan Ren[1], Rui Li[1], Hui Liu[1]

[1]*College of Computer Science, Inner Mongolia University, Hohhot 010021, Inner Mongolia, China*

[2]*National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Hohhot 010021, Inner Mongolia, China*

[3]*Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology, Hohhot 010021, Inner Mongolia, China*

Abstract

Multimodal recommender systems (MRSs) exploit diverse content sources (e.g., text, images) to enhance recommendation accuracy. However, they still face two fundamental challenges: (1) effectively capturing users' diverse interests, and (2) filtering out noisy signals from heterogeneous modalities. To address these issues, we propose DMIMRec, a framework for Disentangled Multi-Interest Modeling in multimodal recommendation. On the item side, we construct modality-specific item–item graphs and introduce a reconstruction-difference guided pruning strategy that evaluates each edge's usefulness by checking how much the reconstruction quality changes when that edge is removed, thereby discarding connections that contribute little or introduce noise. On the user side, we perform clustering over interacted items to initialize multi-interest prototypes, then apply a triple disentanglement module that separates user representations into interest-invariant, effective interest-specific, and ineffective interest-specific components, ensuring clearer semantic boundaries and avoiding representation collapse. Experiments on three bench-mark datasets demonstrate that our methods achieve better performance on different metrics.

Keywords: Multimodal Recommendation; Graph Transformer; Disentangled Representation; Multi-Interest Modeling; User Behavior

1. Introduction

As e-commerce rapidly grows, users increasingly interact with diverse content such as images, texts, and videos, reflecting complex and multifaceted preferences. Traditional recommendation systems, which primarily rely on user-item interactions such as clicks or ratings, often fail to capture these rich signals. Multimodal Recommendation Systems (MRSs) address this limitation by integrating multiple data modalities, enabling a deeper understanding of user interests and delivering more accurate, personalized, and context-aware recommendations.

Previous studies [3, 23, 24, 30] typically follow a common pipeline: extracting multimodal features, modeling user interests, and measuring user-item similarity. Techniques such as Graph

†Corresponding author: Minghua Nuo (Email: nuominghua@163.com; ORCID: 0000-0002-6923-3895)

Convolutional Networks (GCNs) [24, 30] and attention mechanisms [3, 23] have been employed to capture high-order cross-modal interactions. More recently, self-supervised learning [22] have been introduced to further enhance multimodal representations. While these approaches demonstrate improved performance by utilizing multimodal information, they often overlook fine-grained analysis of how different modalities contribute to recommendation effectiveness, limiting their ability to fully exploit the richness of multimodal signals.

User preferences are often driven by multiple latent interests, which are key to improving recommendation accuracy. Existing multi-interest models typically rely either on auxiliary signals, which are often unavailable in practice, or extract interests directly from user behavior using attention or capsule networks. Although flexible, the latter often suffers from representation collapse, where different interest embeddings become overly similar and fail to reflect diverse user intent. To mitigate this, prior works introduce regularization strategies such as decorrelation [21], sparsity constraints [26], and norm regulation [13, 17]. However, these methods remain largely passive, focusing on preventing overlap rather than actively guiding semantically related items into distinct interest spaces. This highlights the need for a more proactive and semantically grounded approach to multi-interest modeling.

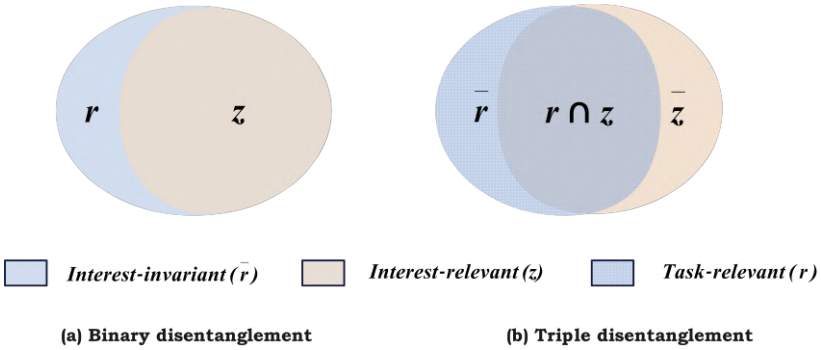

Figure 1: Comparison of binary and triple disentanglement strategies in multi-interest modeling

To tackle interest overlap, we adopt disentangled representation learning in recommendation. Traditional binary methods [12] split features into invariant and specific feature, assuming both are task-relevant which often fails in practice. For example, a user may often browse "streetwear hoodies" and "minimalist basics," but occasionally view "family-themed outfits" as gifts. Clustering these behaviors yields multiple interest vectors. However, traditional methods split features into shared and specific parts without filtering out noisy or irrelevant signals, often resulting in representation collapse and interest overlap.

To address the above issues, we propose **DMIMRec**, a **D**isentangled **M**ulti-**I**nterest Representation Learning for **M**ultimodal **Rec**ommendation Systems. As illustrated in Figure 1, unlike traditional binary disentanglement methods that split features into two parts, our DMIMRec introduces a more fine-grained separation of interest representations into three components: interest-invariant ($\bar{r}$) representations, effective interest-specific ($r \cap z$) representations, and ineffective interest-specific ($\bar{z}$) representations. This enables the model to actively filter out noisy or misleading signals, such as occasional behaviors that do not reflect true user preferences.

Our approach consists of two main stages. On the item side, we build a modality-aware item-item graph to enhance item representations and introduce a reconstruction-difference pruning

strategy that evaluates edge reliability based on both multimodal similarity and reconstruction contribution. On the user side, we cluster interacted items to initialize multi-interest embeddings and apply disentanglement learning within each cluster to ensure semantic clarity and reduce interest overlap. Unlike prior methods that passively enforce separation via regularization, our model promotes interest diversity through structured disentanglement.

- We propose DMIMRec, the first multimodal recommendation framework that disentangles user interests into invariant ($\bar{r}$), effective-specific ($r \cap z$), and ineffective-specific ($\bar{z}$) components, effectively mitigating interest collapse.
- We enhance item representations using modality-specific item-item graphs and introduce a reconstruction-difference guided pruning strategy to remove semantically noisy connections.
- Extensive experiments on three real-world datasets demonstrate the superior performance of DMIMRec over state-of-the-art methods, with further analyses validating the effectiveness of our design choices.

2. RELATED WORK

2.1. Multimodal Recommendation

Multimodal Recommender Systems enhance recommendation performance by integrating multimodal item information with users' historical behaviors. Early studies, such as Visual Bayesian Personalized Ranking (VBPR) proposed by He et al. [7], extended the BPR [14] to model user preferences for visual information. In recent years, Graph Neural Networks (GNNs) have emerged as effective tools for multimodal recommendation, capturing higher-order dependencies between users and items. For instance, LATTICE [29] performs graph convolution on the learned latent graph and explicitly injects higher-order affinities into the item representations; FREEDOM [32] enhances multimodal recommendation by denoising user-item interaction graphs and merging frozen item-item graphs; LGMRec [6] utilizes a modality hypergraph module and graph convolution for hyperedge information propagation, uncovering comprehensive global user interests. In addition, more recent advancements in attention-based methods [27, 23] have been widely applied in multimodal recommender systems to enhance feature extraction. UGT[27] builds a unified graph neural network to jointly fuse the multi-modal user/item representations derived from the output of the multi-way transformer; LightGT[23], a lightweight model, combines GNNs and attention mechanisms, efficiently learning user preferences from item features through attention, improving recommendation accuracy while reducing computational costs.

