Seeing Beyond Noise: Joint Graph Structure Evaluation and Denoising for Multimodal Recommendation

Yuxin Qi [1,2*], **Quan Zhang** [3*], **Xi Lin** [1,2†], **Xiu Su** [4], **Jiani Zhu** [1,2], **Jingyu Wang** [5], **Jianhua Li** [1,2]

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai, China
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[4]Big Data Institute, Central South University, Changsha, China
[5]Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China
{qiyuxin98, linxi234, 3951230458, lijh888}@sjtu.edu.cn, zhangqua22@mails.tsinghua.edu.cn, xiusu1994@csu.edu.cn, wangjingyu3186@stu.ouc.edu.cn

Abstract

Multimodal Recommendation Systems (MRSs) boost traditional user-item interaction-based methods by incorporating multimodal information. However, existing methods ignore the inherent noise brought by (1) noisy semantic priors in multimodal content, and (2) noisy user interactions in history records, therefore diminishing model performance. To fill this gap, we propose to denoise MRSs by jointly EValuating structure Effectiveness and mitigating Noisy links (EVEN). Firstly, for semantic prior noise in multimodal content, EVEN builds item homogeneous consistency and denoises it by evaluating behavior-driven confidence. Secondly, for noise in user interactions, EVEN updates user feedback by denoising observed interactions following implicit contribution evaluation of high-order representations. Thirdly, EVEN performs cross-modal alignment through self-guided structure learning, reinforcing task-specific inter-modal dependency modeling and cross-modal fusion. Through extensive experiments on three widely-used datasets, EVEN achieves an average improvement of 8.95% and 5.90% in recommendation accuracy compared with LGMRec and FREEDOM, respectively, without extending the total training time.

Introduction

Multimodal recommendation systems (MRSs) enhance traditional recommendation methods, which primarily depend on historical user-item interactions, by incorporating multimodal content such as item images and textual descriptions (He and McAuley 2016; Wei et al. 2019, 2020; Wang et al. 2021a; Zhang et al. 2021; Zhou et al. 2023b; Zhou and Shen 2023; Wei et al. 2023a; Yu et al. 2023; Zhou et al. 2023a). MRSs provide richer content, enabling the capture of fine-grained modality-level user preference. This enhances the performance of recommendations and improves user experience (Deldjoo, Schedl, and Knees 2024). For example, items with similar visual styles can be recommended based on visual preferences, or specific types of reading materials can be suggested based on textual preferences.

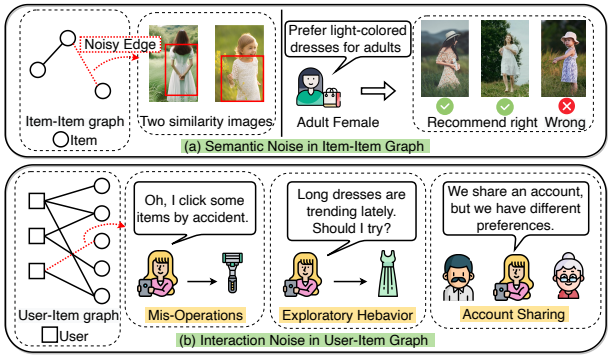

Figure 1: Noise presents challenges for multimodal recommendation systems. (a) Semantic noise introduced by using multimodal features as priors. For example, an adult woman searching for light-colored dresses might mistakenly get children's white dress recommendations because she browsed white dresses before. (b) Interaction Noise in user-item historical records. Users' historical data might contain irrelevant preferences due to mis-operations, exploratory behaviors, or account sharing.

Existing MRSs explore how to effectively integrate multimodal content to enrich recommendations. Conventional methods construct user and item representations by linear fusion the modal feature and ID embedding (He and McAuley 2016; Liu, Wu, and Wang 2017) or adopt the attention mechanism (Chen et al. 2019; Liu et al. 2019). However, the performance is limited because of the constrained information of low-order interactions. To capture high-order connectivity and enhance the recommendation, Graph Neural Network (GNN)-based collaborative filtering frameworks are proposed to incorporate multimodal content in user interest modeling (Wei et al. 2019, 2020; Wang et al. 2021a; Tao et al. 2022; Wei et al. 2023b) or inject multimodal semantic priors into the message-passing mechanism (Zhang et al. 2021; Zhou and Shen 2023; Yu et al. 2023).

However, MRSs are sensitive to inherent noise, which misleads the user preference capture and decreases recom-

*These authors contributed equally.
†Corresponding author.

mendation performance. In Figure 1, we highlight two types of noise: semantic noise among items and interaction noise in user-item feedback. Most existing studies on MRSs suffers from the ignore of above noises. (1) As for semantic noise, models like LATTICE and FREEDOM try to connect items through similarity-based graphs. However, they face challenges related to diverse user preferences (Wei et al. 2023a), the limitations of cosine similarity in high dimensions (Ahn 2008), and discrepancies in modal distribution (Liang et al. 2021). These lead to inaccurate semantic links and point out wrong user interest. (2) As for interaction noise, factors like mis-operations, exploratory behaviors, or account sharing introduce noise in user-item interactions, confusing the user interest modeling. Although few methods realize the noise in MRSs, current methods are based on strong assumptions to detect noisy records. RGCN uses implicit feedback to identify noise, assuming items in false-positive interactions are distinct from user preferences. This overlooks false positives similar to user interest. The challenge we face is *how to identify and mitigate item semantic noise and user-item interaction noise in MRSs without making special assumptions about user behavior, allowing for greater flexibility and robustness across user profiles and interaction patterns.*

In this paper, we propose a EValuating graph structure Effectiveness and mitigating Noisy links (EVEN) method for recommendation, named EVEN, which mitigates the impact of inherent noise. Firstly, to incorporate multimodal raw features while denoising semantic priors, we propose building item consistency into a homogeneous graph, evaluating the graph structure through task-specific relevance analysis, and denoising it via behavior-driven adjustments. Secondly, we introduce an adaptive graph pruning method to denoise observed interactions by evaluating implicit contribution of representations during graph message-passing. This approach mitigates the impact of less-contributed user feedback, enabling the model to focus on more preference-relevant interactions. Thirdly, we propose a self-guided structure learning module to achieve task-specific cross-modal alignment and fusion, boosting MRSs performance against inherent noise and multi-source information. Our contributions can be summarized as follows:

