DHMRec: Collaboration-Guided Multimodal Disentanglement and Hierarchical Fusion for Recommendation

Xiaohan Zhan[1], Yuliang Shi[1,2*], Jihu Wang[3*], Shijun Liu[1,2], Fanyu Kong[1], Zhiyong Chen[1]

[1]School of Software, Shandong University
[2]Dareway Software Co., Ltd
[3]School of Control Science and Engineering, Shandong University
202315284@mail.sdu.edu.cn, {shiyuliang, wangjihu, lsj, fanyukong, chenzy}@sdu.edu.cn

Abstract

Multimodal recommender systems have emerged as a pivotal paradigm for harnessing diverse data modalities to deliver personalized services. Contemporary research predominantly focuses on integrating heterogeneous modality information through graph learning. However, these approaches face two key challenges: (1) the inherent complexity of modalities, characterized by entangled redundant signals and noise; and (2) the challenge of effectively integrating multimodal representations, each of which may exert varying degrees of influence on users' preferences. To address these challenges, we propose a novel Collaboration-Guided Multimodal Disentanglement and Hierarchical Fusion for Recommendation (DHMRec), which simultaneously achieves intra-modal denoising disentanglement and inter-modal hierarchical fusion. Specifically, we introduce a collaboration-related modality disentanglement module to distinguish between modality-common and modality-specific features. Then, through multi-view graph learning to capture both item-item dependencies and user-item interaction patterns. Additionally, we implement hierarchical fusion between the disentangled multimodal features and ID embeddings using a positive-negative attention-aware fusion module and an interaction distribution-based alignment module. Extensive experiments on three benchmarks demonstrate that our DHMRec surpasses various state-of-the-art baselines, highlighting its effectiveness in intra-modal disentanglement and multimodal features fusion.

Extended version — https://github.com/cheer-io/DHMRec

Introduction

Taking advantage of the potential semantics of various modalities (e.g., visual and textual), multimodal recommender systems (MRSs) can deliver more precise and personalized services. For example, in e-commerce, product images capture users' aesthetic preferences while textual descriptions convey functional attributes. Integrating these modalities into the recommender system enables a more comprehensive understanding of users' intent, potentially yielding more accurate and satisfying recommendations.

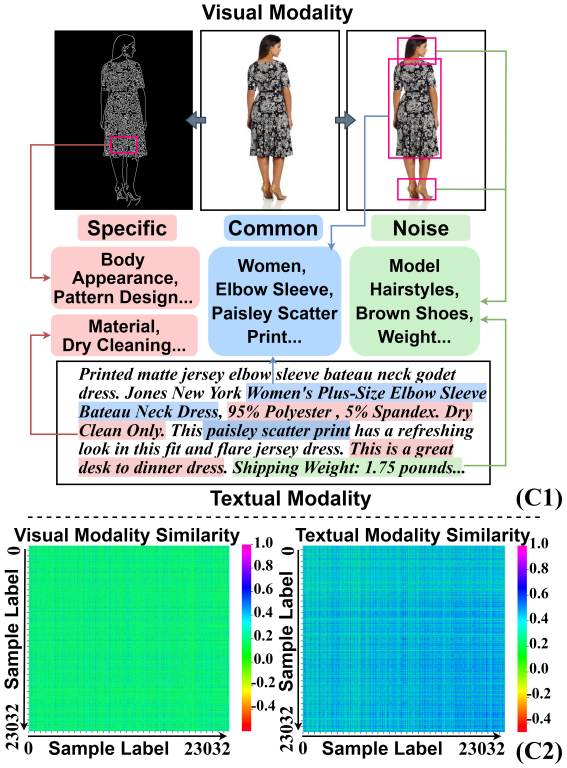


Figure 1: Illustration of existing challenges in MRSs. (C1) An example to illustrate the challenge of intra-modal feature entanglement and noise interference. (C2) Similarity of item pairs in visual modality and textual modality.

Early research in multimodal recommendation usually extends the conventional collaborative filtering (CF) framework by directly fusing multimodal features and ID embeddings (He and McAuley 2016b). Benefiting from the powerful capabilities of graph neural networks (GNNs) in modeling complex relationships, recent studies have increasingly applied graph representation learning to multimodal graphs to capture users' higher-order interests. For example, MMGCN (Wei et al. 2019) and GRCN (Wei et al. 2020) leverage message passing on modality-specific user-item bi-

*Corresponding author.

partite graphs to capture users' preferences across different modalities. Furthermore, LATTICE (Zhang et al. 2021), MICRO (Zhang et al. 2022) and FREEDOM (Zhou and Shen 2023) explore semantic relationships by constructing item-item graphs. LGMRec (Guo et al. 2024) further introduces a hypergraph module to capture modality-aware global representations.

Although these methods have shown impressive performance, there are still two critical challenges to be solved. **C1: Intra-Modal Feature Entanglement and Noise Interference.** Although multimodal data offer rich semantic cues, the inherent heterogeneity across modalities often introduces redundancy and irrelevant noise, hindering effective multimodal representation learning. As shown in Figure 1(C1) using the Amazon Clothing dataset, textual and visual modalities contain both shared features (e.g., color, style) and complementary information (e.g., visual patterns or textual material descriptions). However, these features are often entangled, making it challenging to isolate modality-specific semantics and accurately capture user preferences. Besides, as multimodal feature encoders in MRSs aim to comprehensively capture fine-grained modal details, they may inadvertently introduce noise irrelevant to user interests (e.g., model's hairstyle). Most existing methods fail to disentangle shared and specific features or filter out preference-independent noise, limiting their ability to fully utilize cross-modal information. **C2: Fusion of Differentiated Multimodal Representations.** Another key challenge lies in dynamically fusing multimodal information aligned with user preferences. As shown in Figure 1(C2), item-item similarities differ significantly across modalities, and these diverse modality features impact user preferences to varying degrees across different tasks. For instance, visual features are more critical in clothing recommendation, while textual details are more relevant for baby products. However, most graph-based MRSs treat all modalities equally, overlooking such distinctions. This uniform treatment may amplify irrelevant signals and lead to inaccurate user interest modeling, ultimately degrading recommendation performance.