2.2. Disentangled Representation Learning

Recent studies have introduced dynamic analysis and feature learning within each modality, enabling a more granular approach to multimodal representation learning. These techniques typically disentangles features into two types: task-invariant representations, and task-specific representations. Inspired by the success of disentangled representation learning across various fields, numerous works have focused on learning disentangled representations for users and items in recommendation systems [13, 16, 21, 12]. For instance, MacridVAE[13] isolates user preferences associated with distinct concepts that align with different user intentions. Building on

the disentangled representations derived from user-item interaction modeling, ADDVAE[16] introduces additional disentangled user representations from textual content, aligning them with original user-item representations based on disentangled representations from user-item interaction modeling. DGCF[21] incorporates a graph disentangling module to refine an intent-aware interaction graph and factorial representations, capturing user intent diversity in item selection. Meanwhile, DRML[12] uses an attention module to estimate user attention weights across different modality factors, modeling preferences across various dimensions.

3. METHODOLOGY

In this section, we present a comprehensive description of the proposed DMIMRec framework, with its overall architecture illustrated in Figure 2.

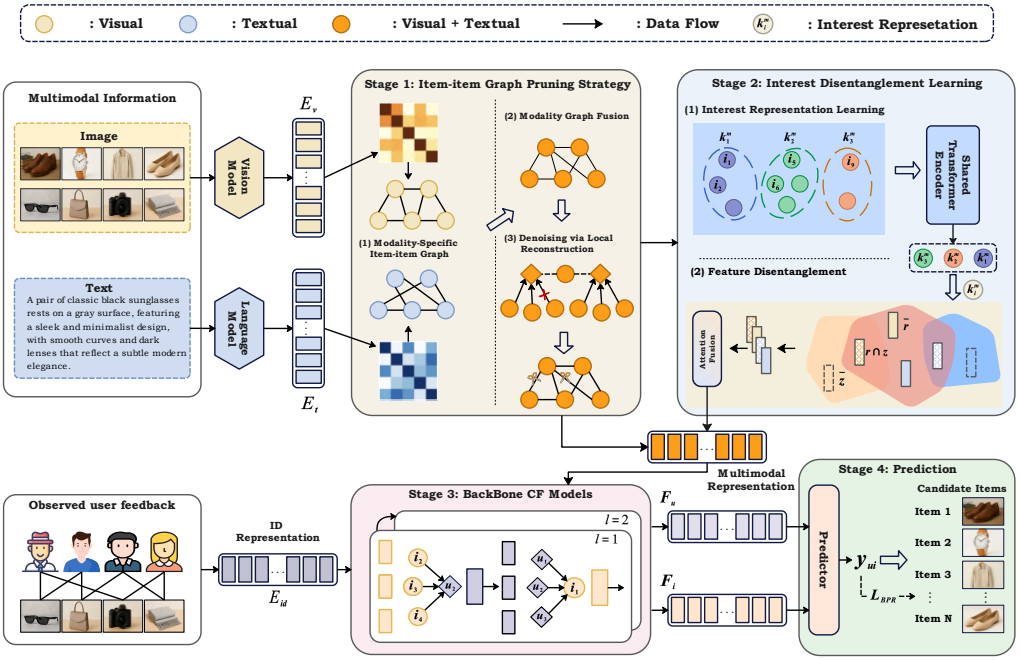

Figure 2: The structure overview of the proposed DMIMRec

3.1. Preliminaries

Let $\mathcal{U} = \{u\}$ denote the user set and $\mathcal{I} = \{i\}$ denote the item set. Then, we denote the features of each modality and the input ID embedding as $E_m = \{E_{u_m} \parallel E_{i_m}\} \in \mathbb{R}^{d_m \times (|\mathcal{U}|+|\mathcal{I}|)}$ and $E_{id} = \{E_{u_{id}} \parallel E_{i_{id}}\} \in \mathbb{R}^{d_m \times (|\mathcal{U}|+|\mathcal{I}|)}$, where $m \in \mathcal{M}$ is the modality, $\mathcal{M}$ is the set of modalities, d_m is the dimension of the features, and $\parallel$ denotes concatenation operation. We only consider visual, and textual modalities denoted by $\mathcal{M} = \{v, t\}$.

3.2. Denoising Multimodal Item Content

To utilize and denoise multimodal features, we first encode multimodal features and construct an item–item graph. We reconstruct each item's modal representation by aggregating features from its neighbors and compare it with the item's modal representation to assess the contribution of each edge. Edges that increase the reconstruction error are removed, preserving structurally consistent connections.

3.2.1. Constructing Item-Item Graph

We employ the KNN algorithm [2] to construct an item-item graph for each modality m, aimed at extracting multimodal relationships between items. Specifically, we calculate the similarity score S_{ij}^m for the item pair $(i, j) \in \mathcal{J}$ by measuring the cosine similarity between their original modality features, X_i^m and X_j^m.

$$S_{i,j}^m = \frac{X_i^m \cdot X_j^m}{\|X_i^m\| \cdot \|X_j^m\|}, \tag{1}$$

We retain only the top-k neighbors with the highest similarity scores to capture the most relevant features:

$$\bar{S}_{i,j}^m = \begin{cases} 1, & \text{if } j \in \text{top-k}(S_{i,p}^m, p \in \mathcal{J}), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where $\bar{S}_{i,j}^m$ represents the edge weight between item i and item j within modality m, and $S_{i,p}^m, p \in \mathcal{J}$ represents the neighbor scores for the i-th item.

3.2.2. Reconstruction-difference guided pruning strategy

To mitigate the inherent semantic noise in $\bar{S}_{i,j}^m$, we propose adjusting by reconstructing $\bar{S}_{i,j}^m$ the item's modal representation and evaluating edge quality by measuring the difference between the original feature and the version reconstructed from its neighbors. A useful neighbor should help preserve the target node's semantic information, while a noisy neighbor may distort it. For node j, the reconstructed representation is computed as:

$$\hat{X}_i = \frac{1}{|\mathcal{N}_i|} \sum_{i' \in \mathcal{N}_i} X_{i'}, \quad \bar{X}_i = \frac{1}{|\mathcal{N}_i - 1|} \sum_{i' \in \mathcal{N}_i, i' \notin k} X_{i'} \tag{3}$$

where $\hat{X}_i$ denotes the reconstructed representation using all neighbors, and $\bar{X}_i$ denote the reconstructed representation obtained by excluding neighbor k. The reconstruction difference I_i is defined as the difference between the $d\left(X_i, \hat{X}_i\right)$ and $d\left(X_i, \bar{X}_i\right)$:

$$I_i = d(X_i, \hat{X}_i) - d(X_i, \bar{X}_i) \tag{4}$$

where $d\,(a, b) = |a - b|^2$ denotes the squared Euclidean distance. If $I_i > 0$, removing neighbor k reduces the reconstruction error, which means this neighbor is noisy and should be pruned; otherwise, the edge is considered reliable. The pruning strategy is applied independently to each modality-specific graph to account for semantic differences.

