Refining Contrastive Learning and Homography Relations for Multi-Modal Recommendation

Shouxing Ma
University of Technology Sydney
Sydney, Australia
shouxingmaa@gmail.com

Yawen Zeng
Hunan University
Changsha, China
yawenzeng11@gmail.com

Shiqing Wu*
Faculty of Data Science
City University of Macau
Macau SAR, China
sqwu@cityu.edu.mo

Guandong Xu*
The Education University of Hong Kong
Hong Kong SAR, China
gdxu@eduhk.hk

Abstract

Multi-modal recommender system focuses on utilizing rich modal information (i.e., images and textual descriptions) of items to improve recommendation performance. The current methods have achieved remarkable success with the powerful structure modeling capability of graph neural networks. However, these methods are often hindered by sparse data in real-world scenarios. Although contrastive learning and homography (i.e., homogeneous graphs) are employed to address the data sparsity challenge, existing methods still suffer two main limitations: 1) Simple multi-modal feature contrasts fail to produce effective representations, causing noisy modal-shared features and loss of valuable information in modal-unique features; 2) The lack of exploration of the homograph relations between user interests and item co-occurrence results in incomplete mining of user-item interplay.

To address the above limitations, we propose a novel framework for **RE**fining multi-mod**A**l cont**R**astive learning and ho**M**ography relations (**REARM**). Specifically, we complement multi-modal contrastive learning by employing meta-network and orthogonal constraint strategies, which filter out noise in modal-shared features and retain recommendation-relevant information in modal-unique features. To mine homogeneous relationships effectively, we integrate a newly constructed user interest graph and an item co-occurrence graph with the existing user co-occurrence and item semantic graphs for graph learning. The extensive experiments on three real-world datasets demonstrate the superiority of REARM to various state-of-the-art baselines. Our visualization further shows an improvement made by REARM in distinguishing between modal-shared and modal-unique features. Code is available here.

CCS Concepts

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

Keywords

Multi-modal Recommendation, Contrastive Learning, Graph Neural Network

ACM Reference Format:
Shouxing Ma, Yawen Zeng, Shiqing Wu, and Guandong Xu. 2025. Refining Contrastive Learning and Homography Relations for Multi-Modal 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/10.1145/3746027.3755779

*Corresponding Authors.

MM '25, October 27–31, 2025,Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755779

1 Introduction

Recommender systems have become indispensable tools in contemporary e-commerce for discovering items of interest to users based on their preferences and behaviors [6, 33]. Technological advances make it easier for users to access a wealth of multi-modal information about items, such as images, texts, and videos. By exploiting these rich contents, multi-modal recommender systems could obtain more accurate inferences about user interests or preferences compared to general recommender systems [44, 49, 57].

Early researchers integrate multi-modal information into the classical collaborative filtering (CF) framework by either concatenating or summing multi-modal information [13, 27]. Recent approaches [8, 56] consider the graph structures for user-item interactions and multi-modal information, and explore potential higher-order connectivity with the help of graph neural networks (GNNs). Current state-of-the-art multi-modal recommender systems require high-quality supervised data to achieve optimal performance. However, observed interactions are extremely sparse in real e-commerce scenarios, compared to the entire interaction space [12]. This sparsity hinders the model's ability to learn efficiently, thereby limiting the performance of recommender systems [20, 24, 45].

Inspired by the success of Self-Supervised Learning (SSL), researchers address data sparsity by creating self-supervised signals with multi-modal information [47, 50]. Unlike in computer vision [47] and natural language processing [19]—where SSL trains models for downstream tasks—most multi-modal recommendations [8, 57] construct self-supervised signals by contrasting features for the final representation. For instance, DiffMM [17] employs cross-modal contrastive learning to align multi-modal information and reduce noise, improving user preference learning. While exploring SSL, other efforts [41, 49, 50] explore multi-modal information

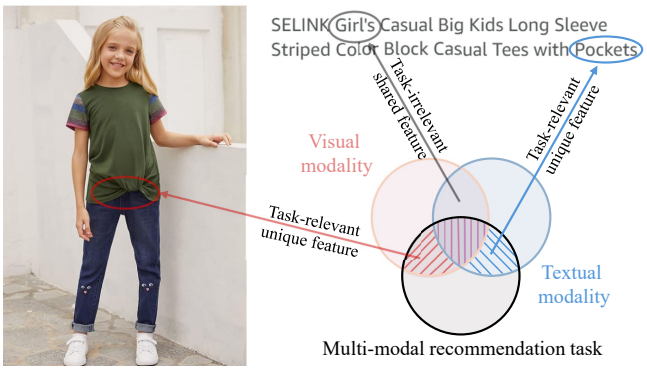

Figure 1: Illustration of contrastive learning for modalities in the multi-modal recommendation. Obtaining modal-shared features only by contrastive learning ignores valuable modal-unique features, and risks modal-shared features containing noise (irrelevant for the multi-modal recommendation task).

and historical interactions to create homogeneous graphs, to alleviate the issue of sparse data. For example, DRAGON [54] combines dual representations from both heterogeneous and homogeneous graphs (i.e., user co-occurrence graph & item semantic graph). Although current studies show promising results in multi-modal recommendations, they still face two significant limitations: **1) Simple contrasting multi-modal features leads to incomplete user and item representations.** Directly fusing multi-modal features fails to take into account unique and shared information is insufficient [7, 22, 34, 39]. On the one hand, unique features may be lost when aligning for consistency with contrastive learning [23, 36]. For the example illustrated in Figure 1, the model can well maintain the consistency features through multi-modal contrastive learning, i.e., modal-shared features, in the overlapping regions of visual and textual modalities. However, there are modal-unique features that are simultaneously relevant to the multi-modal recommendation task. For instance, a fashionable twist design might be evident in the image but not in the text, or "pockets" mentioned in the text might not be clear from the image alone in Figure 1. Missing discriminative modal-unique information leads to suboptimal recommendation performance. Preserving valuable modal-unique features when exploring modal consistency is essential for achieving high-quality recommendations. On the other hand, not all modal-shared features extracted by the model are useful; some may be noisy, irrelevant, or even misleading [22, 39]. For example, an item in Figure 1 is identified as a girl's shirt by both image and text, but it could also be for boys based on the "Big Kids" tag. Insufficient and noisy feature representations may substantially degrade the performance of the model. Hence, the ability to retain salient modal-unique features and eliminate modal-shared noise is vital for the effectiveness of multi-modal self-supervised models. **2) Neglect of user interest and item co-occurrence relationships.** A substantial amount of informative signals are embedded within user interest patterns and item co-occurrence relationships, which can be leveraged to facilitate a deeper understanding of user behaviors and item semantics [4, 21, 28, 51]. However, existing methods primarily focus on user co-occurrence and item semantic relationships while overlooking the associations between user interests and item co-occurrence patterns [50, 54, 56]. Hence, it is crucial to incorporate both co-occurrence and semantic (interest) relationships of users and items to enhance model representation capabilities.