To address these challenges, we propose a framework called Collaboration-Guided Multimodal Disentanglement and Hierarchical Fusion for Recommendation (DHMRec). It introduces a unified disentanglement–fusion–alignment design to realize a collaborative guidance framework for multimodal redundancy reduction and adaptive fusion tailored to recommendation tasks. Specifically, to address **C1**, we introduce a collaboration-related modality disentanglement module that explicitly separates common and specific modality features. It employs four modal disentanglement encoders and a collaborative signal-guided gated network to remove task-irrelevant noise. This module ensures that the learned representations are devoid of redundant and noisy signals, providing pure representations for subsequent preference learning. To address **C2**, we perform multi-view graph representation learning with disentangled multimodal features to capture item-item dependency and user-item interaction patterns. For graph-augmented representations of users and items, we propose a hierarchical fusion strategy to fuse disentangled multimodal features and ID embeddings. Specifically, we employ a designed positive-negative attention-aware features fusion to fuse disentangled multimodal features. Then we implement an interaction distribution-based fusion to align multimodal features with ID embeddings. We conduct comprehensive experiments on three benchmarks to demonstrate the effectiveness of DHMRec.

Related Work

Multimodal Recommendation

Multimodal recommendation aims to leverage diverse modality information to enhance recommendation performance. Early approaches typically extend collaborative filtering (CF) by incorporating modality features and ID embeddings. For example, VBPR (He and McAuley 2016b) combines visual features with ID embeddings to alleviate the cold-start problem. With the growing success of Graph Neural Networks (GNNs) in capturing high-order relations, methods such as MMGCN (Wei et al. 2019) and GRCN (Wei et al. 2020) model users' modal preferences through interaction graphs and graph denoising techniques. Subsequent works (Zhang et al. 2021; Wang et al. 2021; Zhou et al. 2023a) introduce homogeneous graph learning to capture item-item or user-user co-occurrence. Recent studies, including FREEDOM (Zhou and Shen 2023) and TMLP (Huang et al. 2025), focus on enhancing representation robustness via structural pruning, denoising. However, existing methods often overlook fine-grained user-specific modality preferences and lack adaptive fusion strategies aligned with individual interests.

Disentanglement Learning

Disentanglement learning aims to decompose complex representations into independent latent factors, facilitating more effective and interpretable feature extraction. Inspired by its success in domains like computer vision, recent studies have introduced disentanglement techniques into recommender systems. For instance, ADDVAE (Tran and Lauw 2022) learns disentangled user representations from textual data, while DMRL (Liu et al. 2022) separates modality-independent factors and captures users' modality preferences via a shared-attention mechanism. To enhance interpretability, KDR (Mu et al. 2021) incorporates knowledge graphs to guide the disentanglement process, and AD-DRL (Li et al. 2024) fuses multimodal attributes to refine factor-specific representations. Although disentangling multimodal factors enhances robustness, most existing methods fail to eliminate redundancy and irrelevant noise, thus impeding full utilization of cross-modal complementary signals.

Methodology

This section presents the DHMRec architecture and its components. The overall framework is shown in Figure 2.

Problem Definition

Let $\mathcal{U}$ and $\mathcal{I}$ denote the user and item sets. The implicit-feedback matrix $\boldsymbol{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$ indicates interactions. The ID embeddings for users and items are represented as $\boldsymbol{E}_{id} = \{\boldsymbol{E}_{u,id} || \boldsymbol{E}_{i,id}\} \in \mathbb{R}^{d_k \times (|\mathcal{U}|+|\mathcal{I}|)}$. For each

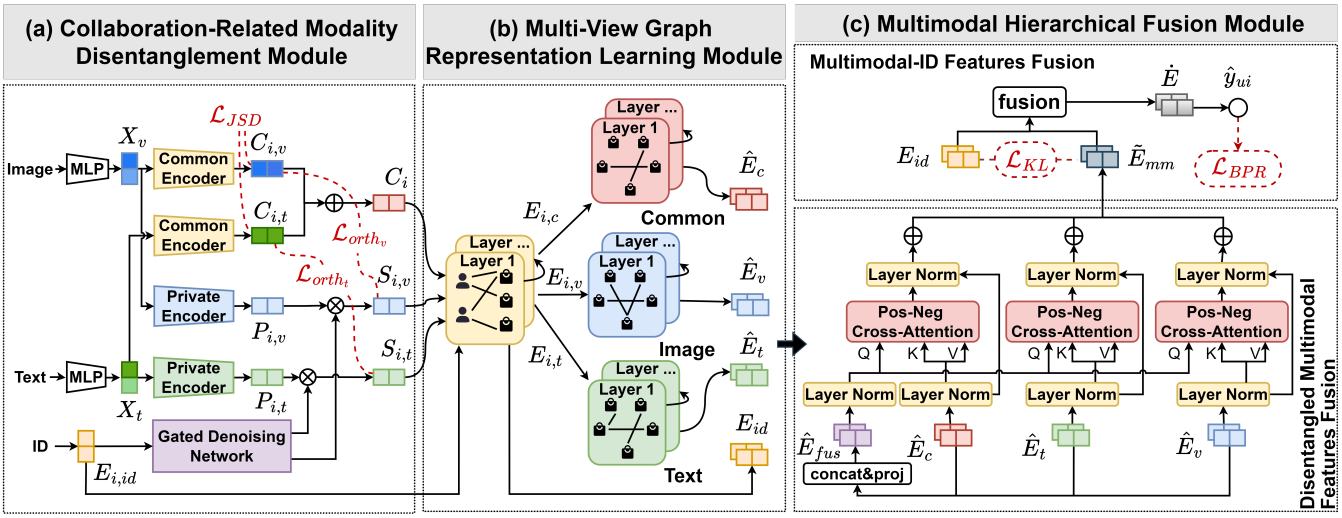


Figure 2: Overall framework of DHMRec. It consists of three major components: (a) Collaboration-related modality disentanglement module; (b) Multi-view graph representation learning module; and (c) Multimodal hierarchical fusion module.

modality $m \in \mathcal{M} = \{t, v\}$, pre-extracted item features are $\boldsymbol{X}_m \in \mathbb{R}^{d_m \times |\mathcal{I}|}$. Given $\boldsymbol{R}$ and $\boldsymbol{X}_m$, the MRS predicts preference scores $\hat{y}_{ui}$ and returns top-k items per user.

