Behavior Conditional Diffusion Model for Multi-Modal Recommendation

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

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

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

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

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

Abstract

Multi-modal recommenders (MRs) focus on leveraging the item modality features to facilitate user preferences modeling. Previous research mainly suffers from two limitations: (1) The pre-trained modality features are usually extracted by the encoders trained on general tasks (e.g., text classification), and thus inevitably contain the recommendation-irrelevant features. (2) Existing modality fusion mechanisms often diminish the contribution of features from weaker modalities, leading to biased fused representations. To address these challenges, we propose a novel **B**ehavior **C**onditional **D**ifussion model for **M**ulti-**M**odal recommendation (BCDMM). Specifically, we first design a Behavior Multi-modal Diffusion (BMD) module to filter the recommendation-irrelevant noise within the pre-trained modality features. Then, we iteratively denoise the modality features with the guidance of user behavior signals to reconstruct the recommendation-related features. Next, we apply a Multi-modal Graph Fusion (MGF) module to explore the item modality latent structures. Moreover, we construct a modality fusion graph to capture the cross-modal complementary features for comprehensively modeling user preferences. Finally, a set of adversarial loss functions is used to balance the preservation of modality-specific and modality-shared features. Extensive experiments on three real-world datasets demonstrate the superiority of our method. We release our code at https://github.com/fanko79/BCDMM2025.

*Corresponding author.

MMAsia '25, Kuala Lumpur, Malaysia

ACM ISBN 979-8-4007-2005-5/25/12
https://doi.org/10.1145/3743093.3770970

CCS Concepts

• **Information systems** → *Recommender systems.*

Keywords

Multi-Modal Recommendation, Diffusion Model, Graph Neural Network

ACM Reference Format:
Mengfan Kong, Chonghao Chen, Zhiqiang Pan, Fei Cai, and Aimin Luo. 2025. Behavior Conditional Diffusion Model for Multi-Modal Recommendation. In *ACM Multimedia Asia (MMAsia '25), December 09–12, 2025, Kuala Lumpur, Malaysia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3743093.3770970

1 Introduction

Recommender systems primarily rely on the user-item interaction data to model user preferences. However, historical interactions are often highly sparse in real-world scenarios, which significantly limits the model ability to learn effective user and item representations. With the rapid development of online multimedia platforms, items are increasingly associated with diverse multi-modal features, e.g., image, text, and audio. Such multi-modal contents provide additional perspectives that can reflect user preferences.

Within the traditional recommendation paradigm, the central focus of multi-modal recommender systems (MRs) is how to efficiently integrate multi-modal features to learn the representations of users and items. Early research [5] incorporates visual features into the matrix factorization models to uncover the latent dimensions that influence user preferences. Subsequently, the attention mechanism is utilized to disentangle user preferences across different modalities [2, 9, 17]. Moreover, inspired by the success of graph neural networks, graph-based methods [10, 21, 28, 29] leverage graph convolutional networks (GCNs) to propagate modality features and explore higher-order connections between users and items. More recently, several methods [16, 20, 22, 30] explore the integration of self-supervised learning framework into MRs. These

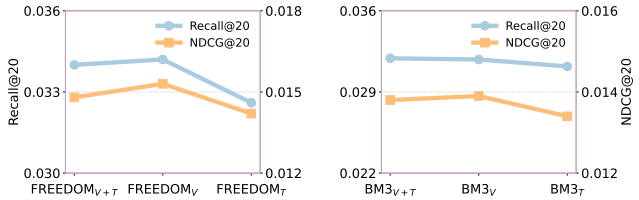


Figure 1: The performance comparison of different variants of FREEDOM and BM3 on TMALL. "V" indicates using the visual modality, while "T" denotes using the textual modality.

approaches utilize contrastive learning techniques to align features across different modalities and produce robust item representations.

Despite the notable progress achieved by recent methods, existing MRs are still hindered by two core issues: (1) **The inherent noise exists in the pre-trained modality features.** The modality features of items are typically generated by the pre-trained encoders [4, 11, 13, 14] trained on general tasks. Hence, they inevitably contain the recommendation-irrelevant features that can mislead user preferences modeling [3]. For instance, two visually distinct clothing items may be mistakenly regarded as similar due to their shared background or image brightness features [24]. Such features captured by the pre-trained models are the recommendation-irrelevant and redundant. (2) **Existing modality fusion mechanisms tend to suppress features from weaker modalities, resulting in biased fused representations.** As shown in Fig. 1, we compare the performance of BM3 and FREEDOM [1] with incorporating different modality-related features, i.e., text and image, text only, and image only. Interestingly, we observe a counter-intuitive performance ranking: FREEDOM$_V$ > FREEDOM$_{V+T}$ > FREEDOM$_T$, BM3$_V$ ≈ BM3$_{V+T}$ > BM3$_T$. That is, the corresponding model using only the visual modality achieves comparable or even better performance than that leveraging multiple modalities. We argue that this phenomenon is caused by the visual modality dominating the modality fusion process, leading to the features from the weaker textual modality being overlooked.

To deal with the above-mentined limitations in MRs, we propose a novel multi-modal recommendation method, termed **B**ehavior **C**onditional **D**iffusion model for **M**ulti-**M**odal recommendation (**BCDMM**). First, to filter out the recommendation-irrelevant features, we propose a behavior multi-modal diffusion module (BMD). Specifically, we progressively inject Gaussian noise into the modality embeddings in the forward process. Then, in the reverse process, we introduce user behavior signals from the interaction graph as guidance information, which explicitly steer the generation of the recommendation-related modality features. Then, to address the issue of biased fusion during multi-modal feature integration, we propose a Multi-Modal Graph Fusion (MGF) module, which applies graph convolutions to both modality-specific graphs and the modality-fusion graph. This allows the model to integrate features from different modalities, preventing any single modality from becoming dominant. Finally, we design a simple yet effective loss function to balance the preservation of modality-specific features and the alignment of modality-shared features. We conduct extensive experiments demonstrating that BCDMM outperforms various competitive baselines, with average gains of 3.07% in Recall@20 and 3.57% in NDCG@20, respectively.