To address the above issues, we propose a novel framework for **RE**fining multi-mod**A**l cont**R**astive learning and latent ho**M**ography relations (**REARM**). Our overall approach (Figure 2) comprises three core components: homography relation learning, heterography relation learning, and refining contrastive learning. First, we utilize the existing interactions and multi-modal information to additionally construct an item co-occurrence graph and user interest graph based on previous studies [47, 54]. The item semantic graph and the user co-occurrence graph, together, form user and item homogeneous graphs, respectively. These graphs enable the model to explore potential relationships among users and items from different perspectives of structure and semantics (interest). Next, we explore potential higher-order interactions from the interaction graph for each modality independently. We follow previous studies [8, 17] to introduce an auxiliary task to alleviate the data sparsity, i.e., we contrast multi-modal features to generate self-supervised signals. To further filter out noises in modal-shared features and preserve the modal-unique and recommendation-relevant information after contrasting, we refine the multi-modal contrastive learning by introducing the meta-network and orthogonal constraint techniques. The former leverages customized transformation matrices to extract recommendation-relevant information from modal-shared features to filter out noise, while the latter utilizes the orthogonal constraint loss function to encourage multi-modal features to retain modal-unique information. We validate the effectiveness of our framework on three public datasets, and our experimental results demonstrate distinct advantages of our model. Furthermore, we visualize the difference in the probability of interaction before and after refining the contrasting multi-modal features, which clearly shows the superiority of our method in distinguishing between modal-shared features and modal-unique features.

Our main contributions are summarized as follows.

- We propose a novel multi-modal contrastive recommendation framework (REARM), which preserves recommendation-relevant modal-shared and valuable modal-unique information through meta-network and orthogonal constraint strategies, respectively.
- We jointly incorporate co-occurrence and similarity graphs of users and items, allowing more effective capturing of the underlying structural patterns and semantic (interest) relationships, thereby enhancing recommendation performance.
- Extensive experiments are conducted on three publicly available datasets to evaluate our proposed method. The experimental results show that our proposed framework outperforms several state-of-the-art recommendation baselines.

2 Related Work

2.1 Multi-modal Recommendation

Due to the sparse interaction of recommender systems in the real world, numerous studies [41, 44] have leveraged multi-modal data about users or items (e.g., images, descriptions, or reviews). Earlier

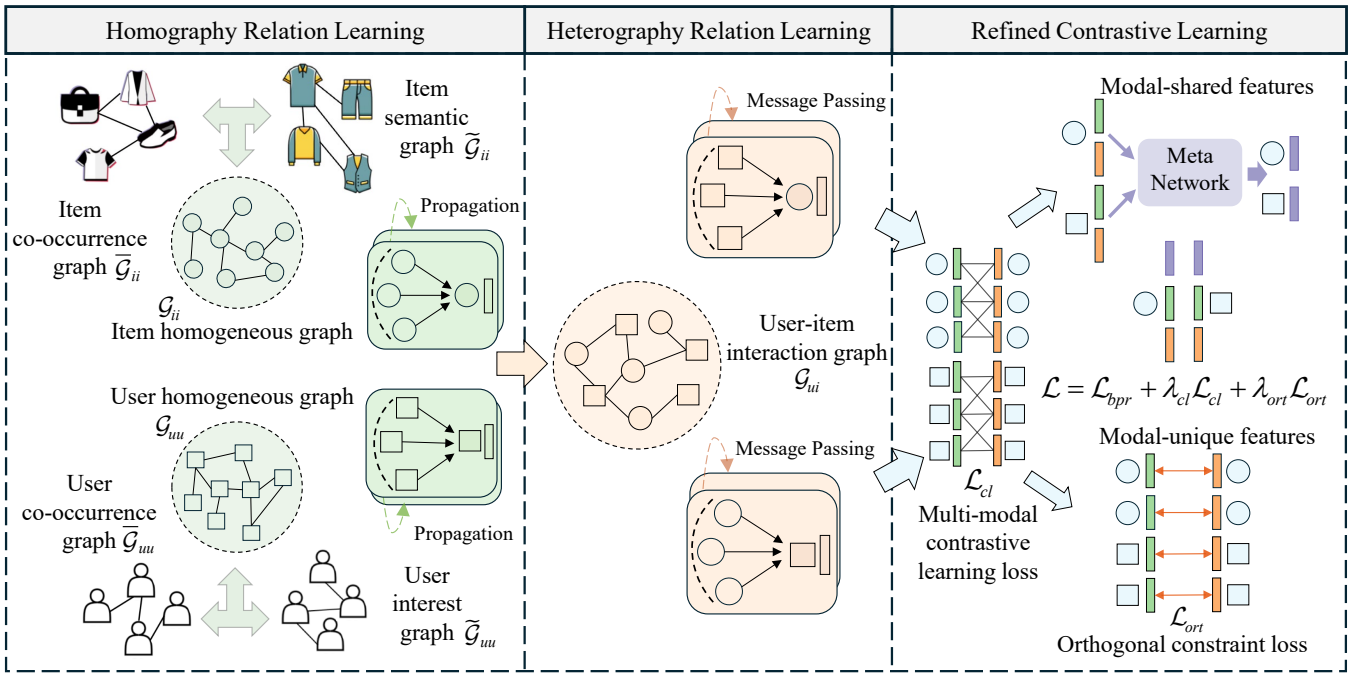

Figure 2: The structure overview of the proposed REARM.

work, such as VBPR [13], directly splices visual features from items into the vanilla CF framework. Later, GNNs shown to be capable of exploring potential higher-order interactions are introduced and largely utilized [44, 47]. Specifically, LGMRec [8] combines local graph and global hypergraph embedding modules to enhance representations. Recently, SSL has received massive attention for the benefits of non-addition of data [46]. For instance, DiffMM [17] improves representations by contrasting multi-modal features.

2.2 GNN-based Recommendation

The powerful higher-order connectivity of GNNs further extends the expressive power of CF models [42]. Along this line, LightGCN [14] further investigates the importance of each component in GNNs, and experimentally removes feature transformation and non-linear activation modules to better fit the recommendation. Later on, GNN-based methods are widely adopted as an essential component of various recommendations [3, 44]. For example, DRAGON [54] proposes a dual graph learning representation, which incorporates the user co-occurrence graph and the item semantic graph.

2.3 Contrastive Learning for Recommendation

Contrastive learning has been widely practiced in various fields [5, 19], since it benefits from generating self-supervised signals based on the original data. This advantage also renders it a powerful tool in tackling the inherent data sparsity issue in recommender systems [3, 24]. More specifically, SLMRec [35] performs dropout and masking operations on multi-modal features to construct additional self-supervised signals. Also, similar multi-modal contrastive learning strategies are employed in DiffMM [17] to capture and utilize the modality-related consistency information.

3 Theoretical Guarantee

3.1 Multi-modal Contrastive Learning

Let X_1 and X_2 denote the different data modalities, while Z and Y denote the latent variable and the task, respectively. Multi-modal contrastive learning as a universal framework yields a latent variable Z by maximizing the mutual information $I(X_1, X_2)$ to satisfy task Y [22]. However, this strategy is only effective under the assumption of multi-view redundancy [7, 23, 39].

