Information Fusion 122 (2025) 103168

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

PEARL: A dual-layer graph learning for multimodal recommendation

Yuzhuo Dang [a], Wanyu Chen [b], Zhiqiang Pan [a], Yuxiao Duan [c], Fei Cai [a,*], Honghui Chen [a]

[a] National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha, Hunan, China
[b] College of Electronic Countermeasures, National University of Defense Technology, Hefei, Anhui, China
[c] Laboratory for Big Data and Decision, National University of Defense Technology, Changsha, Hunan, China

ARTICLE INFO

Keywords:
Multimodal recommendation
Graph purification
Affinity graph learning
Contrastive learning

ABSTRACT

Multimodal recommendation has increasingly become a mainstream information service technology, which transcends traditional user–item interaction-based recommender systems by integrating the multimodal features of items. Although existing works have made notable progress by focusing on user–item interaction graph structures and self-supervised learning to enhance multimodal representation learning, they still exhibit the following two limitations: (1) Performing graph convolution operations on a fixed interaction graph introduces misleading noisy signals caused by user's imbalanced attention on various modalities. (2) Lacking exploration of inherent self-supervised signals in multimodal attributes fails to mitigate the distribution bias introduced during data augmentation. To address these issues, we propose a novel method named **P**urified-int**E**raction and **A**ffinity g**R**aph **L**earning (**PEARL**) for multimodal recommendation, which utilizes a dual-layer graph learning to model the user preference. Specifically, to eliminate misleading noisy signals, we designed a graph purification strategy that constructs purified modality-specific interaction graphs, thereby removing noisy edges from the raw interaction graph. Then, the user and item affinity graphs are constructed based on the user co-occurrences and the item multimodal features, respectively, which are then utilized for message passing to mine the implicit self-supervised signals between similar users or items. After that, we propose an augmentation-free contrastive learning task in the fusion module to improve the quality of ID embeddings and multimodal features, ultimately generating the final representations of users and items. Comprehensive experimental results on three publicly available datasets verify the superiority of PEARL against diverse state-of-the-art baselines, where the improvements are especially noticeable in large-scale datasets. Our experimental code and data utilized in this work can be accessed at https://github.com/Yuzhuo-Dang/PEARL.

1. Introduction

As information explodes on the internet, recommender systems have emerged as an important tool for information filtering and personalized services [1–4]. Traditional recommender systems generally capture the user preference based on the binary interactions between users and items [5–7]. Although these methods can effectively improve the recommendation accuracy, they often face the data sparsity problem when dealing with new emerging users or items. To solve this problem, recent works introduce the multimodal recommender systems (MMRS), which utilize the multimodal data (e.g., image, text and audio) to provide additional information and assist in capturing more aspects of user preferences [8–12].

The mainstreams MMRS methods are generally implemented in a two-step strategy. Specifically, the multimodal features of items are first derived using the advanced pre-trained models [13,14], which are then integrated into traditional recommenders by fusing multimodal features with the user–item interaction signals. Early studies on MMRS either employ the linear fusion between ID embeddings of items and their multimodal features [8,15,16], or adopt a weighted fusion on inter-modal representations of users or items by an attention mechanism [17,18]. Subsequently, based on the effective representation learning of graph neural networks (GNN), researchers introduce GNN into MMRS for mining high-order features of users and items [19–21]. However, the performance of these GNN-based methods are sub-optimal due to the sparsity of interactions between users and items comparing to the entire interaction space [22,23]. Therefore, building on the achievements of self-supervised learning (SSL) [24], multi-views of the interaction graph are constructed by random masking or replacement. Subsequently, contrastive learning is used to maximize

* Corresponding author.
E-mail addresses: dangyuzhuo@nudt.edu.cn (Y. Dang), wanyuchen@nudt.edu.cn (W. Chen), panzhiqiang@nudt.edu.cn (Z. Pan), duanyuxiao19@nudt.edu.cn (Y. Duan), caifei08@nudt.edu.cn (F. Cai), chenhonghui@nudt.edu.cn (H. Chen).

https://doi.org/10.1016/j.inffus.2025.103168
Received 10 November 2024; Received in revised form 24 March 2025; Accepted 27 March 2025
Available online 8 April 2025

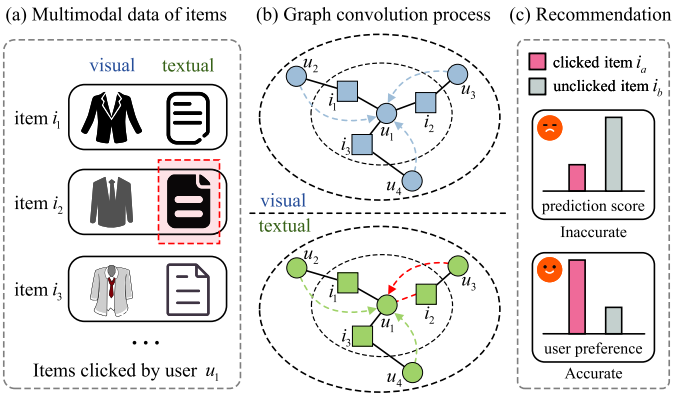


Fig. 1. Illustration of the noise diffusion in the interaction graph, where the textual information (covered by a red box) of item i_2 is not attended by the user u_1. In the textual graph convolution process, information propagates through true-positive ($\langle u_1, i_1\rangle$ and $\langle u_1, i_3\rangle$) and false-positive ($\langle u_1, i_2\rangle$) interactions, which are denoted as green and red dashed curves, respectively. As a result, the trained model recommends item i_a to user u_1 instead of the real preferred item i_b.

the consistency of positive pairs which are constructed from different views of each node and minimize the inconsistency of negative pairs which are constructed from different nodes [25,26], generating accurate representations of users and items.

Despite the remarkable success of existing methods, there still remain two limitations: (1) **Noise in the fixed interaction graph.** Existing methods directly perform graph convolution operations in each modality on the raw historical user–item interaction graph, without identifying and removing noisy edges that arise from the user's imbalanced attention on various modalities [27,28]. This causes misleading noisy signals carried by the unnoticed multimodal features are introduced to the system (e.g., text of item i_2 in Fig. 1). More seriously, these noisy signals propagate and accumulate through the graph convolution process (e.g., red dashed lines like $\langle u_1, u_3\rangle$ in Fig. 1), resulting in the system not recommending accurate items to the user. (2) **Lack of mining for self-supervised signals.** In the data augmentation stage of self-supervised learning, the data generated by random masking and replacement exhibits a distribution bias, introducing the noisy self-supervised signals to the training stage [29–31]. Furthermore, existing methods treat the explicit inter-modal representations of the same user or item as positive samples in contrastive learning [11,12], while ignoring the implicit intra-modal affinity between similar users or items, which limits the representation learning ability.

To address the above limitations, we propose a novel method named **P**urified-int**E**raction and **A**ffinity g**R**aph **L**earning (**PEARL**) for multimodal recommendation. Specifically, we first design a graph purification strategy to eliminate the noisy edges caused by the user's imbalanced attention on various modalities, thereby generating the modality-specific interaction graphs in a hard denoising way. Then, we apply three graph convolutional networks in parallel to process the multimodal features and ID embeddings, generating the high-order representations of users and items. In addition, we perform a message passing on the frozen user co-occurrence and item modality-aware affinity graphs to aggregate features between similar nodes, which mines the implicit intra-modal affinity between users or items to enhance their representations. Finally, we propose an augmentation-free contrastive learning task that leverages the inherent self-supervised signals between the ID embeddings and the multimodal features, avoiding the use of biased data augmentation. The overall contribution of the above techniques is to obtain the accurate multimodal user preference by effectively learning the user–item interaction signals and multimodal features.

To verify the superiority of PEARL, we conduct extensive experiments on three public datasets, i.e., *Baby*, *Sports*, and *Clothing*. The experimental results show that PEARL achieves the state-of-the-art performance in terms of Recall@K and NDCG@K. In addition, the ablation study is designed to validate the effectiveness of various components designed in PEARL and the contribution of different multimodal features utilized in PEARL to recommendation.

In general, our contributions can be summarized as follows:

- We are the first to focus on the noise generated by the user's imbalanced attention on various modalities. To this end, we design a graph purification strategy to identify and eliminate the noisy edges in the historical interaction graph under different modalities, thereby generating the modality-specific interaction graphs.
- We propose a novel multimodal recommender named PEARL, which combines the interaction and affinity graph learning as well as includes an augmentation-free contrastive learning task. It mines the implicit intra-modal affinity between similar users or items and improves the quality of representation learning.
- We conduct sufficient experiments on three publicly available datasets to validate the superiority of PEARL. Further experiments validate the effectiveness of its four components and the contribution of multimodal features to recommendation.

2. Related works

We categorize the related work into three groups, including GCN-based recommendation in Section 2.1, self-supervised recommendation in Section 2.2 and multimodal recommendation in Section 2.3.

2.1. GCN-based recommendation

Graph convolutional network (GCN) [32,33] has demonstrated powerful capabilities in processing graph-structured data, which has greatly promoted the development of recommendations [34,35]. Specifically, GCN can effectively capture the hidden multi-hop relationships between users and items by learning the bipartite graph constructed from the historical user–item interactions, thus enhancing the representations of users and items [36,37]. For example, van den Berg et al. [38] introduce GCN into matrix factorization to solve the data sparse problem by aggregating neighbor information of user and item nodes. In addition, to avoid overfitting caused by redundancy parameters in GCN, He et al. [19] remove the weight transformation matrix and nonlinear activation functions in the vanilla GCN, thus simultaneously improving the model training efficiency.

Then, some studies focus on using the hypergraphs to explore finer-grained relationships between user groups and item categories [39–41]. For instance, Zhang et al. [39] exploit the data from group buying behaviors in shopping to extract complex high-order graph structures by performing multi-view embedding propagation on the constructed directed heterogeneous graph. Moreover, Han et al. [41] propose a hypergraph convolutional network model that effectively identifies high-order relationships and alleviates the data sparsity issue, which simulates the interactions between users and items in different fields and integrates multi-domain features to improve the recommendation performance.

Building on the success of prior researches, our proposed PEARL adopts GCN as the backbone network to propagate and aggregate features of similar nodes. Specifically, we perform GCNs on the modality-specific interaction graphs constructed by a purification strategy and the affinity graphs constructed from the user co-occurrences and the item multimodal features to ensure the aggregation of neighbor information.

2.2. Self-supervised recommendation

Self-supervised learning first gains prominence in the field of computer vision (CV) [42,43], where it employs unlabeled data to train models without the human annotations. Recently, this methodology has been widely implemented in recommendations to solve the problems of data sparsity and cold start [44,45], which provides additional supervisory signals through techniques like data augmentation and contrastive learning, thereby improving the generalization ability and robustness of models [22,46]. For example, S^3-Rec proposed by Zhou et al. [47] is a pioneering application of contrastive learning in recommender systems, which masks different levels of contextual information and mines self-supervised signals in the three views of attributes, items, and sequences, respectively. Moreover, Wu et al. [44] utilize dropout and random walk techniques to create three contrasting views of the user–item graphs, followed by which contrastive learning is applied to enhance the mutual information shared among these views. Then, Xie et al. [48] adopt three data augmentation methods, including crop, mask, and reorder, to construct different views of user behaviors and obtain self-supervised signals from the original user interactions.

Furthermore, Yu et al. [30] point out that the role of graph augmentation is secondary and propose a contrastive learning method without graph augmentation, introducing directed random noise in the embedding space to form contrastive views. In addition, Liu et al. [46] integrate the hyper-relational knowledge graphs with dynamic hypergraph learning, and effectively addresses the issues of over-smoothing and sparse supervised signals through a self-supervised learning framework.

Inspired by the aforementioned researches, we integrate self-supervised learning with MMRS for improving the quality of representation learning. Specifically, we employ an augmentation-free contrastive learning which leverages the inherent self-supervised signals of multimodal attributes to avoid the use of biased data augmentation.