[1]The results are produced according to the public resources from https://github.com/enoche/BM3 and https://github.com/enoche/MMRec

2 APPROACH

In this section, we introduce the proposed DAHM model in detail. The overall framework of our model is shown in Fig. 2.

2.1 PRELIMINARIES

We first provide a formal formulation of the multi-modal recommendation task. Specifically, let $\mathcal{U}$ represent the set of users and $\mathcal{I}$ denote the set of items, respectively. The historical interactions between users and items can be described as a bipartite graph $\mathcal{G} = (\mathcal{U}, \mathcal{I}, \mathcal{E})$, where each edge $(u, i) \in \mathcal{E}$ indicates an observed interaction. These interactions are recorded by an adjacency matrix $\mathbf{R} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $\mathbf{R}_{(u,i)} = 1$ if user u has interacted with item i, and $\mathbf{R}_{(u,i)} = 0$, otherwise. Moreover, we represent the ID embedding matrices of users and items as $\mathbf{E}_u^{id} \in \mathbb{R}^{|\mathcal{U}|\times d}$ and $\mathbf{E}_i^{id} \in \mathbb{R}^{|\mathcal{I}|\times d}$, respectively, where d is the embedding dimension. Let $\mathbf{e}_u^{id}$ and $\mathbf{e}_i^{id}$ denote the ID embedding vectors corresponding to user u and item i, obtained by indexing into $\mathbf{E}_u^{id}$ and $\mathbf{E}_i^{id}$, respectively.

Additionally, each item contains features from both visual and textual modalities, which we denote as $\mathbf{e}_i^m \in \mathbb{R}^{d_m}$. Here, let $M = \{v, t\}$ denote the set of modalities, corresponding to the visual and textual modalities, respectively. We use $m \in M$ to index each modality. Similarly, the modality feature matrix of all items are represented as $\mathbf{E}_i^m \in \mathbb{R}^{|\mathcal{I}|\times d_m}$. The objective of multi-modal recommendation is to learn a function f to predict the potential interactions $\hat{y}_{u,i}$, which can be formulated as $\hat{y}_{u,i} = f(\mathcal{G}, \mathbf{E}_u^{id}, \mathbf{E}_i^{id}, \mathbf{E}_i^m)$.

2.2 Behavior Multi-Modal Diffusion

In this section, we detail BMD, our designed diffusion-based method that incorporates user behavior signals to generate the recommendation-related modality features.

2.2.1 Forward Process. We first align the dimensions of the pre-trained modality features $\mathbf{E}_i^m$ through linear transformations: $\mathbf{E}_i^m = \mathbf{W}_{1,m}\mathbf{E}_i^m$, where $\mathbf{W}_{1,m} \in \mathbb{R}^{d\times d_m}$ are learnable parameters. Subsequently, in the forward process, we define the initial state of the modality m as $\mathbf{X}_0^m = [\mathbf{E}_u^m || \mathbf{E}_i^m]$. Specially, we generate user modality feature matrix $\mathbf{E}_u^m$ by aggregating the features of the items they have interacted with:

$$\mathbf{e}_u^m = \sum_{i\in\mathcal{N}_u} \frac{1}{|\mathcal{N}_u|}\mathbf{e}_i^m, \tag{1}$$

where $\mathcal{N}_u$ denotes the interacted item set of user u.

After that, we consider a single vector $\mathbf{x}_0^m$ as a row sampled from $\mathbf{X}_0^m$, and inject Gaussian noise into it step by step. The same process is applied to all rows. At each sampled diffusion timestep t, the noisy modality embeddings $\mathbf{x}_t^m$ are generated as follows:

$$q(\mathbf{x}_t^m|\mathbf{x}_{t-1}^m) = \mathcal{N}\left(\mathbf{x}_t^m; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}^m, \beta_t\mathbf{I}\right), \tag{2}$$

where $t \in \{1, \ldots, T\}$ is the diffusion timestep, $\beta_t \in (0, 1)$ is the variance schedule controlling the noise injection, and $\mathcal{N}$ is the Gaussian distribution. Moreover, $\sqrt{1-\beta_t}\,\mathbf{x}_{t-1}^m$ represents the mean, and $\beta_t\mathbf{I}$ is the variance term with isotropic Gaussian noise.