- We introduce a novel method EVEN to evaluate and mitigate the inherent semantic and interaction noise on higher-order structure connectivity in MRSs, boosting the performance without additional user assumptions.
- We propose to denoise the multimodal content semantic priors by building and evaluating task-specific item consistency in the homogeneous graph. An implicit contribution-based graph prune method over user-item heterogeneous graph is proposed to evaluate and denoise the observed interactions.
- We present a self-guided structure learning method to achieve task-specific cross-modal alignment and fusion over the denoised two graphs, which enhances fine-grained representation and preference mining.
- Through experiments across Amazon datasets, EVEN demonstrates its effectiveness and achieves an average improvement of 8.95% and 5.90% in recommendation accuracy compared with LGMRec and FREEDOM, without extending the total training time.

Related work

Multimedia Recommendation

MRSs leverage diverse modalities information to enrich recommendations. Existing MRSs can be divided into two types: (1) Item embedding direct fusion-based methods. BPR (He and McAuley 2016) is the first work to integrate visual raw features with ID embeddings to represent items. MAML (Liu et al. 2019) refines users' multimodal preferences via attention mechanisms. With the advent of GNN, models such as MMGCN (Wei et al. 2019), GRCN (Wei et al. 2020), DualGNN (Wang et al. 2021a), SLMRec (Tao et al. 2022),and MGCN (Yu et al. 2023) leverage graph convolutions to inject higher-order dependency into item representations, capturing connectivity through message-passing. Recently, contrastive learning is introduced to fuse multimodal item embeddings in BM3 (Zhou et al. 2023b) and MMSSL (Wei et al. 2023a). (2) Structure-based methods construct another view of Item-Item (I-I) graph to build item dependency. LATTICE (Zhang et al. 2021) designs a learnable layer to construct I-I structure and generate item semantic embedding over the dynamic I-I graph. FREEDOM (Zhou and Shen 2023) freezes the I-I graph and mines item representations over the frozen graph. LGMRec (Guo et al. 2024) builds the many-to-many dependency between item and features into a hypergraph and generate global item embedding over hypergraph message-passing.

Denoise Learning in Recommendation

Denoising efforts in MRSs are limited and can be categorized into two types: (1) Learning-based methods. GRCN (Wei et al. 2020), which use pre-defined functions to assess edges based on relevance to user preferences, mitigating noise through edge confidence scores. (2) Rule-based methods. LayerGCN (Zhou et al. 2022) and FREEDOM (Zhou and Shen 2023), which adjust structures using predetermined rules without relying on training feedback.

For traditional recommendation methods without multimodal information, existing denoise methods can be divided into two categories: (1) Sample selection methods (Ding et al. 2019; Gantner et al. 2012). (2) Sample re-weighting methods (Hu et al. 2021; Wang et al. 2021c). ADT (Wang et al. 2021b) addresses the challenge of integrating false-positive interactions in initial training stages by dynamically balancing truncation and re-weighted losses. SGDL (Gao et al. 2022) guides training with meta-learning by leveraging early training interactions for noise identification. However, its dependency on initial data might lead to misleading denoising when the data is biased, exacerbating noise issues and extending the training time.

To avoid potential biases in early-stage data without relying on additional assumptions about user behavior, we propose a data-independent denoising method. EVEN integrates multimodal denoising feature fusion with user-item dependency exploration from a global perspective.