Modality Disentanglement Module

This module disentangles common and specific modality features to remove redundancy and task-irrelevant noise.

Disentangled Encoder The raw multimodal features for each item are obtained from pre-trained networks with different representational dimensions. We first project them into the same representation space.

$$\boldsymbol{Z}_m = \boldsymbol{W}_1^m \boldsymbol{X}_m, \quad m \in \{t, v\}, \tag{1}$$

where $\boldsymbol{W}_1^m \in \mathbb{R}^{d_k \times d_m}$ transforms each original multimodal features uniformly from d_m-dimension to d_k-dimension.

For each modality, we employ a common encoder and a private encoder to extract modality-common and modality-private representations.

$$\boldsymbol{C}_{i,m} = E_{C_m}\left(\boldsymbol{Z}_m; \theta_{C_m}\right), \boldsymbol{P}_{i,m} = E_{P_m}\left(\boldsymbol{Z}_m; \theta_{P_m}\right), \tag{2}$$

where $\boldsymbol{C}_{i,m}, \boldsymbol{P}_{i,m} \in \mathbb{R}^{d_k \times |\mathcal{I}|}$. Both the common encoders $E_{C_m}(\cdot; \theta_{C_m})$ and the private encoders $E_{P_m}(\cdot; \theta_{P_m})$ are implemented as simple linear network and learn the parameters for each modality.

Collaborative Signal-Guided Denoising To avoid noise interference, inspired by MGCN(Yu et al. 2023), we design a collaborative signal-guided gated denoising network. Since noise tends to appear in private features, we input the modality-private representations $\boldsymbol{P}_{i,m}$ into a gated network. This network is guided by ID embeddings containing rich collaborative information to identify modality-specific feature representations relevant to the recommendation task.

$$\boldsymbol{S}_{i,m} = f_{gate}^m\left(\boldsymbol{P}_{i,m}, \boldsymbol{E}_{i,id}\right) = \boldsymbol{P}_{i,m} \odot \sigma\left(\boldsymbol{W}_2^m \boldsymbol{E}_{i,id} + \boldsymbol{b}_2^m\right), \tag{3}$$

where $\boldsymbol{W}_2^m \in \mathbb{R}^{d_k \times d_k}$ and $\boldsymbol{b}_2^m \in \mathbb{R}^{d_k}$ denote the trainable parameters, $\boldsymbol{E}_{i,id} \in \mathbb{R}^{d_k \times |\mathcal{I}|}$ represents ID embeddings of items, $m \in \{t, v\}$ denotes the modality, σ is the sigmoid funtion and $\odot$ represents the element-wise product.

Modality-Common Representations Alignment Jensen-Shannon Divergence (JSD) provides a symmetric measure for distribution similarity, addressing the bias inherent in KL Divergence. Inspired by previous work (Zheng et al. 2016; Bachman, Alsharif, and Precup 2014; Kannan, Kurakin, and Goodfellow 2018; Hendrycks et al. 2019), to treat the modality-common representations equally, we minimize their distributional differences via the JSD.

$$\mathcal{L}_{JSD} = \frac{1}{2}\left(KL\left(\boldsymbol{P} \parallel \boldsymbol{M}\right) + KL\left(\boldsymbol{Q} \parallel \boldsymbol{M}\right)\right), \tag{4}$$

where $\boldsymbol{P} = \sigma(\boldsymbol{C}_{i,t})$, $\boldsymbol{Q} = \sigma(\boldsymbol{C}_{i,v})$ denote sigmoid function, $\boldsymbol{M} = (\boldsymbol{P} + \boldsymbol{Q})/2$ denotes the mixture of $\boldsymbol{P}$ and $\boldsymbol{Q}$.

To preserve distributional properties, we fuse modality-common features via the logit pooling (Yao et al. 2024):

$$\begin{aligned} \boldsymbol{C}_i &= \sigma^{-1}\left(\frac{\sigma(\boldsymbol{C}_{i,t}) + \sigma(\boldsymbol{C}_{i,v})}{2}\right) \\ &= \log \frac{2\exp(\boldsymbol{C}_{i,t} + \boldsymbol{C}_{i,v}) + \exp(\boldsymbol{C}_{i,t}) + \exp(\boldsymbol{C}_{i,v})}{2 + \exp(\boldsymbol{C}_{i,t}) + \exp(\boldsymbol{C}_{i,v})}. \end{aligned} \tag{5}$$

Disentangled Constraint We employ orthogonal loss by minimizing the cosine similarity between modality-common features $\boldsymbol{C}_{i,m}$ and modality-specific features $\boldsymbol{S}_{i,m}$, encouraging them to capture distinct modal aspects:

$$\mathcal{L}_{orth_m} = \frac{|\langle \boldsymbol{C}_{i,m}, \boldsymbol{S}_{i,m} \rangle|}{\|\boldsymbol{C}_{i,m}\|_2 \cdot \|\boldsymbol{S}_{i,m}\|_2}, \quad m \in \{t, v\}, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between vectors.

Multi-View Graph Learning Module

This module captures item-item dependency and user-item interaction patterns by graph representation learning.

Denoising User-Item Graph We construct four user-item graphs $\mathcal{G} = \{\mathcal{G}_{\hat{m}} \mid \mathcal{G}_{id}, \mathcal{G}_v, \mathcal{G}_t, \mathcal{G}_c\}$ sharing identical structure. All graphs have node set $\mathcal{V} \subseteq \{\mathcal{U}, \mathcal{I}\}$, with adjacency matrix $\boldsymbol{A}^{\hat{m}} = \begin{vmatrix} 0 & \boldsymbol{R} \\ \boldsymbol{R}^T & 0 \end{vmatrix}$. We use the strategy called degree-sensitive edge pruning of FREEDOM (Zhou and Shen 2023) to alleviate over-smoothing in $\mathcal{G}_{id}$. The sampling probability for each edge is $p_k = 1/(\sqrt{w_i}\sqrt{w_j})$, where w_i and w_j are the degrees of nodes i and j. Then, each graph performs multi-hop propagation aggregation:

$$\boldsymbol{E}_{\hat{m}}^{(l)} = \left(\left(\boldsymbol{D}^{\boldsymbol{A}^{\hat{m}}}\right)^{-\frac{1}{2}} \boldsymbol{A}^{\hat{m}} \left(\boldsymbol{D}^{\boldsymbol{A}^{\hat{m}}}\right)^{-\frac{1}{2}}\right) \boldsymbol{E}_{\hat{m}}^{(l-1)}, \quad (7)$$

where $\boldsymbol{D}^{\boldsymbol{A}^{\hat{m}}}$ is the diagonal degree matix of $\boldsymbol{A}^{\hat{m}}$, $\boldsymbol{E}_{\hat{m}}^{(l)}$ represents the embeddings of modality $\hat{m}$ for users and items at the l-th layer of graph convolution. The multimodal embeddings of users $\boldsymbol{E}_{u,\hat{m}}^{(l)}$ are initialized by aggregating from interacted items, while the items $\boldsymbol{E}_{i,\hat{m}}^{(l)}$ are initialized by disentangled representations, as $\boldsymbol{E}_{i,c}^{(0)} = \boldsymbol{C}_i$, $\boldsymbol{E}_{i,v}^{(0)} = \boldsymbol{S}_{i,v}$, $\boldsymbol{E}_{i,t}^{(0)} = \boldsymbol{S}_{i,t}$.

By integrating multi-layer neighbors, we derive multimodal feature representations $\boldsymbol{E}_{\hat{m}}$ capturing collaborative information.

$$\boldsymbol{E}_{\hat{m}} = \frac{1}{L+1} \sum_{i=0}^{L} \boldsymbol{E}_{\hat{m}}^{(l)}, \quad \hat{m} \in \{id, t, v, c\}. \quad (8)$$

Modality Item-Item Graph To capture the potential item semantics, we employ the KNN algorithm (Chen, Fang, and Saad 2009) to construct item-item graphs under the modality-common, text-specific and image-specific views. The edge weight $I_{i,i'}^m$ is computed as cosine similarity of raw multimodal features.

$$I_{i,i'}^m = \frac{\left(\boldsymbol{x}_i^m\right)^\top \boldsymbol{x}_{i'}^m}{\|\boldsymbol{x}_i^m\| \|\boldsymbol{x}_{i'}^m\|}, \quad m \in \{t, v\}. \quad (9)$$

We select only the top-k neighbors that exhibit the highest similarity scores to capture the most relevant features.

$$\hat{\boldsymbol{I}}_{i,i'}^m = \begin{cases} 1 & I_{i,i'}^m \in top\text{-}k(\boldsymbol{I}_i^m) \\ 0 & otherwise \end{cases}. \quad (10)$$

Then, we normalize the adjacency matrix addressing the problems of gradient explosion or vanishing.

$$\tilde{\boldsymbol{I}}^m = \left(\boldsymbol{D}^{\hat{\boldsymbol{I}}^m}\right)^{-\frac{1}{2}} \hat{\boldsymbol{I}}^m \left(\boldsymbol{D}^{\hat{\boldsymbol{I}}^m}\right)^{-\frac{1}{2}}, \quad (11)$$

where $\boldsymbol{D}^{\hat{\boldsymbol{I}}^m}$ is the diagonal degree matrix of $\hat{\boldsymbol{I}}^m$. The modality-common graph is constructed by weighted summation of normalized textual and visual similarity matrices:

$$\tilde{\boldsymbol{I}}^c = \sum_{m \in \mathcal{M}} \beta_m \tilde{\boldsymbol{I}}^m, \quad (12)$$

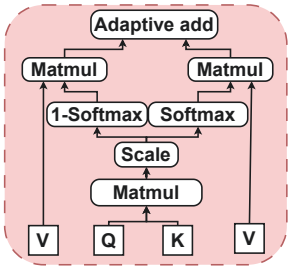


Figure 3: The diagram of the pos-neg attention module.

where β_m weights modality m. We aggregate multilayer neighbors to capture modality-specific item similarities:

$$\hat{\boldsymbol{E}}_{i,\tilde{m}}^{(l)} = \tilde{\boldsymbol{I}}^{\tilde{m}} \hat{\boldsymbol{E}}_{i,\tilde{m}}^{(l-1)}, \quad \tilde{m} \in \{t, v, c\}, \quad (13)$$

where $\hat{\boldsymbol{E}}_{i,\tilde{m}}^{(0)}$ is initialized by the representations $\boldsymbol{E}_{i,\hat{m}}$ obtained after the U-I graph learning, and we take the l-th layer as the final representations of the items, as $\hat{\boldsymbol{E}}_{i,\tilde{m}} = \hat{\boldsymbol{E}}_{i,\tilde{m}}^{(l)}$.

Finally, we obtain users' modality features by combining the modality features of the interacted items.

$$\hat{\boldsymbol{E}}_{u,\tilde{m}} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \hat{\boldsymbol{E}}_{i,\tilde{m}}. \quad (14)$$

Multimodal Hierarchical Fusion Module

This module aims to fuse disentangled multimodal features and ID embeddings through the hierarchical fusion strategy.

Disentangled Modality Features Fusion We propose a positive-negative (pos-neg) attention module, shown in Figure 3. Negative branch captures informative yet low-similarity cues that standard self-attention overlooks.

We first concatenate the modality-common and modality-specific features obtained after graph-based encoding as $\hat{\boldsymbol{E}} = \left[\hat{\boldsymbol{E}}_c, \hat{\boldsymbol{E}}_t, \hat{\boldsymbol{E}}_v\right] \in \mathbb{R}^{3d_k \times (|\mathcal{U}|+|\mathcal{I}|)}$, and project it into d_k-dimension space: $\hat{\boldsymbol{E}}_{fus} = \boldsymbol{W}_3 \hat{\boldsymbol{E}}$, where $\boldsymbol{W}_3 \in \mathbb{R}^{d_k \times 3d_k}$. For each modality view $\tilde{m} \in \{c, v, t\}$, the query $\boldsymbol{Q}_{\tilde{m}}$ is derived from $\hat{\boldsymbol{E}}_{fus}$, while the key $\boldsymbol{K}_{\tilde{m}}$ and the value $\boldsymbol{V}_{\tilde{m}}$ are computed from $\hat{\boldsymbol{E}}_{\tilde{m}}$. Specifically, $\boldsymbol{Q}_{\tilde{m}} = \boldsymbol{W}_q^{\tilde{m}} \hat{\boldsymbol{E}}_{fus}$, $\boldsymbol{K}_{\tilde{m}} = \boldsymbol{W}_k^{\tilde{m}} \hat{\boldsymbol{E}}_{\tilde{m}}$, $\boldsymbol{V}_{\tilde{m}} = \boldsymbol{W}_v^{\tilde{m}} \hat{\boldsymbol{E}}_{\tilde{m}}$, where $\boldsymbol{W}_q^{\tilde{m}}, \boldsymbol{W}_k^{\tilde{m}}, \boldsymbol{W}_v^{\tilde{m}} \in \mathbb{R}^{d_k \times d_k}$. Then we get the positive and negative attention-aware feature representations $\boldsymbol{O}_{\tilde{m}}^{pos}$, $\boldsymbol{O}_{\tilde{m}}^{neg}$ by calculating the positive and negative correlation between the query and the key under each dimension.