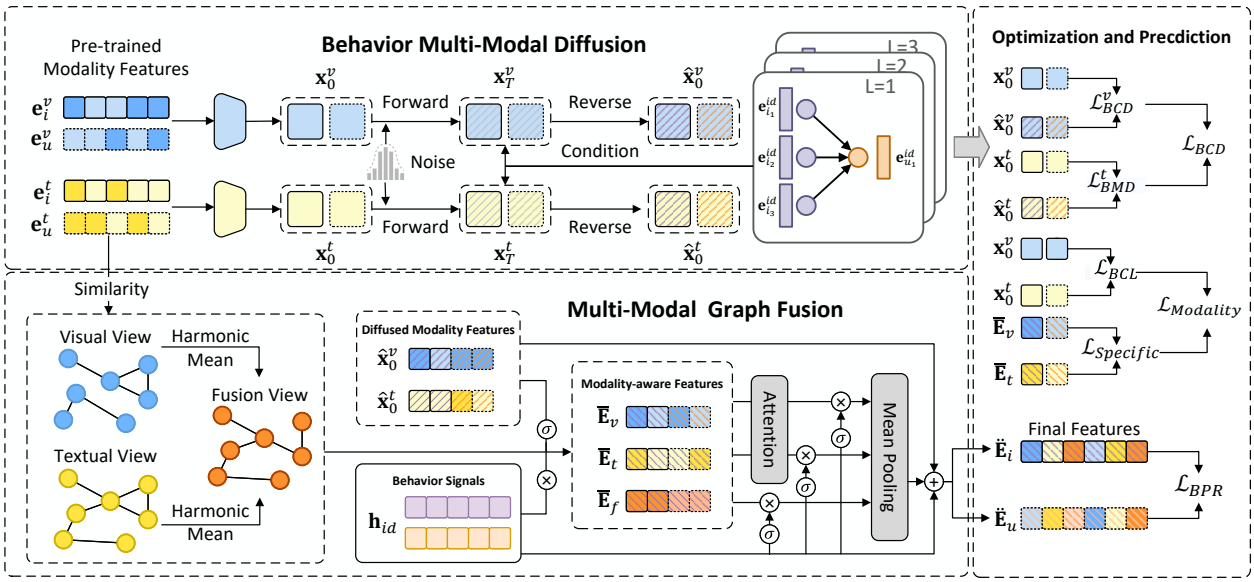


Figure 2: The framework of BCDMM.

In practice, we derive $\mathbf{x}_t^m$ from $\mathbf{x}_0^m$ using the reparameterization [7] and the additivity property of independent Gaussian noises:

$$q(\mathbf{x}_t^m|\mathbf{x}_0^m) = \mathcal{N}\left(\mathbf{x}_t^m; \sqrt{\bar{\alpha}_t}\mathbf{x}_0^m, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{t'=1}^{t} \alpha_{t'}$. Then, we derive that $\mathbf{x}_t^m = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0^m + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Next, we adopt a shared variance schedule β_t to ensure training stability across modalities:

$$\beta_t = [\beta_{min} + \frac{t-1}{T-1}(\beta_{max} - \beta_{min})], t \in \{1, \ldots, T\}, \tag{4}$$

where β_{max} and β_{min} are hyperparameters controlling the minimum and maximum noise levels at each diffusion step, respectively. Therefore, at the final diffusion step $t = T$, we obtain the noised modality features $\mathbf{x}_T^m$ as the result of the forward process.

2.2.2 Reverse Process. After getting the corrupted modality features $\mathbf{x}_T^m$, we iteratively filter out the recommendation-irrelevant features in the reverse process, which can be formulated as:

$$p_\theta(\hat{\mathbf{x}}_{t-1}^m|\hat{\mathbf{x}}_t^m) = \mathcal{N}(\hat{\mathbf{x}}_{t-1}^m; \mu_\theta(\hat{\mathbf{x}}_t^m, t), \Sigma_\theta(\hat{\mathbf{x}}_t^m, t)), \tag{5}$$

where $\hat{\mathbf{x}}_t^m$ is the denoised modality features in the reverse step t, $\mu_\theta(\hat{\mathbf{x}}_t^m, t)$ and $\Sigma_\theta(\hat{\mathbf{x}}_t^m, t))$ are the mean and covariance. θ is the learnable parameters of the neural network to predict the reverse distribution. Then, we execute this process iteratively for T steps to generate the final recommendation-related features $\hat{\mathbf{x}}_0^m$. Next, we provide a detailed description of the modality denoising module and the behavior condition used to guide the reverse process.

• **Modality Denoising Module.** As illustrated in Eq. (5), the main focus of the reverse process is to obtain reliable $\mu_\theta(\hat{\mathbf{x}}_t^m, t)$ and $\Sigma_\theta(\hat{\mathbf{x}}_t^m, t))$ for guiding the denoising process. To achieve this, we align the distribution $p_\theta(\hat{\mathbf{x}}_{t-1}^m|\hat{\mathbf{x}}_t^m)$ with the posterior distribution $q(\mathbf{x}_{t-1}^m|\mathbf{x}_t^m, \mathbf{x}_0^m)$. Naturally, we can use $q(\mathbf{x}_{t-1}^m|\mathbf{x}_t^m, \mathbf{x}_0^m)$ to constrain $p_\theta(\hat{\mathbf{x}}_{t-1}^m|\hat{\mathbf{x}}_t^m)$. This process is derived from Bayes' theorem:

$$q(\mathbf{x}_{t-1}^m|\mathbf{x}_t^m, \mathbf{x}_0^m) = \mathcal{N}(\mathbf{x}_{t-1}^m, \widetilde{\mu}_t(\mathbf{x}_t^m, \mathbf{x}_0^m), \widetilde{\beta}_t I), \quad \text{where}$$
$$\begin{cases} \widetilde{\mu}_t(\mathbf{x}_t^m, \mathbf{x}_0^m) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t^m + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\mathbf{x}_0^m \\ \widetilde{\beta}_t = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \end{cases}, \tag{6}$$

where $\widetilde{\mu}_t(\mathbf{x}_t^m, \mathbf{x}_0^m)$ and $\widetilde{\beta}_t$ are the mean and covariance of $q(\mathbf{x}_{t-1}^m|\mathbf{x}_t^m, \mathbf{x}_0^m)$, respectively. After obtaining the $q(\mathbf{x}_{t-1}^m|\mathbf{x}_t^m, \mathbf{x}_0)$, we can similarly factorize $p_\theta(\hat{\mathbf{x}}_{t-1}^m|\hat{\mathbf{x}}_t^m)$:

$$p_\theta(\hat{\mathbf{x}}_{t-1}^m|\hat{\mathbf{x}}_t^m) = \mathcal{N}(\hat{\mathbf{x}}_{t-1}^m; \mu_\theta(\hat{\mathbf{x}}_t^m, t), \widetilde{\beta}_t I), \quad \text{where}$$
$$\mu_\theta(\hat{\mathbf{x}}_t^m, t) = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{x}}_t^m - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} f_\theta(\hat{\mathbf{x}}_t^m, t)\right). \tag{7}$$