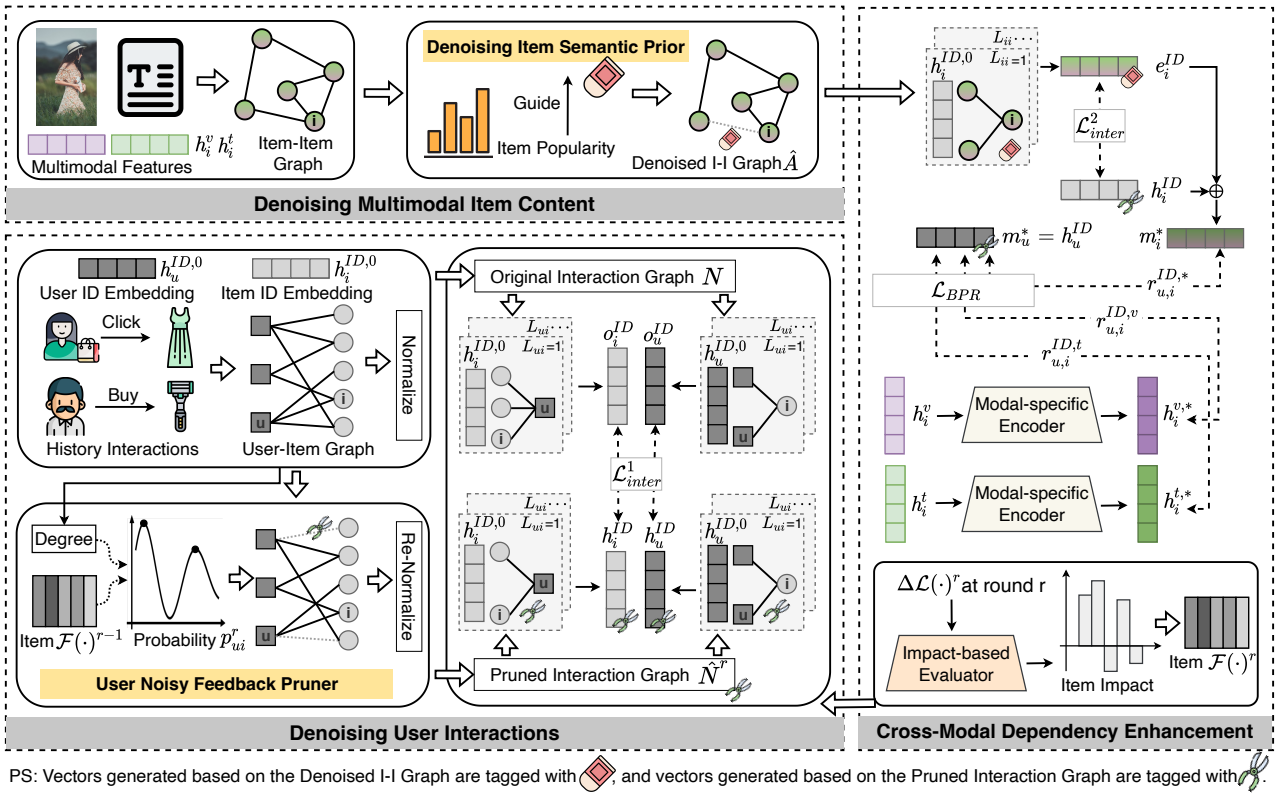

Figure 2: Framework of EVEN. For semantic priors noise in multimodal features, EVEN first builds item content consistency into a homogeneous graph, and then denoises it by integrating task-specific and behavior-driven adjustments. For noise in observed user interactions, we introduce an implicit influence-based graph pruning method, which denoises user feedback by evaluating the impact of high-order representations during message-passing over the user-item heterogeneous graph. After obtained the denoised graphs, EVEN performs task-specific cross-modal alignment through self-guided structure learning. This process enhances robust and fine-grained representations by emphasizing the relative relationships in feedback, rather than being misled by individual noisy samples.

Preliminary

Given user-item interactions, we generate a graph $G = \{(u,i)|u \in \mathcal{U}, i \in \mathcal{I}\}$, where $\mathcal{U}$ and $\mathcal{I}$ are the user and item sets. An edge (u,i) is formed if an interaction exists, incorporating multimodal features. Following (Zhang et al. 2021; Zhou and Shen 2023), we consider visual and textual modalities. $\boldsymbol{h}_i^m, m \in \{v,t\}$ represents item i's visual and textual feature. $\boldsymbol{h}_u^{ID}$ and $\boldsymbol{h}_i^{ID}$ represents user u's and item i's ID embedding (Zhao et al. 2018). Our denoised multi-modal system aims to learn user preferences through collaborative filtering, addressing interaction and semantic noise.

Proposed Method: EVEN

The framework of proposed EVEN is shown in Figure 2. In the following sections, the detailed designs are shown.

Denoising Multimodal Item Content

To utilize and denoise multimodal features, we first extract semantic priors from raw multimodal features and construct a similarity-based item consistency graph. We then incorporate task-specific behavior relationships from user interactions to mitigate noise in the item homogeneous graph. After behavior-driven adjustments, a more comprehensive item-to-item relationship is built, benefiting from considering user preferences in addition to item multimodal features.

Constructing Item Semantic Consistency For each modality m, we construct an item-item graph $\boldsymbol{A}^m$ using raw item multimodal features to extract semantic priors and mine content consistency. $\boldsymbol{A}^m$ captures the inherent consistency between items by establishing edges based on feature semantic similarity. The adjacency matrix is defined as $\boldsymbol{A}_{ij}^m = \frac{(\boldsymbol{h}_i^m)^\top \boldsymbol{h}_j^m}{|\boldsymbol{h}_i^m||\boldsymbol{h}_j^m|}$, where $\boldsymbol{h}_i^m$ is the raw feature of item i in modality m, $\boldsymbol{A}_{ij}^m$ is similarity score between items i and j. To enhance graph sparsity, $\boldsymbol{A}^m$ is transformed into an unweighted graph by retaining edges to each item's top-k_1 similar items as $\boldsymbol{A}_{ij} = 1$, while others are set to 0. This ensures $\sum_{j \in \mathcal{I}} \boldsymbol{A}_{ij} \leq k_1$, focusing on the most relevant connections and eliminating less significant ones. The visual $\boldsymbol{A}^v$ and textual $\boldsymbol{A}^t$ modality similarity graphs are merged as: $\boldsymbol{A}^1 = \alpha_1 \boldsymbol{A}^v + (1-\alpha_1)\boldsymbol{A}^t$, where learnable parameter α_1 controls the influence of the visual modality, efficiently

integrating multimodal contents.

Denoising Item Semantic Priors To mitigate inherent semantic noise in $\boldsymbol{A}^1$, we propose adjusting $\boldsymbol{A}^1$ by incorporating task-specific behavior relationships. An item occurrence matrix $\boldsymbol{C}$ is constructed by measuring the shared interest frequency. This is motivated by the fact that if items i and j are frequently liked by the same user, then i and j are task-specific relevant. $\boldsymbol{C}_{ij}$ is computed as the count of unique users interacting with both items i and j:

$$\boldsymbol{C}_{ij} = \left|\{u \in \mathcal{U} \mid \boldsymbol{M}_{u,i} = 1 \wedge \boldsymbol{M}_{u,j} = 1\}\right|, \tag{1}$$

where $\boldsymbol{M}$ is the adjacency matrix of the user-item interaction graph G, with $\boldsymbol{M}_{u,i} = 1$ denoting an observed interaction between user u and item i.

To avoid the impact of random behaviors, for each item i, we set the behavior occurrence for the top-k_2 items $j \in \mathcal{I}$ in $\boldsymbol{C}_{ij}$ to 1, and others to 0. We then merge the task-specific item occurrence matrix $\boldsymbol{C}$ with the semantic consistency matrix $\boldsymbol{A}^1$ to form a more robust Item-Item (I-I) graph by introducing a learnable parameter α_2 as $\boldsymbol{A}^2 = \alpha_2 \boldsymbol{A}^1 + (1 - \alpha_2)\boldsymbol{C}$, where $\boldsymbol{A}^2$ is the denoised I-I graph. Next, $\boldsymbol{A}^2$ is normalized to mitigate the impact of node degree on aggregation:

$$\widehat{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{A}^2 \boldsymbol{D}^{\frac{1}{2}}, \tag{2}$$

with $\boldsymbol{D}$ as the diagonal degree matrix of $\boldsymbol{A}^2$, where $\boldsymbol{D}_{ii} = \sum_{j \in \mathcal{I}} \boldsymbol{A}^2_{ij}$. Normalization ensures a balanced contribution from each node during message passing.