$$\boldsymbol{O}_{\tilde{m}}^{pos} = \text{SoftMax}\left(\frac{\boldsymbol{Q}_{\tilde{m}} \boldsymbol{K}_{\tilde{m}}{}^T}{\sqrt{N}}\right) \boldsymbol{V}_{\tilde{m}}, \quad (15)$$

$$\boldsymbol{O}_{\tilde{m}}^{neg} = \left(1 - \text{SoftMax}\left(\frac{\boldsymbol{Q}_{\tilde{m}} \boldsymbol{K}_{\tilde{m}}{}^T}{\sqrt{N}}\right)\right) \boldsymbol{V}_{\tilde{m}}, \quad (16)$$

where $N = |\mathcal{U}| + |\mathcal{I}|$ represents the number of users and items. The obtained positive and negative dual-branch outputs are dynamically summed by fusion parameters.

$$\boldsymbol{O}_{\tilde{m}} = \alpha_{pos} \boldsymbol{O}_{\tilde{m}}^{pos} + \alpha_{neg} \boldsymbol{O}_{\tilde{m}}^{neg}, \quad (17)$$

where integration weights $\alpha_{pos}, \alpha_{neg}$ are learnable dynamic parameters. To avoid gradient vanishing, we employ a residual connection around the attention fusion layer, followed by layer normalization.

$$\tilde{\boldsymbol{E}}_{\tilde{m}} = LN\left(LN\left(\boldsymbol{W}_o^{\tilde{m}}\boldsymbol{O}_{\tilde{m}}\right) + \hat{\boldsymbol{E}}_{\tilde{m}}\right) \in \mathbb{R}^{d_k \times (|\mathcal{U}|+|\mathcal{I}|)}, \tag{18}$$

where $\boldsymbol{W}_o^{\tilde{m}} \in \mathbb{R}^{d_k \times d_k}$ and $LN()$ represents layer normalization. The final multimodal fusion features are obtained by summing all modal perspectives.

$$\tilde{\boldsymbol{E}}_{mm} = \sum_{\tilde{m}} \tilde{\boldsymbol{E}}_{\tilde{m}}, \quad \tilde{m} \in \{c, v, t\}. \tag{19}$$

Multimodal-ID Features Fusion To ensure consistency between multimodal fusion and collaborative signals, we align the preference distributions from multimodal and ID embeddings by computing interaction distributions via inner products of user and item embeddings:

$$\boldsymbol{P}_{mm} = \text{SoftMax}\left((\tilde{\boldsymbol{E}}_{u,mm})^T \tilde{\boldsymbol{E}}_{i,mm}\right), \tag{20}$$

where $\tilde{\boldsymbol{E}}_{u,mm}$ and $\tilde{\boldsymbol{E}}_{i,mm}$ are user and item embeddings from the multimodal fusion features. Similarly, $\boldsymbol{P}_{id}$ is derived from $\boldsymbol{E}_{id}$. Align two distributions via KL divergence:

$$\mathcal{L}_{KL} = KL\left(\boldsymbol{P}_{mm} \parallel \boldsymbol{P}_{id}\right). \tag{21}$$

The final user and item embeddings are weighted by the multimodal fusion embeddings and ID embeddings.

$$\dot{\boldsymbol{E}} = \boldsymbol{E}_{id} + \alpha_{mm}\tilde{\boldsymbol{E}}_{mm}, \tag{22}$$

where $\dot{\boldsymbol{E}} \in \mathbb{R}^{d_k \times (|\mathcal{U}|+|\mathcal{I}|)}$ represents the final embeddings and α_{mm} is a hyperparameter that controls the fusion strength of the multimodal embeddings.

Modal Prediction & Optimization

We obtain the users' prediction scores by computing the inner product of the final embeddings of the users and items:

$$\hat{y}_{ui} = (\dot{\boldsymbol{e}}_u)^T \dot{\boldsymbol{e}}_i. \tag{23}$$

We adopt Bayesian Personalized Ranking (BPR) (Rendle et al. 2012) as the recommendation optimization loss:

$$\mathcal{L}_{BPR} = \sum_{(u,i,j)\in\mathcal{D}_{train}} -\ln \sigma(\hat{y}_{ui} - \hat{y}_{uj}), \tag{24}$$

where (u, i, j) denotes a training triplet with i as a positive item and j as a randomly sampled negative item.

The overall objective integrates BPR loss, disentangled loss, KL aligning loss and L_2 regularization:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{BPR} + \gamma_{dis}(\mathcal{L}_{JSD} + \mathcal{L}_{orth_v} + \mathcal{L}_{orch_t}) \\ + \gamma_{kl}\mathcal{L}_{KL} + \gamma_{reg}||\theta||^2\end{aligned}, \tag{25}$$

where $\gamma_{dis}, \gamma_{kl}, \gamma_{reg}$ are weights of loss term.

Experiment

Experimental Settings

Datasets We conduct experiments on three widely used Amazon dataset (He and McAuley 2016a): Baby, Sports, and Clothing. Additional dataset processing details are in Extended version.

