PDF Download
3746027.3754953.pdf
22 January 2026
Total Citations: 0
Total Downloads: 352

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

RESEARCH-ARTICLE

TAMER: Interest Tree Augmented Modality Graph Recommender for Multimodal Recommendation

FANSHEN MENG, Beijing University of Posts and Telecommunications, Beijing, Beijing, China

ZHENHUA MENG, Inner Mongolia University China, Hohhot, Nei Mongol, China

RU JIN, Beijing University of Posts and Telecommunications, Beijing, Beijing, China

YULI CHEN, Beijing University of Posts and Telecommunications, Beijing, Beijing, China

RONGHENG LIN, Beijing University of Posts and Telecommunications, Beijing, Beijing, China

BUDAN WU, Beijing University of Posts and Telecommunications, Beijing, Beijing, China

Open Access Support provided by:

Beijing University of Posts and Telecommunications

Inner Mongolia University China

MM '25: Proceedings of the 33rd ACM International Conference on Multimedia (October 2025)
https://doi.org/10.1145/3746027.3754953
ISBN: 9798400720352

TAMER: Interest Tree Augmented Modality Graph Recommender for Multimodal Recommendation

Fanshen Meng
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
Beijing, China
mengfanshen@bupt.edu.cn

Zhenhua Meng*
College of Computer Science, Inner Mongolia University
Hohhot, China
zhmeng@bupt.edu.cn

Ru Jin
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
Beijing, China
rjin@bupt.edu.cn

Yuli Chen
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
Beijing, China
chenyuli@bupt.edu.cn

Rongheng Lin†
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
Beijing, China
rhlin@bupt.edu.cn

Budan Wu
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
Beijing, China
wubudan@bupt.edu.cn

Abstract

Multimodal recommender systems enhance recommendation performance by integrating information from different modalities (e.g., text and images). A common approach is to link items with high modality similarity in modality graphs, helping users explore their interests more broadly. However, existing methods often introduce noise when enhancing modality graphs, making it challenging to effectively balance performance and accuracy. To address this issue, we propose an Interest **T**ree **A**ugmented **M**odality Graph Recommend**ER** for Multimodal Recommendation **(TAMER)**. In this framework, we first redistribute item modality features using various component analysis methods to ensure more reliable item similarity within modality graphs. Next, we construct interest graphs based on reliable semantic relationships and prune the interest graphs into multiple interest trees. These interest trees are then applied to the multimodal item-item homogeneous graph to extend potential links within the modality homogeneous graph. The interest tree-based enhancement method effectively captures high-order relationships in the modality graph while avoiding noisy links. The effectiveness of the proposed method is demonstrated through comprehensive experiments on three real-world datasets. Compared with the strongest baseline methods, our method achieves an average improvement of 9.98% across four evaluation metrics. The source code is available at https://github.com/Z-last-ONE/TAMER.

CCS Concepts

• **Information systems → Recommender systems**; **Multimedia and multimodal retrieval**.

Keywords

Multimodal Recommendation, Interest Tree, Modality Graph Enhancement

ACM Reference Format:
Fanshen Meng, Zhenhua Meng, Ru Jin, Yuli Chen, Rongheng Lin, and Budan Wu. 2025. TAMER: Interest Tree Augmented Modality Graph Recommender for Multimodal Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746027.3754953

1 Introduction

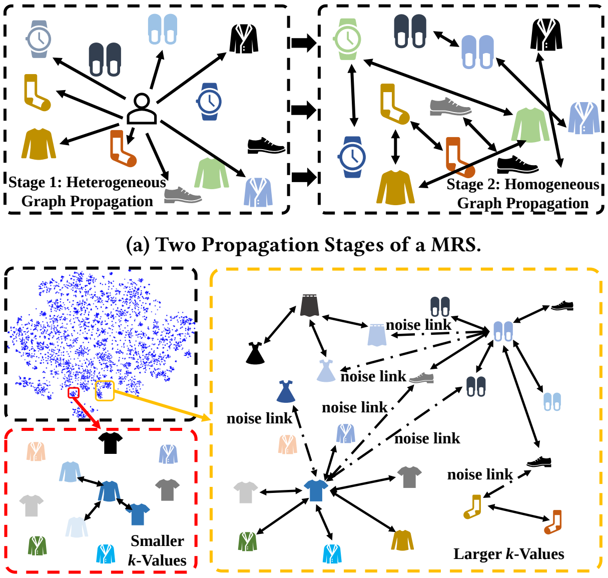
Recommender systems are crucial tools for helping users discover content of interest from massive data [10, 32]. In recent years, increasing attention has been paid to the role of multimedia information (e.g., text, images) in recommendations, making multimodal recommender systems a topic of widespread interest [27, 34, 36, 43]. By modeling multimedia content, multimodal recommender systems capture user interests from multiple perspectives, offering greater potential to accurately learn user preferences [7, 13].

Early approaches, such as VBPR [5], integrated modality signals with item IDs to provide a multifaceted description of item features, thereby improving item representation. With the development of graph neural networks (GNNs) [14, 20, 30], multimodal recommendation methods based on graph convolutional networks (GCNs) have demonstrated promising performance. To better capture the internal relationships between items, LATTICE [37] constructs a learnable adjacency matrix based on modality similarity to identify potential item relationships. However, FREEDOM [44] argues that learnable homogeneous graph structures are inefficient and instead proposes freezing item modality graphs, achieving superior

*This work was completed while he was at Beijing University of Posts and Telecommunications.
†Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754953



(a) Two Propagation Stages of a MRS.

(b) Effect of Different k in Modality Graph.

Figure 1: Figure (a) illustrates the two stages of current MRS: heterogeneous graph propagation and homogeneous graph propagation. Figure (b) depicts the potential impact of different k on propagation within a homogeneous modality graph.

performance. This approach has become a paradigm for multimodal recommendation research in recent years, where learning is performed separately on a heterogeneous interaction graph and a homogeneous frozen modality graph, followed by feature fusion. As shown in Figure 1a, the model first learns user preferences through user-item interactions, then uses item modality similarity for top-k pruning to construct the modality graph, propagating item features through the modality graph. However, the representational capability of the frozen modality graph pruned via top-k completely depends on the original distribution of modality features and the choice of k. As illustrated in Figure 1b, the black box represents the distribution of item features, and a smaller k value in the red box limits the propagation of item features to only a few of the most similar items, increasing the risk of users being trapped in an information bubble. Conversely, a larger k value in the yellow box may lead to peripheral items being connected to irrelevant items from adjacent clusters, introducing noisy connections [33], even if there is no real association between these items. For different items within the same set, it is challenging to determine an optimal k value that ensures as many relevant connections as possible while avoiding the introduction of noise.