Denoising User Interactions

We introduce an implicit impact-based graph pruning method to denoise observed interactions by considering the contribution of high-order item representations during message-passing. EVEN probabilistically drop less contributed and high-degree item edges, which preserves essential feedback and prevents over-smoothing during graph convolution, enhancing recommendation performance by balancing sparsity and information integrity. Then over the pruned interactions, EVEN generates refined high-order representations of users and items by emphasizing relative feedback, minimizing individual sample noises.

Implicit Impact-Based Interaction Evaluator Considering that observed interactions contains user-preference-agnostic records, we propose to iteratively assess item contribution by evaluating its impact through the gradients of previous training round. Gradients capture the extent of loss function changes with specific parameters, aiding in the identification and pruning of less-contributed edges. We denote the impact indicator of item i in round r as $\mathcal{F}(i)^r = \frac{\partial \mathcal{L}}{\partial i}^{r-1}$, where $\mathcal{L}$ is the overall optimization function, as shown in eq. (9). Note that $\mathcal{F}(i)^r$ is obtained from training process without additional user behavior assumption.

User Noisy Feedback Pruner Given $\mathcal{F}(i)^r$ for item i at round r, the edge pruning probability is set as:

$$p^r_{ui} = \frac{\mathcal{F}(i)^{r-1}}{\sum_{j \in \mathcal{I}} \mathcal{F}(j)^{r-1}} \cdot \frac{1}{\sqrt{d_u}\sqrt{d_i}}, \tag{3}$$

where $\mathcal{F}(i)^{r-1}$ is the contribution of item i at previous round $r-1$, $\sqrt{d_u}$ is the degree of u. Then, we construct a symmetric adjacency matrix $\boldsymbol{N} \in R^{|\mathcal{U}|+|\mathcal{I}|}$ from the user-item interaction matrix $\boldsymbol{M}$ as:

$$\boldsymbol{N} = \begin{pmatrix} 0 & \boldsymbol{M} \\ \boldsymbol{M}^\top & 0 \end{pmatrix}, \tag{4}$$

where $|\mathcal{U}|$ and $|\mathcal{I}|$ represents the users and items number.

Then sample edges in $\boldsymbol{N}$ from the multinomial distribution with probability vector $\mathbf{p} = < ...p^r_{ui}... >$ and target data length $\lfloor |G|(1-\rho) \rfloor$, where ρ is the pre-defined dropout probability, $|G|$ is the total edge number of user-item graph. Now items with high contribution and low degree are more likely to be retained. After edge pruning, the symmetric adjacency matrix is constructed following eq. (4) and re-normalizated following eq. (2), obtaining refined matrix $\widehat{\boldsymbol{N}}^r$ for round r.

The feedback pruner preserves important interactions and reduces noisy clicks, enhancing model robustness. By iteratively refining $\widehat{\boldsymbol{N}}^r$, pruning probabilities adapt based on previous round impacts. In the inference phase, the original Laplacian normalization of $\boldsymbol{N}$ is used directly.

Interaction Dependency Denoising For Interaction-Graph ID Embedding (IG-ID Embedding) generation, convolutional aggregations are performed on pruned Interaction-Graph $\widehat{\boldsymbol{N}}^r$ of round r. The IG-ID Embedding of user $\boldsymbol{h}^{ID}_u$ and item $\boldsymbol{h}^{ID}_i$ are obtained through the stacked of all the hidden latent with a differentiable function as:

$$\begin{aligned} \boldsymbol{h}^{ID}_u &= \text{READOUT}(\boldsymbol{h}^{ID,0}_u, \boldsymbol{h}^{ID,1}_u, ..., \boldsymbol{h}^{ID,L_{ui}}_u), \\ \boldsymbol{h}^{ID}_i &= \text{READOUT}(\boldsymbol{h}^{ID,0}_i, \boldsymbol{h}^{ID,1}_i, ..., \boldsymbol{h}^{ID,L_{ui}}_i), \end{aligned} \tag{5}$$

where $\boldsymbol{h}^{ID,0}_u$ and $\boldsymbol{h}^{ID,0}_i$ denote the initial ID embeddings of user u and item i, $\boldsymbol{h}^{ID,l}_u$ denotes the IG-ID embedding of user u at l-th layer, and L_{ui} is the total convolutional operation layers over $\widehat{\boldsymbol{N}}^r$. Following (He et al. 2020; Zhou and Shen 2023), we use the default mean function as readout.

To further capture user preference in interactions, we leverage the potential of self-supervised learning to mine inter-modal deep-layer feedback relationships. In particular, $\widehat{\boldsymbol{N}}^r$ in every round r can be regarded as contrastive views of structural perturbations for original noisy user-item graph $\boldsymbol{N}$. Therefore, we also perform L_{ui} convolutional aggregations over original interaction graph $\boldsymbol{N}$ and stack of all the hidden representations with a differentiable function as

$$\begin{aligned} \boldsymbol{o}^{ID}_u &= \text{READOUT}(\boldsymbol{o}^{ID,0}_u, \boldsymbol{o}^{ID,1}_u, ..., \boldsymbol{o}^{ID,L_{ui}}_u), \\ \boldsymbol{o}^{ID}_i &= \text{READOUT}(\boldsymbol{o}^{ID,0}_i, \boldsymbol{o}^{ID,1}_i, ..., \boldsymbol{o}^{ID,L_{ui}}_i), \end{aligned} \tag{6}$$