Definition 1 (Multi-view redundancy). $\exists \varepsilon > 0$, such that $I(X_1; Y|X_2) \leq \varepsilon$ and $I(X_2; Y|X_1) \leq \varepsilon$.

This assumption indicates that most of the task-relevant information is shared and that the unique information is as small as ε. While multi-view redundancy is correct for particular types of multi-modal data, valuing unique information as well as ignoring redundant irrelevant information is also critical [23, 34, 39]. Furthermore, Liang et al. [22] factorize task-relevant information to shared and unique information with separate optimizations. The same applies to multi-modal recommendations, and simple contrasting multi-modal features leads to incomplete representations. Hence, considering multi-modal contrastive learning, we further filter out recommendation-irrelevant information in modal-shared and retain recommendation-relevant information in modal-unique.

3.2 Orthogonal Constraint

A matrix, $\mathbf{W}$, is orthogonal if $\mathbf{W}^\top\mathbf{W} = \mathbf{W}\mathbf{W}^\top = \mathbf{I}$. Enforcing an orthogonality constraint between matrices could penalize and reduce the redundant information between them [32, 38, 40]. Moreover, soft orthogonal constraint loss is defined to approximate orthogonality efficiently [1, 9, 26]. Similar to previous work [31, 52], we

utilize orthogonal constraint loss to encourage non-redundancy retention between modals, and further incorporate classical recommendation loss to preserve valuable modal-unique information.

4 Methodology

4.1 Preliminaries

We conceptualize the user-item interactions as a bipartite graph $\mathcal{G}_{ui} = \{(u, i, e)|u \in \mathcal{U}, i \in \mathcal{I}, e \in \mathcal{E}\}$, where $\mathcal{U}$ denotes the set of N users, $\mathcal{I}$ denotes the set of M items, and $\mathcal{E}$ denotes the set of edges representing the observed interactions. Furthermore, we denote the multi-modal information derived for each item in pre-trained models as $m \in \mathcal{M}$. The modality feature of the item i is represented as $\mathbf{i}_m \in \mathbb{R}^{d_m}$, where d_m denotes the feature dimension. The modal features (interest preferences) of the user u can be easily calculated as the mean of the modal features of all items observed interacting with him/her, i.e., $\mathbf{u_m} = \frac{1}{|\mathcal{I}_u|}\sum_{i\in\mathcal{I}_u}\mathbf{i}_m$, where $\mathcal{I}_u$ denotes the set of all items that the observed the user u has interacted with. In this work, we consider only two mainstream modalities, visual modality v and textual modality t, i.e., $\mathcal{M} = \{v, t\}$. Nevertheless, our model could be easily extended to scenarios with more than two modalities. The goal of this work is to predict unobserved interactions between users and items given multi-modal and interaction information.

4.2 Homography Relation Learning

Prior work [46, 47, 50] has shown that mining potential inter-item associations could further enrich item representations. Moreover, other research [41, 54] verifies the effectiveness of leveraging user-side co-occurrence for performance gains. However, potential associations between user interests and item co-occurrence patterns remain under-explored. Hence, we jointly explore co-occurrence and similarity relations from both user and item perspectives.

For clarity, we illustrate the process using item examples, as user and item relationships are built similarly. Notably, homography is constructed before training, thus avoiding additional training costs. Moreover, data sparsity further alleviates computational overhead.

4.2.1 Item Co-occurrence Graph Construction. Similar to establishing user co-occurrence relationships [54], we construct the item co-occurrence graph $\bar{\mathcal{G}}_{ii} = \{\mathcal{I}, \bar{\mathcal{C}}_{ii}\}$, where $\bar{\mathcal{C}}_{ii} = \{\bar{e}_{i,i'}|i, i' \in \mathcal{I}\}$ denotes the set of edges $\bar{e}_{i,i'}$ between item i and item i' in $\bar{\mathcal{G}}_{ii}$. We reserve the items with top-k common interactions among items and take the number as the weight value, which can be denoted as

$$\bar{e}_{i,i'} = \begin{cases} \bar{e}_{i,i'}, & \text{if } i' \in top\text{-}k(i), \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Furthermore, as with [41] handling co-occurrence, the softmax function is adopted to differentiate among item contributions.

4.2.2 Item Semantic Graph Construction. Analogous to studies [50, 54], we construct the modality-aware item semantic graph $\tilde{\mathcal{G}}_{ii}^m = \{\mathcal{I}, \tilde{\mathcal{C}}_{ii}^m\}$ for each modality m, where $\tilde{\mathcal{C}}_m^i = \{\tilde{e}_{i,i'}^m|i, i' \in \mathcal{I}\}$ denotes the set of edges between item nodes in $\tilde{\mathcal{G}}_m$. And we calculate the similarity $s_{ii'}$ between the two items (i and i') with cosine similarity, and take it as the value of weight between them. Following the experience of [50], we also perform the sparsification [2] of the item semantic graph, and retain only the edges with high similarity. Specifically, the edge weights of the items could be formulated as

$$\tilde{e}_{i,i'}^m = \begin{cases} s_{i,i'}^m, & \text{if } s_{i,i'}^m \in top\text{-}k(s_i^m), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

To mitigate potential gradient explosion or vanishing [18], degree normalization is applied as in [50]. Then, all modal semantic graphs are fused by summation using modal importance coefficients α_m^{item}, which sum up to 1. Finally, we obtain the item semantic graph $\tilde{\mathcal{G}}_{ii} = \{\mathcal{I}, \tilde{\mathcal{C}}_{ii}\}$, where $\tilde{\mathcal{C}}_{ii} = \{\tilde{e}_{i,i'}|i, i' \in \mathcal{I}\}$, $\tilde{e}_{i,i'} = \sum_{m\in\mathcal{M}} \alpha_m^{item}\tilde{e}_{i,i'}^m$.

4.2.3 Homogeneous Relations Learning. Given differences between the item co-occurrence graph $\bar{\mathcal{G}}_{ii}$ and the item semantic graph $\tilde{\mathcal{G}}_{ii}$ in various scenarios, we introduce the item co-occurrence factor α_{co}^{item} to distinguish the final contribution, $\mathcal{G}_{ii} = \alpha_{co}^{item}\bar{\mathcal{G}}_{ii} + (1-\alpha_{co}^{item})\tilde{\mathcal{G}}_{ii}$.