Baselines We compare DHMRec with two types of SOTA recommendation methods. (1) General CF Models: **BPR** (Rendle et al. 2012), **LightGCN** (He et al. 2020), **LayerGCN** (Zhou et al. 2023b). (2) Multimodal models: **VBPR** (He and McAuley 2016b), **MMGCN** (Wei et al. 2019), **GRCN** (Wei et al. 2020), **DualGNN** (Wang et al. 2021), **LATTICE** (Zhang et al. 2021), **MICRO** (Zhang et al. 2022), **BM3** (Zhou et al. 2023c), **FREEDOM** (Zhou and Shen 2023), **LGMRec** (Guo et al. 2024), **DA-MRS** (Xv et al. 2024), **DiffMM** (Jiang et al. 2024), **TMLP** (Huang et al. 2025), **CMDL** (Lin et al. 2025).

Evaluation Protocols To ensure fair comparison, we utilize two commonly adopted metrics: Recall@K (R@K) and NDCG@K (N@K). We report the average metrics of all users in the test dateset under both K = 10 and K = 20. Following the standard evaluation protocol outlined in (Zhou and Shen 2023), we apply an 8:1:1 random split of the dataset for training, validation, and testing.

Implementation Details We implement our proposed DHMRec with MMRec (Zhou 2023). For the general settings, we set the embedding sizes for both users and items at 64, initializing the modal parameters with the Xavier method (Glorot and Bengio 2010), and employ the Adam optimizer (Kingma 2014) for model training. To achieve a fair evaluation, we perform a complete grid search for each baseline method following its published paper to find the optimal setting. For our proposed DHMRec, we perform the grid search on hyperparameters γ_{dis} and γ_{kl} in $\{1e^{-3}, 1e^{-2}, 1e^{-1}, 0.5, 1\}$, γ_{reg} in $\{0, 1e^{-6}, 1e^{-5}, ..., 1e^{-1}\}$, α_{mm} in $\{0, 0.1, 0.2, ..., 1.5\}$. The k for top-k in the construction of item-item graph sets at 10, and β_m is taken to be 0.1. We use Recall@20 on the validation data as the training-stopping criterion and fix the early stopping epoch at 20.

Performance Comparison

The performance comparison of all methods on the three datasets is summarized in Table 1. We have the following key observations: (1) **DHMRec model achieved outstanding performance across multiple metrics.** In R@20, it achieves average improvements of 44.67% (Baby), 125.94% (Clothing), and 41.60% (Sports) over traditional models, and 17.93%, 39.50% and 19.86% over multimodal models. Even compared to the best baseline, DHMRec achieved 2.63%, 1.09% and 2.44% performance improvements on three datasets in R@20 respectively. We attribute the performance gains to redundancy reduction via modality disentanglement and adaptive modality weighting through pos-neg attention. (2) **Multimodal features significantly enhance accuracy.** Compared to the best traditional method (LayerGCN), all multimodal methods achieved average performance improvements of 32.64% (R@10), 32.17% (R@20), 32.11% (N@10), and 31.99% (N@20) on Clothing. This highlights their ability to alleviate data sparsity by enriching representations with multimodal data. (3) **Homogeneous graph learning further enhances performance.** FREEDOM and MICRO benefit from adding homogeneous item graph. The improvement of DHMRec further emphasizes

Datasets	Baby				Clothing				Sports			
Metrics	**R@10**	**R@20**	**N@10**	**N@20**	**R@10**	**R@20**	**N@10**	**N@20**	**R@10**	**R@20**	**N@10**	**N@20**
BPR	0.0357	0.0575	0.0192	0.0249	0.0187	0.0279	0.0103	0.0126	0.0432	0.0653	0.0241	0.0298
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0340	0.0526	0.0188	0.0236	0.0569	0.0864	0.0311	0.0387
LayerGCN	0.0516	0.0829	0.0279	0.0359	0.0372	0.0557	0.0202	0.0249	0.0609	0.0940	0.0330	0.0415
VBPR	0.0423	0.0663	0.0223	0.0285	0.0281	0.0410	0.0157	0.0190	0.0561	0.0855	0.0307	0.0383
MMGCN	0.0409	0.0671	0.0220	0.0287	0.0231	0.0365	0.0121	0.0155	0.0389	0.0632	0.0202	0.0264
GRCN	0.0523	0.0840	0.0286	0.0367	0.0438	0.0665	0.0230	0.0288	0.0599	0.0910	0.0322	0.0403
DualGNN	0.0508	0.0808	0.0277	0.0354	0.0474	0.0698	0.0254	0.0311	0.0590	0.0898	0.0324	0.0403
LATTICE	0.0537	0.0845	0.0287	0.0366	0.0468	0.0690	0.0256	0.0312	0.0624	0.0955	0.0337	0.0422
MICRO	0.0575	0.0894	0.0320	0.0404	0.0520	0.0769	0.0284	0.0347	0.0666	0.0991	0.0368	0.0454
BM3	0.0574	0.0876	0.0300	0.0378	0.0424	0.0637	0.0230	0.0284	0.0639	0.0981	0.0347	0.0435
FREEDOM	0.0619	0.0990	0.0325	0.0420	0.0615	0.0921	0.0331	0.0409	0.0723	0.1087	0.0387	0.0481
LGMRec	0.0633	0.0976	**0.0347**	0.0435	0.0556	0.0822	0.0300	0.0368	0.0724	0.1083	0.0395	0.0488
DA-MRS	0.0619	0.0959	0.0335	0.0423	0.0621	0.0916	0.0336	0.0412	0.0703	0.1079	0.0382	0.0479
DiffMM	0.0596	0.0940	0.0311	0.0396	0.0489	0.0740	0.0262	0.0325	0.0673	0.1014	0.0369	0.0453
TMLP	0.0644	0.0978	0.0344	0.0430	0.0602	0.0910	0.0330	0.0408	0.0721	0.1105	0.0390	0.0489
CMDL	0.0598	0.0947	0.0318	0.0408	0.0560	0.0833	0.0305	0.0375	0.0611	0.0940	0.0330	0.0415
DHMRec	**0.0646***	**0.1016***	0.0343*	**0.0436***	**0.0629***	**0.0931***	**0.0340***	**0.0417***	**0.0738***	**0.1132***	**0.0401***	**0.0502***

Table 1: Performance comparison of different recommendation models. The best and the second results are in boldface and underlined, respectively. * denotes that the improvement is significant with $p < 0.05$ based on paired t-test.

that modality disentanglement prior to graph learning removes redundancy, enhancing key features capture.