After removing noisy connections, we obtain denoised $\bar{S}^m_{i,j}$ for each modality m. we merge all modality-specific item-item graph to improve representation learning:

$$\tilde{S} = \sum_{m \in \mathcal{M}} \alpha \bar{S}^m_{i,j}, \quad A^{(l)}_m = \sum_{j \in \mathcal{N}^i} \tilde{S} \cdot A^{(l-1)}_{j_m}, \tag{5}$$

where A_{j_m} is the embedding of item j in modality m. Following [32], we freeze the item-item graphs after initialization to avoid additional computation during training. The final enhanced embedding E_m for modality m is obtained as: $E_m = E_m + A^{(l)}_m$.

3.2.3. User-Item Behavioral Graph

We can treat ID embeddings as a unique representation of another modality. To capture high-order features, we initialize learnable embeddings for users and items and construct the user-item interaction graph $\mathcal{G}$. Formally, the user and item representations at l-th graph convolution layer can be formulated as:

$$E^{(l)}_{u_{id}} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} E^{(l-1)}_{i_{id}}, \quad E^{(l)}_{i_{id}} = \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} E^{(l-1)}_{u_{id}}. \tag{6}$$

where $\mathcal{N}_u$ and $\mathcal{N}_i$ denote the one-hop neighbors of z and i in $\mathcal{G}$, respectively. Meanwhile, feature propagation is performed on the constructed item-item similarity matrix $\tilde{S}_{i,j}$, adopting the same method as modality feature propagation to capture deep relations between items. The feature propagation is as follows:

$$A^{(l)}_{i_{id}} = \sum_{j \in \mathcal{N}^i} \tilde{S} \cdot A^{(l-1)}_{i_{id}}, \quad E_i = E_i + A^{(l)}_{i_{id}} \tag{7}$$

The final E_{id} is calculated as $E_{id} = \{E_{u_{id}} || E_{i_{id}}\}$, where $||$ denotes the concatenation operation.

3.3. Multi-Interest Representation Learning

In multimodal recommendation systems, leveraging multimodal information enhances item representations. However, understanding users' interests in detail remains a critical task. To better capture the diverse aspects of users' interests, we refine the analysis of their interaction data and apply clustering techniques to extract multiple preference profiles for each user.

3.3.1. Interest Representation Learning

We first convert the user history behavior sequences $\mathcal{S}_u = (i^1_u, i^2_u, ..., i^{|S_u|-1}_u)$ into a fixed-length sequence $\mathcal{S}_u = (i^1_u, i^2_u, ..., i^n_u)$, where n is the maximum sequence length and a hyperparameter. If the number of past interactions for user u exceeds n, we retain the most recent n items as the user sequence $\mathcal{S}_u$; otherwise, we pad the sequence length to n.

To better capture the diverse preferences of users, we extract multimodal embeddings for each item in $\mathcal{S}_u$, denoted as $H_u = [h_1, h_2, ..., h_n] \in \mathbb{R}^{n \times d}$. Before training, we perform clustering on these item embeddings as a preprocessing step to generate the initial user interest representation.

Since K-Means is sensitive to initialization, often leading to unstable clustering and slow convergence, we adopt K-Means++ [1] to improve initialization quality and accelerate convergence. K-Means++ partitions the item embeddings into K distinct groups. Each cluster centroid is then treated as a latent interest vector, resulting in a multi-interest representation of the user.

Algorithm 1 K-Means++ User Interest Clustering

Input: user history interaction $\mathbf{H}_u \in \mathbb{R}^{n \times d}$, number of interest clusters K, iteration number T
Output: user interest representation $f_{u_i}^m \in \mathbb{R}^{K \times d}$

Initialize K cluster centers $\{c_k^0\}_{k=1}^K$ using K-Means++ on $\mathbf{H}^u$
$t \leftarrow 0$
while $t < T$ **do**
  Compute similarity scores $S = \mathbf{H}_u(C^t)^\top$
  Assign items: $C = \arg\max(S, \text{axis} = \{-1\})$
  for $k = 1$ to K **do**
    $\mathbf{H}^k = \{h_i \in \mathbf{H}_u \mid C[i] = k\}$
    $c_k^{t+1} = \frac{1}{|\mathbf{H}^k|} \sum_{h_i \in \mathbf{H}^k} h_i$
  end for
  $t \leftarrow t + 1$
end while
return $f_u^m = [c_1^{u_i}, \ldots, c_k^{u_i}, \ldots, c_K^{u_i}]$

The complete clustering procedure used to construct the initial user interest representation is outlined in Algorithm 1, which serves as the core preprocessing step before training.

Because user interaction data often involves multiple modalities, it is essential to effectively extract and fuse multimodal features when modeling user interests. Transformer-based methods have recently shown strong capabilities in this area, but their complexity and parameter count grow rapidly with more modalities.

To address this, recent studies [9] have emphasized unified architectures that use shared cross-modal representations to enable efficient fusion and reduce redundancy [5]. Building on these insights, we introduce a shared modality Transformer for feature extraction, as shown in Figure 2. Using N layers, the model projects features $f_{u_i}^m$ into a unified embedding space, while preserving modality-specific characteristics and enabling implicit alignment. Compared to traditional multimodal Transformers [18], it significantly reduces parameter count. To support disentanglement, its output dimension is aligned with that of the interest-specific encoder, enabling efficient feature extraction and simplifying the overall model.

3.3.2. Disentanglement module

As stated in the introduction, the goal is to disentangle each interest representation into three components: interest-invariant representation $\bar{r}$, effective interest-specific representation $r \cap z$, and ineffective interest-specific representation $\bar{z}$, by removing $\bar{z}$ to preserve the independence between different interests. It is well known that the cross-attention mechanism can capture relationships between two vectors, helping the model understand where two representations overlap. Based on this, we design a dual-path cross-attention mechanism, as illustrated in Figure 3, to disentangle each user's interest representation.

We map the interest features to r and z, modeling them as task-relevant representations and interest-specific representations, respectively. Then, by applying the cross-attention mechanism, we allow r and z to attend to the relevant parts of each representation, enabling the model to precisely capture the intricate relationships between r and z. For the input features r, the query vector Q_r is generated as follows:

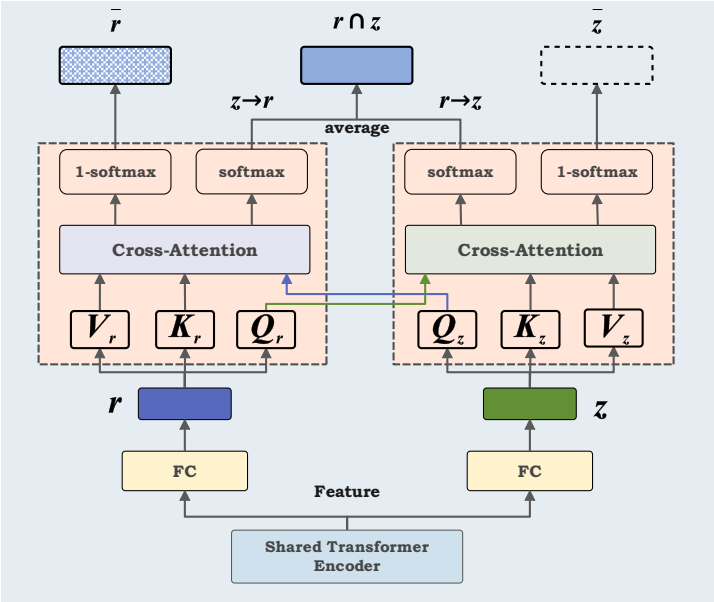

Figure 3: The interests disetanglenment module

$$Q_r = rW_{Q_r} \tag{8}$$

Simultaneously, the key and value vectors K_r and V_r are generated from r as:

$$K_r = rW_{K_r}, \quad V_r = rW_{V_r}. \tag{9}$$

where W_{Q_r}, W_{K_r}, and W_{V_r} are learnable parameters. Then, the intersection component from the perspective of r attending to z:

$$f^m_{r\to z} = \text{softmax}\left(\frac{Q_r K_z^\top}{\sqrt{d}}\right) V_z, \tag{10}$$

where d represents the dimension of the feature vectors. Assuming r lies in a unit space, the interest-invariant representation $\bar{r}$ can be derived using the complementary attention weights:

$$f^m_{\bar{r}} = (1 - softmax\left(\frac{Q_z K_r^\top}{\sqrt{d}}\right))V_r, \tag{11}$$