where $\boldsymbol{o}^{ID,0}_u = \boldsymbol{h}^{ID,0}_u$ and $\boldsymbol{o}^{ID,0}_i = \boldsymbol{h}^{ID,0}_i$. The inter-modal user-item graph structure loss of round r is defined as:

$$\begin{aligned} \mathcal{L}^1_{inter} = &\sum_{u \in \mathcal{U}} -\log \frac{\exp\left(\boldsymbol{h}^{ID}_u \cdot \boldsymbol{o}^{ID}_u / \tau\right)}{\sum_{v \in \mathcal{U}} \exp\left(\boldsymbol{h}^{ID}_u \cdot \boldsymbol{o}^{ID}_v / \tau\right)} \\ &+ \sum_{i \in \mathcal{I}} -\log \frac{\exp\left(\boldsymbol{h}^{ID}_i \cdot \boldsymbol{o}^{ID}_i / \tau\right)}{\sum_{j \in \mathcal{I}} \exp\left(\boldsymbol{h}^{ID}_i \cdot \boldsymbol{o}^{ID}_j / \tau\right)}, \end{aligned} \tag{7}$$

where τ is the temperature hyper-parameter.

Cross-Modal Dependency Enhancement

To align and fuse cross-modal contents, we propose a self-guided structure learning module to better distinguish positive and negative interactions.

Multi-modal Item Dependency Modeling Considering the modal-specific semantic differences between item multi-modal contents and item interactions, we explore item consistent dependency through the alignment of multimodal features and ID embeddings. For multimodal item embedding generation, we first perform graph convolution over the denoised I-I graph $\widehat{\boldsymbol{A}}$ using ID embeddings as $\boldsymbol{e}_i^{ID,l} = \sum_{j \in \mathcal{N}(i)} \widehat{\boldsymbol{A}}_{ij} \boldsymbol{e}_i^{ID,l-1}, \quad \boldsymbol{e}_i^{ID,0} = \boldsymbol{h}_i^{ID,0}$, where $\boldsymbol{e}_i^{ID,l}$ is the l-layer Semantic-Graph ID Embedding (SG-ID Embedding) of item i, $\mathcal{N}(i)$ is the neighbor set of i. SG-ID Embedding integrates the ID embeddings into the multimodal consistency graph, ensuring that the embeddings are aligned and consistent across different modalities.

Given the IG-ID Embedding $\boldsymbol{h}_i^{ID}$ of item i in eq. (5), the final item representation $\boldsymbol{m}_i^*$ is obtained as: $\boldsymbol{m}_i^* = \boldsymbol{h}_i^{ID} + \boldsymbol{e}_i^{ID}, \quad \boldsymbol{e}_i^{ID} = \boldsymbol{e}_i^{ID,L_{ii}}$, where L_{ii} is the total layers on $\widehat{\boldsymbol{A}}$. To further align item i's SG-ID Embeddings $\boldsymbol{e}_i^{ID}$ and IG-ID Embedding $\boldsymbol{h}_i^{ID}$, we define the multi-modal enhancement loss as:

$$\mathcal{L}_{inter}^2 = \sum_{t \in \mathcal{I}} -\log \frac{\exp(\boldsymbol{h}_t^{ID} \cdot \boldsymbol{e}_t^{ID})/\tau}{\sum_{k \in \mathcal{I}} \exp(\boldsymbol{h}_t^{ID} \cdot \boldsymbol{e}_k^{ID})/\tau}, \tag{8}$$

where τ is the temperature hyper-parameter. It encourages the cross-modal item feature to be at the same latent space.

To explore the raw multi-modal item feature $\boldsymbol{h}_i^m, m \in \{v, t\}$ in user preference mining, modal-specific encoder $ME^m(\cdot)$ is designed to map $\boldsymbol{h}_i^v$ and $\boldsymbol{h}_i^t$ into the near latent space as $\boldsymbol{h}_i^{m,*} = ME^m(\boldsymbol{h}_i^m), m \in \{v, t\}$. Here we use the widely used MLPs as the encoder.

Joint Optimization over Two Graphs The final representation $\boldsymbol{m}_u^*$ of user u is set as the IG-ID Embedding obtained in eq. (5), that is $\boldsymbol{m}_u^* = \boldsymbol{h}_u^{ID}$. Given $\boldsymbol{m}_u^*$ and $\boldsymbol{m}_i^*$, the inner product of $\boldsymbol{m}_u^*$ and $\boldsymbol{m}_i^*$ is used to calculate the core preference score $r_{u,i}^{ID,*}$. Similar, the side preference scores $r_{u,i}^{ID,v}$ and $r_{u,i}^{ID,t}$ are obtained by the inner product of $\boldsymbol{m}_u^*$ to the encoded multi-modal item feature $\boldsymbol{h}_i^{v,*}$ and $\boldsymbol{h}_i^{t,*}$, separately.

We adopt the pairwise Bayesian Personalized Ranking (BPR) (Rendle et al. 2009) to encourage model chose positive user-item pair by the given core and side preference scores $r_{u,i}^{ID,*}, r_{u,i}^{ID,v}$ and $r_{u,i}^{ID,t}$. Then the cross-modal dependency optimization over two graphs is as follows:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda(\mathcal{L}_{inter}^1 + \mathcal{L}_{inter}^2), \tag{9}$$

where $\mathcal{L}_{BRP}$ is the BRP loss, λ is hyper-parameter to guide optimization ratio.

Experiments

Experimental Settings

Datasets We conduct experiments on three widely used and public Amazon dataset: (i) Baby, (ii) Sports and outdoors, and (iii) Clothing, Shoes and Jewelry.

Evaluation Protocols and Hyperparameters Settings To make fair comparison, we adopt the same evaluation setting of baselines. The grid search results of hyperparameters are reported. Detailed description is shown in Appendix.