Before exploiting higher-order associations by performing message passing on graph $\mathcal{G}_{ii}$, we first convert the item modality features $\mathbf{i}_m^{raw}$ to align with the ID embedding dimension. Formally,

$$\mathbf{i}_m = \mathbf{W}_m^i\mathbf{i}_m^{raw} + \mathbf{b}_m^i, \tag{3}$$

where $\mathbf{W}_m^i \in \mathbb{R}^{d\times d_m}$ and $\mathbf{b}_m^i \in \mathbb{R}^{d\times 1}$ are the transformation and bias parameters, and d is the ID embedding dimension. With the consideration that the item IDs contain information that is distinct from the multi-modal features, we concatenate them to obtain the item representation $\mathbf{h}_i$, and perform homography relation learning,

$$\mathbf{h}_i^{(l+1)} = \sum_{i'\in\mathcal{N}_i} e_{i,i'}\mathbf{h}_{i'}^{(l)}, \tag{4}$$

where $\mathcal{N}_i$ denotes the set of neighbors of item i in the item homogeneous graph $\mathcal{G}_{ii}$. $\mathbf{h}_i^{(0)}$ is initialized with spliced IDs and multi-modal features, $\mathbf{h}_i^{(0)} = \mathbf{i}_{id}||\mathbf{i}_v||\mathbf{i}_t$. And we take the final layer output as the item homography representation. The item ID $\mathbf{i}_{id}^{hom}$ and multi-modal features $\mathbf{i}_m^{hom}$ learned through homography relations are richer in semantic and co-occurrence information.

Similarly, we derive the user co-occurrence graph $\bar{\mathcal{G}}_{uu}$, the interest graph $\tilde{\mathcal{G}}_{uu}$, and the homogeneous graph $\mathcal{G}_{uu}$ in sequence. Following Equation (4), the operation on $\mathcal{G}_{uu}$ yields the user representations ($\mathbf{u}_{id}^{hom}$ and $\mathbf{u}_m^{hom}$) carrying more user's internal relations.

4.2.4 Item Feature Attention Integration. Since the multi-modal features are not tailored for the recommendation task [46, 53] and GNNs may further amplify the noise in the modality [25], we further refine by introducing self-attention and cross-attention modules.

The self-attention module better adapts to the downstream recommendation task by adaptively adjusting the magnitudes of the values of the dimensions within the modality. Formally,

$$\mathbf{i}_m^{h-self} = softmax(\frac{(\mathbf{i}_m^{hom}\mathbf{W}_m^Q)^\top(\mathbf{i}_m^{hom}\mathbf{W}_m^K)}{\sqrt{d}})\mathbf{i}_m^{hom}\mathbf{W}_m^V, \tag{5}$$

where $\mathbf{W}_m^Q, \mathbf{W}_m^K, \mathbf{W}_m^V \in \mathbb{R}^{d\times d}$ are parameter matrices, $softmax(\cdot)$ is the softmax function. Furthermore, layer normalization and residual connection are applied to enhance the robustness and generalization of the model [11]. More specifically, the multi-modal feature of the item after self-attention is represented as

$$\mathbf{i}_m^{hs} = ly\_norm(\mathbf{i}_m^{hom} + \mathbf{i}_m^{h-self}), \tag{6}$$

where $ly\_norm(\cdot)$ denotes the layer normalization function.

The cross-attention module is designed to explore the inter-modal influences, and the cross-attention multi-modal features differ from various viewpoints. Therefore, the item's cross-attention visual features and textual features are calculated separately as

$$\mathbf{i}_v^{h-cross} = softmax(\frac{(\mathbf{i}_t^{hs}\mathbf{W}_t^Q)^\top(\mathbf{i}_v^{hs}\mathbf{W}_v^K)}{\sqrt{d}})\mathbf{i}_v^{hs}\mathbf{W}_v^V, \tag{7}$$

$$\mathbf{i}_t^{h-cross} = softmax(\frac{(\mathbf{i}_v^{hs}\mathbf{W}_v^Q)^\top(\mathbf{i}_t^{hs}\mathbf{W}_t^K)}{\sqrt{d}})\mathbf{i}_t^{hs}\mathbf{W}_t^V. \tag{8}$$

As in the self-attention post-processing step, residual connection and layer normalization are employed. Formally,

$$\mathbf{i}_m^{hsc} = ly\_norm(\mathbf{i}_m^{hs} + \mathbf{i}_m^{h-cross}). \tag{9}$$

Note that we use the dropout mechanism in all attention modules to enhance the expressiveness of the model based on the experience of [37]. Since users' multi-modal preferences are not directly obtained, we do not fine-tune them, as evidenced by the experiments.

4.3 Heterography Relation Learning

Consistent with previous work [8, 15, 44], we conduct graph convolution operations by leveraging LightGCN [14]. Furthermore, the same interaction graph structure $\mathcal{G}_{ui}$ could be employed by different multi-modal and ID information for $\mathcal{G}_{ui}^m$ and $\mathcal{G}_{ui}^{id}$ individually, which operates to ensure non-interference with each other.

More specifically, the user and item multi-modal features at layer $l+1$ are represented in $\mathcal{G}_{ui}^m$ separately as

$$\mathbf{u}_m^{(l+1)} = \sum_{i\in\mathcal{N}_u}\lambda_n\mathbf{i}_m^{(l)}, \quad \mathbf{i}_m^{(l+1)} = \sum_{u\in\mathcal{N}_i}\lambda_n\mathbf{u}_m^{(l)}, \tag{10}$$

where $\mathbf{u}_m^{(0)} = \mathbf{u}_m^{hom}$, $\mathbf{i}_m^{(0)} = \mathbf{i}_m^{hsc}$, $\mathcal{N}_u$ and $\mathcal{N}_i$ denote the set of neighbors of the user u and the item i in $\mathcal{G}_{ui}^m$, respectively. The symmetric normalization factor $\lambda_n = \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}}$ is employed to prevent the problem of gradient vanishing or explosion as the number of layers of the graph convolution layer stacks up [18]. By combining the neighbor information aggregated across all layers, we obtain the multi-modal representation of the user and the item,

$$\bar{\mathbf{u}}_m = \frac{1}{L+1}\sum_{l=0}^{L}\mathbf{u}_m^{(l)}, \quad \bar{\mathbf{i}}_m = \frac{1}{L+1}\sum_{l=0}^{L}\mathbf{i}_m^{(l)}, \tag{11}$$

where L denotes the number of layers of LightGCN. It is the same way that we obtain the representation of the user ID $\bar{\mathbf{u}}_{id}$ and the item ID $\bar{\mathbf{i}}_{id}$ after message passing in $\mathcal{G}_{ui}^{id}$ by leveraging GNNs.

4.4 Refined Contrastive Learning

Through multi-modal contrastive learning, not only could the self-supervised signals be introduced to alleviate the data sparsity, but the consistency of modalities in the user-item interaction mode could also be enhanced [8, 17]. However, simply contrasting modalities causes the loss of recommendation-relevant information in modal-unique features, and may introduce noise in modal-shared features [7, 22, 39], thereby impairing recommendation performance. To resolve the above concerns, we introduce meta-network and orthogonal constraint techniques after contrastive learning to further tackle the modal-shared and modal-unique features.