To address this issue, some studies have proposed optimization methods for item modality graphs. For example, DA-MRS [33] reduces noisy connections in the original visual and textual modality graphs by leveraging behavior graph. SOIL [19] constructs an interest graph by aggregating item-wise modality-specific similarity connections, and shares these connections across modalities to enhance structural consistency. However, the above methods still have certain limitations. DA-MRS only models low-order semantic relationships. Although it performs well in noise reduction, it lacks the capability to directly perceive high-order semantic information and is therefore limited in expanding the diversity of item connections. While SOIL effectively perceives more potential item connections, its approach of sharing connections between textual and visual modalities may introduce the issue described in DA-MRS, where highly related items in the visual modality are connected with completely unrelated items in the textual modality (e.g., a carpet with a Pikachu print and a painting). This could inadvertently introduce new noise into the modality graph.

Therefore, we propose a novel Interest **T**ree **A**ugmented **M**odality Graph Recommend**ER** for Multimodal Recommendation **(TAMER)**. Specifically, to capture high-order item interests while avoiding the introduction of additional noise, we introduce Interest Trees to enhance modality representation. In our approach, we define the set of items interacted with by a user as an interest set. By computing the relationships among all interest sets, we construct an interest graph. Based on this graph, we prune it into multiple Interest Trees using a breadth-first search (BFS) strategy. Each node in an Interest Tree is assigned a weight representing its confidence score relative to the root node. By incorporating these confidence scores as coefficients into the modality similarity matrix, we further enhance modality relationships. However, this method is partially dependent on the original distribution of modality features. WhitenRec [39] pointed out that pre-trained text embeddings exhibit an average cosine similarity as high as 0.8, which not only weakens the model's ability to differentiate features within the embedding space but also increases the risk of gradient explosion, compromising the stability and effectiveness of the recommendation model. To address this issue, we integrate various component analysis methods into multimodal recommendation. Specifically, we employ Independent Component Analysis (ICA) [8] and other analysis techniques to optimize the distribution of items in the modality feature space, eliminating redundant features and noise, ensuring a more uniform distribution of items while preserving key information as much as possible. This approach enhances the discriminative power of item features, ensuring more reliable multimodal representations.

In summary, our contributions are as follows:

- We propose a Multi-component Analysis Collaborative Processing (MACP) method that integrates multiple component analysis techniques to optimize modality feature distribution, enhancing item similarity discriminability within the homogeneous graph and mitigating the risk of gradient explosion during Interest Tree enhancement.
- We construct multiple Interest Trees based on user interest information and then utilize a multi-stage homogeneous graph enhancement method to provide high-order information for the modality homogeneous graph, improving its expressiveness while effectively avoiding noise introduction.
- We conduct extensive experiments on three real-world datasets, comparing our method against 20 baseline models. The results demonstrate that our approach achieves an average performance improvement of 9.98% over the best baseline model.

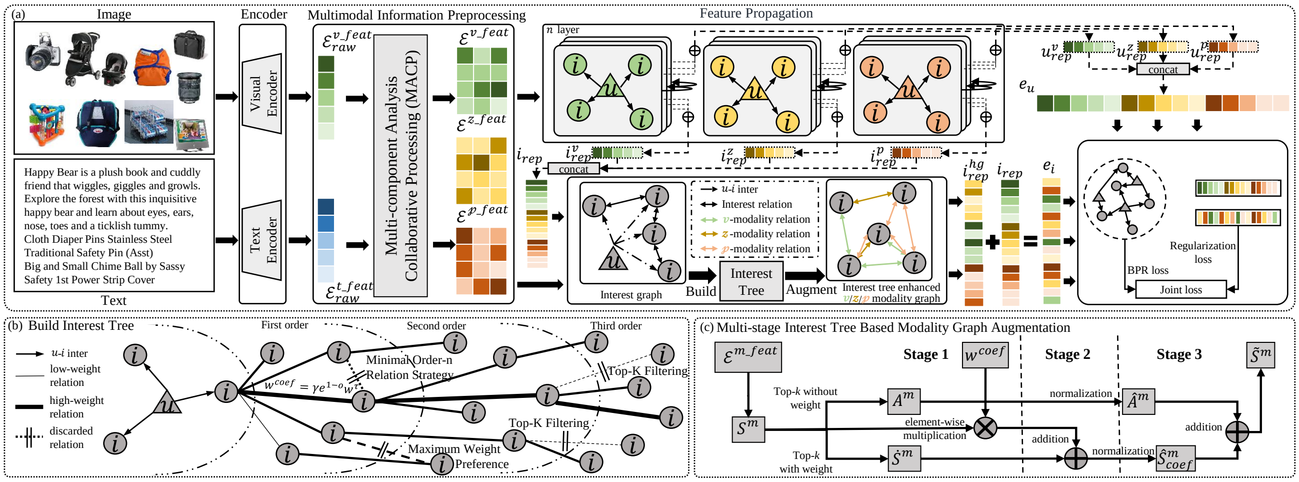


Figure 2: Overall framework of TAMER: (a) The overall workflow of the proposed method: TAMER first encodes multimodal features and then jointly learns on both homogeneous and heterogeneous graphs. (b) The process of constructing the Interest Tree. (c) The enhancement of the modality graph based on the Interest Tree.