2.3. Multimodal recommendation

With the increasing amount of multimodal data on online platforms, there has been a growing research trend on MMRS [12,49]. Early studies effort in using the collaborative filtering (CF) [19] paradigm to process the multimodal information. For example, He and McAuley [8] first propose to fuse the visual features with item's ID embeddings to enrich the item representations. Moreover, Chen et al. [50] consider the varying degrees of user's attention on different regions of an image and allocate personalized attention to the pre-segmented sub-images, which mitigates the influence of noise in the visual features.

In addition, researchers employ GCN on multimodal user–item interaction graphs for graph learning [9]. This approach is different from multi-source data graph learning, which focuses on integrating nodes and edges of graph structures from different sources into a global graph Zhan et al. [51]. The multimodal graph learning operates as a graph-structured node representation learning paradigm that enhances the modal features of nodes through information propagation and aggregation mechanisms. Specifically, Tao et al. [18] propose to assign weights in the information propagation process on the user–item interaction graph for modeling the fine-grained user preferences. Then, Zhang et al. [10] construct the item–item graphs based on the similarity of item modal features to mine the latent structural information. Recently, self-supervised learning has been widely introduced to the researches of multimodal recommendation [28,45], which involves creating new views for contrastive learning through data augmentation [11], thus learning more accurate representations for users and items. For instance, Yi et al. [25] devise edge dropout and modality masking to generate multi-views of the user–item interaction graph, and employ the negative sampling to ensure the valid contribution of each modality. Moreover, Guo et al. [12] design a local graph and hypergraph learning to model the local and global user preferences for addressing the sparse interaction problems faced by local interest modeling.

While the above methods have achieved obvious improvements in MMRS, they fail to remove the noisy multimodal features and mine the self-supervised signals between similar users or items. Thus, we design a purification strategy to remove the noisy edges in the interaction graph and propose the affinity graph learning to mine the implicit signals between similar nodes.

Table 1
Main notations employed in PEARL.

Notation	Description
$\mathcal{U}$	A set of all users
$\mathcal{I}$	A set of all items
$\mathbf{e}_u^{id}, \mathbf{e}_i^{id}$	The ID embeddings of users and items
$\mathcal{M}$	A set of modalities
$\mathbf{e}_u^m$	The multimodal preferences of users
$\mathbf{e}_i^m, \tilde{\mathbf{e}}_i^m$	The original and projected multimodal features of items
$\mathbf{G}$	A sparse matrix recording the historical user–item interactions
$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}, \mathcal{G}^m = \{\mathcal{V}, \mathcal{E}\}$	The historical and modality-specific user–item interaction graphs
g_{ui}^m	The attention scores of user u to item i under modality m
$p_{\langle u,i\rangle}^m$	A pruning probability of the edge $\langle u,i\rangle$ under modality m
ρ	A certain pruning percentage of edges in the graph purification
$\mathcal{A} = \{\mathcal{U}, \mathcal{C}\}, \mathcal{A}^m = \{\mathcal{U}, \mathcal{C}^m\}$	The user co-occurrence and item modality-aware affinity graphs
n_u, n_i	The top-n value in the user and item affinity graphs
$\mathbf{X}_{\mathcal{U},\mathcal{I}}^{id(0)}, \mathbf{X}_{\mathcal{U},\mathcal{I}}^{m(0)}$	The high-order ID embeddings and multimodal features of users and items
α	A hyper-parameter as the importance score of visual modality in fusing item features
$\mathbf{Z}_{\mathcal{U}}^{id}, \mathbf{Z}_{\mathcal{U}}^{m}, \mathbf{Z}_{\mathcal{I}}^{id}, \mathbf{Z}_{\mathcal{I}}^{m}$	The advanced ID embeddings and multimodal features of users and items
β	A trainable parameter as the attention weight of visual modality in fusing user features
$\mathbf{H}_{\mathcal{U}}, \mathbf{H}_{\mathcal{I}}$	The final representations of users and items
L_{in}, L_{af}	The layers of convolution in the interaction graphs and affinity graphs
$\mathcal{L}, \mathcal{L}_{BPR}, \mathcal{L}_{CL}$	The joint, BPR, and contrastive loss
λ_1, λ_2	The trade-off hyper-parameters for balancing $\mathcal{L}_{CL}$ and L_2 regularization
K	The number of items recommended to each user

3. Approach

In this section, we first describe the task of multimodal recommendation and outline the overview of PEARL in Section 3.1. Then, we provide a detailed explanation of four components in PEARL, including initialization and graph purification in Section 3.2, interaction graph learning in Section 3.3, affinity graph learning in Section 3.4, and fusion and recommendation in Section 3.5.

3.1. Preliminary and overview

The ID embeddings of each user $u \in \mathcal{U}$ and each item $i \in \mathcal{I}$ are initialized as $\mathbf{e}_u^{id}$ and $\mathbf{e}_i^{id}$ within the d-dimensional space $\mathbb{R}^d$, respectively. Moreover, we extract the original multimodal features of items through the pre-trained models as $\mathbf{e}_i^m \in \mathbb{R}^{d_m}$ for each modality $m \in \mathcal{M}$, where d_m denotes the dimension of multimodal features. In addition, we randomly initialize the multimodal user preferences $\mathbf{e}_u^m \in \mathbb{R}^d$. In this work, we deal with the visual and textual modalities, i.e., $\mathcal{M} = \{v, t\}$ following [52]. The historical interactions between users and items are represented as a sparse matrix $\mathbf{G} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, which is used to construct a historical interaction graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ (❄). Here, $\mathcal{V} = \{\mathcal{U} \cup \mathcal{I}\}$ is the node set containing users and items, $\mathcal{E} = \{\langle u,i\rangle \,|\mathbf{G}_{u,i} = 1\}$ is the edge set that represents the interactions between users and items. The main notations employed in PEARL are listed in Table 1.

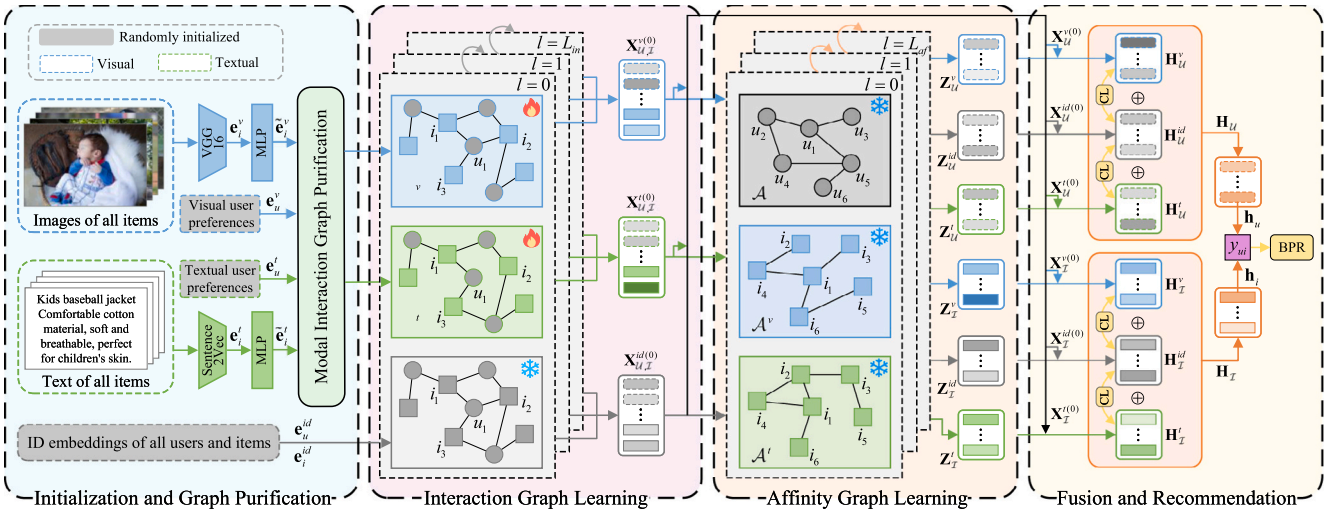


Fig. 2. Overview of the PEARL framework. The frozen graph and the tunable graph are labeled as snowflake (❄) and spark (🔥), respectively.

Task definition. The objective of multimodal recommendation is to predict the likelihood of user u selecting item i by learning accurate representations of users and items from the user behavior data and the multimodal information of items.

Overview. As shown in Fig. 2, PEARL comprises four components: (i) **Initialization and Graph Purification**, which employs the pre-trained models to extract the multimodal features while randomly initializing the ID embeddings and multimodal user preferences, and performs the edge denoising to construct the modality-specific interaction graphs (🔥) based on the proposed graph purification strategy; (ii) **Interaction Graph Learning**, which applies three GCNs in parallel on the historical and modality-specific interaction graphs to learn the high-order ID embeddings and multimodal features, respectively; (iii) **Affinity Graph Learning**, which aggregates features of similar nodes in the user co-occurrence and item modality-aware affinity graphs (❄) that are both frozen and pre-constructed using the original multimodal features. (iv) **Fusion and Recommendation**, which fuses the high-order and advanced representations, as well as optimizes the output through an augmentation-free contrastive learning. Finally, items are recommended to each user based on the preference scores calculated from the final representations which are concatenated by the ID embeddings and multimodal features.

3.2. Initialization and graph purification

Different from the soft denoising or the feature-level denoising in previous works [25,28], we prune the historical interaction graph for each modality to prevent noisy signals caused by the user's imbalanced attention from propagating in graph learning. Specifically, we design a purification strategy based on the modality difference in user's attention to items as shown in Fig. 3.

Considering the difference in dimensions and semantics of the item features generated by the pre-trained models for different modalities, we project the item features of all modalities into a unified embedding space $\mathbb{R}^d$ using the Multilayer Perceptron (MLP):

$$\tilde{\mathbf{e}}_i^m = \sigma_1\left(\mathbf{e}_i^m \mathbf{W}^m + \mathbf{b}^m\right), \tag{1}$$

where $\tilde{\mathbf{e}}_i^m \in \mathbb{R}^d$ is the projected features of modality m, $\mathbf{W}^m \in \mathbb{R}^{d_m \times d}$ and $\mathbf{b}^m \in \mathbb{R}^d$ denote the transformation matrix and bias vector, respectively. The function $\sigma_1(\cdot)$ is the LeakyRelu activation function. After that, we calculate the similarity sim_{ui}^m between user u and item i under modality m using the inner product operation combined with the degrees of nodes as follows:

$$sim_{ui}^m = \frac{\left(\mathbf{e}_u^m\right)^{\mathrm{T}} \tilde{\mathbf{e}}_i^m}{\sqrt{d(u)}\sqrt{d(i)}}, \tag{2}$$

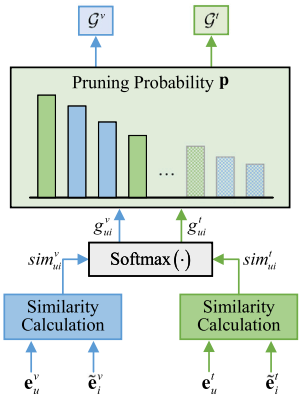


Fig. 3. Illustration of the graph purification.

where $d(\cdot)$ denotes the degree of nodes in graph $\mathcal{G}$. Then, the attention scores g_{ui}^m of user u to item i are obtained by employing the softmax function as follows:

$$g_{ui}^m = \frac{\exp(sim_{ui}^m)}{\sum_{i' \in \mathcal{N}_u} \exp(sim_{ui'}^m)}, \tag{3}$$

where $\mathcal{N}_u$ denotes the item set clicked by user u and $\exp(\cdot)$ is the exponential operation.