4.4.1 Multi-modal Contrastive Learning. Consistent with previous work [8, 17], we maximize the mutual information between the two modal feature views with the help of InfoNCE [29]. Formally, the item multi-modal contrastive learning loss is defined as

$$\mathcal{L}_{cl}^i = \sum_{i\in\mathcal{I}} -\log\frac{\exp(sim(\bar{\mathbf{i}}_v,\bar{\mathbf{i}}_t)/\tau)}{\sum_{i'\in\mathcal{I}}\exp(sim(\bar{\mathbf{i}}_v,\bar{\mathbf{i}}_t')/\tau)}, \tag{12}$$

where $sim(\cdot,\cdot)$ is the cosine similarity function, and τ is the temperature coefficient. Similarly, we can define the contrastive learning loss on the user side $\mathcal{L}_{cl}^u$. At last, the total multi-modal contrastive learning loss could be obtained as $\mathcal{L}_{cl} = \mathcal{L}_{cl}^i + \mathcal{L}_{cl}^u$.

4.4.2 Modal-shared Meta-network. Simply utilizing features after multi-modal contrasts ignores noise in modal-shared features, such as the "girl" in Figure 1, further harming recommendation accuracy. To filter out the noise in modal-shared and derive recommendation-relevant features, we design a meta-network to extract valuable information in modal-shared features, which is inspired by [3, 58].

We first splice the contrasting multi-modal features and align the dimensions to obtain the modal-shared meta-knowledge. Formally, the meta-knowledge modal-shared feature of the item is denoted as

$$\mathbf{i}_{share} = \mathbf{W}_{share}^i(\bar{\mathbf{i}}_v||\bar{\mathbf{i}}_t) + \mathbf{b}_{share}^i, \tag{13}$$

where $\mathbf{W}_{share}^i \in \mathbb{R}^{d\times 2d}$, $\mathbf{b}_{share}^i \in \mathbb{R}^d$, and the notation $||$ denotes the concatenation operation. Next, we employ a meta-network to extract recommendation-relevant knowledge and parametrically preserve knowledge into customized transformation matrices. Our proposed meta-neural network could be represented as

$$\mathbf{W}_{share}^{i_1} = g^{i_1}(\mathbf{i}_{share}), \quad \mathbf{W}_{share}^{i_2} = g^{i_2}(\mathbf{i}_{share}), \tag{14}$$

where $g^{i_1}(\cdot), g^{i_2}(\cdot)$ are meta knowledge learners that comprise a two-layer feed-forward neural network with PReLU activation function. And, $\mathbf{W}_{share}^{i_1} \in \mathbb{R}^{d\times k}, \mathbf{W}_{share}^{i_2} \in \mathbb{R}^{k\times d}$ are customized transformation matrices where modal-shared features are extracted knowledge through the meta-network to filter out noise. The hyperparameter k is the rank of the transformation matrices with restriction to $k < d$, which effectively controls the trainable parameters while enhancing the robustness of the model. To further combine the extracted knowledge by the meta-network with the item representation, inspired by the bridge function [3, 58], we migrate them to the item ID information. Formally, the transferred modal-shared feature of the item $\bar{\mathbf{i}}_{share}$ is expressed as

$$\bar{\mathbf{i}}_{share} = \mathbf{W}_{share}^{i_1}\mathbf{W}_{share}^{i_2}\bar{\mathbf{i}}_{id}. \tag{15}$$

Since item ID is learned after homography and heterography relationships, we believe it may contain semantic and behavioral information. Hence, we sum it $\bar{\mathbf{i}}_{id}$ with the obtained modal-shared features $\bar{\mathbf{i}}_{share}$ to get the item modal-shared representation $\mathbf{i}_{share}^f$,

$$\mathbf{i}_{share}^f = \delta(\bar{\mathbf{i}}_{share}) + \bar{\mathbf{i}}_{id}. \tag{16}$$

where $\delta(\cdot)$ is the PReLU activation function, enhancing feature expressiveness [10]. Similarly, we get the user's meta-knowledge modal-shared preference $\mathbf{u}_{share}$ and its transferred counterpart $\bar{\mathbf{u}}_{share}$. By fusing ID, we obtain the final user modal-shared representation $\mathbf{u}_{share}^f$. Note that the complexity of this component is on par with that of the vanilla GNN due to the typically small k [3].

4.4.3 Modal-unique Orthogonal Constraint. As shown in Figure 1, "the fashionable twist design" in the image, the combined utilization of different modal-unique and recommendation-relevant information further highlights the complementary nature of multi-modal.

Considering orthogonal constraint loss achieves non-overlapping information between features by encouraging individual features to retain unique information [26]. Inspired by [1, 9, 52], we impose an orthogonal constraint between multi-modal features to distinguish modal-unique information. Formally, the item multi-modal orthogonal constraint loss between modalities could be denoted as

$$\mathcal{L}^{i}_{ort} = \sum_{i \in \mathcal{I}} \left\| \bar{\mathbf{i}}_{v}^{\top} \bar{\mathbf{i}}_{t} \right\|_F^2, \tag{17}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm. We represent item visual and textual features after orthogonal constraints with $\bar{\mathbf{i}}_{v-uni}$ and $\bar{\mathbf{i}}_{t-uni}$, respectively. The same operation is taken to compute the orthogonal loss of the user $\mathcal{L}^{u}_{ort}$. At last, we sum to get the total multi-modal orthogonal constraint loss $\mathcal{L}_{ort} = \mathcal{L}^{u}_{ort} + \mathcal{L}^{i}_{ort}$.

After obtaining the modal-unique information, we optimize it by utilizing Bayesian Personalized Ranking (BPR) [30] to retain the recommendation-relevant information. We align with the treatment of transferred modal-shared features by summing the ID with the modal-unique information. The fusion process is represented as

$$\mathbf{i}^{f}_{m-uni} = \delta(\bar{\mathbf{i}}_{m-uni}) + \bar{\mathbf{i}}_{id}. \tag{18}$$

where $m \in \{v, t\}$, and $\delta(\cdot)$ denotes the PReLU activation function. Then, we splice each modal-unique feature to get the final item modal-unique feature $\mathbf{i}^{f}_{uni} = \mathbf{i}^{f}_{v-uni} \| \mathbf{i}^{f}_{t-uni}$. Similarly, we can obtain the final user modal-unique preference $\mathbf{u}^{f}_{uni}$.

4.5 Prediction and Optimization

Finally, we splice modal-shared and modal-unique features separately to form the final representation, as

$$\mathbf{u}^{*} = \mathbf{u}^{f}_{share} \| \mathbf{u}^{f}_{uni}, \quad \mathbf{i}^{*} = \mathbf{i}^{f}_{share} \| \mathbf{i}^{f}_{uni}. \tag{19}$$

We predict the likelihood of interaction between the user u and the item i by calculating the value of inner product, i.e., $\hat{y}_{ui} = \mathbf{u}^{*\top} \mathbf{i}^{*}$.