2 Related work

2.1 Multimodal Recommendation Methods

The initial multimodal recommendation approaches directly applied visual signals to recommender systems. For example, VBPR [5] enriches item representation by extracting visual feature matrices through convolutional neural networks and concatenating these features directly with item IDs. With the rise of GNNs, some studies began leveraging GCNs to capture user preferences. For instance, MMGCN [29] constructs multiple modality-specific bipartite graphs to learn user preferences from different modalities and fuses these results for recommendations, achieving significant effects. GRCN [28] optimizes user-item interaction graphs using modality information, mitigating the negative impact of false positive edges in implicit feedback on recommendation performance. SLMRec [21] integrates self-supervised learning into GNN recommendation models, designing three training tasks from modality-agnostic and modality-specific perspectives to generate powerful representations. BM3 [45] uses contrastive learning to align item ID embeddings with latent representations of item modality features, resulting in more efficient recommendations. MGCN [35] purifies each modality's information using item IDs and constructs a gating unit for each modality to perceive user behaviors and integrate with modality information. AlignRec [12] proposes a multi-task alignment scheme that aligns multimodal feature representations through content alignment, content-category alignment, and user-item alignment, thereby bridging the semantic gap between multimodal content and item ID representations. SMORE [16] utilizes the spectral space for feature fusion, employing an adaptive filter to suppress noise while integrating a graph learning module to capture semantic relationships between items. MENTOR [31] enhances model robustness through a cross-modal alignment task and a feature augmentation strategy, ensuring better consistency and adaptability across different modalities. By integrating multimodal information into the recommendation system, significant performance improvements have been achieved.

2.2 Homogeneous Graph-Based Recommendation with Enhancement or Denoising

DualGNN [25] constructs a user-item bipartite graph and establishes a GCN for each modality to learn user preferences. Then, DualGNN builds a user co-occurrence homogeneous graph, propagating user preferences through the homogeneous graph to enhance the expressiveness of user features. LATTICE [37] and MICRO [38] propose a modality-aware structural learning network to dynamically discover item relationships within modalities and construct item homogeneous graphs based on these relationships. FREEDOM [44] suggests that using frozen item homogeneous graphs could provide better performance, using these frozen graphs to denoise the user-item bipartite graph, thereby significantly improving model performance. Building on DualGNN, DRAGON [40] proposes an attention concatenation mechanism that effectively integrates user preferences with item features and constructs similarity-based homogeneous graphs for each modality to learn dual representations of users and items. LGMRec [4] captures modality-related and collaborative user interests through a local graph embedding module, and models hyperedges in each modality's homogeneous graph using a global hypergraph embedding module, further exploring modality-specific item dependencies. SOIL [19] addresses the limitations of the "best deal matching principl" in certain scenarios by developing a secondary interest learning framework that utilizes the complementary relationships between different modalities to extend user interest connections in the user-item bipartite graph. GUME [11] introduces a user-item bipartite graph enhancement strategy based on modality similarity, using homogeneous graphs to improve the connectivity of long-tail items, effectively mitigating the long-tail issue and achieving excellent recommendation results. DA-MRS [33] suppresses noise by constructing a cross-modality consistent item graph and enhances representations through a dual alignment strategy, which incorporates both user preferences and hierarchical relationships. These methods effectively improve recommendation performance.

3 Method

In a typical multimodal recommendation system, the user set and item set are represented by $\mathcal{U} = \{u_1, u_2, ..., u_n\}$ and $\mathcal{I} = \{i_1, i_2, ..., i_k\}$, respectively. The modality information $m \in \mathcal{M} = \{v, t\}$ represents different modalities, where v and t denote visual and textual modalities, respectively. The historical interactions between users and items can be represented by an interaction matrix $R \in \{0, 1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where matrix element $r_{u,i} \in \{0, 1\}$ indicates the interaction between user u and item i: 1 indicates an interaction exists, and 0 indicates no interaction. Based on these interactions, we construct a bipartite graph $\mathcal{G} = \{V, E\}$, where $V = \{\mathcal{U} \cup \mathcal{I}\}$ and $E = \{(u, i) \mid r_{u,i} = 1\}$. For each user and item under modality m, we randomly initialize embedding vectors $\mathcal{E}^m_{id} \in \mathbb{R}^d$ and $\mathcal{E}^m_u \in \mathbb{R}^d$, where d is the embedding dimension, representing item ID and user embeddings, respectively. In addition, the original modality feature of an item is represented as $\mathcal{E}^{m\_feat}_{raw} \in \mathbb{R}^{d_m}$, where d_m is the feature dimension of modality m.

3.1 MACP

The overall framework is shown in Figure 2. First, we preprocess the distribution of multimodal data to prevent gradient explosion during Interest Tree enhancement. Currently, ZCA-based whitening methods for redistributing item features have demonstrated strong performance [39]. To further enhance the distinctiveness of nodes in the item-item graph and enable the Interest Tree to strengthen homogeneous relationships from different perspectives and weights, we employ the PCA [9], ICA [8] and ZCA [1] methods. Specifically, we process modality vectors to obtain $\mathcal{E}^{m\_feat}$, where $m \in \{v, p, z\}$:

$$\begin{aligned}
\mathcal{E}^{p\_feat} &= \mathrm{ICA}(\mathrm{PCA}(\mathcal{E}^{t\_feat}_{raw})), \\
\mathcal{E}^{z\_feat} &= \mathrm{ZCA}(\mathcal{E}^{t\_feat}_{raw}), \\
\mathcal{E}^{v\_feat} &= \mathcal{E}^{v\_feat}_{raw}.
\end{aligned} \tag{1}$$

3.2 Interest Tree Enhanced Homogeneous Graph

3.2.1 Modality Graph Construction. First, we initialize a similarity matrix S^m, where each element S^m_{ij} represents the similarity between item i and item j in modality $m \in \mathcal{M}' = \{v, p, z\}$. We use cosine similarity to calculate S^m_{ij} [44], as follows:

$$S^m_{ij} = \frac{(e^{m\_feat}_i)^\top (e^{m\_feat}_j)}{||(e^{m\_feat}_i)|| \, ||(e^{m\_feat}_j)||}, \tag{2}$$

where $e^{m\_feat}_i$ and $e^{m\_feat}_j$ represent the modality feature representations of item i and item j, respectively, in $\mathcal{E}^{m\_feat}$. We employ a k-nearest neighbors method to construct a modality-specific adjacency matrix $\mathcal{A}^m$ and a similarity matrix $\dot{S}^m$:

$$\mathcal{A}^m_{ij} = \begin{cases} 1, & \text{if} \quad S^m_{ij} \in \text{top-}k(S^m_i), \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

$$\dot{S}^m_{ij} = \begin{cases} S^m_{ij}, & \text{if} \quad S^m_{ij} \in \text{top-}k(S^m_i), \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

where $\mathcal{A}^m_{i,j} = 1$ indicates the presence of a connection between items i and j in the adjacency matrix $\mathcal{A}^m$, and $\dot{S}^m_{ij}$ retains the similarity value between items i and j. To ensure the robustness of graph convolution learning, we typically normalize the adjacency matrix $\mathcal{A}^m$ to obtain a normalized matrix $\hat{\mathcal{A}}^m$ as follows:

$$\hat{\mathcal{A}}^m = D^{-\frac{1}{2}} \mathcal{A}^m D^{-\frac{1}{2}}. \tag{5}$$