Similarly, from the perspective of z attending to r, we can obtain $f^m_{z\to r}$ and $f^m_{\bar{z}}$:

$$f^m_{z\to r} = \text{softmax}\left(\frac{Q_z K_r^\top}{\sqrt{d}}\right) V_r, \quad f^m_{\bar{z}} = (1 - softmax\left(\frac{Q_r K_z^\top}{\sqrt{d}}\right))V_z \tag{12}$$

The final intersection $f^m_{r\cap z}$ is then computed as the average of the two directional attentions:

$$f^m_{r\cap z} = \frac{f^m_{r\to z} + f^m_{z\to r}}{2}, \tag{13}$$

To avoid losing useful information through excessive disentangling, we add a residual connection with the input feature c_k^u, thereby preserving and enhancing important features, as detailed below:

$$\tilde{f}^m_{r\cap z} = f^m_{r\cap z} + c^u_k, \tag{14}$$

After disentangling each interest representation for the user, we first remove $f^m_{\bar{z}}$, and then aggregate the components $\tilde{f}^m_{r\cap z}$ and $f^m_{\bar{r}}$ for each interest. The specific aggregation process is as follows:

$$\bar{f}^m_{r\cap z} = \text{MeanPool}(\tilde{f}^{m,c_1}_{r\cap z}, ..., \tilde{f}^{m,c_k}_{r\cap z}, ..., \tilde{f}^{m,c_K}_{r\cap z}), \tag{15}$$

$$\bar{f}^m_{\bar{z}} = \text{MeanPool}(f^{m,c_1}_{\bar{z}}, ..., f^{m,c_k}_{\bar{z}}, ..., f^{m,c_K}_{\bar{z}}) \tag{16}$$

Given $\bar{f}^m_{r\cap z}$ and $\bar{f}^m_{\bar{z}}$ are highly relevant to the task, we employ a multi-head attention mechanism to model their interactions, thereby deriving the ultimate user representation. The detailed procedure is outlined as follows:

$$f^m_u = \text{Attention}([\bar{f}^m_{r\cap z}, \bar{f}^m_{\bar{r}}]), \tag{17}$$

Subsequently, we utilize the updated representations to enhance the representations of modal items through graph convolution, aiming to capture higher-order modality features:

$$F^{(l)}_{i_m} = \sum_{u\in\mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} F^{(l-1)}_{u_m}, \tag{18}$$

3.4. Fusion and Prediction

We obtain the final representations E' of users and items by fusing their two types of modality embeddings, $E_{id} = [E_{u_{id}} \| E_{i_{id}}]$ and $F_m = [F_{u_m} \| F_{i_m}]$, as follows:

$$E' = E_{id} + \sum_{m\in\mathcal{M}} Norm(F_m), \tag{19}$$

where Norm$(\cdot)$ is a normalization function that helps mitigate the scale differences among embeddings. To predict the interactions between users and items, we calculate the preference score $r_{u,i}$ using the inner product, as follows:

$$r_{u,i} = {e'}_u \cdot {e'}^T_i, \tag{20}$$

where $r_{u,i}$ represents the predicted score, estimating the likelihood that user u prefers item i based on historical behaviors. To optimize the model parameters, we use the Bayesian Personalized Ranking (BPR) loss [14], which is defined as follows:

$$\mathcal{L} = \sum_{(u,i^+,i^-)\in\mathcal{R}} -\ln\varphi(r_{u,i^+} - r_{u,i^-}) + \lambda\|\Theta\|_2^2. \tag{21}$$

where $\mathcal{R} = \{(u, i^+, i^-) \mid (u, i^+) \in \mathcal{V}, (u, i^-) \notin \mathcal{V}\}$ represents a triplet consisting of a user, an item i^+, that the user interacted with, and an item i^- that the user did not interact with. Additionally, $\varphi(\cdot)$ denotes the sigmoid function, and λ and Θ denote the regularization coefficient and model parameters, respectively. [ht]

Table 1: Statistics of the evaluation datasets.

Dataset	#User	#Item	#Interaction	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.96%
Clothing	39,387	23,033	278,677	99.97%

Table 2: Overall performance comparison between our model and the baselines on three datasets.

Baseline	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
LightGCN	0.0464	0.0732	0.0251	0.0320	0.0553	0.0829	0.0307	0.0379	0.0331	0.0514	0.0181	0.0227
SGL	0.0532	0.0820	0.0289	0.0363	0.0620	0.0945	0.0339	0.0423	0.0392	0.0586	0.0216	0.0266
NCL	0.0538	0.0836	0.0292	0.0369	0.0616	0.0940	0.0339	0.0421	0.0410	0.0607	0.0228	0.0275
LATTICE	0.0536	0.0858	0.0287	0.0370	0.0618	0.0950	0.0337	0.0423	0.0459	0.0702	0.0253	0.0306
MMGCL	0.0521	0.0790	0.0283	0.0352	0.0617	0.0913	0.0351	0.0428	0.0410	0.0607	0.0227	0.0277
SLMRec	0.0540	0.0810	0.0296	0.0361	0.0676	0.1007	0.0374	0.0462	0.0452	0.0675	0.0247	0.0303
Freedom	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0629	0.0941	0.0341	0.0420
LGMRec	0.0639	0.0989	0.0337	0.0430	0.0719	0.1068	0.0387	0.0477	0.0555	0.0828	0.0302	0.0371
DMRL	0.0543	0.0847	0.0322	0.0405	0.0672	0.1008	0.0393	0.0484	0.0549	0.0791	0.0311	0.0373
LightGT	0.0544	0.0867	0.0294	0.0381	0.0652	0.0953	0.0347	0.0440	0.0514	0.0755	0.0299	0.0353
UGT	0.0602	0.0930	0.0330	0.0325	0.0705	0.1034	0.0391	0.0477	0.0603	0.0922	0.0330	0.0402
DMIMRec	**0.0660**	**0.1025**	**0.0354**	**0.0453**	**0.0741**	**0.1128**	**0.0405**	**0.0502**	**0.0667**	**0.0973**	**0.0360**	**0.0437**
%Improv	3.29%	3.33%	5.05%	5.35%	3.06%	3.58%	3.05%	3.72%	6.04%	3.40%	5.57%	4.04%

4. EXPERIMENTS

In this section, our goal is to validate the effectiveness of the proposed method, DMIMRec. Specifically, we conducted a series of comprehensive experiments to investigate the following research questions:

- **RQ1:** Whether the proposed DMIMRec outperforms state-of-the-art multimodal recommendation methods?
- **RQ2:** How the different modules in DMIMRec affect the model's performance?
- **RQ3:** Which feature extraction methods are effective for disentanglement modules?
- **RQ4:** How different degrees of disentanglement influence our model's performance?
- **RQ5:** What is the effectiveness and efficiency of the modules in feature extraction?