BPR is employed as our principal loss, popularly utilized in recommendation tasks [8, 44]. And the BPR loss enables REARM to better retain more recommendation-relevant information in multi-modal-unique features. Specifically, we construct a triplet (u, i, i') with a user u and two items, where one item i is observed to have an interaction with the user u and the other item i' is not,

$$\mathcal{R} = \{(u, i, i') | u \in \mathcal{U}, i \in \mathcal{I}, i' \notin \mathcal{I}\}. \tag{20}$$

Formally, the BPR loss function could be expressed as

$$\mathcal{L}_{bpr} = \sum_{(u,i,i') \in \mathcal{R}} -ln\psi(\hat{y}_{ui} - \hat{y}_{ui'}), \tag{21}$$

where $\psi(\cdot)$ is the sigmoid function. For better optimization of the model, combining the multi-modal contrastive loss and the orthogonal constraint loss, our final loss function is denoted as

$$\mathcal{L} = \mathcal{L}_{bpr} + \lambda_{cl}\mathcal{L}_{cl} + \lambda_{ort}\mathcal{L}_{ort} + \lambda_{p}\|\Theta\|_2^2, \tag{22}$$

where λ_{cl} and λ_{ort} are hyperparameters that control contrastive and orthogonal loss, respectively. Θ denotes all model parameters, and hyperparameter λ_p controls the impact of $L2$ regularization.

Table 1: Statistics of the three datasets.

Dataset	#User	#Item	#Interaction	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	278,677	99.97%

5 Evaluation

5.1 Experimental Settings

5.1.1 Evaluation Datasets. Consistent with previous studies [8, 15, 47, 56], we conduct experiments on three publicly available and widely used Amazon datasets: Baby, Sports, and Clothing. We filter users and items by utilizing a 5-core as the threshold for each dataset. These three datasets provide rich multi-modal feature data (visual modality and textual modality). In this work, we use the 4,096-dimensional visual features and 384-dimensional textual features that have been preprocessed and released in prior work [55]. The statistics for the three datasets are presented in Table 1.

5.1.2 Baseline Methods. To evaluate the effectiveness of REARM, we compare it with state-of-the-art models. Based on whether or not multi-modal utilized, they could be categorized into the general recommendation models (BPR [30], LightGCN [14], ApeGNN [48], and MGDN [16]), and multi-modal recommendation models (VBPR [13], MMGCN [44], DualGNN [41], GRCN [43], LATTICE [49], BM3 [57], SLMRec [35], MICRO [50], MGCN [47], DiffMM [17], FREEDOM [56], LGMRec [8], DRAGON [54], and MIG-GT [15]).

5.1.3 Evaluation Protocols. To obtain fair results for the experimental evaluation, we adopt the same settings as in previous studies [15, 54, 56]. Specifically, we assess recommendation performance with the two most common evaluation metrics, Recall and Normalized Discounted Cumulative Gain (NDCG). Regarding the dataset partition, we randomly split the history interactions according to 8 : 1 : 1 to obtain the train, validate, and test sets. Finally, we evaluate the top-K recommendation with the all-ranking protocol, and report the average performance over all users in the test set for $K = 10$ and $K = 20$. For simplicity, the evaluation metrics can be represented separately as R@K and N@K, where $K \in \{10, 20\}$.

5.1.4 Implementation Details. To be fair, we set both the user and item embedding dimensions to 64 and the training batch size to 2,048 to optimize all models. We search for optimal parameter ranges in the validation set as reported by their original baseline models for all models, and report the corresponding test set results. To ensure convergence, early stopping and the total epochs are set to 20 and 2,000, respectively. In line with previous work [8, 15, 47], we employ R@20 as the training stop indicator on the validation set. For our model, we simply adjust the number of LightGCN layers from 1 to 10, the user and item modal importance coefficients from 0.1 to 1, the user and item co-occurrence coefficients from 0.1 to 1, the transformation matrix rank from 1 to 10, and the dropout rates of all attention modules from 0 to 1.

5.2 Overall Performance

The performance comparison results of all models on three datasets are shown in Table 2. We observe the following findings,

Table 2: Top-k recommendation performance comparison of different models. The best and the second-best performances are marked with boldface and underlining, respectively.

Dataset	Baby				Sports				Clothing			
Metric	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0206	0.0303	0.0114	0.0138
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0361	0.0544	0.0197	0.0243
ApeGNN	0.0501	0.0775	0.0267	0.0338	0.0608	0.0892	0.0333	0.0407	0.0378	0.0538	0.0204	0.0244
MGDN	0.0495	0.0783	0.0272	0.0346	0.0614	0.0932	0.0340	0.0422	0.0362	0.0551	0.0199	0.0247
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
MMGCN	0.0421	0.0660	0.0220	0.0282	0.0401	0.0636	0.0209	0.0270	0.0227	0.0361	0.0154	0.0154
DualGNN	0.0513	0.0803	0.0278	0.0352	0.0588	0.0899	0.0324	0.0404	0.0452	0.0675	0.0242	0.0298
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
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
SLMRec	0.0521	0.0772	0.0289	0.0354	0.0663	0.0990	0.0365	0.0450	0.0442	0.0659	0.0241	0.0296
MICRO	0.0584	0.0929	0.0318	0.0407	0.0679	0.1050	0.0367	0.0463	0.0521	0.0772	0.0283	0.0347
MGCN	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0641	0.0945	0.0347	0.0428
DiffMM	0.0623	0.0975	0.0328	0.0411	0.0671	0.1017	0.0377	0.0458	0.0522	0.0791	0.0288	0.0354
FREEDOM	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0629	0.0941	0.0341	0.0420
LGMRec	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0555	0.0828	0.0302	0.0371
DRAGON	0.0662	0.1021	0.0345	0.0435	0.0752	0.1139	0.0413	0.0512	0.0671	0.0979	0.0365	0.0443
MIG-GT	0.0665	0.1021	0.0361	0.0452	0.0753	0.1130	0.0414	0.0511	0.0636	0.0934	0.0347	0.0422
REARM	**0.0705**	**0.1105**	**0.0377**	**0.0479**	**0.0836**	**0.1231**	**0.0455**	**0.0553**	**0.0700**	**0.0998**	**0.0377**	**0.0454**

- Our model significantly outperforms all baselines (both general and multi-modal recommendation models) on every dataset. We attribute the remarkable improvement to: 1) We refine current multi-modal contrastive learning with meta-network and orthogonal constraint methods to filter out noise from modal-shared features while preserving modal-unique and recommendation-relevant information. 2) We further mine the user interest graph and item co-occurrence graph based on previous homogeneous graphs, which enable the model to explore their potential structural and semantic relationships in detail with the help of GNNs.
- The majority of multi-modal recommendation models (e.g., MIG-GT, DRAGON, LGMRec, and FREEDOM) significantly outperform general recommendation models (e.g., BPR, MGDN, LightGCN, and ApeGNN), which suggests that the multi-modal information can help the models better to understand the items and the users' multi-modal interests. In addition, DRAGON shows better performance by integrating the user co-occurrence graph and the item semantic graph. It reveals that mining homogeneous relationships among users and items is quite effective, based on which we further explore the user interest graph and item co-occurrence graph to refine the homogeneous relationships.