3.2.2 Interest Graph Construction. First, we initialize an item-item co-occurrence matrix S^c, where each element S_{ij} represents the number of times item i and item j appear in the same interest set. To sparsify this matrix while retaining valuable information, we prune the matrix S^c_{ij} using a top-k method to remove weakly related relationships, thereby avoiding noise introduced by random behaviors. Specifically, for each item i, we select the top-k items with the highest co-occurrence counts to generate the top-k list S^c_i. In the pruned interest matrix $\dot{S}^c$, each element $\dot{S}^c_{ij}$ is represented as:

$$\dot{S}^c_{ij} = \begin{cases} S^c_{ij}, & \text{if} \quad S^c_{ij} \in \text{top-}k(S^c_i), \\ 0, & \text{otherwise}. \end{cases} \tag{6}$$

3.2.3 Interest Tree Construction. The construction of the Interest Tree is described in Algorithm 1. Unlike other methods, we focus only on the high-order relationships and weights between items [26]. Its inputs include a weighted graph G, which in this paper corresponds to $\dot{S}^c_{ij}$, a starting node i, and a maximum order n, where n is set to 3 in this paper. The algorithm is based on a BFS strategy. It iteratively explores nodes layer by layer, storing nodes at each order o. The exploration process involves pruning based on a combination of the weight list and the top-k list, ensuring that only the most relevant nodes with optimal weights are retained as related nodes for node i. After computing up to n-th order relationships, the algorithm returns d_{T_i}, a weighted tree $T_i = \{N, w, o\}$ in dictionary form, where: N represents the indices of nodes related to i, w denotes the corresponding weights, o indicates the order of each relationship.

3.2.4 Multi-Stage Homogeneous Graph Augmentation. Previous works typically relied on modality-specific adjacency matrices $\hat{\mathcal{A}}^m$ for feature propagation in homogeneous item graphs. However, these approaches fail to capture semantic information effectively. Therefore, we first perform multi-stage homogeneous graph enhancement. Specifically, as shown in Figure 2(c), in the stage 1, for each item i, we generate a interest tree $T_i = \{N, w, o\}$ to enhance modality relationships. For each item $j \in N$, the weight coefficient is defined as:

$$w^{coef}_{ij} = \gamma e^{(1-o_{ij})} w^{\tau}_{ij}. \tag{7}$$

Here, γ and τ are hyperparameters, e is Euler's number. As a result, we obtain the confidence matrix $w^{coef} = [w^{coef}_{ij}] \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}_{\geq 0}$. In

Algorithm 1: Find Weighted n-order Relationships

```
Input: Weighted graph $G$, starting node $i$, maximum order $n$
Output: Dictionary $d_{T_i}$ storing nodes and weights for each
        order
$order\_dict \leftarrow \emptyset$,   $visited \leftarrow \{i\}$,
  $current\_level\_nodes \leftarrow [(i, 0)]$;
for $order \leftarrow 1$ to $n$ do
    $next\_level\_nodes \leftarrow \emptyset$;
    foreach $(node, \_) \in current\_level\_nodes$ do
        foreach $(neighbor, weight) \in G[node]$ do
            if $neighbor \notin visited$ then
                $visited \leftarrow visited \cup \{neighbor\}$;   Append
                  $(neighbor, weight)$ to $next\_level\_nodes$;
    if $next\_level\_nodes \neq \emptyset$ then
        if $order > 1$ then
            Sort $next\_level\_nodes$ by weight (desc);   Keep
              top $\lfloor \frac{|current\_level\_nodes|}{(order-1)\times 2} \rfloor$ elements;
        $order\_dict[order] \leftarrow next\_level\_nodes$;
    $current\_level\_nodes \leftarrow next\_level\_nodes$;
return $order\_dict$;
```

the stage 2, we then multiply the confidence-weighted coefficient with the similarity matrix of each modality:

$$S^m_{coef} = w^{coef} \odot S^m + \frac{\dot{S}^m}{2}, \tag{8}$$

where $\odot$ denotes element-wise multiplication.

Subsequently, we normalize S^m_{coef} to obtain $\hat{S}^m_{coef}$. Finally, in the stage 3, we fuse $\hat{S}^m_{coef}$ with $\hat{\mathcal{A}}^m$ to generate the interest tree-enhanced adjacency matrix:

$$\tilde{S}^m = \hat{S}^m_{coef} + \frac{\hat{\mathcal{A}}^m}{2}. \tag{9}$$

To further capture semantic relationships, we explicitly model the interest graph $\dot{S}^c$. Specifically, we reset each element in $\dot{S}^c_{ij}$ to obtain $\mathcal{A}^c_{ij} = \{1 \mid \dot{S}^c_{ij} > 0\}$. We then compute its normalized representation: $\tilde{\mathcal{S}}^c = D^{-\frac{1}{2}} \mathcal{A}^c D^{-\frac{1}{2}}$.

To unify these homogeneous graphs, we introduce a set of hyperparameters α_m to balance the contributions of each graph in the homogeneous graph. The final adjacency matrix of the multimodal homogeneous graph is computed as: $\tilde{S} = \sum_{hg \in \mathcal{M}''} \alpha_{hg} \times \tilde{S}^{hg}$, where $\mathcal{M}'' = \mathcal{M}' \cup \{c\}$.

3.3 Dual Graph Learning

Similar to existing methods, we perform dual representation learning on both homogeneous and heterogeneous graphs [40].

3.3.1 Heterogeneous Graph Learning. First, we embed the original modality features using modality-specific embedding functions to obtain embeddings for different modalities, as represented by the following equation:

$$\hat{\mathcal{E}}^{m\_feat} = W_2^m \left(leaky\_relu(W_1^m \mathcal{E}^{m\_feat} + b_1)\right) + b_2, \tag{10}$$

where $W_1^m \in \mathbb{R}^{4d \times d_m}$ and $W_2^m \in \mathbb{R}^{d \times 4d}$ are linear transformation matrices, and $b_1 \in \mathbb{R}^{4d}$ and $b_2 \in \mathbb{R}^d$ are bias terms.