4.1. Experimental Setting

4.1.1. Dataset

To evaluate the proposed model, we followed previous studies and conducted comprehensive experiments on three widely used Amazon datasets, including Baby, Sports and Outdoors(Sports), and Clothing, Shoes and Jewelry(Clothing). The three datasets consist of both visual and textual modal features. In this paper, we utilize the 4096-dimensional original visual features and the 384-dimensional original textual features, which were extracted and published in previous work[33]. The statistics of the three datasets are summarized in Table 1.

4.1.2. Experimental Details

All experiments are conducted on an Nvidia Tesla V100 GPU with 32GB memory. For each dataset, historical interactions are randomly split into training, validation, and test sets with an 8:1:1 ratio. Following [6], we adopt Recall@K and Normalized Discounted Cumulative Gain (NDCG@K) with $K \in \{10, 20\}$, and report the average results on the test set. The proposed DMIMRec model is implemented in PyTorch based on MMRec [31]. To ensure fair comparisons, all models are trained with a batch size of 2048, embedding dimension $d = 64$, and for graph-based methods, a propagation layer count $L = 2$. Model parameters are initialized using Xavier initialization [4] and optimized with Adam [10]. The learning rate and regularization weight are selected via grid search over$\{0.0001, 0.0005, 0.001, 0.005\}$, and early stopping is applied if R@20 on the validation set shows no improvement for 30 consecutive epochs.

4.1.3. Baselines

To evaluate the effectiveness of our model, we compare DMIMRec with CF-based recommendation methods, popular GNN-based models, and state-of-the-art multimodal recommender systems.

1. **GNN-based Recommendation Models**
   - **LightGCN** [8]: It simplifies GCN by removing feature transformation and nonlinear activation, efficiently capturing high-order interactions to compute user and item representations.
   - **SGL** [25]: This model enhances graph collaborative filtering by using contrastive learning signals and data augmentation techniques like node/edge dropout and random walks, maximizing agreement between a node's multiple views.
   - **NCL** [11]: It generates contrastive views by using EM-based clustering to identify semantic and structural neighboring nodes, forming positive contrastive pairs.
2. **Multimodal Recommender Systems**
   - **MMGCL** [28]: It incorporates the graph contrastive learning into recommender through modality edge dropout and masking.
   - **LATTICE** [29]: This model learns the item-item structure for each modality and aggregates them to form a semantic item-item graph, in order to obtain better item representations.
   - **SLMRec** [15]: It designs data augmentation on multimodal content with two components, i.e., noise perturbation over features and multi-modal pattern uncovering augmentation.
   - **Freedom** [32]: This model demonstrates that fixing the latent item-item graph eliminates the need for dynamically generating graph structures, which are both unnecessary and memory-intensive.
   - **LGMRec** [6]: This multimodal method jointly models local and global user interests by learning representations from both local and global graphs.
3. **Graph Transformer and Disentanglement Models**
   - **DMRL** [12]: This model disentangles representations within individual modalities and employs an attention module to determine users' preferences for these representations, thereby enhancing recommendation accuracy.

- **LightGT** [23]: This Transformer-based recommendation model uses interest-specific embeddings for similarity measurement and a lightweight self-attention module for efficiency, integrating user preferences from item features to improve user-item interaction prediction.
- **UGT** [27]: This model leverages a multi-way transformer to extract aligned multimodal features from raw data and build a unified graph neural network to jointly fuse the multimodal user/item representations derived from the output of the multi-way transformer.

4.2. Performance Comparison

The performance comparison of our proposed model and baseline methods across three datasets is shown in Table 2, from which we have the following key observations:

- In all evaluation metrics {Recall, NDCG}@{10, 20}, DMIMRec consistently delivers excellent performance across all datasets. In detail, compared to the best performance of the baselines, our proposed model achieves an average improvement of 3.77%, 3.35%, and 4.71% in the three datasets, respectively. We attribute these results to the incorporation of temporal information into the item-item graph, which effectively reduces noise in the original item-item graph. It can be observed that our method outperforms the baseline methods in every metric for each dataset. This also demonstrates the effectiveness of our proposed approach.
- Although some multimodal methods, such as MMGCL and SLMRec, leverage modality information to enhance contrastive learning for data augmentation, they still exhibit certain limitations. For example, directly masking modality features in MMGCL may lead to the loss of important information, thereby weakening the completeness of representation learning. In addition, SLMRec generates augmented views based on predefined hierarchical correlations among different modalities, which may fail to adapt effectively to the diverse modality structures in various multimodal recommendation datasets, thus compromising the effectiveness of self-supervised signals. On the other hand, our proposed model employs disentangled representation learning to effectively remove interest-irrelevant information while retaining interest-invariant and interest-specific features, thereby enhancing the expressiveness and discriminative power of user representations.
- Recent models such as LightGT and UGT leverage Transformers to enhance the representation capabilities of users and items. However, they still exhibit certain limitations. For instance, LightGT introduces redundant modality information during the learning of user and item representations, failing to effectively eliminate noisy features. Meanwhile, UGT tends to cause interference between modalities when modeling multiple types of modality information simultaneously, making it difficult for different modalities to collaborate effectively, which in turn hinders further improvement in recommendation performance. In contrast, DMIMRec employs disentangled representation learning to effectively separate redundant and interfering information among different user interests across modalities, thereby enhancing the clarity of representations and the independence between modalities.
- Comparing the performance of DMIMRec across three datasets, it is evident that the model performs best in highly sparse interaction scenarios (e.g., the Clothing dataset), demonstrating its strong adaptability and robustness in handling data sparsity and long-tail effects.

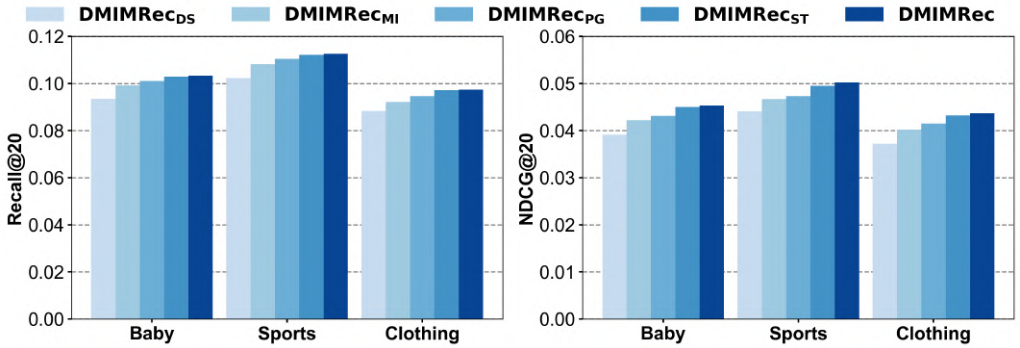

Figure 4: Ablation of different modules on DMIMRec.