Datasets	Baby		Sports		Clothing	
Metrics	R@20	N@20	R@20	N@20	R@20	N@20
DHMRec	0.1016	0.0436	0.1132	0.0502	0.0931	0.0417
w/o DIS	0.1016	0.0445	0.1111	0.0495	0.0901	0.0402
w/o AF	0.0980	0.0425	0.1033	0.0455	0.0824	0.0365
w/o PNF	0.0992	0.0434	0.1121	0.0502	0.0865	0.0388
w/o KL	0.0986	0.0430	0.1096	0.0487	0.0874	0.0392
w/o II	0.0887	0.0389	0.1012	0.0462	0.0698	0.0314

Table 2: Ablation of different components on DHMRec.

Ablation Study

We evaluate key components via five DHMRec variants. Table 2 shows: (1) **w/o DIS** removes the multimodal disentanglement module. The performance drop confirms the presence of redundant and noisy features in raw multimodal features. (2) **w/o AF** adopts averaging instead of attention-aware fusion. DHMRec outperforms this variant by 3.67%, 9.58%, and 12.99% in R@20 on three datasets respectively, demonstrating that our attention-aware fusion effectively suppresses irrelevant modalities and emphasizes those aligned with user preferences. (3) **w/o PNF** replaces pos-neg attention with standard attention. The results show that our pos-neg design better captures complementary signals often ignored by traditional attention, especially on Clothing, where it yields a 7.47% gain in N@20. (4) **w/o KL** removes the KL loss for multimodal-ID features alignment, confirming its role in enhancing fusion consistency. (5) **w/o II** removes the I-I gragh learning, validating the importance of modeling latent inter-item similarities to mitigate sparsity.

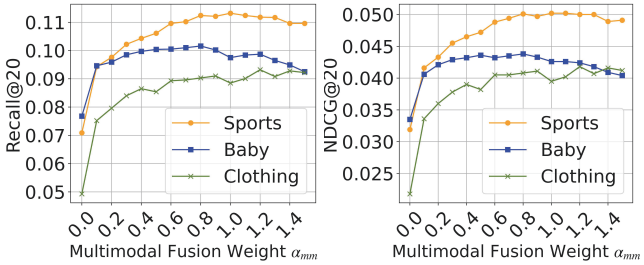
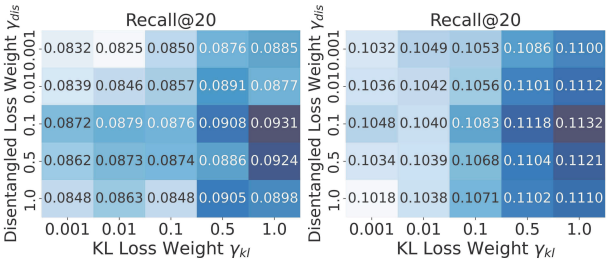


Figure 4: Performance of DHMRec with different α_{mm}.

Impact of Hyperparameters

The Balancing Hyperparameter α_{mm} Figure 4 shows the impact of the fusion weight α_{mm} on DHMRec's performance. Performance initially increases with α_{mm}, plateaus at optimal values. All datasets exhibit minimal performance at $\alpha_{mm} = 0$ (ID embedding-only prediction), confirming the necessity of multimodal information. Furthermore, the optimal α_{mm} varies by dataset: Baby peaks at 0.8, Sports at 1.0, and the sparser Clothing dataset at 1.2. This demonstrates the adaptability of our fusion strategy in balancing multimodal signals under different data sparsity levels.

The Pair of Hyperparameters γ_{dis} and γ_{kl} Figure 5 presents DHMRec's performance under different combinations of γ_{dis} and γ_{kl} on Clothing and Sports. The best results on all datasets are achieved at $\gamma_{dis} = 0.1$ and $\gamma_{kl} = 1$. We analyze that small γ_{dis} may insufficiently disentangle coupled modality information, while too large can hinder the optimization of the main recommendation task.

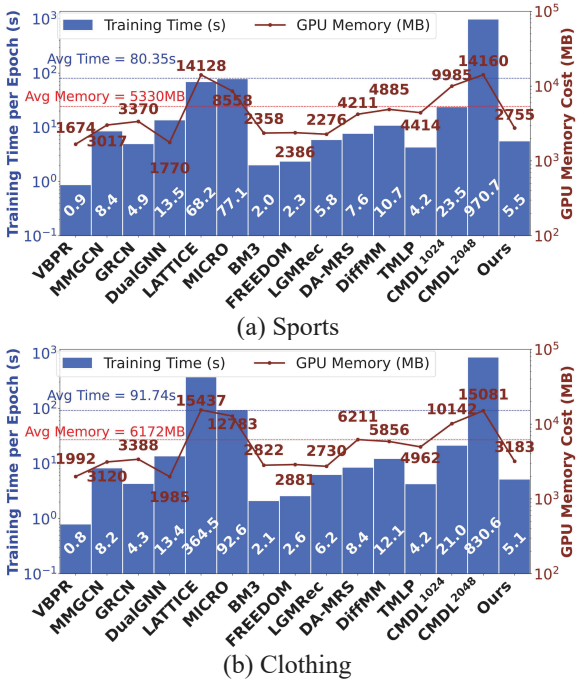


(a) Clothing (b) Sports

Figure 5: Performance of DHMRec with respect to different pairs of γ_{dis} and γ_{kl} on Clothing and Sports in Recall@20.

Complexity Analysis

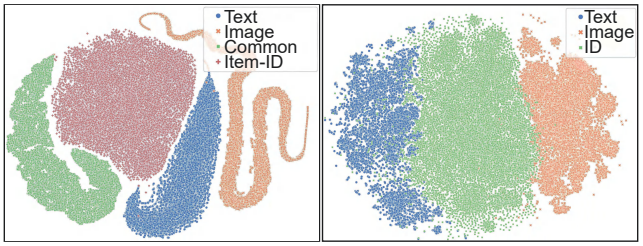
We report the average training time per epoch and GPU memory cost for multimodal models, trained on an NVIDIA GeForce RTX 4080 with a batch size of 2048. Under this setting, CMDL (CMDL^{2048}) nearly exhausts GPU memory, potentially increasing runtime. To control for this, we evaluate CMDL with a reduced batch size of 1024 (CMDL^{1024}). Efficiency results on Sports and Clothing are shown in Figure 6. The result shows that DHMRec achieves strong accuracy with moderate computational cost, maintaining below-average running time and GPU memory cost among baselines. Notably, DHMRec surpasses DualGNN (the graph-based model with minimal GPU cost) by 26.06% and 33.38% in R@20 on Sports and Clothing, with only ≈1GB additional memory. Compared to other disentangled methods, ours also exhibits superior training efficiency. Even if CMDL reduces its batch size, its GPU cost remains higher than ours. Theory analysis is provided in Extended version.