Next, we purify the modality features using item IDs to obtain the embedded modality features:

$$\mathcal{E}^m_i = \mathcal{E}^m_{id} \odot \hat{\mathcal{E}}^{m\_feat}, \tag{11}$$

where $\odot$ denotes element-wise multiplication. We then use a bipartite graph $\mathcal{G}_m$ to learn representations in each modality m, preserving the graph structure $\mathcal{G} = \{V, E\}$ in each modality. Due to the computational efficiency of LightGCN [6], we apply it to heterogeneous graph learning, as represented by the following equations:

$$\begin{aligned} \mathcal{E}^{m(l+1)}_u &= \sum_{i \in V_u} \frac{1}{\sqrt{|V_u|}\sqrt{|V_i|}} \mathcal{E}^{m(l)}_i, \\ \mathcal{E}^{m(l+1)}_i &= \sum_{u \in V_i} \frac{1}{\sqrt{|V_u|}\sqrt{|V_i|}} \mathcal{E}^{m(l)}_u. \end{aligned} \tag{12}$$

where $V_u = \{i \in \mathcal{I} \mid r_{u,i} = 1\}$ and $V_i = \{u \in \mathcal{U} \mid r_{u,i} = 1\}$ represent the first-order neighbors of user u and item i in graph $\mathcal{G}$, respectively, and l denotes the propagation layer.

After L layers of propagation and aggregation, we fuse the representations from each layer using element-wise summation to obtain the final user representation $u^m_{rep} = \sum_{l=0}^{L} \mathcal{E}^{m(l)}_u$ and item representation $i^m_{rep} = \sum_{l=0}^{L} \mathcal{E}^{m(l)}_i$ for heterogeneous graph learning. We concatenate the modality-specific representations to generate the final representations for users and items in heterogeneous graph learning [40]:

$$\begin{aligned} u_{rep} &= [W^p \times u^p_{rep} \mid W^z \times u^z_{rep} \mid W^v \times u^v_{rep}], \\ i_{rep} &= [\, i^p_{rep} \mid i^z_{rep} \mid i^v_{rep} \,]. \end{aligned} \tag{13}$$

where [|] denotes concatenation, and $W^m, m \in \mathcal{M}'$, are learnable parameters used to weight different modalities.

3.3.2 Homogeneous Graph Learning. We employ graph convolution operations to capture relationships between items [44]:

$$i^{hg}_{rep} = \sum_{j \in \mathcal{N}_i} \tilde{S}_{ij} \times i_{rep}. \tag{14}$$

where $\mathcal{N}_i = \{ j \mid \tilde{S}_{ij} \neq 0 \}$ represents the set of neighboring items j for item i in the matrix $\tilde{S}^{hg}$.

3.3.3 Fusion and Prediction. We obtain the final user representation e_u and item representation e_i:

$$e_u = u_{rep}, \quad e_i = i_{rep} + i^{hg}_{rep}. \tag{15}$$

We compute the user's preference score for item i by using the inner product between the user representation e_u and the item representation e_i. To optimize the model parameters, we adopt Bayesian Personalized Ranking (BPR) [18] loss, which is formulated as:

$$\mathcal{L} = \sum_{u, i^+, i^- \in D} (-\ln \sigma(z_{u,i^+} - z_{u,i^-})) + \beta \, \|\Theta\|_2^2, \tag{16}$$

where $z_{u,i} = e_u \cdot e_i^\top$, each triplet (u, i^+, i^-) satisfies $(u, i^+) \in E$, $(u, i^-) \notin E$. β denotes the L2 regularization coefficient, and Θ represents the model parameters.

Table 1: Comparison of performance with general and multimodal recommendation systems. The best performance is highlighted in bold, while the second-best performance is indicated with an underline. "improv." represents the improvement of TAMER over the best baseline for each metric on each dataset. "-"indicates the model cannot be fitted into a NVIDIA 4090.

Dataset	Baby				Sports				Electronics			
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(UAI'09)	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0235	0.0367	0.0127	0.0161
LightGCN(SIGIR'20)	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0313	0.0387	0.0363	0.0540	0.0204	0.0250
LayerGCN(ICDE'23)	0.0529	0.0820	0.0281	0.0355	0.0594	0.0916	0.0323	0.0406	0.0391	0.0581	0.0220	0.0269
VBPR(AAAI'16)	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0293	0.0458	0.0159	0.0202
MMGCN(MM'19)	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0207	0.0331	0.0109	0.0141
DualGNN(TMM'21)	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0363	0.0541	0.0202	0.0248
GRCN(MM'20)	0.0532	0.0824	0.0282	0.0358	0.0559	0.0877	0.0306	0.0389	0.0349	0.0529	0.0195	0.0241
LATTICE(MM'21)	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	–	–	–	–
SLMRec(TMM'22)	0.0540	0.0810	0.0285	0.0357	0.0676	0.1017	0.0374	0.0462	0.0422	0.0630	0.0237	0.0291
BM3(WWW'23)	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0437	0.0648	0.0247	0.0302
MICRO(TKDE'22)	0.0584	0.0929	0.0318	0.0407	0.0679	0.1050	0.0367	0.0463	–	–	–	–
FREEDOM(MM'23)	0.0624	0.0985	0.0324	0.0416	0.0710	0.1077	0.0382	0.0476	0.0396	0.0601	0.0220	0.0273
MGCN(MM'23)	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0439	0.0643	0.0245	0.0298
DRAGON(ECAI'23)	0.0662	0.1021	0.0345	0.0435	0.0752	0.1139	0.0413	0.0512	0.0450	0.0678	0.025	0.0309
LGMRec(AAAI'24)	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0440	0.0665	0.0244	0.0303
SOIL(MM'24)	0.0680	0.1028	0.0365	0.0454	0.0786	0.1155	0.0435	0.0530	0.0492	0.0718	0.0279	0.0337
GUME(CIKM'24)	0.0673	0.1042	0.0365	0.0460	0.0778	0.1165	0.0427	0.0527	0.0458	0.0680	0.0253	0.0310
AlignRec(CIKM'24)	0.0674	0.1046	0.0363	0.0458	0.0758	0.1160	0.0414	0.0517	0.0472	0.0700	0.0262	0.0321
SMORE(WSDM'25)	0.0680	0.1035	0.0365	0.0457	0.0762	0.1142	0.0408	0.0506	0.0437	0.0650	0.0244	0.0299
MENTOR(AAAI'25)	0.0678	0.1048	0.0362	0.0450	0.0763	0.1139	0.0409	0.0511	0.0439	0.0655	0.0244	0.0300
TAMER	**0.0722**	**0.1123**	**0.0395**	**0.0498**	**0.0867**	**0.1272**	**0.0481**	**0.0585**	**0.0553**	**0.0808**	**0.0313**	**0.0379**
improv.	6.17%	7.15%	8.21%	8.26%	10.30%	9.18%	10.57%	10.37%	12.39%	12.53%	12.18%	12.46%

Table 2: Statistical overview of the datasets.