To prevent the pruning process from removing the connections between user u and item i across all modalities due to uneven distribution of user attention scores, we define a pruning probability $p_{\langle u,i \rangle}^m$ that determines whether an edge $\langle u, i \rangle$ in graph $\mathcal{G}$ is pruned to obtain the modality-specific interaction graph $\mathcal{G}^m$ under modality m, which can be formulated as follows:

$$p_{\langle u,i \rangle}^m = \sigma_2\left(\frac{1}{|\mathcal{M}|} \sum_{m' \in \mathcal{M}} g_{ui}^{m'} - g_{ui}^m\right), \tag{4}$$

where $\sigma_2(\cdot)$ denotes the Sigmoid activation function for normalization. The pruning probabilities of edges in the modality-specific interaction graphs under all modalities form a probability distribution $\mathbf{p} = \{p_0, p_1, \ldots, p_{|\mathcal{M}||\mathcal{E}|-1}\}$, and the pruning probabilities of edges $\langle u, i \rangle$ in different modality-specific graphs are distributed in each $\frac{1}{|\mathcal{M}|}$ segment of $\mathbf{p}$. For instance, this study only considers visual and textual modalities, so $p_{\langle u,i \rangle}^v$ and $p_{\langle u,i \rangle}^t$ are located in the first half and the second half of the probability distribution P, respectively.

Subsequently, we prune a certain percentage ρ ($0.1 \leq \rho \leq 1$) of edges in the modality-specific interaction graph. Here, we follow the

assumption that every edge in $\mathcal{G}$ appears in at least one modality-specific interaction graph, i.e., each user is affected by at least one modality. Specifically, we prune $\lfloor \rho(|\mathcal{M}|-1)|\mathcal{E}| \rfloor$ edges according to the probability distribution $\mathbf{p}$, where $\lfloor \cdot \rfloor$ denotes the downward rounding. In this way, the interactions with lower attention score in each modality has a higher probability of being pruned and edge $\langle u, i \rangle$ is retained in at least one modality. After that, the modality-specific interaction graph $\mathcal{G}^m = \{\mathcal{V}, \mathcal{E}^m\}$ is constructed, and the purified interaction matrix under modality m is $\mathbf{G}^m$, where $\sum_{m \in \mathcal{M}} |\mathcal{E}^m| = |\mathcal{M}|\,|\mathcal{E}| - \lfloor \rho(|\mathcal{M}|-1)|\mathcal{E}| \rfloor$. Since we use the projected features in Eq. (2) to calculate the similarity between users and items under different modalities, this value is continuously optimized as the model is trained, resulting in the modality-specific interaction graphs are tunable (🔥).

3.3. Interaction graph learning

To capture the high-order representations of users and items from the historical and modality-specific interaction graphs, we employ LightGCN He et al. [19] to perform the convolution operations on $\mathcal{G}$ and $\mathcal{G}^m$. This architecture is a lightweight graph convolutional network and significantly reduces the model complexity by removing the weight transformation matrices and activation functions, thus it is widely adopted in recommender systems Zhang et al. [10],Guo et al. [12],Yu et al. [21],Liu et al. [28]. Specifically, the convolution operation on the lth layer of graph $\mathcal{G}$ can be formulated as follows:

$$\mathbf{E}_{\mathcal{U},\mathcal{I}}^{id(l)} = \left((\mathbf{D})^{-\frac{1}{2}} \mathbf{M} (\mathbf{D})^{-\frac{1}{2}} \right) \mathbf{E}_{\mathcal{U},\mathcal{I}}^{id(l-1)}, \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} 0 & \mathbf{G} \\ \mathbf{G}^{\mathrm{T}} & 0 \end{bmatrix}, \tag{5}$$

where $\mathbf{M} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times(|\mathcal{U}|+|\mathcal{I}|)}$ is a symmetric adjacency matrix derived from $\mathbf{G}$, and $\mathbf{D}$ denotes a diagonal degree matrix where the diagonal element $\mathbf{D}_{aa}$ is equal to $\sum_b \mathbf{M}_{ab}$ (i.e., the sum of elements in the ath row of $\mathbf{M}$). The embeddings matrix $\mathbf{E}_{\mathcal{U},\mathcal{I}}^{id(l)} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times d}$ of the lth layer is updated only from the $(l-1)$th layer, where the initial matrix in the 0th layer is set to $\mathbf{E}_{\mathcal{U},\mathcal{I}}^{id(0)} = \{\mathbf{e}_{u_1}^{id}, \ldots, \mathbf{e}_{u_{|\mathcal{U}|}}^{id}, \mathbf{e}_{i_1}^{id}, \ldots, \mathbf{e}_{i_{|\mathcal{I}|}}^{id}\}$.

Given the absence of self-loops in the interaction graph, information of the same node from the previous layer is not retained during the message passing. Consequently, we employ a summation to integrate the ID embeddings from each layer as follows:

$$\mathbf{X}_{\mathcal{U},\mathcal{I}}^{id(0)} = \sum_{l=0}^{L_{in}} \mathbf{E}_{\mathcal{U},\mathcal{I}}^{id(l)}, \tag{6}$$

where $\mathbf{X}_{\mathcal{U},\mathcal{I}}^{id(0)}$ denotes the high-order ID embeddings of users and items, L_{in} is the number of convolutional layers in the interaction graph.

Similarly, we perform the convolution operation as in Eq. (5) on the modality-specific interaction graph $\mathcal{G}^m$ as follows:

$$\mathbf{E}_{\mathcal{U},\mathcal{I}}^{m(l)} = \left((\mathbf{D}^m)^{-\frac{1}{2}} \mathbf{M}^m (\mathbf{D}^m)^{-\frac{1}{2}} \right) \mathbf{E}_{\mathcal{U},\mathcal{I}}^{m(l-1)}, \tag{7}$$

where $\mathbf{M}^m$ is derived from $\mathbf{G}^m$ and the initial features matrix in the 0th layer is set to $\mathbf{E}_{\mathcal{U},\mathcal{I}}^{m(0)} = \{\mathbf{e}_{u_1}^m, \ldots, \mathbf{e}_{u_{|\mathcal{U}|}}^m, \tilde{\mathbf{e}}_{i_1}^m, \ldots, \tilde{\mathbf{e}}_{i_{|\mathcal{I}|}}^m\}$. Then, the high-order multimodal features $\mathbf{X}_{\mathcal{U},\mathcal{I}}^{m(0)}$ is calculated as follows:

$$\mathbf{X}_{\mathcal{U},\mathcal{I}}^{m(0)} = \sum_{l=0}^{L_{in}} \mathbf{E}_{\mathcal{U},\mathcal{I}}^{m(l)}. \tag{8}$$

So far, the user–item interaction signals and the multimodal features are encoded into the high-order representations of users and items via the above interaction graph learning.

3.4. Affinity graph learning

To further mine the hidden feature of implicit intra-modal affinity among similar users or items, we perform a graph learning on the user co-occurrence graph (❄) and the item modality-aware graph (❄) which are both frozen. These graphs are constructed using the raw multimodal features $\mathbf{e}_i^m$ and the top-n strategy prior to model training, preserving essential neighborhood relationships while maintaining computational efficiency.

User co-occurrence graph. Inspired by [53], we propose to consider the existing co-occurrence correlation between users who interact with the same items, and then construct a user co-occurrence graph to learn the similar preference for such users. Specifically, we first calculate the user co-occurrence matrix based on the user–item interactions as follows:

$$\mathbf{S} = \mathbf{G} \cdot \mathbf{G}^{\mathrm{T}}, \tag{9}$$

where the elements in $\mathbf{S} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{U}|}$ denote the count of co-occurrences between users. Considering that most users have limited co-occurrences, we employ the top-n strategy to retain users with large co-occurrences as follows:

$$\tilde{\mathbf{S}}_{a,b} = \begin{cases} 1, & \mathbf{S}_{a,b} \in \text{top-}n_u(\mathbf{S}_{a,\cdot}), \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $\tilde{\mathbf{S}}_{a,b}$ denotes the transformed co-occurrence between user u_a and user u_b, and $\mathbf{S}_{a,\cdot}$ is the elements of the ath raw in matrix $\mathbf{S}$. Consequently, we obtain the user co-occurrence graph $\mathcal{A} = \{\mathcal{U}, C\}$ (❄) based on the matrix $\tilde{\mathbf{S}}$, where $C = \{\langle u_a, u_b \rangle \,|\, u \in \mathcal{U}, \tilde{\mathbf{S}}_{a,b} = 1\}$ is an edge set that denotes the user co-occurrences. Then, we perform the graph convolution on graph $\mathcal{A}$ using the users' high-order ID embeddings $\mathbf{X}_{\mathcal{U}}^{id(0)} \in \mathbb{R}^{|\mathcal{U}|\times d}$ (obtained from Eq. (6)) and multimodal features $\mathbf{X}_{\mathcal{U}}^{m(0)} \in \mathbb{R}^{|\mathcal{U}|\times d}$ (obtained from Eq. (8)). Taking the convolution on the ID embeddings as an example, it can be formulated as follows:

$$\mathbf{X}_{\mathcal{U}}^{id(l)} = \left((\mathbf{N})^{-\frac{1}{2}} \tilde{\mathbf{S}} (\mathbf{N})^{-\frac{1}{2}} \right) \mathbf{X}_{\mathcal{U}}^{id(l-1)}, \tag{11}$$

where $\mathbf{N} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{U}|}$ is a diagonal degree matrix derived from $\tilde{\mathbf{S}}$, and its diagonal elements are equal to n_u. Unlike the interaction graph, the affinity graph constructed through the top-n strategy inherently includes the self-loops (i.e., each node is connected to itself). For each convolutional layer, the representation of a node not only aggregates the neighbor information through the message passing but also retains its own information from the previous layer via self-loops. Therefore, to avoid the information redundancy, we directly use the output of the last layer as the advanced representations of users:

$$\begin{cases} \mathbf{Z}_{\mathcal{U}}^{id} = \mathbf{X}_{\mathcal{U}}^{id(L_{af})}, \\ \mathbf{Z}_{\mathcal{U}}^{m} = \mathbf{X}_{\mathcal{U}}^{m(L_{af})}, \end{cases} \tag{12}$$

where L_{af} is the convolutional layer number in the affinity graph, $\mathbf{Z}_{\mathcal{U}}^{id}$ and $\mathbf{Z}_{\mathcal{U}}^{m}$ denote users' advanced ID embeddings and multimodal features.

Item modality-aware graph. In addition, to learn the implicit semantics between the item features, we construct the item modality-aware graphs based on the original multimodal features $\mathbf{e}_i^m$. Specifically, we first compute the similarity $\mathbf{S}_{a,b}^m$ between item i_a and i_b under modality m as follows:

$$\mathbf{S}_{a,b}^m = \cos\left(\mathbf{e}_{i_a}^m, \mathbf{e}_{i_b}^m\right) = \frac{(\mathbf{e}_{i_a}^m)^{\mathrm{T}} \mathbf{e}_{i_b}^m}{\left\|\mathbf{e}_{i_a}^m\right\| \cdot \left\|\mathbf{e}_{i_b}^m\right\|}, \tag{13}$$

where $\mathbf{S}^m \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ is a fully connected matrix, $\cos(\cdot,\cdot)$ is the cosine function, and $\|\cdot\|$ denotes the length of vectors. To reduce the computational cost and avoid the noise introduced by unnecessary connections, we similarly adopt the top-n strategy as in Eq. (10), which can be formulated as follows:

$$\tilde{\mathbf{S}}_{a,b}^m = \begin{cases} 1, & \mathbf{S}_{a,b}^m \in \text{top-}n_i(\mathbf{S}_{a,\cdot}^m), \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where $\tilde{\mathbf{S}}^m$ is a sparse matrix. Thus, we can obtain the item modality-aware graph $\mathcal{A}^m = \{\mathcal{I}, C^m\}$ (❄), where $C^m = \{\langle i_a, i_b \rangle \,|\, i \in \mathcal{I}, \tilde{\mathbf{S}}_{a,b}^m = 1\}$ is an edge set that denotes the semantic relation between items.