4.3. Ablation Study

We perform ablation studies to investigate the compositional effects of DMIMRec. From the results reported in Figure 4, we found:

- **w/o DMIMRec$_{DS}$**: We conducted an ablation study by removing the interest disentanglement module and directly using the simple multi-interest features, which resulted in the worst performance. This emphasizes the critical issue of modality noise in multimodal recommendations and underscores the important contribution of our disentanglement module.
- **w/o DMIMRec$_{MI}$**: To compare the importance of the multi-interest module, we replaced it with the traditional user representation. The comparison results show a significant drop in model performance when using the traditional user representation. This is likely due to the separation of multiple interests, which diminishes the model's ability to comprehend user preferences, thereby validating the superiority of our multi-interest module.
- **w/o DMIMRec$_{PG}$**: After removing the pruned item-item graph module, the model showed a consistent drop in recommendation performance across all three datasets. This indicates that modality-based similarity graphs without pruning tend to preserve semantically irrelevant or misleading edges, which act as "false neighbors" and hinder the learning of expressive item representations. In contrast, our reconstruction-difference guided pruning strategy effectively eliminates such low-quality edges, improving the structural reliability and informativeness of the item graph, thereby enhancing the accuracy and robustness of recommendations.
- **w/o DMIMRec$_{ST}$**: To validate the rationale behind the shared Transformer module, we replaced it with a modality-specific Transformer structure and conducted a comparative analysis. The results show that the shared-parameter Transformer effectively reduces the number of model parameters without compromising performance. Furthermore, the shared Transformer not only improves the efficiency and consistency across different modalities, but also implicitly achieves cross-modal alignment.

4.4. In-Depth Analysis

4.4.1. Effect of Different Feature Extraction in Disentanglement

Table 3 compares the performance of three feature extractors used in the disentanglement module: MLP-Dis, GCN-Dis, and Attention-Dis. Across all datasets and evaluation metrics, Attention-Dis consistently achieves the best performance, with improvements up to 18.7% in R@20 and 26.7% in N@20 over MLP-Dis on the Baby dataset. While GCN-Dis performs better than MLP-Dis by leveraging graph structural information, it still falls short of Attention-Dis. These results demonstrate that the attention mechanism is more effective in capturing complex feature dependencies and generating more discriminative representations for disentanglement.

Table 3: Performance comparison and analysis of different feature extraction methods for disentanglement.

Methods	Baby		Sports		Clothing	
	R@20	N@20	R@20	N@20	R@20	N@20
MLP-Dis	0.0864	0.0353	0.0886	0.0387	0.0829	0.0271
GCN-Dis	0.0986	0.0432	0.1072	0.0478	0.0937	0.0335
Attention-Dis	**0.1025**	**0.0453**	**0.1128**	**0.0502**	**0.0973**	**0.0437**

4.4.2. Effect of Different Disentanglement Degrees

We investigated the impact of different levels of disentanglement on model performance, as shown in Figure 5. The results consistently show that the model achieves optimal performance when the disentanglement degree is set to 1.0. This indicates a complete separation between invariant, effective, and ineffective interest components. As the disentanglement degree decreases from 1.0 to 0.6, we observe a clear performance drop across all datasets. This trend suggests that insufficient separation between latent interest types weakens the model's ability to suppress noisy signals, ultimately degrading recommendation precision. The consistent decline in both R@20 and N@20 further confirms the importance of high disentanglement. These findings highlight the necessity of a strong disentangling strategy to preserve the purity and diversity of user preference representations.

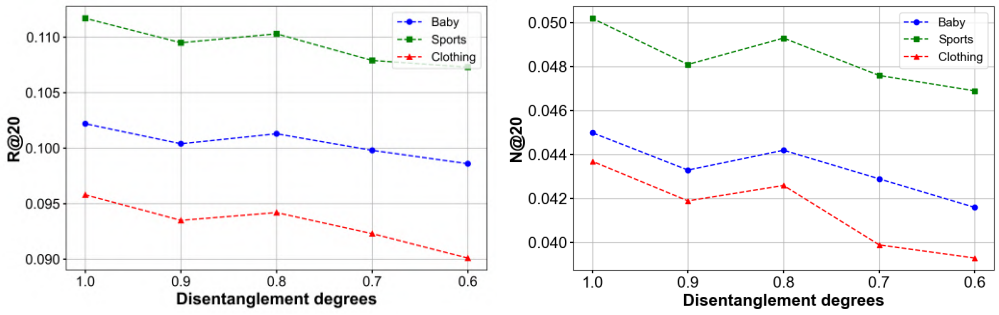

Figure 5: Effect of different disentanglement degrees

4.4.3. Effect of Interest Cluster Numbers

As shown in Figure 6, we evaluate the impact of varying the number of interest clusters $k \in \{2, 3, 4\}$ on model performance across the Baby, Sports, and Clothing datasets. The results

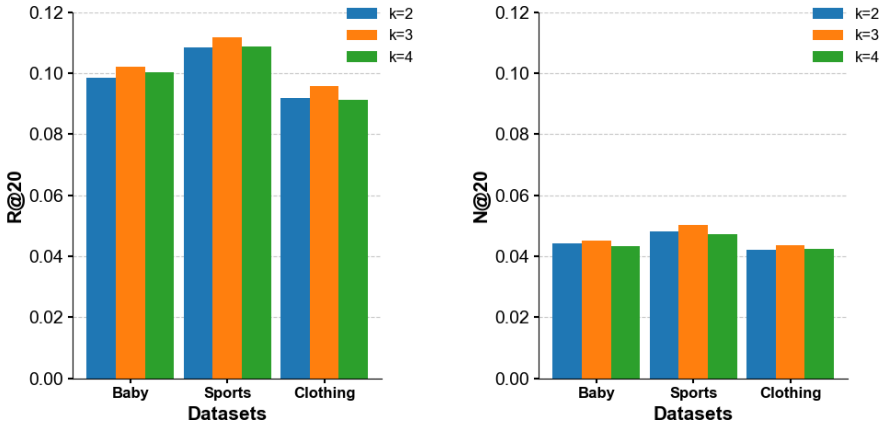

Figure 6: Effect of interest cluster numbers k.

show that setting $k = 3$ consistently achieves the best performance in terms of R@20 and N@20. When k is too small, the model may underrepresent the diversity of user interests; when k is too large, it may lead to interest redundancy and noisy representations, especially on smaller or sparser datasets. These findings suggest that using a moderate number of clusters provides a good trade-off between expressive power and generalization ability.

4.4.4. *Effect of Shared Transformer Layers*

As shown in Figure 7, we investigate the impact of varying the number of shared Transformer layers from 1 to 4. The model consistently achieves the best performance at 2 layers across all datasets in both R@20 and N@20. With only 1 layer, the model may lack sufficient capacity to capture deeper interest semantics. Increasing to 2 layers significantly improves performance. However, using 3 or 4 layers results in performance degradation, likely due to overfitting, especially on smaller datasets such as Clothing. These results demonstrate that using two shared Transformer layers strikes an effective balance between expressiveness and generalization.