Dataset	Users	Items	Interactions	Sparsity
Baby	19445	7050	160792	99.88%
Sports	35598	18357	296337	99.95%
Electronics	192403	63001	1689188	99.98%

4 Experiment

We perform extensive empirical studies on three public datasets in order to investigate the following Research Questions (**RQ**):

- **RQ1**: How does the proposed method recommendation performance compared to existing SOTA approaches?
- **RQ2**: How does each module impact performance?
- **RQ3**: How do hyper-parameters affect performance?
- **RQ4**: How does the proposed method compare to others in expanding user interests and reducing noise?
- **RQ5**: Why does using MACP and Interest Tree lead to better recommendation performance?

4.1 Experimental Settings

Our method is evaluated through extensive testing on three well-known datasets from Amazon reviews datasets [15]. The selected categories are: (a) Baby, (b) Sports and Outdoors (denoted as Sports), and (c) Electronics. Each of these datasets contains item-related visual images and textual descriptions. We employ pre-extracted features: 4,096-dimensional features for visual data, and textual features are extracted using a pre-trained sentence-transformers [17]. In our study, all items and users are filtered using a 5-core setting. The statistics of the datasets are summarized in Table 2.

4.1.1 Baselines. To evaluate the effectiveness of our proposed model, we compare it with two categories of baseline models: General Models, which rely solely on interaction data, and Multimedia Models, which leverage multi-modal features.

General Models: Matrix factorization model BPR [18] and graph-based model LightGCN [6], LayerGCN [42].

Multi-modal Models: Several STOA method have been selected for comparison, including VBPR [5], MMGCN [29], DualGNN [25], GRCN [28], LATTICE [37], SLMRec [21], BM3 [45], MICRO [38], FREEDOM [44], MGCN [35], DRAGON [40], LGMRec [4], SOIL [19], GUME [11], AlignRec [12], SMORE [16], MENTOR [31].

4.1.2 Evaluation Metrics. To ensure fair and consistent comparisons, we adopt the evaluation setup used in prior work [19], randomly partitioning each user's interaction history into training, validation, and test sets in an 8:1:1 ratio. The optimal model is chosen according to the highest Recall@20 achieved on the validation set. The average performance across all users in the test set is evaluated using Recall@10, Recall@20, NDCG@10, and NDCG@20.

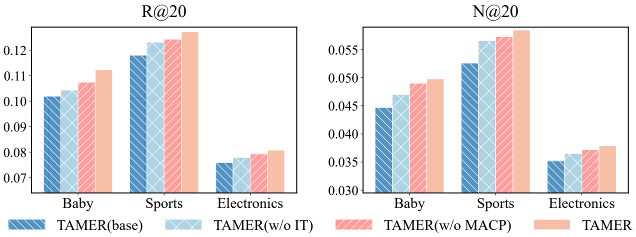


Figure 3: Results of the ablation study.

4.1.3 Implementation Details. We implement our proposed method based on the MMRec [41] framework. To ensure a fair comparison, we adopt the optimal hyperparameter configurations as reported in the original baseline studies. For general settings, we use the Adam [2] optimizer to optimize all methods and initialized embedding vectors of size 64 using Xavier [3] initialization. The number of GCN layers L is set to 2 for user-item graphs and 1 for item-item graphs. The k for top-k is set to 10. Training is conducted on a single NVIDIA RTX 4090 GPU. The hyper-parameter learning rate and regularization weight are chosen from {1e-1, 1e-2, 1e-3, 1e-4, 1e-5}. The values of γ and τ are selected from {0.1, 0.2, 0.3} and {0.6, 0.7, 0.8}, respectively.

4.2 Experiment Result

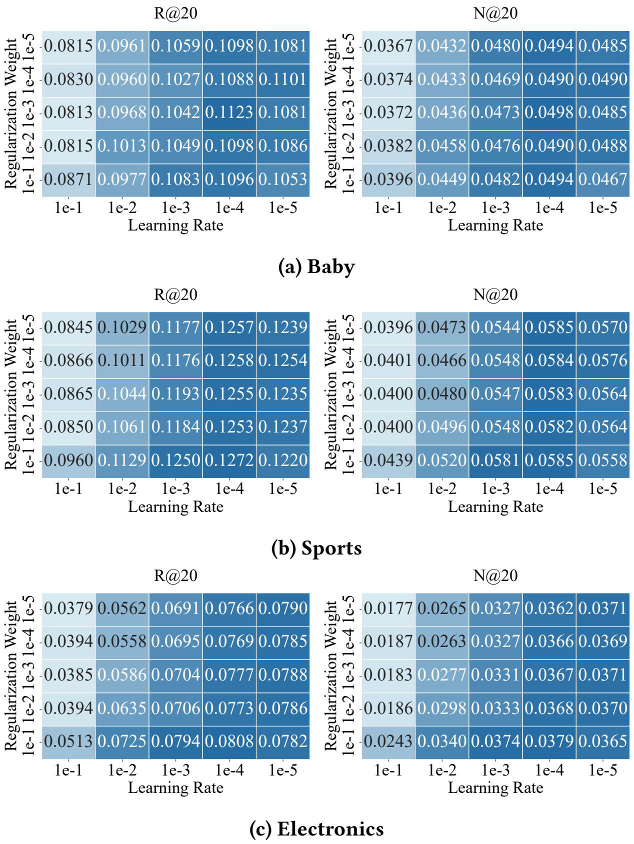
4.2.1 Overall Performance (RQ1). To evaluate the effectiveness of our proposed method, we perform comparative experiments against multiple traditional and multimodal models on three widely used real-world datasets. The experimental results are reported in Table 1, where our method outperforms all previous models across all evaluation metrics, achieving SOTA performance.