Compared Baseline Models

We compare our proposed method with both traditional and multimodal recommendation methods. For traditional methods, the competitor is graph neural network method **LightGCN** (He et al. 2020). For multimodal recommendation methods, the competitors include (1) without I-I graph: **VBPR** (He and McAuley 2016), **DualGNN** (Wang et al. 2021a), **SLMRec** (Tao et al. 2022), **BM3** (Zhou et al. 2023b), **MMSSL** (Wei et al. 2023a),**LGMRec** (Guo et al. 2024) and (2) with I-I graph: **LATTICE** (Zhang et al. 2021), **FREEDOM** (Zhou and Shen 2023). We also compare EVEN against existing typical denoising models in both traditional recommendation systems and MRSs, including **LayerGCN** (Zhou et al. 2022), **GRCN** (Wei et al. 2020), and **ADT** (Wang et al. 2021b).

Performance Comparison

Effectiveness Comparison. In Table 1, we compare the performance of all methods across three datasets. The results show that EVEN outperforms the baselines on all datasets. Specifically, EVEN boosts R@10 by 3.57%/5.42%/5.58% on Baby, Sports, and Clothing, respectively, compared to the second-best method. These consistent gains indicate that EVEN increases the number of relevant recommendations and improves their ranking by noise reduction and refined higher-order connectivity mining.

We also compare EVEN with existing denoising methods, including traditional GNN-based collaborative filtering approaches like LayerGCN and multimodal approaches like RGCN. Additionally, we adapt the traditional non-GNN denoising method ADT for MRSs, using FREEDOM as the backbone—selected because it does not introduce extra structural components, ensuring a fair comparison—referred to as FREEDOM-ADT. As shown in the last five rows of Table 1, EVEN outperforms these methods, achieving a 25% improvement across all metrics on the Baby dataset. FREEDOM-ADT's underperformance highlights the limitations of traditional denoising approaches in addressing the semantic noise in multimodal features.

Efficiency Comparison. To investigate the training efficiency, we report the model performance changes on metrics R@10 and R@20 as the training epochs increase in Figure 3. The point where the curve flattens indicates the model has converged. EVEN shows a more stable and faster convergence rate. Compared to LGMRec, EVEN's recommendation performance steadily increases during optimization until convergence. Additionally, EVEN requires the fewest epochs to reach convergence compared to FREEDOM, LATTICE, and BM3. This shows that EVEN can expedite model convergence by eliminating potential noise in MRSs.

We also analyze the average training time per epoch (Train-T), total convergence time (T-Conv), and average inference time (Infer-T) for each model on Baby. As shown in Table 2, although the Train-T is higher, T-Conv remains

Method	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
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0340	0.0526	0.0188	0.0236
SLMRec	0.0529	0.0775	0.0290	0.0353	0.0663	0.0990	0.0365	0.0450	0.0452	0.0675	0.0247	0.0303
LATTICE	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
FREEDOM	0.0626	0.0985	0.0327	0.0420	0.0717	0.1089	0.0385	0.0481	0.0627	0.0940	0.0336	0.0415
MMSSL	0.0613	0.0971	0.0326	0.0420	0.0673	0.1013	0.0380	0.0474	0.0531	0.0797	0.0291	0.0359
LGMRec	0.0644	0.1002	0.0347	0.0440	0.0720	0.1068	0.0390	0.0480	0.0555	0.0828	0.0302	0.0371
EVEN	**0.0667**	**0.1031**	**0.0355**	**0.0448**	**0.0759**	**0.1143**	**0.0411**	**0.0510**	**0.0662**	**0.0978**	**0.0356**	**0.0436**
improv.	↑ 3.57%	↑ 2.90%	↑ 2.31%	↑ 1.81%	↑ 5.42%	↑ 4.96%	↑ 5.38%	↑ 6.03%	↑ 5.58%	↑ 4.04%	↑ 5.95%	↑ 5.06%
LayerGCN	0.0529	0.0820	0.0281	0.0355	0.0594	0.0916	0.0323	0.0406	0.0371	0.0566	0.0200	0.0247
GRCN	0.0532	0.0824	0.0282	0.0358	0.0599	0.0919	0.0330	0.0413	0.0421	0.0657	0.0224	0.0284
REEEDOM-ADT	0.0492	0.0769	0.0262	0.0333	0.0467	0.0738	0.0257	0.0327	0.0374	0.0578	0.0199	0.0251
EVEN	**0.0667**	**0.1031**	**0.0353**	**0.0448**	**0.0759**	**0.1143**	**0.0411**	**0.0510**	**0.0662**	**0.0978**	**0.0356**	**0.0436**
improv.	↑ 25.94%	↑ 26.21%	↑ 26.60%	↑ 24.02%	↑ 23.54%	↑ 23.39%	↑ 21.81%	↑ 21.79%	↑ 57.24%	↑ 48.86%	↑ 58.93%	↑ 53.52%

Table 1: Recommendation Performance Comparison: Overall performance on three datasets in terms of R@K and N@K are reported. Last 5 lines show the comparison with existing denoising methods. The top performers are highlighted in **bold** and the next best are underlined. *improv.* is calculated by comparing the best and second best one.

Methods	Train-T (s/epoch)	T-Conv Time (s)	Infer-T (s/epoch)	R@20
BM3	0.79	79.79	2.85	0.0883
LATTICE	2.35	277.30	2.84	0.0850
LGMRec	3.42	277.02	2.67	0.1002
FREEDOM	1.78	370.24	2.64	0.0985
EVEN	3.46	238.74	2.60	0.1031

Table 2: Efficiency Comparison: Average training time for each epoch (Train-T), total training converge time (T-Conv), and average inference time (Infer-T) on Baby.

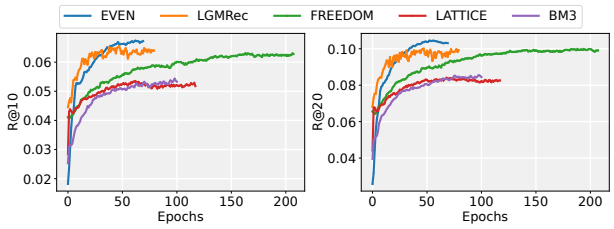

Figure 3: Training Convergence Comparison: Epochs versus R@10 and R@20 on Baby.

comparable and even shorter to most baselines, all while achieving the best performance. Since the structure denoising of the I-I and U-I Graphs occurs only during the training phase, then the denoised structure is frozen and used during the inference phase. Therefore, EVEN does not incur any additional time costs in Infer-T.