Here, $p_\theta(\hat{\mathbf{x}}_{t-1}^m|\hat{\mathbf{x}}_t^m)$ serves as a parameterized approximation of Eq. (6). Since the true $\widetilde{\mu}_t(\mathbf{x}_t^m, \mathbf{x}_0^m)$ depends on the $\mathbf{x}_0^m$, which is unknown in the reverse process, we construct $f_\theta(\cdot)$ by the multi-layer perceptron(MLP) to guide the generation of the recommendation-related modality features $\hat{\mathbf{x}}_0^m$.

• **Behavior Condition.** As shown in Eq. (7), we use the sinusoidal function in [19] to encode the timestep t into a high-dimensional vector $\mathbf{s}_t \in \mathbb{R}^{d_t}$. However, our goal is not to reconstruct the original pre-trained features $\mathbf{x}_0^m$ [25]. Instead, we aim to only recover the recommendation-relevant parts that truly reflect user preferences. Therefore, we incorporate user behavior signals $\mathbf{H}_{id}$ as an additional conditional signal. As a result, Eq. (7) can be rewritten as:

$$\mu_\theta(\hat{\mathbf{x}}_t^m, t) = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{x}}_t^m - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} f_\theta(\hat{\mathbf{x}}_t^m, \mathbf{h}_{id}, \mathbf{s}_t)\right). \tag{8}$$

Here, $\mathbf{h}_{id}$ denotes the row in $\mathbf{H}_{id}$ corresponding to the given index. Specifically, $\mathbf{H}_{id}$ is obtained from the ID-corresponding user

interaction graph via a simplified graph neural network [6]:

$$\begin{aligned}\mathbf{H}_{id} &= \frac{1}{L+1}\sum_{l=0}^{L}\mathbf{E}_l^{id},\\ \mathbf{E}_l^{id} &= \left(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{E}_{l-1}^{id},\end{aligned} \tag{9}$$

where $\mathbf{A} = \begin{bmatrix} 0 & \mathbf{R} \\ \mathbf{R}^T & 0 \end{bmatrix}$, $\mathbf{D}$ is the degree matrix of $\mathbf{A}$, and L denotes the number of propagation layers. The initial matrix $\mathbf{E}_0^{id} = [\mathbf{E}_u^{id}||\mathbf{E}_i^{id}] \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times d}$ is the concatenation of user and item ID embedding matrices. Thus, $\mathbf{H}_{id}$ provides strong preference signals derived from user-item interactions, which steers the generation of the recommendation-relevant features.

2.2.3 Reconstruction Optimization. The optimization of the diffusion model focuses on minimizing the distribution difference between $q(\mathbf{x}_{t-1}^m|\mathbf{x}_t^m, \mathbf{x}_0^m)$ and $p_\theta(\hat{\mathbf{x}}_{t-1}^m|\hat{\mathbf{x}}_t^m)$. Therefore, the evidence lower bound of the $\hat{\mathbf{x}}_0^m$ can be optimized via KL divergence. In detail, we refer to this objective as the behavior condition diffusion loss $\mathcal{L}_{BCD}$, which can be formulated as follows:

$$\mathcal{L}_{BCD} = \frac{1}{|M|}\sum_{m\in M}^{|M|}\mathbb{E}_q[||\mathbf{x}_0^m - f_\theta(\hat{\mathbf{x}}_1^m, \mathbf{h}_{id}, \mathbf{s}_1)||_2^2], \tag{10}$$

where $|M|$ denotes the number of modalities. Therefore, we obtain the denoised modality features $\hat{\mathbf{x}}_0^m = f_\theta(\hat{\mathbf{x}}_1^m, \mathbf{h}_{\text{id}}, \mathbf{s}_1)$, and $\hat{\mathbf{X}}_0^m \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times d}$ is constructed by stacking all $\hat{\mathbf{x}}_0^m$ vectors as rows.

However, such an objective $\mathcal{L}_{BCD}$ may not be appropriate, since $\mathbf{x}_0^m$ inevitably contains the recommendation-irrelevant noise. Our goal is to push $\hat{\mathbf{x}}_0^m$ not to the noisy $\mathbf{x}_0^m$, but to an ideal version $\bar{\mathbf{x}}_0^m$ that better reflects user-intent semantics. Therefore, we approximate $\bar{\mathbf{x}}_0^m$ by contrasting different modality views that are grounded in the same user-item interaction context. Specifically, we refer to this objective as the behavior-aligned contrastive loss $\mathcal{L}_{BCL}$, which can be computed as follows:

$$\begin{aligned}&\mathcal{L}_{BCL} = \mathcal{L}_{BCL}^u + \mathcal{L}_{BCL}^i, \quad \text{where}\\ &\begin{cases}\mathcal{L}_{BCL}^u = -\sum_{m\in M}^{|M|}\log\frac{\exp(\text{sim}(\mathbf{Z}_u^m, \mathbf{Z}_u^{m'})/\tau)}{\sum_{v\in\mathcal{U}}\exp(\text{sim}(\mathbf{Z}_u^m, \mathbf{Z}_u^{m'})/\tau)}\\ \mathcal{L}_{BCL}^i = -\sum_{m\in M}^{|M|}\log\frac{\exp(\text{sim}(\mathbf{Z}_i^m, \mathbf{Z}_i^{m'})/\tau)}{\sum_{v\in\mathcal{I}}\exp(\text{sim}(\mathbf{Z}_i^m, \mathbf{Z}_i^{m'})/\tau)}\\ \mathbf{Z}_u^m = \mathbf{A}_{u,*}\mathbf{E}_u^{id},\ \mathbf{Z}_i^m = \mathbf{A}_{*,i}\mathbf{E}_i^m,\end{cases}\end{aligned} \tag{11}$$