Among all results, our method surpasses previous SOTA multimodal models by relative performance gains of 7.44%, 10.10%, and 12.39% (mean 9.98%) on the Baby, Sports, and Electronics datasets, respectively. These results strongly validate the effectiveness of our approach.

Among all comparison methods, MENTOR and TAMER share a similar fundamental network structure. However, SOIL demonstrates superior performance over MENTOR, as the ability to perceive user interest information effectively expands the range of potentially relevant items, thereby yielding improved recommendation results. Our method surpasses MENTOR by 8.35%, 14.34%, and 25.98%, and improves over SOIL by 8.33%, 10.34%, and 12.39% on three datasets. This shows our approach effectively enhances modality features with higher-order user interests and reduces noise, significantly boosting performance.

4.2.2 Ablation Study (RQ2). To investigate the contributions of individual components in our framework, we conduct ablation experiments on three datasets. The experimental results are illustrated in Figure 3. We design the following model variants:

- TAMER (base): Both the modality graph enhanced by the Interest Tree and the MACP module are removed.
- TAMER (w/o IT): The modality graph enhanced by the Interest Tree is removed, retaining only the MACP module.



(a) Baby

(b) Sports

(c) Electronics

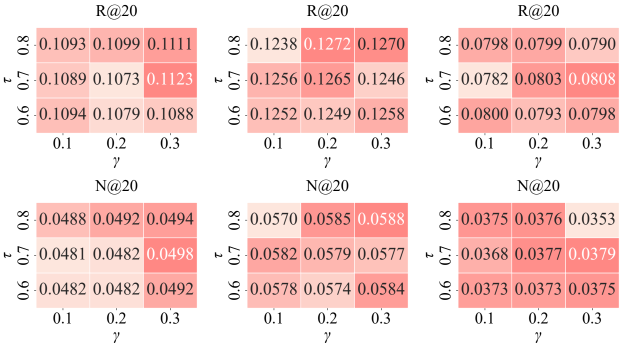
Figure 4: The performance impact of learning rate and regularization weight (darker hues denote better performance).

- TAMER (w/o MACP): The MACP module is removed, retaining only the Interest Tree-enhanced modality graph.
- TAMER: The full model with all components included.

Experimental results show that the removal of any single component from TAMER leads to a performance drop, indicating that each component plays a vital role in enhancing the model's recommendation effectiveness.

4.2.3 Effect of Learning Rate and Regularization Weight (RQ3). We set the search range for TAMER's learning rate and regularization weight to {1e-1, 1e-2, 1e-3, 1e-4, 1e-5}, and the experimental results on the three datasets are presented in Figure 4. For the Baby, Sports, and Electronics datasets, the optimal learning rate and regularization weight are found to be (1e-4, 1e-3), (1e-4, 1e-1) and (1e-4, 1e-1), respectively. These results indicate that, across most datasets, the optimal learning rate is 1e-4, while the choice of regularization weight depends on the learning rate. Specifically, as the learning rate decreases, reducing the regularization weight tends to yield better performance.

4.2.4 Effect of γ and τ (RQ3). The hyperparameters γ and τ jointly control the enhancement coefficient of the Interest Tree. Excessively large values of γ and τ may lead to gradient explosion, while overly small values may weaken the enhancement effect of the



(a) Baby **(b) Sports** **(c) Electronics**

Figure 5: The performance impact of γ and τ.

Interest Tree. To balance these effects, we set the search range for γ to $\{0.1, 0.2, 0.3\}$ and for τ to $\{0.6, 0.7, 0.8\}$, and conducted hyperparameter tuning. As shown in Figure 5. The results across the three datasets suggest that the optimal values for γ and τ are 0.3 and 0.7 for the Baby and Electronics datasets, whereas for the Sports dataset, the best performance is achieved when $\gamma = 0.2$ and $\tau = 0.8$.

4.2.5 Performance Comparison with Integrated Methods (RQ4). As previously discussed, while SOIL is capable of perceiving user interest preferences, the shared candidate item sets from different modality graphs can lead to the incorrect linking of items with similar visual features but completely different usages, resulting in noise. DA-MRS addresses this issue through a denoising approach. A straightforward solution to this problem is to combine the two models. To evaluate this, we conduct a set of comparative experiments. As shown in Table 3, we create several variants, where TAMER (base) serves as the baseline model from the ablation experiments, and * indicates a model component rather than the full model. SOIL* refers to the interest-aware graph component in SOIL that extends user interests, while DA-MRS* represents the denoising item-item graph component from DA-MRS. We conduct experiments by adding SOIL*, DA-MRS*, and both SOIL* and DA-MRS* to the baseline model.

The experimental results on the Baby and Sports datasets indicate that both SOIL* and DA-MRS* effectively enhance recommendation performance, suggesting the importance of perceiving user interests and the presence of noise connections in the interest modality graph. After combining the two methods, the performance of the model is further improved, but TAMER still demonstrates stronger overall performance. This highlights the effectiveness of our proposed approach, which not only mines users' higher-order interests through the user interest graph but also effectively avoids noisy connections.

4.2.6 Visualization Analysis (RQ5). We randomly select 1,000 data points from $\mathcal{E}^{t\_feat}_{raw}$, $\mathcal{E}^{z\_feat}$, and $\mathcal{E}^{p\_feat}$ in the Baby dataset and project their representations into a 2D space using t-SNE [23]. We then visualize the 2D feature distribution using Gaussian Kernel Density Estimation (KDE) [22], as shown in Figure 6. The original modality representations on the leftmost side exhibit multiple uneven unimodal patterns, which reduce item distinguishability and consequently degrade recommendation preferences [35]. MACP enhances the uniformity of text representations, thereby influencing recommendation performance [24].

Table 3: Comparison experiment between TAMER and TAMER(base)+SOIL*+DA-MRS*.