Then, we apply the operation as in Eq. (11) to the graph $\mathcal{A}^m$, which can be formulated as follows:

$$\mathbf{X}_{\mathcal{I}}^{m(l)} = \left((\mathbf{N}^m)^{-\frac{1}{2}}\tilde{\mathbf{S}}^m(\mathbf{N}^m)^{-\frac{1}{2}}\right)\mathbf{X}_{\mathcal{I}}^{m(l-1)}, \tag{15}$$

where $\mathbf{N}^m \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ is a diagonal degree matrix derived from $\tilde{\mathbf{S}}^m$, and its diagonal elements are equal to n_i. In addition to the multimodal features, we also propagate the ID embeddings $\mathbf{X}_{\mathcal{I}}^{id(0)} \in \mathbb{R}^{|\mathcal{I}|\times d}$ (from Eq. (6)) on the item modality-aware graph. Specifically, we perform the above operation under different item modality-aware graphs to obtain $\mathbf{X}_{\mathcal{I}}^{id,m(l)}$ using $\mathbf{X}_{\mathcal{I}}^{id(0)}$ as input. After that, the advanced representations of items are generated as follows:

$$\begin{cases}\mathbf{Z}_{\mathcal{I}}^{id} = \sum\limits_{m\in\mathcal{M}} \alpha_m \mathbf{X}_{\mathcal{I}}^{id,m(L_{af})}, \\ \mathbf{Z}_{\mathcal{I}}^{m} = \mathbf{X}_{\mathcal{I}}^{m(L_{af})},\end{cases} \tag{16}$$

where α_m denotes the importance score of the item features under modality m, $\mathbf{Z}_{\mathcal{I}}^{id}$ and $\mathbf{Z}_{\mathcal{I}}^{m}$ are the items' advanced ID embeddings and multimodal features. In this work, we introduce a hyper-parameter α as the importance score of visual modality α_v (i.e., $\alpha_v = \alpha$), and the importance score of textual modality is correspondingly set to $\alpha_t = (1-\alpha)$.

3.5. Fusion and recommendation

Upon learning the high-order and advanced representations of users and items from the interaction and affinity graph, we perform the summation to effectively capture the user–item interaction signals and their implicit intra-modal affinity. Specifically, the operation on the ID embeddings of users is formulated as follows:

$$\mathbf{H}_{\mathcal{U}}^{id} = \mathbf{X}_{\mathcal{U}}^{id(0)} + \mathbf{Z}_{\mathcal{U}}^{id}, \tag{17}$$

where $\mathbf{H}_{\mathcal{U}}^{id} \in \mathbb{R}^{|\mathcal{U}|\times d}$ denotes the global ID embeddings of users. Similarly, we can obtain $\mathbf{H}_{\mathcal{U}}^{m}$, $\mathbf{H}_{\mathcal{I}}^{id}$ and $\mathbf{H}_{\mathcal{I}}^{m}$ by performing the above operation on the multimodal features of users, the ID embeddings of items and the multimodal features of items, respectively.

To avoid the distribution bias in traditional self-supervised learning, we propose an augmentation-free contrastive learning task for multimodal recommendation, which is predicated on the inherent self-supervised signals between the user–item interaction and the multimodal features. Specifically, we use the Noise Contrastive Estimation (InfoNCE) loss [54] to maximize the similarity between ID embeddings and multimodal features of the identical node and simultaneously minimize the similarity between those belonging to different nodes. For instance, the contrastive loss $\mathcal{L}_{CL,\mathcal{U}}$ from the user side is constructed as follows:

$$\mathcal{L}_{CL,\mathcal{U}}^{m} = -\sum_{u\in\mathcal{U}'} \log \frac{\exp\left(\cos(\mathbf{h}_u^{id}, \mathbf{h}_u^{m})/\tau\right)}{\sum_{u'\in\mathcal{U}'}\exp\left(\cos(\mathbf{h}_u^{id}, \mathbf{h}_{u'}^{m})/\tau\right)}, \tag{18}$$

where $\mathcal{U}'$ denotes a set of users drawn from a batch of training data and τ is a temperature coefficient. Similarly, we compute the contrastive loss $\mathcal{L}_{CL,\mathcal{I}}^{m}$ from the item side. Summing the above results from all modalities yields the total contrastive loss $\mathcal{L}_{CL}$ as follows:

$$\mathcal{L}_{CL} = \sum_{m\in\mathcal{M}}\left(\mathcal{L}_{CL,\mathcal{U}}^{m} + \mathcal{L}_{CL,\mathcal{I}}^{m}\right). \tag{19}$$

The aforementioned method captures the interactions between the ID embeddings and modal features by maximizing their mutual information, thereby effectively enhancing the accuracy of user preference and item feature representations. Subsequently, we use the concatenation operation to fuse the ID embeddings and the multimodal features, thereby obtaining the final representations of users and items. Specifically, we integrate the concept of attention mechanisms by introducing a learnable attention weight β to regulate the proportion of visual user preferences and textual user preferences. For the item nodes, we employ the visual importance score α defined in Eq. (16) to control the fusion of the item features. This can be formulated as follows:

$$\begin{cases}\mathbf{H}_{\mathcal{U}} = \mathbf{H}_{\mathcal{U}}^{id} \oplus \beta\mathbf{H}_{\mathcal{U}}^{v} \oplus (1-\beta)\mathbf{H}_{\mathcal{U}}^{t}, \\ \mathbf{H}_{\mathcal{I}} = \mathbf{H}_{\mathcal{I}}^{id} \oplus \alpha\mathbf{H}_{\mathcal{I}}^{v} \oplus (1-\alpha)\mathbf{H}_{\mathcal{I}}^{t},\end{cases} \tag{20}$$

where $\mathbf{H}_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{U}|\times 3d}$ and $\mathbf{H}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|\times 3d}$ denotes the final representations of users and items, respectively, $\oplus$ denotes the concatenation operation, β is a trainable parameter. Since $\mathbf{H}_{\mathcal{I}}$ and Eq. (16) both fuse the item representations learned under different modalities, we use the same hyper-parameter α to control the importance score of visual modality. This design enables the model to learn the interactions between the modal features, thereby capturing the complex relationships well among the multimodal features.

After the above operations, following the previous recommender systems [20,21], we calculate the preference scores using the inner product operation as follows:

$$y_{ui} = (\mathbf{h}_u)^{\mathrm{T}}\mathbf{h}_i, \tag{21}$$

where y_{ui} denotes the prediction score between user u and item i, and item i has not been interacted by user u. Then, we recommend items with the top-K preference scores for each user.

To further optimize the parameters of PEARL, we employ the Bayesian Personalized Ranking (BPR) loss [55] to learn from the historical user–item interactions, which aims to generate a higher score for the positive user–item pair than the negative pair. Specifically, the loss function $\mathcal{L}_{BPR}$ is formulated as:

$$\mathcal{L}_{BPR} = -\sum_{(u,i_a,i_b)\in\mathcal{D}} \log\sigma_2\left(y_{ui_a} - y_{ui_b}\right), \tag{22}$$

where $\mathcal{D} = \{(u,i_a,i_b)|\langle u,i_a\rangle \in \mathcal{E}, \langle u,i_b\rangle \notin \mathcal{E}\}$ is a triplet set and item i_b is randomly sampled that has not been interacted by user u. The function $\sigma_2(\cdot)$ is the Sigmoid activation function.

Finally, we jointly combine the contrastive loss $\mathcal{L}_{CL}$ and the BPR loss $\mathcal{L}_{BPR}$ as the final optimization loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda_1\mathcal{L}_{CL} + \lambda_2\|\Theta\|_2^2, \tag{23}$$

where $\|\cdot\|_2^2$ indicates the L_2 regularization, Θ denotes the model parameters, λ_1 and λ_2 are hyper-parameters that balance the contrastive loss and L_2 regularization term, respectively.

4. Experiments

To validate the effectiveness of PEARL, we conduct sufficient experiments on the three publicly available datasets to address the following five research questions:

- **RQ1**: Can PEARL achieve the state-of-the-art performance compared with the diverse baselines on the multimodal recommendation task? (See Section 5.1)
- **RQ2**: How is the contribution of each component designed in PEARL and each modality utilized in PEARL to the improvement of recommendation performance? (See Section 5.2)
- **RQ3**: How is the sensitivity of PEARL to the contained hyper-parameters including the pruning percentage, top-n value, visual importance score, and loss weight? (See Section 5.3)
- **RQ4**: Why can our proposed multimodal learning framework PEARL achieve better performance than the baselines? (See Section 5.4)

4.1. Datasets and evaluation metrics

Following [10,12,21], we perform experiments on three publicly available Amazon datasets[1]: (a) Baby, (b) Sports and Outdoors, and

[1] http://jmcauley.ucsd.edu/data/amazon/links.html.

Table 2
Statistics of the experimental datasets.

Datasets	#Users	#Items	#Interactions	Density
Baby	19,445	7,050	160,792	0.117%
Sports	35,598	18,357	296,337	0.045%
Clothing	39,387	23,033	278,677	0.031%

Table 3
Hyper-parameter values for three datasets.

Datasets	ρ	α	n_u	n_i	λ_1	λ_2
Baby	0.3	0.2	30	10	1e−03	1e−04
Sports	0.3	0.2	40	5	1e−04	1e−05
Clothing	0.4	0.2	30	5	1e−03	1e−02

(c) Clothing, Shoes, and Jewelry, which we simply refer to as *Baby*, *Sports* and *Clothing*, respectively. Each of these three datasets contain pictures and text data for describing the items. Moreover, We employ the 5-core setting on the raw data to ensure that each user and item has a minimum of 5 interactions, and the statistics of the processed datasets are shown in Table 2. Furthermore, we directly utilize the 4096-dimensional visual features and 384-dimensional textual features adopted in [52], which are extracted by the pre-trained model VGG16 [13] and the sentence-transformers Sentence2Vec [14], respectively. In addition, we randomly split the historical interactions in each dataset to the training, validation, and test sets with the proportions of 8:1:1.

In addition, we employ the commonly used metrics Recall@K (denoted as R@K) and NDCG@K (denoted as N@K) to evaluate the recommendation performance of different models. Specifically, we select $K \in \{10, 20\}$ and present the mean scores across all users in the test set.

4.2. Baselines summary

To verify the superiority of our proposed model, we compare PEARL with various recommendation baselines which are categorized into the following three groups:

(i) **General recommenders**:

- **BPR** [55] is a matrix factorization (MF) based method that performs personalized ranking by comparing the item pairs for each user and uses the Bayesian methods for model optimization.
- **LightGCN** [19] simplifies the vanilla graph convolutional network by removing the layers including the nonlinear activation and the feature transformation.

(ii) **Self-supervised recommenders**:

- **SGL** [44] utilize the node dropout, random walk and edge dropout to generate three views for the user–item interaction graph to improve the robustness of representation learning.
- **NCL** [22] adopts an expectation–maximization clustering to identify nodes that are adjacent in semantics and graph structure, and utilizes these nodes to generate contrastive views that are considered as the positive sample pairs.

(iii) **Multimodal recommenders**:

- **VBPR** [8] augments the BPR method by incorporating the visual features derived from the item images, thereby uncovering the visual dimensions that influence the user preferences.
- **MMGCN** [9] performs information propagation on the user–item interaction graph using multimodal features to refine modality-aware representations of users.
- **GRCN** [20] identifies false-positive edges in the user–item interaction graph caused by accidental user clicks, and prune the edges according to the similarity scores between users and items.
- **DualGNN** [53] constructs a user-user correlation graph and enhances the user representations by mining the hidden information between users with similar preferences.
- **LATTICE** [10] contracts the item–item graphs under different modalities and mines the latent semantic information of items using information propagation.
- **SLMRec** [11] introduces self-supervised learning into multimodal recommendation by generating contrasting views through feature dropout and feature masking to obtain additional supervision signals.
- **MGCL** [28] leverages the self-supervised signals from different modality views to optimize the model with contrastive learning, thereby improving the representations of users and items.
- **MGCN** [21] denoises at the multimodal feature level using the user–item interactions and learning the relative importance of different modalities for adaptive feature fusion.
- **LGMRec** [12] designs a global hypergraph structure to capture the many-to-many dependencies between items and their attributes.