Ablation Study

We design the following four variants of EVEN to evaluate the contribution of each denoising components:

- **EVEN w/o IID** ignores the effect of semantic noise in multimodal item prior (without I-I graph Denoising).
- **EVEN w/o UID** ignores the effect of interaction noise in user-item history records (without U-I graph Denoising).
- **EVEN w/o IID & UID** ignores both semantic noise in multimodal item priors and interaction noise in user-item history records (without I-I and U-I graph Denoising).

Dataset	Variants	R@10	R@20	N@10	N@20
Baby	EVEN w/o IID	0.0655	0.1016	0.0349	0.0443
	EVEN w/o UID	0.0637	0.0981	0.0344	0.0433
	EVEN w/o IID & UID	0.0631	0.0967	0.0340	0.0427
	EVEN w/o MAF	0.0511	0.0807	0.0273	0.0350
	EVEN	**0.0667**	**0.1031**	**0.0355**	**0.0448**
Sports	EVEN w/o IID	0.0741	0.1130	0.0401	0.0501
	EVEN w/o UID	0.0739	0.1129	0.0398	0.0499
	EVEN w/o IID & UID	0.0730	01120	0.0387	0.0487
	EVEN w/o MAF	0.0644	0.0959	0.0353	0.0453
	EVEN	**0.0759**	**0.1143**	**0.0411**	**0.0510**
Clothing	EVEN w/o IID	0.0648	0.0968	0.0349	0.0430
	EVEN w/o UID	0.0643	0.0953	0.0346	0.0425
	EVEN w/o IID & UID	0.0634	0.0950	0.0343	0.0423
	EVEN w/o MAF	0.0413	0.0617	0.0223	0.0275
	EVEN	**0.0662**	**0.0978**	**0.0356**	**0.0436**

Table 3: Ablation studies of four variants on all datasets.

- **EVEN w/o MAF** ablates the contribution of multimodal feature (without Multimodal Alignment and Fusion).

In Table 3, we present comparative results on all datasets. The gap between variants and EVEN shows the performance gain achieved by each components. **EVEN w/o MAF** performs the worst, indicating that proposed multimodal information alignment and fusion method contributes the MRSs performance. The gap between **EVEN w/o IID & UID** and **EVEN** shows that although multimodal content contains useful information, the inherent noise in semantic prior and observed interactions effects the user preference mining. **EVEN w/o IID** and **EVEN w/o UID** shows the effectiveness of EVEN in the user preference-related denoising over I-I graph and U-I graph. Denoising the interaction noise gains more performance compared with semantic noise.

Performance under Noisy Settings

To further verify performance against noise, we perform two noisy settings: adding extra noise to multimodal raw feature and user-item interactions. For multimodal semantic noise, we randomly set some dimensions of raw features as zero. For noisy interactions, we inject random links into observed user feedback. The perturbed ratio varies among $\{0.05, 0.10, 0.20, 0.30\}$, with 0.00 indicating no extra added noise. Con-

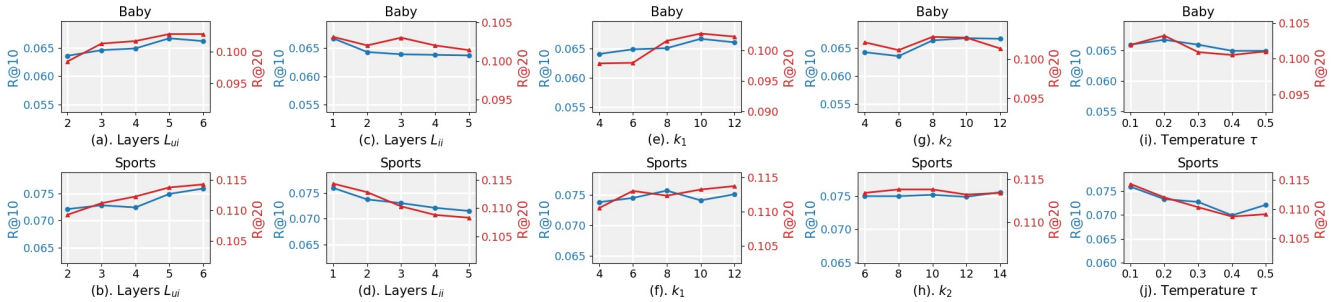

Figure 4: Performance under different settings of hyperparameters of EVEN on Baby and Sports.

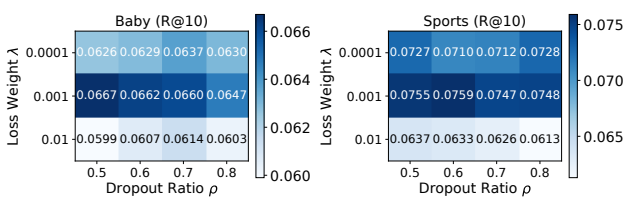

Figure 5: Performance under different loss weight λ and dropout ratio β of EVEN on Baby and Sports.

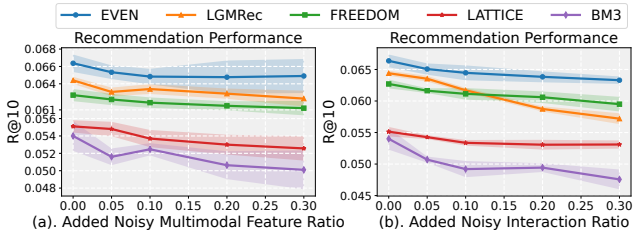

Figure 6: Performance comparison between EVEN and other four baselines under two noisy settings on Baby.

sidering the noise randomness, we conduct four experiments per setting, presenting the mean and variance in Figure 6.