Dataset	Baby		Sports	
Metrics	R@20	N@20	R@20	N@20
TAMER(base)	0.1019	0.0447	0.1179	0.0526
+SOIL*	0.1027	0.0461	0.1187	0.0531
+DA-MRS*	0.1044	0.0468	0.1186	0.0539
+SOIL*+DA-MRS*	0.1051	0.0472	0.1217	0.0545
TAMER	**0.1123**	**0.0498**	**0.1272**	**0.0585**

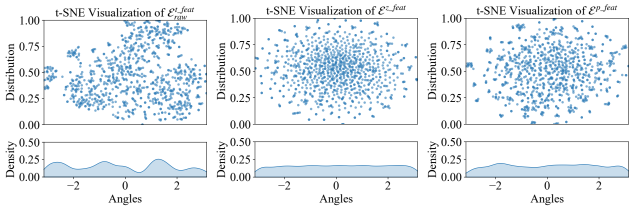


Figure 6: The distribution of representations in text modality.

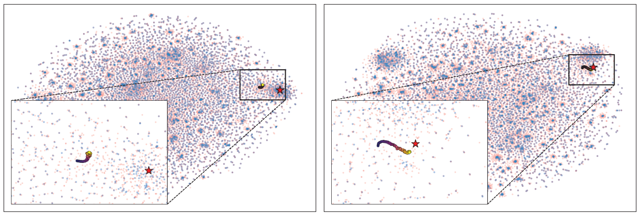


Figure 7: Visualizing embeddings during the training process.

Furthermore, we visualize the embedding trajectory of a specific item during training, as shown in Figure 7. User nodes are depicted in Soft Salmon and item nodes in Lake Blue. The trajectory illustrates the evolution direction of the item's embedding during training. The pentagram denotes the target users who interacted with this item in the test set. The left side shows the case without Interest Tree, while the right side includes Interest Tree. It can be observed that with Interest Tree, the item embedding effectively perceives the preferences learned by users, guiding its learning direction toward user embeddings and thereby facilitating more effective recommendations.

5 Conclusion

This work proposes TAMER, an Interest Tree Augmented Modality Graph Recommender for multimodal recommendation. It captures higher-order user interests in homogeneous graphs while reducing noise. An Interest Tree expands potential user interests, and a MACP module refines modality feature distribution. Experiments on real-world datasets show that TAMER significantly outperforms SOTA methods.

Acknowledgments

Support for this project was provided by the Science and Technology Project of State Grid Corporation of China, grant number 5108-202218280A-2-380-XG (The project's ERP system number is 520625220031).

References

[1] Anthony J Bell and Terrence J Sejnowski. 1997. The "independent components" of natural scenes are edge filters. *Vision research* 37, 23 (1997), 3327–3338.

[2] Kingma Diederik. 2014. Adam: A method for stochastic optimization. *(No Title)* (2014).

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

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

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

[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] 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.

[8] Christian Jutten and Jeanny Herault. 1991. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal processing* 24, 1 (1991), 1–10.

[9] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. 2018. Optimal whitening and decorrelation. *The American Statistician* 72, 4 (2018), 309–314.

[10] Guohui Li, Zhiqiang Guo, Jianjun Li, and Chaoyang Wang. 2022. Mdgcf: Multi-dependency graph collaborative filtering with neighborhood-and homogeneous-level dependencies. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 1094–1103.

[11] 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.

[12] Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. 2024. AlignRec: Aligning and Training in Multimodal Recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1503–1512.

[13] Haokai Ma, Yimeng Yang, Lei Meng, Ruobing Xie, and Xiangxu Meng. 2024. Multimodal conditioned diffusion model for recommendation. In *Companion Proceedings of the ACM Web Conference 2024*. 1733–1740.

[14] Kelong Mao, Jieming Zhu, Xi Xiao, Biao Lu, Zhaowei Wang, and Xiuqiang He. 2021. UltraGCN: ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 1253–1262.

[15] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[16] Rongqing Kenneth Ong and Andy WH Khong. 2024. Spectrum-based Modality Representation Fusion Graph Convolutional Network for Multimodal Recommendation. *arXiv preprint arXiv:2412.14978* (2024).

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

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

[19] Hongzu Su, Jingjing Li, Fengling Li, Ke Lu, and Lei Zhu. 2024. SOIL: Contrastive Second-Order Interest Learning for Multimodal Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5838–5846.

[20] Jianing Sun, Yingxue Zhang, Wei Guo, Huifeng Guo, Ruiming Tang, Xiuqiang He, Chen Ma, and Mark Coates. 2020. Neighbor interaction aware graph convolution networks for recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1289–1298.

[21] 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.

[22] George R Terrell and David W Scott. 1992. Variable kernel density estimation. *The Annals of Statistics* (1992), 1236–1265.

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

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

[25] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* 25 (2021), 1074–1084.

[26] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

[27] Wei Wei, Jiabin Tang, Lianghao Xia, Yangqin Jiang, and Chao Huang. 2024. Promptmm: Multi-modal knowledge distillation for recommendation with prompt-tuning. In *Proceedings of the ACM Web Conference 2024*. 3217–3228.

[28] 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.

[29] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 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.

[30] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. 2022. Graph neural networks in recommender systems: a survey. *Comput. Surveys* 55, 5 (2022), 1–37.

[31] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C-H Ngai. 2024. Mentor: multi-level self-supervised learning for multimodal recommendation. *arXiv preprint arXiv:2402.19407* (2024).

[32] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. 2025. A Survey on Multimodal Recommender Systems: Recent Advances and Future Directions. *arXiv preprint arXiv:2502.15711* (2025).

[33] Guipeng Xv, Xinyu Li, Ruobing Xie, Chen Lin, Chong Liu, Feng Xia, Zhanhui Kang, and Leyu Lin. 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.

[34] 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.

[35] 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.

[36] Jinghao Zhang, Guofan Liu, Qiang Liu, Shu Wu, and Liang Wang. 2024. Modality-Balanced Learning for Multimedia Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7551–7560.

[37] 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.

[38] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9154–9167.

[39] Lingzi Zhang, Xin Zhou, Zhiwei Zeng, and Zhiqi Shen. 2024. Are id embeddings necessary? whitening pre-trained text embeddings for effective sequential recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 530–543.

[40] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. 2023. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In *ECAI 2023*. IOS Press, 3123–3130.

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

[42] Xin Zhou, Donghui Lin, Yong Liu, and Chunyan Miao. 2023. Layer-refined graph convolutional networks for recommendation. In *2023 IEEE 39th international conference on data engineering (ICDE)*. IEEE, 1247–1259.

[43] Xin Zhou and Chunyan Miao. 2024. Disentangled graph variational auto-encoder for multimodal recommendation with interpretability. *IEEE Transactions on Multimedia* (2024).

[44] 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.

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