4.3. Hyper-parameter settings

To ensure fairness in the experimental comparisons, we implement PEARL and baselines with a unified framework named MMRec [49]. For all models, we assign a consistent embedding size of 64 dimensions for both users and items, and adopt the Xavier method [56] to initialize the model parameters. All experimental implementations are conducted using Pytorch and performed on an NVIDIA RTX3090Ti GPU with 24 GB of memory. And we utilize the Adam [57] optimizer for learning the model parameters with a batch size of 2048 and a learning rate of 0.001. Moreover, we adopt a maximum of 1000 epochs for training, and an early stop strategy is employed to stop training when the R@20 metric does not improve over 20 consecutive epochs on the validation set. Following [21], we set the number of GCN layers L_{in} and L_{af} as 2 and 1, respectively. And the temperature coefficient τ in the contrastive learning is set to 0.2 as in [12]. In addition, the optimal values of other hyper-parameters are determined through the grid search. Specifically, the pruning percentage ρ and the importance score α are tuned in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, the top-$n$ values n_u and n_i are searched in $\{10, 20, 30, 40, 50, 60\}$ and $\{5, 10, 15, 20, 25, 30\}$, respectively. And both the weight of contrastive loss λ_1 and the regularization coefficient λ_2 are tuned in $\{1e^{-1}, 1e^{-2}, \ldots, 1e^{-5}\}$. Table 3 summarizes the optimal hyperparameter values obtained from the experiments on the three datasets.

5. Result and discussion

In this section, we present the findings from our experimental studies and conduct analytical discussion to address the four research questions posed in Section 4. Specifically, we first discuss the overall performance of PEARL and the baselines in Section 5.1. Then, we explore the contribution of each component designed in PEARL and each modality utilized in PEARL in Section 5.2. After that, we test the sensitivity of PEARL to the hyper-parameters in Section 5.3. Finally, we provide a visualization to intuitively explain the advantages of our representation learning approach in Section 5.4.

5.1. Overall performance

To answer RQ1, we compare the performance of PEARL with the baselines on three datasets, i.e., *Baby*, *Sports* and *Clothing*. Table 4 reports the overall performance in terms of R@10, R@20, N@10, and N@20.

As for the baselines, it is evident that introducing self-supervised learning can enhance the recommendation performance. For instance, on the *Baby* dataset, the self-supervised recommender SGL improves the performance of BPR by 26.46% in terms of R@20. In addition, the introduction of multimodal features brings more advantages to the

Table 4
Overall performance of PEARL and baselines in terms of Recall and NDCG on three datasets. The optimal and suboptimal results in each column are boldfaced and underlined, respectively. Improvement percentage (improv.) indicates the ratio of the metric increment from the best baseline to PEARL. To verify the stability of the PEARL and the significance of the improvement, we take the average of the results under five random seeds. The statistical significance of improvements from the best baseline to PEARL is confirmed with a t-test.

Models	*Baby*				*Sports*				*Clothing*			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0187	0.0279	0.0103	0.0126
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0340	0.0526	0.0188	0.0236
SGL	0.0461	0.0727	0.0247	0.0316	0.0548	0.0832	0.0299	0.0373	0.0327	0.0508	0.0180	0.0226
NCL	0.0484	0.0761	0.0263	0.0331	0.0575	0.0871	0.0313	0.0340	0.0344	0.0532	0.0191	0.0239
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0280	0.0414	0.0159	0.0193
MMGCN	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0197	0.0328	0.0101	0.0135
GRCN	0.0532	0.0824	0.0282	0.0358	0.0559	0.0877	0.0306	0.0389	0.0424	0.0650	0.0225	0.0283
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
LATTICE	0.0545	0.0849	0.0289	0.0368	0.0618	0.0950	0.0334	0.0419	0.0465	0.0709	0.0258	0.0312
SLMRec	0.0538	0.0809	0.0284	0.0357	0.0660	0.0988	0.0359	0.0441	0.0441	0.0658	0.0238	0.0293
MGCL	0.0578	0.0885	0.0309	0.0386	0.0676	0.1025	0.0364	0.0452	0.0457	0.0699	0.0246	0.0310
MGCN	0.0616	0.0961	0.0337	0.0426	0.0726	0.1102	0.0392	0.0491	0.0638	0.0941	0.0343	0.0422
LGMRec	0.0641	0.0995	0.0342	0.0431	0.0712	0.1061	0.0383	0.0476	0.0557	0.0837	0.0303	0.0374
PEARL	**0.0668***	**0.1068***	**0.0355***	**0.0459***	**0.0782***	**0.1168***	**0.0423***	**0.0523***	**0.0695***	**0.1017***	**0.0375***	**0.0457***
improv.	4.21%	7.34%	3.80%	6.50%	7.71%	5.99%	7.91%	6.52%	8.93%	8.08%	9.33%	8.29%

* For p-value < 0.05.

recommender systems. And VBPR which merely linearly combines the visual features with ID embeddings shows superior performance to BPR. Moreover, by applying GCN to more effectively model the multimodal features, such as GRCN. However, MMGCN fails to distinguish the uni-modality unique features, resulting in insufficient utilization of multimodal information. Differently, DualGNN and LATTICE mine the hidden semantic information in the item–item graph to enhance the representations of items. In addition, it can be observed that recent studies (i.e., SLMRec, MGCL) integrate self-supervised learning into GCN-based multimodal recommendation, further enhancing the recommendation performance. Among these baselines, MGCN performs best on the *Sports* and *Clothing* datasets, while LGMRec outperforms on the *Baby* dataset. The reason for this performance difference lies in the feature-level denoising method in MGCN is more efficient when processing large-scale datasets, while the hypergraph structure used in LGMRec is relatively complex and more suitable for small-scale datasets.

Next, we conduct a comprehensive comparison between the baselines and PEARL. The experimental results demonstrate that PEARL significantly outperforms all baselines across different datasets. Specifically, PEARL achieves the performance improvements of 7.34%, 5.99%, and 8.08% in terms of R@20 over the suboptimal model on the three datasets, respectively. These performance gains can be attributed to two key aspects of our motivation: effective denoising based on users' imbalanced attention and the self-supervised signals mined through dual graph learning. In detail, on the one hand, PEARL employs a hard denoising strategy to eliminate the noisy edges in the fixed interaction graph, thus achieving superior performance compared to GRCN, which relies on soft denoising, and MGCN, which performs feature-level denoising. On the other hand, we develop an affinity graph learning method that more effectively mines the implicit intra-modal affinity than SLMRec and MGCN, and use an augmentation-free contrastive task to enhance the quality of representation learning. While LGMRec utilizes a complex hypergraph learning approach, PEARL obtains more effective multimodal representation by employing dual graph learning to simultaneously capture the historical interaction information and affinity relationships. According to the statistics of experimental datasets shown in Table 2, we observe that the *Baby* dataset is the smallest in scale, while the *Clothing* dataset is the largest. Notably, the results in Table 4 indicate that PEARL achieves a more significant performance improvement on the *Clothing* dataset, suggesting that PEARL holds a greater advantage when applied to large-scale datasets.

5.2. Ablation study

To answer RQ2, we perform an exhaustive ablation study to assess the impact of different components designed in PEARL and the influence of various modalities utilized in PEARL.

5.2.1. Impact of the components

We investigate the impact of the key components in PEARL by comparing the recommendation performance with its five model variants:

- PEARL $w/o\ GP$, which abandons the graph purification, i.e., the impact of noisy multimodal information is not considered in the interaction graph learning.
- PEARL $w/\ SP$, which replaces the graph purification in PEARL with the soft purification strategy in GRCN [20].
- PEARL $w/o\ AL$, which removes the affinity graph learning, i.e., the implicit intra-modal affinities between similar nodes are not aggregated.
- PEARL $w/o\ AL_{user}$, which removes the user co-occurrence graph learning and retains only the item modality-aware graph in the affinity graph learning.
- PEARL $w/o\ AL_{item}$, which removes the item modality-aware graph learning and retains only the user co-occurrence graph in the affinity graph learning.
- PEARL $w/o\ CL$, which eliminates the augmentation-free contrastive learning, i.e., the self-supervised signals between ID embeddings and multimodal features are ignored.

Based on the results presented in Table 5, which compares the performance of PEARL and the above variants across the three datasets, we can obtain the following observations:

(1) Comparing PEARL and PEARL $w/o\ GP$, we can see that abandoning the GP module results in a performance decline. This indicates that constructing accurate modality-specific user–item graphs in the interaction graph learning is crucial, and our designed graph purification strategy is effective for eliminating the noisy edges. Furthermore, comparing PEARL and PEARL $w/\ SP$, we find that replacing the hard purification strategy of GP module with the soft purification strategy in GRCN reduces the recommendation performance, especially on sparse datasets (e.g., *Clothing* dataset). This phenomenon is also presented in Table 4, where GRCN performs well on the denser datasets but poorly on the sparse datasets. In addition, it can be observed that our hard purification strategy improves the recommendation performance stably, which reflects the robustness of our purification strategy.

Table 5
Performance of variants that remove different components in PEARL. "w/o" denotes the removal of the corresponding component from PEARL. "$w/$" denotes the new component replaced in PEARL. ↓ introduces the percentage reduction in the variant's performance over PEARL.

Model variant	*Baby*		*Sports*		*Clothing*	
	R@20	N@20	R@20	N@20	R@20	N@20
PEARL	**0.1068**	**0.0459**	**0.1168**	**0.0523**	**0.1017**	**0.0457**
$w/o\ GP$	$0.1032_{\downarrow 3.4\%}$	$0.0443_{\downarrow 3.5\%}$	$0.1125_{\downarrow 3.7\%}$	$0.0504_{\downarrow 3.6\%}$	$0.0980_{\downarrow 3.6\%}$	$0.0439_{\downarrow 3.9\%}$
$w/\ SP$	$0.1049_{\downarrow 1.8\%}$	$0.0451_{\downarrow 1.7\%}$	$0.1136_{\downarrow 2.7\%}$	$0.0511_{\downarrow 2.3\%}$	$0.0982_{\downarrow 3.4\%}$	$0.0441_{\downarrow 3.5\%}$
$w/o\ AL$	$0.0958_{\downarrow 10.3\%}$	$0.0417_{\downarrow 9.1\%}$	$0.1029_{\downarrow 11.9\%}$	$0.0459_{\downarrow 12.3\%}$	$0.0782_{\downarrow 23.1\%}$	$0.0348_{\downarrow 23.8\%}$
$w/o\ AL_{user}$	$0.1049_{\downarrow 1.8\%}$	$0.0451_{\downarrow 1.7\%}$	$0.1151_{\downarrow 1.5\%}$	$0.0515_{\downarrow 1.5\%}$	$0.1000_{\downarrow 1.7\%}$	$0.0449_{\downarrow 1.8\%}$
$w/o\ AL_{item}$	$0.0995_{\downarrow 6.9\%}$	$0.0433_{\downarrow 5.6\%}$	$0.1075_{\downarrow 8.0\%}$	$0.0480_{\downarrow 8.2\%}$	$0.0821_{\downarrow 19.3\%}$	$0.0366_{\downarrow 19.9\%}$
$w/o\ CL$	$0.1029_{\downarrow 3.7\%}$	$0.0442_{\downarrow 3.7\%}$	$0.1135_{\downarrow 2.8\%}$	$0.0509_{\downarrow 2.7\%}$	$0.0982_{\downarrow 3.4\%}$	$0.0443_{\downarrow 3.1\%}$

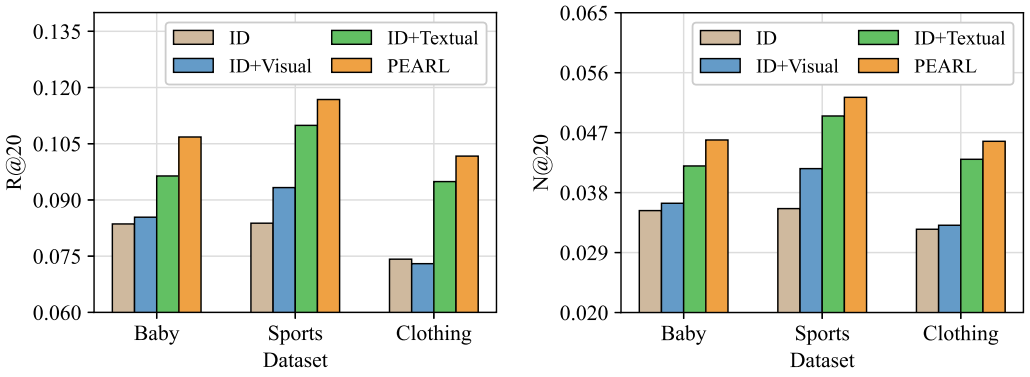


Fig. 4. Performance comparison of incorporating different modalities on three datasets.