Performance under Noisy Multimodal Raw Features. As shown in Figure 6 (a), EVEN consistently outperforms baselines across all noise levels, demonstrating its superior robustness. EVEN's performance is the least impacted as the noise intensity increased, showing its stability against noisy settings. The gap from 0.00 to 0.05 across all models highlights the necessity of denoising multimodal content.

Performance under Noisy User-Item Interactions. As shown in Figure 6 (b), EVEN outperforms other baselines across all noise levels, highlighting its effectiveness in evaluating structural contributions and managing noisy interactions. As the noisy interaction ratio increases, EVEN shows the most stable performance, while the second-best model, LGMRec, is the heavily affected by noisy interactions. The performance drop from 0.00 to 0.05 across all models highlights the importance of denoising observed interactions.

Hyperparameters Sensitivity Analysis

Effect of Graph Convolution Layers on L_{ui} and L_{ii}. As shown in Figure 4 (a)-(d), we adjust the message passing layers L_{ui} pruned interaction graph $\hat{\boldsymbol{N}}$ from 2 to 6 and layers L_{ui} on denoised I-I graph $\hat{\boldsymbol{A}}$ from 1 to 5. Compared to LGMRec, which achieves optimal performance with $L_{ui} = 2$ on Baby (Guo et al. 2024), EVEN performs best with $L_{ui} = 5$. This suggests that EVEN captures deeper inter-layer associations, effectively countering over-smoothing and uncovering intricate patterns previously concealed by noise.

Effect of Graph Structure Parameter k_1 and k_2. As shown in Figure 4 (e)-(h), we vary k_1 from 4 to 12 on Baby and Sports, and k_2 from 4 to 12 on Baby and 6 to 14 on Sports. k_1 controls the semantically similar neighbor numbers to connect in the original I-I graph $\boldsymbol{A}^1$. k_2 balances the trade-off between task-related behavior and random behavior in $\boldsymbol{C}$. A small k_1 may lose item consistency, while a large k_1 may introduce edges with no information gain. Similarly, a small k_2 may under-utilize task-specific information for denoising $\boldsymbol{A}^1$, whereas a large k_2 may introduce negative impacts due to noise in interactions.

Effect of Temperature τ. τ controls the sharpness of similarity scores, influencing how strongly positive and negative pairs are separated in the representation space. We empirically set τ=0.2 on Baby and τ=0.1 on Sports.

Effect of Dropout Ratio ρ and Loss Weight λ. We vary ρ from 0.5 to 0.8, and λ in {1e-4, 1e-3, 1e-2}. Figure 5 reports the performance for various combinations of ρ and λ, showing a general decline as the ρ increases. This trend indicates that a higher ρ may over-denoise valuable information due to increased sparsity in the graph.

Conclusion

In this paper, we propose a novel model EVEN that evaluate and denoise information in multimodal content and observed interactions for MRSs. EVEN first builds item content consistency into a homogeneous graph and performs task-specific, behavior-driven denoising. For observed interaction noise, we propose an implicit contribution-based graph pruning method to assess and refine user feedback during message-passing. Then EVEN enhances fine-grained representations through self-guided structural learning-based cross-modal alignment and fusion. Experiments show EVEN's effectiveness across multiple datasets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62202302 and U20B2048.

References

Ahn, H. J. 2008. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information sciences*, 178(1): 37–51.

Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; and Zha, H. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of SIGIR*, 765–774.

Deldjoo, Y.; Schedl, M.; and Knees, P. 2024. Content-driven music recommendation: Evolution, state of the art, and challenges. *Computer Science Review*, 51: 100618.

Ding, J.; Yu, G.; He, X.; Feng, F.; Li, Y.; and Jin, D. 2019. Sampler design for bayesian personalized ranking by leveraging view data. *IEEE transactions on knowledge and data engineering*, 33: 667–681.

Gantner, Z.; Drumond, L.; Freudenthaler, C.; and Schmidt-Thieme, L. 2012. Personalized Ranking for Non-Uniformly Sampled Items. In *KDD Cup*.

Gao, Y.; Du, Y.; Hu, Y.; Chen, L.; Zhu, X.; Fang, Z.; and Zheng, B. 2022. Self-Guided Learning to Denoise for Robust Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. ACM.

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. 2016. 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.

Hu, K.; Li, L.; Xie, Q.; Liu, J.; and Tao, X. 2021. What is next when sequential prediction meets implicitly hard interaction? In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 710–719.

Liang, T.; Lin, G.; Feng, L.; Zhang, Y.; and Lv, F. 2021. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8148–8156.

Liu, F.; Cheng, Z.; Sun, C.; Wang, Y.; Nie, L.; and Kankanhalli, M. 2019. User diverse preference modeling by multimodal attentive metric learning. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1526–1534.

Liu, Q.; Wu, S.; and Wang, L. 2017. Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, 841–844.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 452–461.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T.-S. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*.

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

Wang, W.; Feng, F.; He, X.; Nie, L.; and Chua, T.-S. 2021b. Denoising Implicit Feedback for Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21. ACM.

Wang, Y.; Xin, X.; Meng, Z.; He, X.; Jose, J.; and Feng, F. 2021c. Probabilistic and variational recommendation denoising. *arXiv preprint arXiv:2105.09605*.

Wei, W.; Huang, C.; Xia, L.; and Zhang, C. 2023a. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

Wei, W.; Huang, C.; Xia, L.; and Zhang, C. 2023b. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of WWW*.

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.

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*, MM '23. ACM.

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.

Zhao, K.; Li, Y.; Shuai, Z.; and Yang, C. 2018. Learning and transferring ids representation in e-commerce. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1031–1039.

Zhou, H.; Zhou, X.; Zhang, L.; and Shen, Z. 2023a. Enhancing Dyadic Relations with Homogeneous Graphs for Multimodal Recommendation.

Zhou, X.; Lin, D.; Liu, Y.; and Miao, C. 2022. Layer-refined Graph Convolutional Networks for Recommendation. arXiv:2207.11088.

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. 2023b. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, 845–854.