(a) Sports

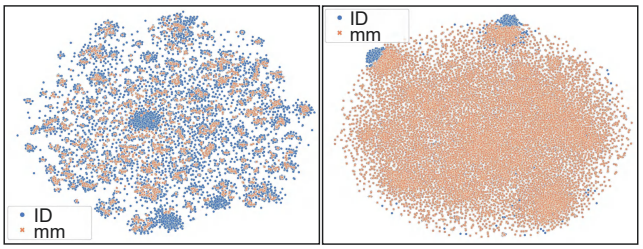
(b) Clothing

Figure 6: Training Time and GPU Memory Cost.



(a) DHMRec (b) w/o DIS

Figure 7: Comparison of multimodal and ID embedding distributions with and without disentanglement on Clothing.



(a) DHMRec (b) w/o II

Figure 8: Comparison of fused multimodal and ID embedding distributions with and without multi-view graph.

Visualization

To further verify the effectiveness of disentanglement learning, we visualize the modality-common, modality-specific, and ID embeddings of items on the Clothing with t-SNE in Figure 7(a), and compare them with non-disentangled embeddings in Figure 7(b). The result shows the distinct separation through the disentanglement learning our designed. We further visualize the final multimodal and ID embeddings on Clothing. As shown in Figure 8(b), features without multi-view learning are scattered and overlapping. In contrast, Figure 8(a) presents a cluster-like distribution, indicating that our multi-view graph learning effectively captures intra-modal semantics and complementary cross-modal information, leading to enhanced recommendation performance.

Conclusion and Future Work

In this work, we propose DHMRec, a novel multimodal recommendation model, which addresses the challenges of intra-modal features disentanglement and inter-modal hierarchical fusion. Specifically, we introduce a collaboration-related multimodal disentanglement module to independently mine the latent semantic relations of items. Furthermore, we implement hierarchical fusion to dynamically merge disentangled multimodal featurs. Extensive experiments conducted on three benchmarks demonstrate the effectiveness of our method. For future work, we aim to explore more advanced techniques for modeling the interplay between different modalities and incorporate additional data sources to further enhance recommendation performance.

Acknowledgements

This research is supported, in part, by the National Natural Science Foundation of China (No. 62376135, No. 62406176); the China Postdoctoral Science Foundation under Grant Number 2024M751810; the Shandong Provincial Natural Science Foundation (No. ZR2024QF091). This research is also administered by Shandong Key Laboratory of Artificial Intelligence Application for Livelihood Services (No. PKL2024A45) and Shandong Higher Education Institution Laboratory of General Artificial Intelligence for Future Industries.

References

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27.

Chen, J.; Fang, H.-r.; and Saad, Y. 2009. Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *Journal of Machine Learning Research*, 10(9).

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; and Ruan, B. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8454–8462.

He, R.; and McAuley, J. 2016a. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, 507–517.

He, R.; and McAuley, J. 2016b. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. 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*, 639–648.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*.

Huang, J.; Qin, J.; Yu, Y.; and Zhang, W. 2025. Beyond Graph Convolution: Multimodal Recommendation with Topology-aware MLPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11808–11816.

Jiang, Y.; Xia, L.; Wei, W.; Luo, D.; Lin, K.; and Huang, C. 2024. Diffmm: Multi-modal diffusion model for recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7591–7599.

Kannan, H.; Kurakin, A.; and Goodfellow, I. 2018. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, Z.; Liu, F.; Wei, Y.; Cheng, Z.; Nie, L.; and Kankanhalli, M. 2024. Attribute-driven Disentangled Representation Learning for Multimodal Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9660–9669.

Lin, X.; Liu, R.; Cao, Y.; Zou, L.; Li, Q.; Wu, Y.; Liu, Y.; Yin, D.; and Xu, G. 2025. Contrastive Modality-Disentangled Learning for Multimodal Recommendation. *ACM Transactions on Information Systems*.

Liu, F.; Chen, H.; Cheng, Z.; Liu, A.; Nie, L.; and Kankanhalli, M. 2022. Disentangled multimodal representation learning for recommendation. *IEEE Transactions on Multimedia*, 25: 7149–7159.

Mu, S.; Li, Y.; Zhao, W. X.; Li, S.; and Wen, J.-R. 2021. Knowledge-guided disentangled representation learning for recommender systems. *ACM Transactions on Information Systems (TOIS)*, 40(1): 1–26.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Tran, N.-T.; and Lauw, H. W. 2022. Aligning dual disentangled user representations from ratings and textual content. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1798–1806.

Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25: 1074–1084.

Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T.-S. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*, 3541–3549.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, 1437–1445.

Xv, G.; Li, X.; Xie, R.; Lin, C.; Liu, C.; Xia, F.; Kang, Z.; and Lin, L. 2024. 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*, 3645–3656.

Yao, W.; Yin, K.; Cheung, W. K.; Liu, J.; and Qin, J. 2024. DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16416–16424.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B.-K. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6576–6585.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, 3872–3880.

Zhang, J.; Zhu, Y.; Liu, Q.; Zhang, M.; Wu, S.; and Wang, L. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 9154–9167.

Zheng, S.; Song, Y.; Leung, T.; and Goodfellow, I. 2016. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 4480–4488.

Zhou, H.; Zhou, X.; Zhang, L.; and Shen, Z. 2023a. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In *ECAI 2023*, 3123–3130. Amsterdam, Netherlands: IOS Press.

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

Zhou, X.; Lin, D.; Liu, Y.; and Miao, C. 2023b. Layer-refined graph convolutional networks for recommendation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 1247–1259. IEEE.

Zhou, X.; and Shen, Z. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 935–943.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023c. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, 845–854.