5.3 Ablation Study

5.3.1 Effect of Different Components of REARM. To analyze the impact of various homographs on the model performance, we categorize them into: no user homograph (w/o uu), no item homograph (w/o ii), no co-occurrence homographs (w/o co), no similarity homographs (w/o sim), and no homographs at all (w/o hom). For the refinement of multi-modal contrastive learning, we split it into: no meta-networks (w/o meta), no orthogonal constraints (w/o ort), and no refinement (w/o ref).

Table 3: Ablation study on different modules of the REARM.

Dataset	Baby		Sports		Clothing	
Metric	R@20	N@20	R@20	N@20	R@20	N@20
w/o uu	0.1055	0.0458	0.1209	0.0544	0.0984	0.0446
w/o ii	0.0972	0.0420	0.1011	0.0453	0.0676	0.0298
w/o co	0.1057	0.0458	0.1158	0.0511	0.0968	0.0434
w/o sim	0.0903	0.0402	0.0982	0.0445	0.0708	0.0325
w/o hom	0.0926	0.0403	0.0975	0.0437	0.0653	0.0290
w/o meta	0.1078	0.0471	0.1214	0.0547	0.0995	0.0451
w/o ort	0.1079	0.0470	0.1201	0.0548	0.0983	0.0447
w/o ref	0.1070	0.0467	0.1190	0.0545	0.0982	0.0446
REARM	**0.1105**	**0.0479**	**0.1231**	**0.0553**	**0.0998**	**0.0454**

- The variant w/o hom achieves the worst results on both datasets, which indicates that the introduction of homographs can greatly improve performance. Furthermore, w/o uu is better than w/o ii, and w/o co is better than w/o sim. This illustrates the greater importance of item homogeneous graphs and semantic relations.
- For refining the contrastive learning module, removing either the meta-learning or the orthogonal constraint module results in performance degradation. This further points to the importance of filtering modal-shared feature noise and focusing on recommendation-related information in modal-unique features.

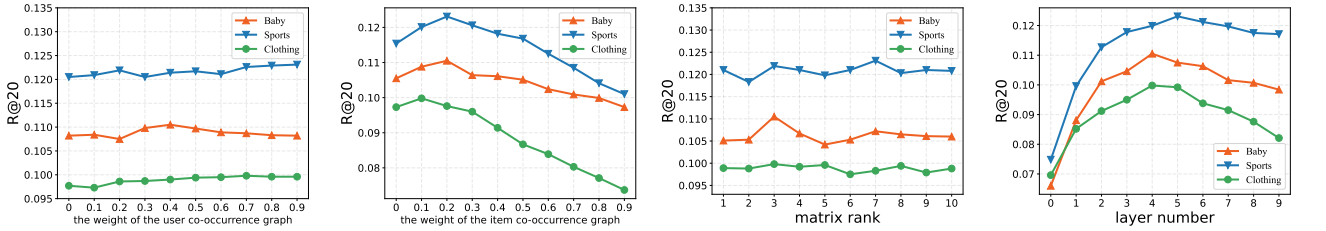

Figure 3: Performance of REARM in terms of R@20 w.r.t the different hyperparameter settings.

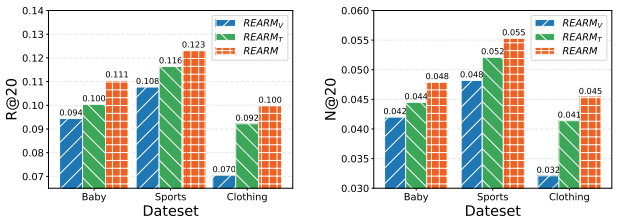

Figure 4: Comparison of different feature modalities.

5.3.2 Effect of Different Modalities of REARM. To study the contribution of diverse modalities to our proposed model, we conduct separate experiments for visual and textual modalities, denoted as $REARM_V$ and $REARM_T$, respectively. The experiment results are shown in Figure 4, and we could make the following findings.

- Neither $REARM_V$ nor $REARM_T$ achieves the desired results under both metrics, which verifies that each modality is critical since both of them contain different and unique information.
- $REARM_T$ is better than $REARM_V$ in all datasets, especially in the Clothing. This is reasonable, as texts contain more refined information and less noise compared to images, which could help users to better select clothes with satisfactory sizes and styles.

5.4 Hyperparameter Analysis

5.4.1 Effect of the Weight of the User Co-occurrence Graph. We search for the proportion of the user co-occurrence graph in the final user homograph in $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. As shown in Figure 3, we find that compared to the user co-occurrence graph, the user interest graph with a certain weight could better improve the model effect, which is consistent with our motivation.

5.4.2 Effect of the Weight of the Item Co-occurrence Graph. We explore the contribution of the item co-occurrence graph by searching in $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The optimal weight values for various datasets are different, as seen in Figure 3, which is expected and demonstrates that mining inter-item co-occurrence relationships could enrich the representation of users and items.

5.4.3 Effect of Matrix Rank in Modal-shared Meta-network. We effectively control the number of model parameters and enhance the model robustness by the rank of the matrix, searched within $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. From Figure 3, we observe that the rank of the matrix for the Sports dataset is 7, which is the largest. A possible reason is that it has the most interactions in three datasets, which may require more model capacity to capture richer interactions.

5.4.4 Effect of the Number of Layers in the Interaction Graph. In the user-item interaction graph, we search for the number of propagation layers from 0 to 9. We find that the optimal number of layers for all datasets is greater than 3, as illustrated in Figure 3, which is inconsistent with previous studies, where most of them are at 2 layers. We speculate that it is related to the attention mechanism, which retains valuable information by exploring intra- and inter-modal relationships, thereby allowing information to be propagated to a greater extent via higher-order GNNs.

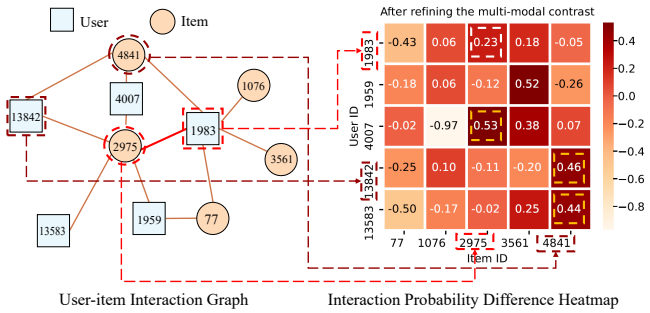

Figure 5: Difference heatmap of interaction probabilities before and after refining multi-modal contrastive learning.

5.5 Visualization Analysis

For further investigation of the importance of refining the multi-modal contrastive learning, we select a portion of users and items from the Baby dataset and calculate the probability of interactions before and after refining contrastive learning, respectively. By differentiating, we get a heatmap of the interaction probability difference.