where $\mathbf{A}$ is the adjacency matrix in Eq. (9), $\mathbf{A}u, *$ and $\mathbf{A}*, i$ denote its first $|\mathcal{U}|$ and last $|\mathcal{I}|$ rows, respectively, $\text{sim}(\cdot)$ is the similarity function, and τ is a hyperparameter. Thus, $\mathcal{L}_{BCL}$ provides structural supervision for $\mathbf{x}_0^m$ by maximizing modality-shared features.

2.3 Multi-Modal Graph Fusion

In this section, we explicitly feed the denoised modality features into the multi-modal graph fusion (MGF) module to generate final user and item representations.

2.3.1 Multi-Modal Graph Construction. Following [26, 27], we build the modality-specific item-item affinity graphs $\mathbf{S}_m \in \mathbf{R}^{|\mathcal{I}|\times|\mathcal{I}|}$:

$$\mathbf{S}_{i,j}^m = \frac{(\mathbf{e}_i^m)^T\mathbf{e}_j^m}{||\mathbf{e}_i^m||\,||\mathbf{e}_j^m||}. \tag{12}$$

Here, $\mathbf{e}_i^m$ and $\mathbf{e}_j^m$ denote the i-th and j-th rows of items modality feature matrix $\mathbf{E}_i^m$, respectively. Moreover, following [26, 29], we perform kNN sparsification [1] to prune unimportant edges:

$$\ddot{\mathbf{S}}_{i,j}^m = \begin{cases}\mathbf{S}_{i,j}^m, & \mathbf{S}_{i,j}^m \in \text{top-}k_m(\mathbf{S}_{i,*}^m),\\ 0, & \text{otherwise.}\end{cases} \tag{13}$$

Here, for each row of the modality-specific matrix $\mathbf{S}^m$, we retain the top-k_m largest values and set the remaining entries to zero.

Moreover, to mitigate the issue of dominant modalities suppressing weaker ones, we construct a modality fusion graph that adaptively balances the contribution of modality-specific graphs using a harmonic mean. In detail, the process is formulated as follows:

$$\ddot{\mathbf{S}}_{i,j}^f = \frac{|M|}{\sum_{m\in M}^{|M|}\frac{1}{\ddot{\mathbf{S}}_{i,j}^m}}. \tag{14}$$

Therefore, $\ddot{\mathbf{S}}_{i,j}^f$ preserves the inherent characteristics of each modality while preventing any single one from dominating the others.

2.3.2 Multi-Modal Graph Convolution. After building the modality-specific and fusion graphs, we distill the recommendation related features $\hat{\mathbf{x}}_0^m$ with the guidance of ID embeddings as follows[24]:

$$\begin{aligned}\mathbf{E}_{i,m} &= f_{gate}^m(\mathbf{E}_i^{id}, \hat{\mathbf{X}}_{0,i}^m) = \mathbf{E}_i^{id}\odot\sigma(\mathbf{W}_{2,m}\hat{\mathbf{X}}_{0,i}^m + \mathbf{b}_{2,m}),\\ \mathbf{E}_{i,f} &= f_{gate}^f(\mathbf{E}_i^{id}, \hat{\mathbf{X}}_{0,i}^f) = \mathbf{E}_i^{id}\odot\sigma(\mathbf{W}_{3,f}\hat{\mathbf{X}}_{0,i}^f + \mathbf{b}_{3,f}),\\ \hat{\mathbf{X}}_{0,i}^f &= \sum_{m\in M}^{|M|}\gamma_m\hat{\mathbf{X}}_{0,i}^m.\end{aligned} \tag{15}$$

Here, $\mathbf{W}_* \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_* \in \mathbb{R}^d$ are learnable parameters, and σ is the sigmoid nonlinearity. Then, $\hat{\mathbf{X}}_{0,i}^m$ denotes the last $|\mathcal{I}|$ rows of $\hat{\mathbf{X}}_0^m$, and γ_m is a set of learnable parameterized vectors. Next, we employ a aggregation operation on $\mathbf{E}_{i,m}$ and $\mathbf{E}_{i,f}$ as follows:

$$\bar{\mathbf{E}}_{i,m} = \ddot{\mathbf{S}}_m\mathbf{E}_{i,m}, \quad \bar{\mathbf{E}}_{i,f} = \ddot{\mathbf{S}}_f\mathbf{E}_{i,f}. \tag{16}$$