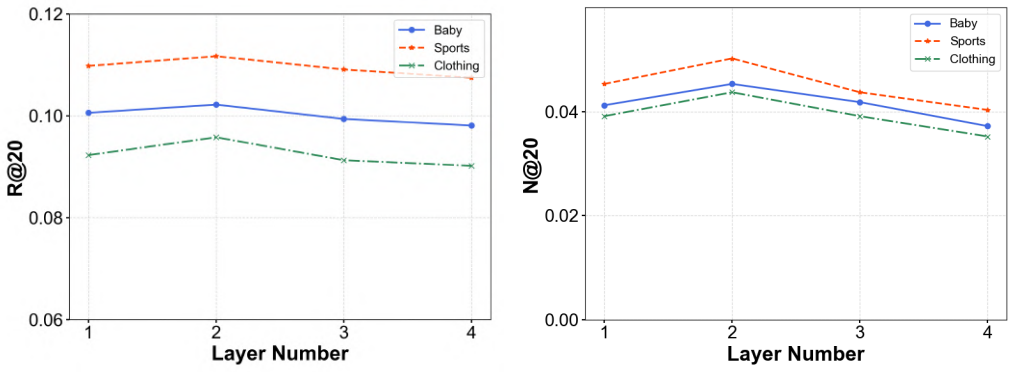

Figure 7: Effect of shared transformer layers.

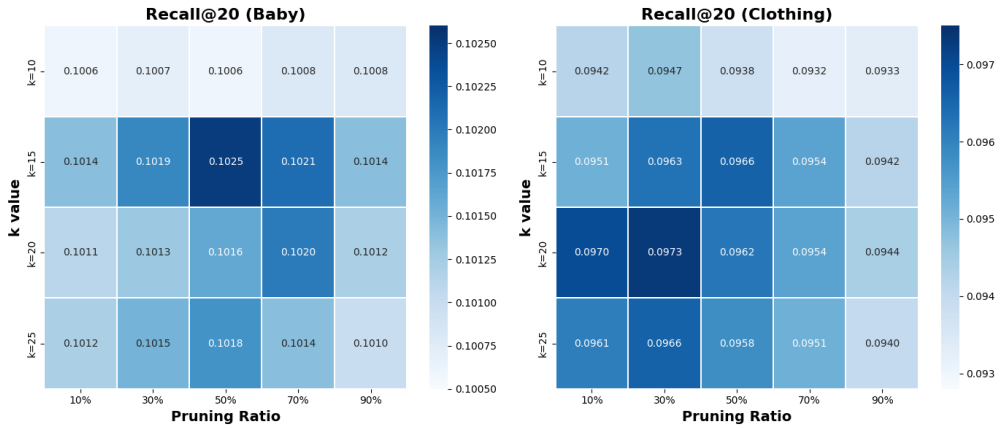

Figure 8: Effect of pruning ratio and neighbor size k.

4.4.5. Effect of Pruning Ratio and Neighbor Size k

We conduct experiments on two datasets with different sparsity levels to investigate the impact of the pruning ratio and the number of neighbors k on model performance. Specifically, we vary the number of neighbors $k \in \{10, 15, 20, 25\}$ and the pruning ratio $[90\%, 70\%, 50\%, 30\%, 10\%]$, and evaluate performance using Recall@20, as shown in Figure 8. On the denser Baby dataset, the model achieves its highest performance at $k = 15$ with a pruning ratio of 50%. This suggests that moderate pruning effectively removes redundant edges while preserving essential structural information. In contrast, for the sparser Clothing dataset, the best performance is observed at $k = 20$ with a pruning ratio of 70%, indicating that in sparse scenarios, increasing the neighborhood size combined with more aggressive pruning facilitates the construction of a more meaningful item-item graph by filtering out noisy relations.

4.5. Efficiency Study

As illustrated in Figur 9, we compare the training efficiency of DMIMRec with baseline models including LATTICE, Freedom, and LGMRec in terms of memory usage and training time per epoch. DMIMRec achieves superior performance while keeping both memory usage and training time within an acceptable range. Compared with LGMRec and Freedom, DMIMRec introduces only a slight increase in resource usage across all datasets. Notably, DMIMRec reduces memory and time costs compared to LATTICE, highlighting its lightweight and efficient design. These results demonstrate that DMIMRec maintains a strong efficiency-performance trade-off, making it well-suited for practical deployment.

4.6. Visualization Analysis

To further demonstrate the effectiveness of our proposed disentangled multi-interest modeling strategy, we visualize the distribution of user representations learned by DMIMRec and compare it with the baseline model LGMRec. As described in Section 4.3.2, we randomly sample 2000 user representations and apply t-SNE [19] to project them into a two-dimensional space. The visualization results, shown in Figure 10, reveal that DMIMRec produces more uniformly distributed user embeddings, while LGMRec exhibits relatively scattered and clustered

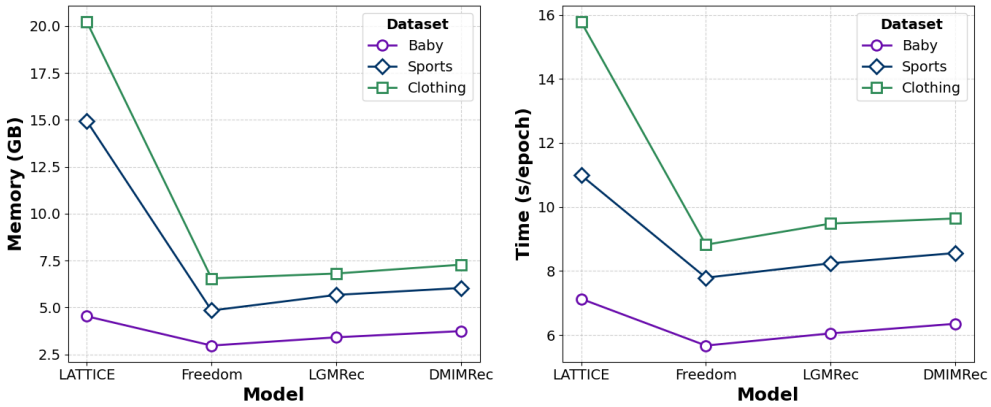

Figure 9: Efficiency analysis of different models. The left figure shows the memory usage across datasets. The right figure presents the training time per epoch.

patterns. Previous research [20] has demonstrated that uniformity in the embedding space is highly correlated with recommendation effectiveness, which explains the improved performance of DMIMRec in our experiments. This suggests that DMIMRec, by incorporating an effective disentanglement strategy, better captures the diversity of user preferences and effectively alleviates representation collapse.

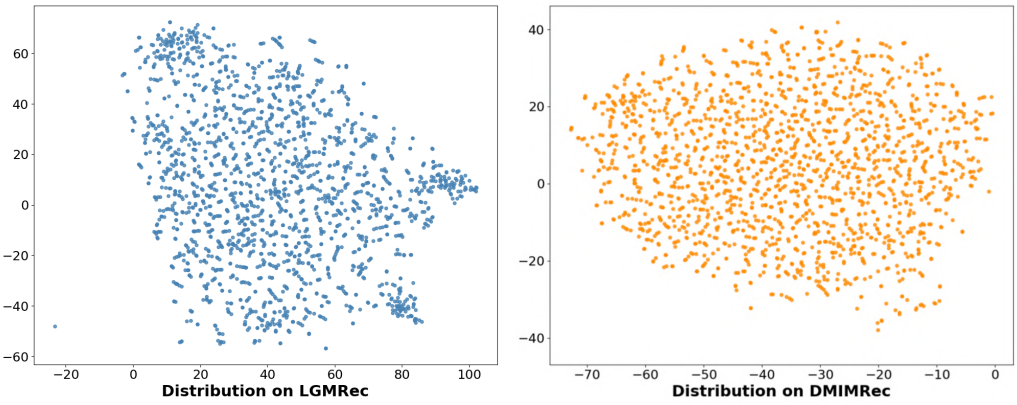

Figure 10: The distribution of user representations. The left part shows LGMRec, while the right part shows DMIMRec with disentanglement.