(2) Utilizing affinity graph learning obviously enhances the model performance, especially on the *Clothing* dataset, demonstrating that datasets with a larger number of users and items contain more implicit intra-modal affinity. Moreover, $w/o\ AL_{item}$ generally underperforms $w/o\ AL_{user}$, which indicates that the item modality-aware graph learning is the primary reason for the performance improvement of affinity graph learning. This is because the user co-occurrence graph is constructed based on user–item interactions, and some user relational information has already been extracted through the interaction graph learning. Differently, the item modality-aware graphs are constructed based on the original multimodal features, capturing the hidden semantic information between items beyond the collaborative signals.

(3) Comparing PEARL and PEARL $w/o\ CL$, it is evident that contrastive learning effectively improves the recommendation performance, which reflects that self-supervised learning without data augmentation is feasible in the multimodal field. Specifically, the representation learning of the ID embeddings and the multimodal features can be further enhanced by maximizing the mutual information between them. In addition, a notable phenomenon is that contrastive learning performs relatively poorly on the *Sports* dataset, which may be due to that the *Sports* dataset contains a more diverse range of product types with obvious feature differences, making the representation learning more challenging.

5.2.2. Impact of the modalities

To investigate the influence of each modality utilized in PEARL on the recommendation accuracy, we incorporate different unimodal features into the ID embeddings in Eq. (20), constructing the following variants: (1) **ID**, which contains only the ID embeddings (i.e., $\mathbf{H}_{\mathcal{U}} = \mathbf{H}^{id}_{\mathcal{U}}$); (2) **ID+Visual**, which includes the ID embeddings and the visual features (i.e., $\mathbf{H}_{\mathcal{U}} = \mathbf{H}^{id}_{\mathcal{U}} \oplus \mathbf{H}^{v}_{\mathcal{U}}$); (3) **ID+Textual**, which includes the ID embeddings and the textual features (i.e., $\mathbf{H}_{\mathcal{U}} = \mathbf{H}^{id}_{\mathcal{U}} \oplus \mathbf{H}^{t}_{\mathcal{U}}$). The results are presented in Fig. 4.

As shown in Fig. 4, the performance of PEARL is superior to the above variants in terms of R@20 and N@20, demonstrating the effectiveness of our designed multimodal fusion framework. In addition, it can be observed that both visual and textual modalities can enhance the recommendation performance, where the effect of the textual modality is more significant. Particularly, the R@20 metric decreases after incorporating the visual features on the *Clothing* dataset, which is attributed to the larger propensity of users to make purchasing decisions based on the text descriptions. Moreover, the existing visual pre-trained models cannot accurately identify the parts of images that stimulate user's purchasing behavior, resulting in the visual features containing a large amount of noise and offering limited assistance in the performance enhancement. In addition, to verify this hypothesis, we conduct further research in Section 5.3.3.

5.3. Hyper-parameter sensitivity analysis

To answer RQ3, we conduct the hyper-parameter sensitivity experiments on the pruning percentage ρ, the top-n value in the affinity graphs, the visual importance score α, as well as the loss weights λ_1 and λ_2.

5.3.1. Impact of the pruning percentage ρ

In the graph purification module, ρ controls the number of edges in the modality-specific interaction graphs $\mathcal{G}^m$. Based on the provision in Section 3.2, the maximum value of ρ is 1. Therefore, we vary the ρ value from 0.1 to 1.0 with a step of 0.1.

As shown in Fig. 5, the experimental results demonstrate that the model performance on three datasets initially increases and then declines as the value of ρ raises. Specifically, increasing the value of ρ in a certain range (i.e., 0.1 to 0.3 on the *Baby* and *Sports* dataset, while 0.1 to 0.4 on the *Clothing* dataset) can effectively prune noisy edges in the graph structure to enhance the model performance. However, the effective connections between users and items are truncated once ρ exceeds a certain threshold, reducing the necessary training data. To be precise, the recommendation accuracy reaches the optimal value with $\rho = 0.3$ on the *Baby* and *Sports* datasets, whereas with $\rho = 0.4$ on the *Clothing* dataset. This distinction may arise because the visual and textual feature differences are more obvious in the *Clothing* dataset,

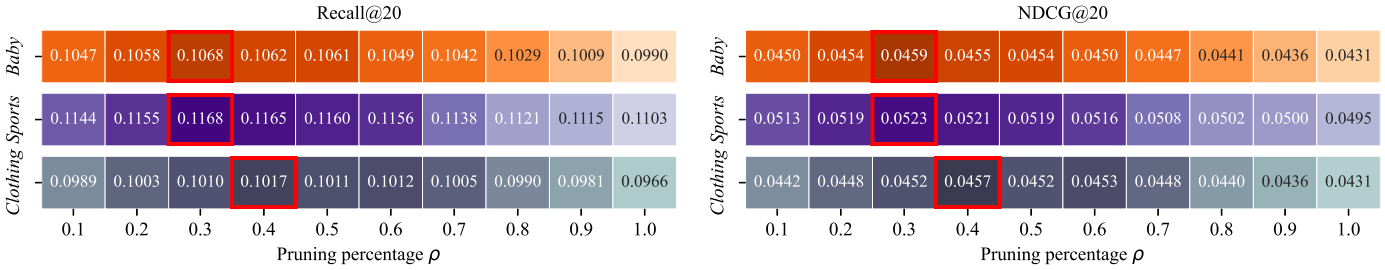
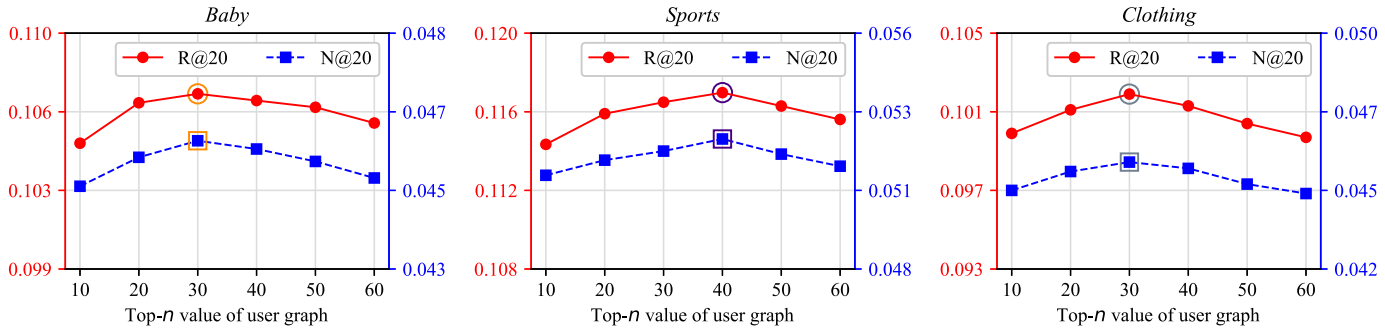
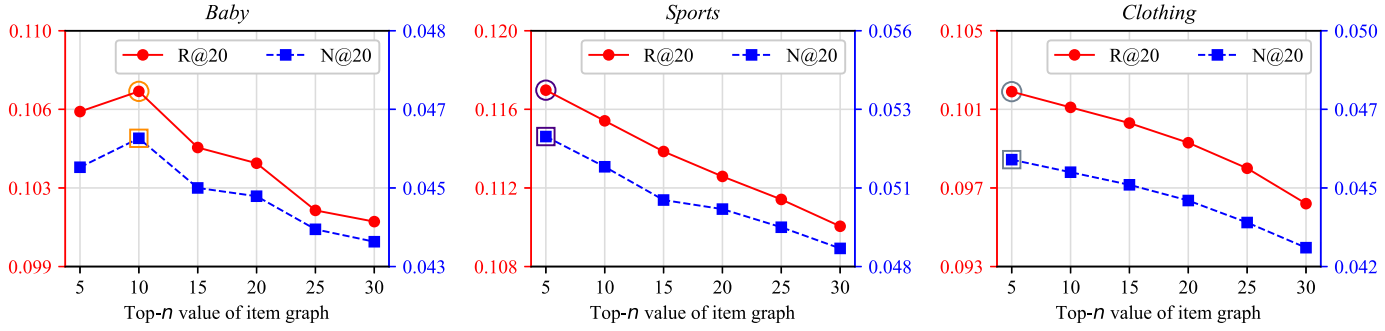


Fig. 5. Model performance under the pruning percentage ρ ranging from 0.1 to 1.0 on three datasets. The optimal value is highlighted by a red box.



(a) In the user con-occurrence graph



(b) In the item modality-aware graph

Fig. 6. Model performance under different top-n values of user graph and item graph on three datasets. The optimal value is highlighted by a colored box.

with more users being influenced by only one modality to inducing the purchasing behavior. This results in a higher number of noisy edges in the modality-specific interaction graphs, thereby requiring a higher value of ρ for pruning the noisy edges.

5.3.2. Impact of the top-n value in affinity graph

To reduce the memory consumption and prevent the message propagation between irrelevant nodes in the affinity graph learning module, we employ the top-n strategy to ensure that every user or item is only connected to the n most similar users or items. In this section, we investigate the optimal values of n_u and n_i on three datasets in the ranges of $\{10, 20, 30, 40, 50, 60\}$ and $\{5, 10, 15, 20, 25, 30\}$, respectively. The results are presented in Fig. 6.

First, Fig. 6(a) shows the impact of n_u in the user co-occurrence graph learning, where it can be observed that $n_u = 30$ is the optimal value of the neighbor number on the *Baby* and *Clothing* datasets, while this value is 40 on the *Sports* dataset. We analyze this is due to that the larger number of interactions in the *Sports* dataset leads to more similar behaviors among users [52]. Moreover, Fig. 6(b) shows the impact of n_i in the item modality-aware learning. It is evident that PEARL achieves the optimal performance with $n_i = 10$ on the *Baby* dataset and $n_i = 5$ on the other two datasets. Overall, a small value of n_i can reduce noise from the irrelevant neighbors. In addition, it is noticeable that the model performance is less sensitive to the setting of n_u than that of n_i. This observation aligns with the conclusion drawn from the phenomenon of $w/o\ AL_{item} < w/o\ AL_{user}$ in Section 5.2.1.