Next, we can obtain the user modality features $\bar{\mathbf{e}}_{u,m}$ and $\bar{\mathbf{e}}_{u,f}$ by aggregating neighborhood item modality features as follows:

$$\bar{\mathbf{e}}_{u,m} = \sum_{i\in\mathcal{N}_u}\frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}}\bar{\mathbf{e}}_{i,m}, \quad \bar{\mathbf{e}}_{u,f} = \sum_{i\in\mathcal{N}_u}\frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}}\bar{\mathbf{e}}_{i,f}, \tag{17}$$

where $\mathcal{N}u$ and $\mathcal{N}i$ denote the neighborhoods of u and i, respectively. We stack $\bar{\mathbf{e}}_{u,m}$ and $\bar{\mathbf{e}}_{u,f}$ row-wise to obtain $\bar{\mathbf{E}}_{u,m}$ and $\bar{\mathbf{E}}_{u,f}$, and then form $\bar{\mathbf{E}}_m = [\bar{\mathbf{E}}_{u,m}||\bar{\mathbf{E}}_{i,m}]$ and $\bar{\mathbf{E}}_f = [\bar{\mathbf{E}}_{u,f}||\bar{\mathbf{E}}_{i,f}]$. Further, we distill modality features using behavior signals $\mathbf{H}_{id}$ via a gate function:

$$\begin{aligned}\mathbf{P}_m &= \sigma(\mathbf{W}_{4,m}\mathbf{H}_{id} + \mathbf{b}_{4,m}),\\ \mathbf{P}_f &= \sigma(\mathbf{W}_{5,f}\mathbf{H}_{id} + \mathbf{b}_{5,f}),\end{aligned} \tag{18}$$

where $\mathbf{W}_* \in \mathbb{R}^{d\times d}$ and $\mathbf{b} \in \mathbb{R}^d$ are learnable parameters, and σ denotes the sigmoid nonlinearity. Additionally, we compute the weight scores of $\bar{\mathbf{E}}_m$ through attention mechanism [19, 24]:

$$\begin{aligned}\bar{\mathbf{E}}_m &= \sum_{m\in M}^{|M|}\alpha_m\bar{\mathbf{E}}_m\\ \alpha_m &= \text{Softmax}(\mathbf{q}_m^{\mathrm{T}}\tanh(\mathbf{W}_{4,m}\bar{\mathbf{E}}_m + \mathbf{b}_{5,m})),\end{aligned} \tag{19}$$

Table 1: Statistics of experimented datasets.

Dataset	Baby	Sports	TMALL
User	19445	35598	13104
Item	7050	18357	7848
Interaction	160792	23033	151928
Density	0.117%	0.045%	0.148%

where $\mathbf{q}_m \in \mathbb{R}^d$ is the attention vector. Finally, we obtain the overall multi-modal features $\ddot{\mathbf{E}}_m$ as follows:

$$\ddot{\mathbf{E}}_m = \frac{1}{|M|} \sum_{m \in M}^{|M|} \left(\bar{\mathbf{E}}_m \odot \mathbf{P}_m\right) + \left(\bar{\mathbf{E}}_f \odot \mathbf{P}_f\right) + \mathbf{H}_{id}. \tag{20}$$

2.3.3 Modality-specific Optimization. Before the final fusion stage, we maximizes the difference between modalities by the modality-specific loss $\mathcal{L}_{Specific}$, which can be formulated as:

$$\mathcal{L}_{Specific} = -||\delta \odot |\bar{\mathbf{E}}_m - \bar{\mathbf{E}}_{m'}|||, \quad m \neq m', \tag{21}$$

where δ is a node-level learnable parameter that adaptively controls the contribution of each modality pair.

2.4 Optimization and Prediction

Based on the denoising diffusion features $\hat{\mathbf{X}}_0^m$ from the BDM module and the multi-modal features $\ddot{\mathbf{E}}_m$ from the MGF module, we form the final representations of users and items as follows:

$$\ddot{\mathbf{E}}_u = \ddot{\mathbf{E}}_{u,m} + \alpha \hat{\mathbf{X}}_{0,u}^m, \quad \ddot{\mathbf{E}}_i = \ddot{\mathbf{E}}_{i,m} + \alpha \hat{\mathbf{X}}_{0,i}^m, \tag{22}$$

where α is a hyperparameter. Next, to strike an approximate balance between modality-specific features and modality-shared features, we introduce an adversarial optimization objective by combining $\mathcal{L}_{Specific}$ and $\mathcal{L}_{BCL}$:

$$\mathcal{L}_{Modality} = \lambda_1 \mathcal{L}_{Specific} + \lambda_2 \mathcal{L}_{BCL}, \tag{23}$$

where λ_1 and λ_2 are hyperparameters. Specifically, $\mathcal{L}_{Modality}$ introduces a soft adversarial regularization that achieves a better trade-off between modality-specific and modality-shared features. Next, we use the inner product to predict the likelihood of interaction, i.e., $\hat{y} = \ddot{\mathbf{e}}_u \cdot \ddot{\mathbf{e}}_i^\top$, where $\ddot{\mathbf{e}}_u \in \ddot{\mathbf{E}}_u$ and $\ddot{\mathbf{e}}_i \in \ddot{\mathbf{E}}_i$. Moreover, we employ the BPR Loss [15] as follows:

$$\mathcal{L}_{BPR} = \sum_{(u,i^+,i^-)} -\log\left(\text{Sigmoid}\left(\hat{y}_{u,i^+} - \hat{y}_{u,i^-}\right)\right), \tag{24}$$

where i^+ and i^- denote the positive and negative samples for user u, respectively. The final loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{BPR} + \mathcal{L}_{Modality} + \lambda_3 \mathcal{L}_{BCD} + \lambda_4 \|\Theta\|_2^2, \tag{25}$$

where λ_3 and λ_4 are hyperparameters. Additionally, the last term $\|\Theta\|_2^2$ represents the weight decay regularization against overfitting.

3 Experiments

We validate the effectiveness of our proposed BCDMM by addressing the following research questions: 1) **RQ1**: How dose the performance of BCDMM compare with various competitive baselines? 2) **RQ2**: What is the contribution of different components within the BCDMM? 3) **RQ3**: What is the impact of the recommendation-related modality features generated by BMD?

3.1 Experimental Settings

3.1.1 Datasets. Experiments are conducted on three real-world multi-modal recommendation datasets: Baby, Sports and TMALL. The items in these datasets contain both textual and visual modality features. The statistics of three datasets are summarized in Table 1.