5. Case Study

To intuitively verify the effectiveness of our disentanglement module in eliminating ineffective interest-specific representations and refining user interest modeling, we conduct a case study on two representative users (e.g., *u2387* and *u4036*), and compare their recommendation results under two settings: Disentangled and Non-Disentangled. The visualized comparison is shown

in Figure 11. As observed, in the non-disentangled representation, similar items (e.g., *i5872* and *i1241* , *i3364* and *i5524*) are mistakenly assigned to different interest clusters, indicating interest overlap and a lack of representation independence. In contrast, the disentangled representation successfully groups these similar items into the same interest cluster, demonstrating that the module can identify and filter out ineffective interest-specific representations (e.g.,*i303* and *i464*)), resulting in more distinct and coherent interest representations. The disentangled model produces recommendations more aligned with user preferences, showing better noise suppression and a clearer focus on core interests compared to the non-disentangled model.

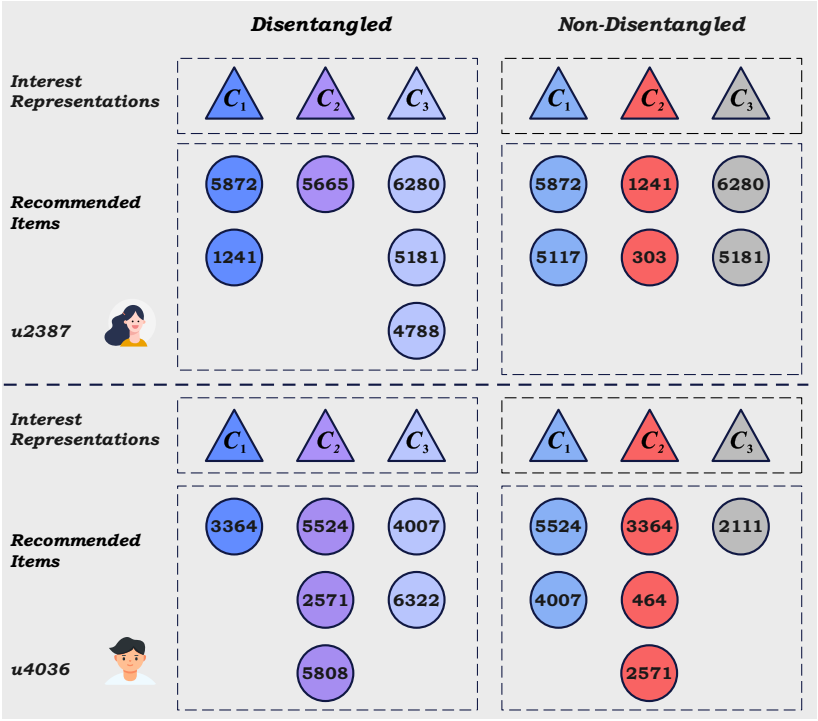

Figure 11: Case study comparing user interest clustering and recommendation results under Disentangled and Non-Disentangled settings.

6. CONCLUSION

In this work, we proposed DMIMRec, a novel multimodal recommendation framework that integrates reconstruction-error-based graph pruning and proactive multi-interest disentanglement. Our method enhances item representations by refining modality-specific item-item graphs and improves user modeling through interest-specific decomposition. Experiments on multiple Amazon datasets demonstrate that DMIMRec effectively captures diverse user preferences and alleviates semantic noise, leading to consistently superior recommendation performance.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No.62366038, No.61966025), Natural Science Foundation of Inner Mongolia (No.2023MS06010, No.2024MS06013). We also would like to express our sincere gratitude to the editor and anonymous reviewers for their valuable comments, which have greatly improved this paper.

AUTHOR CONTRIBUTIONS

C. Z., M. N., and Y. R. conceived the core idea of the paper and designed the overall methodology. C. Z. and M. N. implemented the model and conducted the experiments. Y. R. and R. L. contributed to the analysis of experimental results. C. Z., M. N., and H. L. co-wrote the manuscript and prepared the figures, tables, and supplementary materials. All authors discussed the results and reviewed the final manuscript.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[2] Jie Chen, Haw-ren Fang, and Yousef Saad. Fast approximate knn graph construction for high dimensional data via recursive lanczos bisection. *Journal of Machine Learning Research*, 10(9), 2009.

[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 335–344, 2017.

[4] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[5] Yuan Gong, Alexander H Liu, Andrew Rouditchenko, and James Glass. Uavm: Towards unifying audio and visual models. *IEEE Signal Processing Letters*, 29:2437–2441, 2022.

[6] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. Lgmrec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8454–8462, 2024.

[7] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[8] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 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*, pages 639–648, 2020.

[9] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[10] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM web conference 2022*, pages 2320–2329, 2022.

[12] Fan Liu, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, and Mohan Kankanhalli. Disentangled multimodal representation learning for recommendation. *IEEE Transactions on Multimedia*, 25:7149–7159, 2022.

[13] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. Learning disentangled representations for recommendation. *Advances in neural information processing systems*, 32, 2019.

[14] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[15] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:5107–5116, 2022.

[16] Nhu-Thuat Tran and Hady W Lauw. Aligning dual disentangled user representations from ratings and textual content. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1798–1806, 2022.

[17] Nhu-Thuat Tran and Hady W Lauw. Multi-representation variational autoencoder via iterative latent attention and implicit differentiation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2462–2471, 2023.

[18] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[19] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[20] Chenyang Wang, Yuanqing Yu, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. Towards representation alignment and uniformity in collaborative filtering. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1816–1825, 2022.

[21] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1001–1010, 2020.

[22] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 790–800, 2023.

[23] Yinwei Wei, Wenqi Liu, Fan Liu, Xiang Wang, Liqiang Nie, and Tat-Seng Chua. Lightgt: A light graph transformer for multimedia recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1508–1517, 2023.

[24] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.

[25] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735, 2021.

[26] Yueqi Xie, Jingqi Gao, Peilin Zhou, Qichen Ye, Yining Hua, Jae Boum Kim, Fangzhao Wu, and Sunghun Kim. Rethinking multi-interest learning for candidate matching in recommender systems. In *Proceedings of the 17th ACM conference on recommender systems*, pages 283–293, 2023.

[27] Zixuan Yi and Iadh Ounis. A unified graph transformer for overcoming isolations in multi-modal recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 518–527, 2024.

[28] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1807–1811, 2022.

[29] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3872–3880, 2021.

[30] Feng Zhao and Donglin Wang. Multimodal graph meta contrastive learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3657–3661, 2021.

[31] Xin Zhou. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pages 1–2, 2023.

[32] Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 935–943, 2023.

[33] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 845–854, 2023.