As shown in Figure 5, the deeper the color of the heatmap, the more it indicates that REARM predicts a higher probability of interactions. The yellow boxes in the heatmap indicate that they have been observed to interact in the training set, e.g., user u_{4007} and item i_{2975}. Compared to the existing methods, REARM considers the noise in modal-shared and modal-unique and recommended information, rendering it a better predictor. For example, REARM further predicts the high likelihood of interaction between user u_{1983} and item i_{2975}, which is confirmed to exist in the test set.

6 Conclusion

In this paper, we propose a novel framework for **RE**fining multi-mod**A**l cont**R**astive learning and ho**M**ography relations (**REARM**). A meta-network is designed to denoise the modal-shared features, and the orthogonal constraint loss is utilized to retain the modal-unique and recommendation-relevant information to refine the current multi-modal contrastive learning methods. We also explore potential structural and semantic relationships in user interest and item co-occurrence graphs to enrich representations based on existing homogeneous graphs. Extensive evaluation on three real-world datasets shows that REARM outperforms state-of-the-art baselines.

References

[1] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *NeurIPS* 29 (2016).

[2] Jie Chen, Haw-ren Fang, and Yousef Saad. 2009. Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *JMLR* 10, 9 (2009).

[3] Mengru Chen, Chao Huang, Lianghao Xia, Wei Wei, Yong Xu, and Ronghua Luo. 2023. Heterogeneous graph contrastive learning for recommendation. In *WSDM*. 544–552.

[4] Ming Chen, Tianyi Ma, and Xiuze Zhou. 2022. CoCNN: Co-occurrence CNN for recommendation. *ESWA* 195 (2022), 116595.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR, 1597–1607.

[6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *RecSys*. 191–198.

[7] Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. 2025. What to align in multimodal contrastive learning?. In *ICLR*.

[8] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *AAAI*, Vol. 38. 8454–8462.

[9] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *MM*. 1122–1131.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*. 1026–1034.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[12] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*. 507–517.

[13] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*, Vol. 30.

[14] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*. 639–648.

[15] Jun Hu, Bryan Hooi, Bingsheng He, and Yinwei Wei. 2025. Modality-Independent Graph Neural Networks with Global Transformers for Multimodal Recommendation. In *AAAI*.

[16] Jun Hu, Bryan Hooi, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2024. MGDCF: Distance learning via Markov graph diffusion for neural collaborative filtering. *TKDE* (2024).

[17] Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. 2024. Diffmm: Multi-modal diffusion model for recommendation. In *MM*. 7591–7599.

[18] Thomas N Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

[19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR*.

[20] Zhiwei Li, Guodong Long, and Tianyi Zhou. 2024. Federated Recommendation with Additive Personalization. In *ICLR*.

[21] Dawen Liang, Jaan Altosaar, Laurent Charlin, and David M Blei. 2016. Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence. In *ACM RecSys*. 59–66.

[22] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2024. Factorized contrastive learning: Going beyond multi-view redundancy. *NeurIPS* 36 (2024).

[23] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *NeurIPS* 2021, DB1 (2021).

[24] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *WWW*. 2320–2329.

[25] Kang Liu, Feng Xue, Dan Guo, Peijie Sun, Shengsheng Qian, and Richang Hong. 2023. Multimodal graph contrastive learning for multimedia-based recommendation. *TMM* 25 (2023), 9343–9355.

[26] Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang. 2017. Adversarial Multi-task Learning for Text Classification. In *ACL*. 1–10.

[27] Qiang Liu, Shu Wu, and Liang Wang. 2017. Deepstyle: Learning user preferences for visual recommendation. In *SIGIR*. 841–844.

[28] Yaokun Liu, Xiaowang Zhang, Minghui Zou, and Zhiyong Feng. 2024. Attribute simulation for item embedding enhancement in multi-interest recommendation. In *WSDM*. 482–491.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[30] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*. 452–461.

[31] Sebastian Ruder and Barbara Plank. 2018. Strong Baselines for Neural Semi-Supervised Learning under Domain Shift. In *ACL*. 1044–1054.

[32] Mathieu Salzmann, Carl Henrik Ek, Raquel Urtasun, and Trevor Darrell. 2010. Factorized orthogonal latent spaces. In *AISTATS*. JMLR Workshop and Conference Proceedings, 701–708.

[33] Brent Smith and Greg Linden. 2017. Two decades of recommender systems at Amazon. com. *IC* 21, 3 (2017), 12–18.

[34] Qi Song, Tianxiang Gong, Shiqi Gao, Haoyi Zhou, and Jianxin Li. 2024. QUEST: Quadruple Multimodal Contrastive Learning with Constraints and Self-Penalization. *NeurIPS* 37 (2024), 28889–28919.

[35] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *TMM* 25 (2022), 5107–5116.

[36] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *NeurIPS* 33 (2020), 6827–6839.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* (2017).

[38] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. 2015. Deep multimodal hashing with orthogonal regularization.. In *IJCAI*, Vol. 367. 2291–2297.

[39] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. 2022. Rethinking minimal sufficient representation in contrastive learning. In *CVPR*. 16041–16050.

[40] Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. 2010. Semi-supervised hashing for scalable image retrieval. In *CVPR*. IEEE, 3424–3431.

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

[42] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *SIGIR*. 165–174.

[43] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *MM*. 3541–3549.

[44] 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 *MM*. 1437–1445.

[45] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *SIGIR*. 726–735.

[46] 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 *KDD*. 3645–3656.

[47] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-view graph convolutional network for multimedia recommendation. In *MM*. 6576–6585.

[48] Dan Zhang, Yifan Zhu, Yuxiao Dong, Yuandong Wang, Wenzheng Feng, Evgeny Kharlamov, and Jie Tang. 2023. ApeGNN: node-wise adaptive aggregation in GNNs for recommendation. In *WWW*. 759–769.

[49] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *MM*. 3872–3880.

[50] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *TKDE* 35, 9 (2022), 9154–9167.

[51] Xiaokun Zhang, Bo Xu, Fenglong Ma, Chenliang Li, Liang Yang, and Hongfei Lin. 2023. Beyond co-occurrence: Multi-modal session-based recommendation. *TKDE* 36, 4 (2023), 1450–1462.

[52] Lecheng Zheng, Zhengzhang Chen, Jingrui He, and Haifeng Chen. 2024. MULAN: multi-modal causal structure learning and root cause analysis for microservice systems. In *WWW*. 4107–4116.

[53] Shanshan Zhong, Zhongzhan Huang, Daifeng Li, Wushao Wen, Jinghui Qin, and Liang Lin. 2024. Mirror Gradient: Towards Robust Multimodal Recommender Systems via Exploring Flat Local Minima. In *WWW*. 3700–3711.

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

[55] Xin Zhou. 2023. Mmrec: Simplifying multimodal recommendation. In *MM*. 1–2.

[56] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *MM*. 935–943.

[57] 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 *WWW*. 845–854.

[58] Yongchun Zhu, Zhenwei Tang, Yudan Liu, Fuzhen Zhuang, Ruobing Xie, Xu Zhang, Leyu Lin, and Qing He. 2022. Personalized transfer of user preferences for cross-domain recommendation. In *WSDM*. 1507–1515.