3.1.2 Baselines Methods. To comprehensively evaluate the superiority of our proposed BCDMM, we compared it with various competitive baselines from different research lines, including: (1) GCN-based General Recommender: LightGCN [6]; (2) GCN-based Multi-modal Recommenders: MMGCN [23], LATTICE [26], BM3[30], FREEDOM[29], MGCN [24] and LGMRec [28]; (3) Diffusion-based Mulitimodal Recommenders: MCDRec [12] and DiffMM [8].

3.1.3 Evalution Protocols & Parameter Settings. Due to the length of the article limitation, the details are shown in the released code.

3.2 Overall performance(RQ1)

Overall comparison results between BCDMM and the baselines are summarized in Table 2, from which we find some key observations:

(1) **The leverage of multi-modal features can improve performance in most cases.** In general, multi-modal methods outperform recommenders that solely rely on interaction data (e.g., LightGCN), indicating that the introduction of item modality features can boost performance. However, some GCN-based methods, such as MMGCN, underperform compared to LightGCN, especially on the TMALL dataset. These results indicate the existence of the recommendation-irrelevant features in pre-trained modality features. Although FREEDOM and LGMRec reduce modality noise by mining item-modality latent connections and modeling user global interests, their performance remains inferior to our proposed BCDMM, as they do not directly denoise modality features.

(2) **The learning of modality fusion view has always been overlooked.** Specifically, most GCN-based methods carefully construct modality-specific graphs to explore modality-specific user preferences, which to some extent overlooks the learning of cross-modal complementary features. In multi-modal recommendation scenarios, modeling user preferences requires integrating multiple modalities for a more comprehensive understanding. Although there are methods like DiffMM that capture such cross-modal user preferences through contrastive learning, they still lack direct learning of the modality fusion view. In contrast, our proposed BCDMM constructs a modality fusion graph through a harmonic mean operation, directly mining the high-order cross-modal connections.

(3) **BCDMM outperforms all competitive baselines in three datasets in terms of all metrics.** Specifically, compared to the best-performing baseline, BCDMM achieves improvements of 3.40%, 2.44%, and 3.36% in Recall@20 on Baby, Sports, and TMALL, respectively, and improvements of 2.97%, 2.62%, and 5.11% in NDCG@20.

3.3 Ablation Study (RQ2)

In this section, we validate the effectiveness of key components in BCDMM. The results are presented in Table 3.

(1) The w/o BMD variant removes the behavioral multi-modal diffusion module, leading to a notable performance drop compared to BCDMM, highlighting the need to filter recommendation-irrelevant noise from pre-trained modality features. Prior works apply graph

Table 2: Overall performances of BCDMM and other baselines on three datasets in terms of Recall@10, Recall@20, NDCG@10 and NDCG@20. Our method and the best-performed baseline are highlighted with bold and underlined, respectively.

Models	Baby				Sports				TMALL			
	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	0.0463	0.0750	0.0253	0.0327	0.0557	0.0853	0.0303	0.0379	0.0233	0.0349	0.0130	0.0162
MMGCN	0.0413	0.0649	0.0217	0.0276	0.0382	0.0625	0.0200	0.0263	0.0094	0.0159	0.0046	0.0064
LATTICE	0.0550	0.0848	0.0292	0.0369	0.0651	0.0966	0.0354	0.0436	0.0234	0.0350	0.0136	0.0168
BM3	0.0531	0.0859	0.0283	0.0367	0.0638	0.0967	0.0347	0.0432	0.0191	0.0311	0.0105	0.0137
FREEDOM	0.0631	0.1000	0.0333	0.0428	0.0724	0.1096	0.0394	0.0489	0.0212	0.0340	0.0113	0.0148
MGCN	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0252	0.0387	0.0139	0.0176
LGMRec	0.0643	0.0987	0.0349	0.0438	0.0724	0.1075	0.0392	0.0483	0.0252	0.0381	0.0136	0.0171
MCDRec	0.0556	0.0871	0.0299	0.0381	0.0621	0.0957	0.0341	0.0428	0.0198	0.0322	0.0105	0.0140
DiffMM	0.0602	0.0932	0.0324	0.0408	0.0703	0.1076	0.0381	0.0477	0.0229	0.0368	0.0118	0.0156
BCDMM	**0.0661**	**0.1034**	**0.0355**	**0.0451**	**0.0756**	**0.1133**	**0.0412**	**0.0509**	**0.0267**	**0.0400**	**0.0148**	**0.0185**
Improv.	2.72%	3.40%	1.72%	2.97%	3.70%	2.44%	3.79%	2.62%	5.95%	3.36%	6.47%	5.11%

Table 3: Ablation study results on different datasets.

Datasets	Variants	R@10	R@20	N@10	N@20
Baby	w/o BMD	0.0605	0.0963	0.0327	0.0419
	w/o FG	0.0519	0.0835	0.0280	0.0361
	w/o $\mathcal{L}_{Modality}$	0.0648	0.1010	0.0346	0.0438
Sports	w/o BMD	0.0616	0.0965	0.0331	0.0421
	w/o FG	0.0548	0.0867	0.0302	0.0384
	w/o $\mathcal{L}_{Modality}$	0.0724	0.1099	0.0395	0.0492
TMALL	w/o BMD	0.0226	0.0345	0.0117	0.0149
	w/o FG	0.0238	0.0365	0.0131	0.0166
	w/o $\mathcal{L}_{Modality}$	0.0241	0.0375	0.0126	0.0162