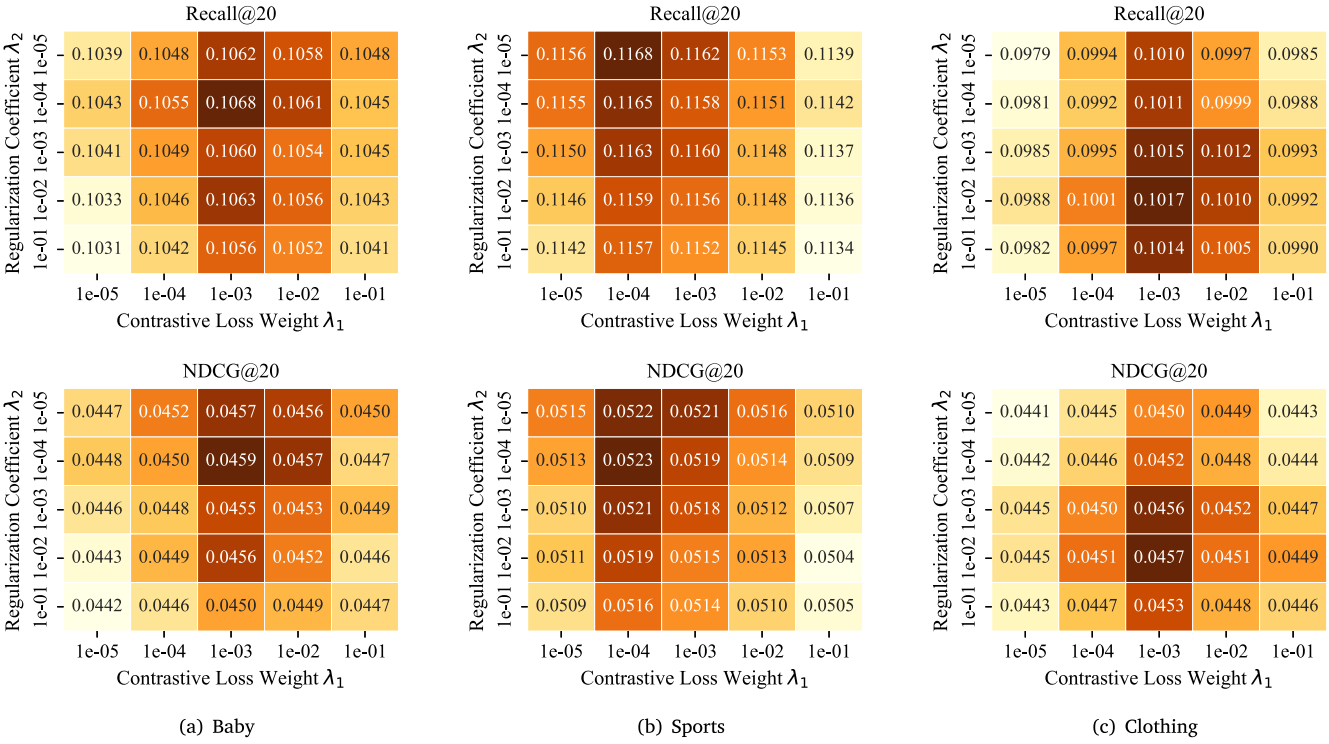
5.3.3. Impact of the visual importance score α

In Eqs. (16) and (20), the visual importance score α controls the trade-off weights between the visual and textual features, while the corresponding coefficient for the textual features is $1 - \alpha$. Building on the findings from Section 5.2.2, to further investigate the degree of contribution from each modality, we adjust the value of α from 0.1 to 0.9 with a step of 0.1.

The experimental results are shown in Table 6, where we can observe the same trend across the three datasets. Specifically, on both the metrics of R@20 and N@20, the optimal result is achieved when $\alpha = 0.2$, i.e., $\alpha_v = 0.2$ and $\alpha_t = 0.8$. The larger value of α_t reflects that

Table 6
Model performance under the visual importance score α ranging from 0.1 to 0.9 on three datasets. The optimal value is boldfaced.

Dataset	Metrics	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Baby	R@20	0.1059	**0.1068**	0.1062	0.1046	0.1042	0.1029	0.1003	0.0962	0.0944
	N@20	0.0456	**0.0459**	0.0457	0.0454	0.0452	0.0448	0.0440	0.0428	0.0418
Sports	R@20	0.1167	**0.1168**	0.1163	0.1142	0.1128	0.1114	0.1090	0.1049	0.1024
	N@20	0.0522	**0.0523**	0.0521	0.0510	0.0502	0.0496	0.0483	0.0472	0.0458
Clothing	R@20	0.1012	**0.1017**	0.1014	0.0983	0.0975	0.0965	0.0940	0.0907	0.0864
	N@20	0.0455	**0.0457**	0.0456	0.0448	0.0441	0.0436	0.0421	0.0412	0.0391



(a) Baby (b) Sports (c) Clothing

Fig. 7. Model performance under the contrastive loss weight λ_1 and regularization coefficient λ_2 in $\{1e^{-1}, 1e^{-2}, \ldots, 1e^{-5}\}$.

the textual features contain more information than the visual features. Furthermore, the accuracy of recommendations shows a continuous decline when the value of α exceeds 0.2, with this trend being particularly obvious on the *Clothing* dataset. Relating this observation to the conclusion obtained in Section 5.2.2, we are more convinced that the disparity in importance between the textual and visual modalities is more significant in the *Clothing* dataset.

5.3.4. Impact of the loss weights λ_1 and λ_2

In the final loss, λ_1 and λ_2 control the weights of the contrastive loss and the L_2 regularization, respectively. To investigate the impact of λ_1 and λ_2, we perform a grid search in $\{1e^{-1}, 1e^{-2}, \ldots, 1e^{-5}\}$ to find the optimal values of λ_1 and λ_2 on three datasets.

As shown in Fig. 7, the performance of PEARL reflects different distributions across different datasets. Specifically, PEARL achieves the best results with $\lambda_1 = 1e^{-3}$ and $\lambda_2 = 1e^{-4}$ on the *Baby* dataset, $\lambda_1 = 1e^{-4}$ and $\lambda_2 = 1e^{-5}$ on the *Sports* dataset, $\lambda_1 = 1e^{-3}$ and $\lambda_2 = 1e^{-2}$ on the *Clothing* dataset, respectively. The optimal contrastive loss weight is smaller on the *Sports* dataset compared to the other two datasets, which is consistent with the conclusion in Section 5.2 that contrastive learning performs poorly on the *Sports* dataset. In addition, we observe that optimizing the regularization weight makes little contributes to the performance improvement of PEARL. However, for the weight of contrastive loss, an appropriate value of λ_1 can reduce the value scale gap between contrastive loss and BPR loss.

5.4. Visualization analysis

To answer RQ4, we compare PEARL with the advanced multi-modal recommenders that incorporate contrastive learning, including SLMRec, MGCL, and MGCN, by visualizing the item representations generated by these methods on the *Baby* dataset. To reduce the computation time, we randomly select 1500 items from the *Baby* dataset and use t-SNE [58] to map their learned feature representations of items to a 2D space. Following [21], we utilize Gaussian Kernel density estimation (KDE) [59] to plot the circular heatmaps that depict the distribution of the item representations, and further draw the corresponding "Angles-Density" graphs to more clearly depict the density at different angles.

Prior research [30] has suggested that maintaining a uniform distribution of representations helps preserve the inherent features of nodes and enhances their generalization ability, thereby improving the recommendation performance. As shown in Fig. 8, both SLMRec with data augmentation and MGCL with contrastive learning exhibit multiple dense areas in the visualization. Although MGCN mitigates distribution aggregation through feature-level denoising, its representations remain less uniform compared to those learned by PEARL. PEARL achieves a more uniform distribution by addressing two key challenges: (1) eliminating noise propagation through the graph purification, and (2) mining implicit intra-modal affinity relationships between similar users or items. These designs collectively enhance the uniformity of feature

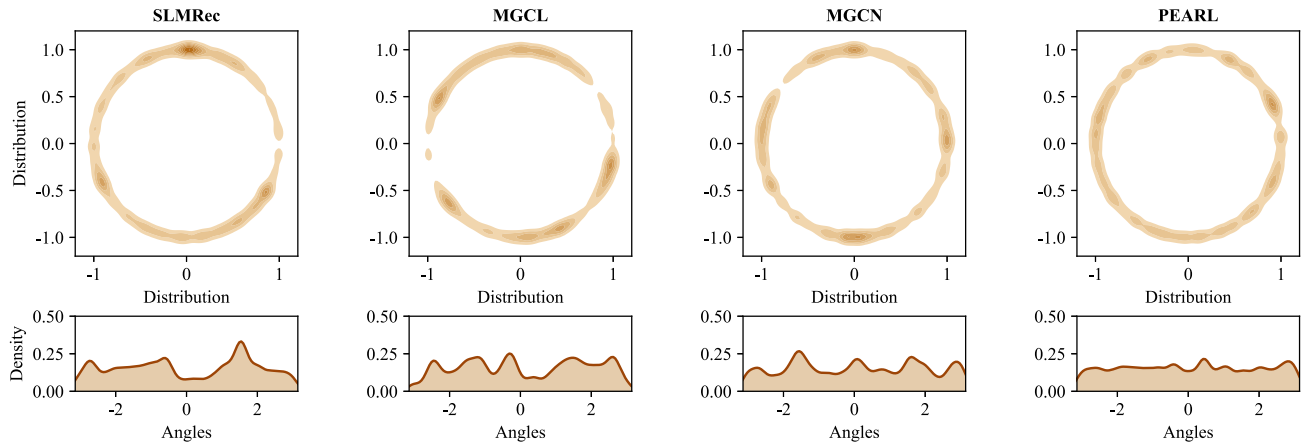


Fig. 8. The distribution of item representations learned by different models on the *Baby* dataset. The upper circular heatmaps display the distribution in $\mathbb{R}^2$, and the distribution density in different angular directions is shown in the lower figures.

representations, which in turn explains why PEARL outperforms other multimodal recommendation methods.

6. Conclusion and future work

In this work, we propose a novel method named PEARL for multimodal recommender systems (MMRS). Specifically, we first design a graph purification strategy to eliminate the misleading noisy signals and obtain the modality-specific interaction graphs. Then, an interaction graph learning module is further designed to mine the high-order representations of users and items. In addition, we construct the user co-occurrence and item modality-aware affinity graphs to learn the implicit self-supervised signals between similar nodes. Finally, we propose an augmentation-free contrastive learning task in the fusion module to assist the model training. Through sufficient experiments conducted on three publicly available datasets, PEARL achieve the obvious performance improvements compared to diverse state-of-the-art baselines in terms of R@K and N@K.

Nevertheless, we find that the visual modality is obviously less important than the textual modality in PEARL and baselines. We speculate that it may be due to the naive feature extraction and processing in the visual modality. Thus in the future works, we aim to design a multimodal feature extraction method, which eliminates the semantic ambiguities between different modalities with the help of multimodal large language models [60,61]. In addition, multimodal graph learning can be widely applied to various tasks, such as multimodal fake news detection [62], cross-modal information retrieval [63], and multimodal medical diagnosis [64]. We hope that the potential of our graph learning approach can be further demonstrated in other tasks.

CRediT authorship contribution statement

Yuzhuo Dang: Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Wanyu Chen:** Project administration, Methodology. **Zhiqiang Pan:** Writing – review & editing, Validation, Investigation, Conceptualization. **Yuxiao Duan:** Investigation, Formal analysis, Data curation. **Fei Cai:** Writing – review & editing, Supervision, Investigation. **Honghui Chen:** Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the support by the COSTA: complex system optimization team of the College of System Engineering at NUDT. This work is supported National Natural Science Foundation of China, No.: 62302511, Scientific Research Project of National University of Defense Technology, No.: ZK22-11, Independent Innovation Science foundation project of National University of Defense Technology, No.: 23-ZZCX-JDZ-43.

Data availability

Data will be made available on request.