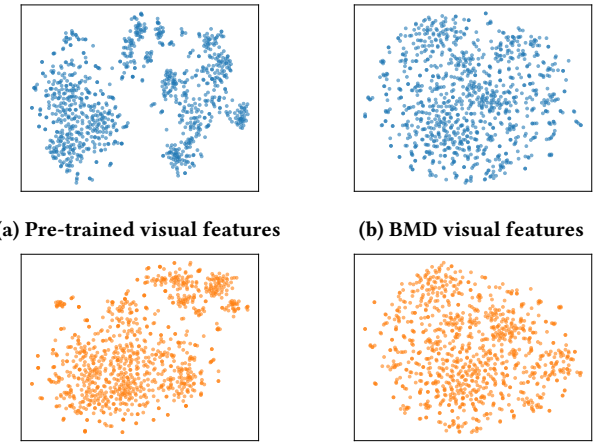
convolution directly to these features, allowing noise to spread across neighboring nodes. In contrast, BMD generates denoised recommendation-related features, preventing noise amplification in subsequent graph learning.

(2) The w/o FG variant, which removes the modality fusion graph and only retains the modality-specific graphs, also suffers a consistent performance drop across all datasets. This suggests that the fusion graph plays a critical role in capturing the user preferences of cross-modal complementary features. Such features serve as complementary signals to the modality-specific features and enhancing the robustness of user preferences modeling.

(3) The variant without $\mathcal{L}_{\text{modality}}$ experiences an average drop of 3.86% in Recall@20 and 6.22% in NDCG@20. This confirms the necessity of jointly learning modality-specific and modality-shared features. Overemphasizing either aspect may distort the other, leading to suboptimal representations and degraded performance.

3.4 Effect of BMD (RQ3)

To validate the contribution of the denoised modality features generated by BMD, we conduct a t-SNE [18] visualization on the TMALL dataset. Specifically, we extract the modality features produced by the BMD module, and project them into a two-dimensional space. As shown in Fig. 3, the pre-trained modality features present multiple distinct and fragmented clusters, especially in the image modality, indicating the existence of substantial recommendation-irrelevant noise. In contrast, the denoised features generated by BMD exhibit a more compact and coherent distribution, with smoother global structures and less pronounced inter-cluster boundaries. This suggests that BMD effectively filters noise from the pre-trained modality features and enhances the intra-modality consistency.



(a) Pre-trained visual features (b) BMD visual features

(c) Pre-trained textual features (d) BMD textual features

Figure 3: The t-SNE distribution of raw and BMD-generated modality features in TMALL dataset.

4 Conclusion

In this paper, we proposed a Behavior-Conditional Diffusion model for Multi-Modal recommendation (BCDMM). Specifically, our model introduces the diffusion mechanism to filter the recommendation-irrelevant noise from the pre-trained modality features. Furthermore, we construct modality-specific and fusion view graphs to explore the latent structure of various item modalities.

References

[1] Jie Chen, Haw-ren Fang, and Yousef Saad. 2009. Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *Journal of Machine Learning Research* 10 (2009), 1989–2012.

[2] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. 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*. 335–344.

[3] Qile Fan, Penghang Yu, Zhiyi Tan, Bing-Kun Bao, and Guanming Lu. 2025. BeFA: A General Behavior-driven Feature Adapter for Multimedia Recommendation. *arXiv preprint arXiv:2406.00323*.

[4] Rohit Girdhar, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Deepak Pathak, Laurens van der Maaten, and Ishan Misra. 2023. ImageBind: One Embedding Space to Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 33598–33608.

[5] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 144–150.

[6] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 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.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Proceedings of the 34th Conference on Neural Information Processing Systems*. 6840–6851.

[8] Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. 2024. DiffMM: Multi-Modal Diffusion Model for Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7591–7599.

[9] Yungi Kim, Taeri Kim, Won-Yong Shin, and Sang-Wook Kim. 2024. MONET: Modality-Embracing Graph Convolutional Network and Target-Aware Attention for Multimedia Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 332–340.

[10] Guojiao Lin, Meng Zhen, Dongjie Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. 2024. GUME: Graphs and User Modalities Enhancement for Long-Tail Multimodal Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1400–1409.

[11] Shuying Liu and Weihong Deng. 2015. Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size. In *Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition*. 730–734.

[12] Haokai Ma, Yimeng Yang, Lei Meng, Ruobing Xie, and Xiangxu Meng. 2024. Multimodal Conditioned Diffusion Model for Recommendation. In *Proceedings of the ACM Web Conference 2024*. 1733–1740.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. CLIP: Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. 8748–8763.

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 3982–3992.

[15] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. 452–461.

[16] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-Supervised Learning for Multimedia Recommendation. *IEEE Transactions on Multimedia* 25 (2022), 5107–5116.

[17] Zhulin Tao, Yin wei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. MGAT: Multimodal Graph Attention Network for Recommendation. *Information Processing & Management*. 57 (2020), 102277.

[18] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*. 6000–6010.

[20] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.

[21] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3541–3549.

[22] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith CH Ngai. 2025. Mentor: multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 12908–12917.

[23] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. 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*. 1807–1811.

[24] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-View Graph Convolutional Network for Multimedia Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6576–6585.

[25] Hansheng Zeng, Yuqi Li, Ruize Niu, Chuanguang Yang, and Shiping Wen. 2025. Enhancing spatiotemporal prediction through the integration of Mamba state space models and Diffusion Transformers. *Knowledge-Based Systems* 316 (2025), 113347.

[26] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[27] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2023. Latent Structure Mining With Contrastive Modality Fusion for Multimedia Recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35 (2023), 9154–9167.

[28] Guo Zhiqiang, Li Jianjun, Li Guohui, Wang Chaoyang, Shi Si, and Ruan Bin. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8454–8462.

[29] Xin Zhou and Zhiqi Shen. 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.

[30] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multimodal Recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.