References

[1] L. Lü, M. Medo, C.H. Yeung, Y. Zhang, Z. Zhang, T. Zhou, Recommender systems, 2012,
[2] P. Covington, J. Adams, E. Sargin, Deep neural networks for YouTube recommendations, in: RecSys, ACM, 2016, pp. 191–198, http://dx.doi.org/10.1145/2959100.2959190.
[3] Y. Guo, Y. Ling, H. Chen, A neighbor-guided memory-based neural network for session-aware recommendation, IEEE Access 8 (2020) 120668–120678, http://dx.doi.org/10.1109/ACCESS.2020.3006360.
[4] N. Khan, Z. Ma, A. Ullah, K. Polat, Similarity attributed knowledge graph embedding enhancement for item recommendation, Inf. Sci. 613 (2022) 69–95, http://dx.doi.org/10.1016/J.INS.2022.08.124.
[5] R. Xie, Z. Qiu, J. Rao, Y. Liu, B. Zhang, L. Lin, Internal and contextual attention network for cold-start multi-channel matching in recommendation, in: IJCAI, ijcai.org, 2020, pp. 2732–2738, http://dx.doi.org/10.24963/IJCAI.2020/379.
[6] Y. Zhu, K. Ge, F. Zhuang, R. Xie, D. Xi, X. Zhang, L. Lin, Q. He, Transfer-meta framework for cross-domain recommendation to cold-start users, in: SIGIR, ACM, 2021, pp. 1813–1817, http://dx.doi.org/10.1145/3404835.3463010.
[7] H. Liu, L. Wang, P. Li, C. Qian, P. Zhao, X. Wu, Relation-propagation meta-learning on an explicit preference graph for cold-start recommendation, Knowl.-Based Syst. 272 (2023) 110579, http://dx.doi.org/10.1016/J.KNOSYS.2023.110579.
[8] R. He, J.J. McAuley, VBPR: Visual Bayesian personalized ranking from implicit feedback, in: AAAI, AAAI Press, 2016, pp. 144–150, http://dx.doi.org/10.1609/AAAI.V30I1.9973.
[9] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T. Chua, MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video, in: ACM Multimedia, ACM, 2019, pp. 1437–1445, http://dx.doi.org/10.1145/3343031.3351034.
[10] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, L. Wang, Mining latent structures for multimedia recommendation, in: ACM Multimedia, ACM, 2021, pp. 3872–3880, http://dx.doi.org/10.1145/3474085.3475259.
[11] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, T. Chua, Self-supervised learning for multimedia recommendation, IEEE Trans. Multimed. 25 (2023) 5107–5116, http://dx.doi.org/10.1109/TMM.2022.3187556.
[12] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, B. Ruan, LGMRec: Local and global graph learning for multimodal recommendation, in: AAAI, AAAI Press, 2024, pp. 8454–8462, http://dx.doi.org/10.1609/AAAI.V38I8.28688.

[13] J. Ni, J. Li, J.J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: EMNLP/IJCNLP, vol. 1, Association for Computational Linguistics, 2019, pp. 188–197, http://dx.doi.org/10.18653/V1/D19-1018.

[14] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: EMNLP/IJCNLP, vol. 1, Association for Computational Linguistics, 2019, pp. 3980–3990, http://dx.doi.org/10.18653/V1/D19-1410.

[15] Q. Xu, F. Shen, L. Liu, H.T. Shen, GraphCAR: Content-aware multimedia recommendation with graph autoencoder, in: SIGIR, ACM, 2018, pp. 981–984, http://dx.doi.org/10.1145/3209978.3210117.

[16] J. Tang, X. Du, X. He, F. Yuan, Q. Tian, T. Chua, Adversarial training towards robust multimedia recommender system, IEEE Trans. Knowl. Data Eng. 32 (5) (2020) 855–867, http://dx.doi.org/10.1109/TKDE.2019.2893638.

[17] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, M.S. Kankanhalli, User diverse preference modeling by multimodal attentive metric learning, in: ACM Multimedia, ACM, 2019, pp. 1526–1534, http://dx.doi.org/10.1145/3343031.3350953.

[18] Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, T. Chua, MGAT: Multimodal graph attention network for recommendation, Inf. Process. Manag. 57 (5) (2020) 102277, http://dx.doi.org/10.1016/J.IPM.2020.102277.

[19] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, LightGCN: Simplifying and powering graph convolution network for recommendation, in: SIGIR, ACM, 2020, pp. 639–648, http://dx.doi.org/10.1145/3397271.3401063.

[20] Y. Wei, X. Wang, L. Nie, X. He, T. Chua, Graph-refined convolutional network for multimedia recommendation with implicit feedback, in: ACM Multimedia, ACM, 2020, pp. 3541–3549, http://dx.doi.org/10.1145/3394171.3413556.

[21] P. Yu, Z. Tan, G. Lu, B. Bao, Multi-view graph convolutional network for multimedia recommendation, in: ACM Multimedia, ACM, 2023, pp. 6576–6585, http://dx.doi.org/10.1145/3581783.3613915.

[22] Z. Lin, C. Tian, Y. Hou, W.X. Zhao, Improving graph collaborative filtering with neighborhood-enriched contrastive learning, in: WWW, ACM, 2022, pp. 2320–2329, http://dx.doi.org/10.1145/3485447.3512104.

[23] W. Wei, C. Huang, L. Xia, C. Zhang, Multi-modal self-supervised learning for recommendation, in: WWW, ACM, 2023, pp. 790–800, http://dx.doi.org/10.1145/3543507.3583206.

[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: NeurIPS, 2020, URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

[25] Z. Yi, X. Wang, I. Ounis, C. MacDonald, Multi-modal graph contrastive learning for micro-video recommendation, in: SIGIR, ACM, 2022, pp. 1807–1811, http://dx.doi.org/10.1145/3477495.3532027.

[26] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, F. Jiang, Bootstrap latent representations for multi-modal recommendation, in: WWW, ACM, 2023, pp. 845–854, http://dx.doi.org/10.1145/3543507.3583251.

[27] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, K. Zheng, Multi-modal knowledge graphs for recommender systems, in: CIKM, ACM, 2020, pp. 1405–1414, http://dx.doi.org/10.1145/3340531.3411947.

[28] K. Liu, F. Xue, D. Guo, P. Sun, S. Qian, R. Hong, Multimodal graph contrastive learning for multimedia-based recommendation, IEEE Trans. Multimed. 25 (2023) 9343–9355, http://dx.doi.org/10.1109/TMM.2023.3251108.

[29] W. Yang, Z. Fang, T. Zhang, S. Wu, C. Lu, Modal-aware bias constrained contrastive learning for multimodal recommendation, in: ACM Multimedia, ACM, 2023, pp. 6369–6378, http://dx.doi.org/10.1145/3581783.3612568.

[30] J. Yu, H. Yin, X. Xia, T. Chen, L. Cui, Q.V.H. Nguyen, Are graph augmentations necessary?: Simple graph contrastive learning for recommendation, in: SIGIR, ACM, 2022, pp. 1294–1303, http://dx.doi.org/10.1145/3477495.3531937.

[31] Y. Zhang, H. Zhu, Z. Song, P. Koniusz, I. King, COSTA: Covariance-preserving feature augmentation for graph contrastive learning, in: KDD, ACM, 2022, pp. 2524–2534, http://dx.doi.org/10.1145/3534678.3539425.

[32] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: ICLR (Poster), OpenReview.net, 2017, URL https://openreview.net/forum?id=SJU4ayYgl.

[33] Z. Pan, F. Cai, X. Liu, H. Chen, Distance-aware learning for inductive link prediction on temporal networks, IEEE Trans. Neural Netw. Learn. Syst. (2023).

[34] Z. Pan, F. Cai, W. Chen, H. Chen, M. de Rijke, Star graph neural networks for session-based recommendation, in: CIKM, ACM, 2020, pp. 1195–1204, http://dx.doi.org/10.1145/3340531.3412014.

[35] C. Chen, F. Cai, X. Hu, W. Chen, H. Chen, HHGN: A hierarchical reasoning-based heterogeneous graph neural network for fact verification, Inf. Process. Manag. 58 (5) (2021) 102659, http://dx.doi.org/10.1016/J.IPM.2021.102659.

[36] S. Wang, L. Hu, Y. Wang, X. He, Q.Z. Sheng, M.A. Orgun, L. Cao, F. Ricci, P.S. Yu, Graph learning based recommender systems: A review, in: IJCAI, ijcai.org, 2021, pp. 4644–4652, http://dx.doi.org/10.24963/IJCAI.2021/630.

[37] M. Gan, H. Zhang, VIGA: A variational graph autoencoder model to infer user interest representations for recommendation, Inf. Sci. 640 (2023) 119039, http://dx.doi.org/10.1016/J.INS.2023.119039.

[38] R. van den Berg, T.N. Kipf, M. Welling, Graph convolutional matrix completion, 2017, CoRR abs/1706.02263. URL http://arxiv.org/abs/1706.02263.

[39] J. Zhang, C. Gao, D. Jin, Y. Li, Group-buying recommendation for social E-commerce, in: ICDE, IEEE, 2021, pp. 1536–1547, http://dx.doi.org/10.1109/ICDE51399.2021.00136.

[40] H. Wu, J. Long, N. Li, D. Yu, M.K. Ng, Adversarial auto-encoder domain adaptation for cold-start recommendation with positive and negative hypergraphs, ACM Trans. Inf. Syst. 41 (2) (2022) 33:1–33:25, http://dx.doi.org/10.1145/3544105.

[41] Z. Han, X. Zheng, C. Chen, W. Cheng, Y. Yao, Intra and inter domain HyperGraph convolutional network for cross-domain recommendation, in: WWW, ACM, 2023, pp. 449–459, http://dx.doi.org/10.1145/3543507.3583402.

[42] K. He, H. Fan, Y. Wu, S. Xie, R.B. Girshick, Momentum contrast for unsupervised visual representation learning, in: CVPR, Computer Vision Foundation / IEEE, 2020, pp. 9726–9735, http://dx.doi.org/10.1109/CVPR42600.2020.00975.

[43] J. Sun, D. Wei, K. Ma, L. Wang, Y. Zheng, Boost supervised pretraining for visual transfer learning: Implications of self-supervised contrastive representation learning, in: AAAI, AAAI Press, 2022, pp. 2307–2315, http://dx.doi.org/10.1609/AAAI.V36I2.20129.

[44] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, X. Xie, Self-supervised graph learning for recommendation, in: SIGIR, ACM, 2021, pp. 726–735, http://dx.doi.org/10.1145/3404835.3462862.

[45] L. Xia, C. Huang, Y. Xu, J. Zhao, D. Yin, J.X. Huang, Hypergraph contrastive collaborative filtering, in: SIGIR, ACM, 2022, pp. 70–79, http://dx.doi.org/10.1145/3477495.3532058.

[46] Y. Liu, H. Xuan, B. Li, M. Wang, T. Chen, H. Yin, Self-supervised dynamic hypergraph recommendation based on hyper-relational knowledge graph, in: CIKM, ACM, 2023, pp. 1617–1626, http://dx.doi.org/10.1145/3583780.3615054.

[47] K. Zhou, H. Wang, W.X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, J. Wen, S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization, in: CIKM, ACM, 2020, pp. 1893–1902, http://dx.doi.org/10.1145/3340531.3411954.

[48] X. Xie, F. Sun, Z. Liu, S. Wu, J. Gao, J. Zhang, B. Ding, B. Cui, Contrastive learning for sequential recommendation, in: ICDE, IEEE, 2022, pp. 1259–1273, http://dx.doi.org/10.1109/ICDE53745.2022.00099.

[49] X. Zhou, MMRec: Simplifying multimodal recommendation, in: MMAsia (Workshops), ACM, 2023, pp. 6:1–6:2, http://dx.doi.org/10.1145/3611380.3628561.

[50] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha, Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation, in: SIGIR, ACM, 2019, pp. 765–774, http://dx.doi.org/10.1145/3331184.3331254.

[51] K. Zhan, C. Zhang, J. Guan, J. Wang, Graph learning for multiview clustering, IEEE Trans. Cybern. 48 (10) (2018) 2887–2895, http://dx.doi.org/10.1109/TCYB.2017.2751646.

[52] H. Zhou, X. Zhou, Z. Zeng, L. Zhang, Z. Shen, A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions, 2023, http://dx.doi.org/10.48550/ARXIV.2302.04473,

[53] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, L. Nie, DualGNN: Dual graph neural network for multimedia recommendation, IEEE Trans. Multimed. 25 (2023) 1074–1084, http://dx.doi.org/10.1109/TMM.2021.3138298.

[54] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: AISTATS, in: JMLR Proceedings, vol. 9, JMLR.org, 2010, pp. 297–304, URL http://proceedings.mlr.press/v9/gutmann10a.html.

[55] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in: UAI, AUAI Press, 2009, pp. 452–461, URL https://www.auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf.

[56] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: AISTATS, in: JMLR Proceedings, vol. 9, JMLR.org, 2010, pp. 249–256, URL http://proceedings.mlr.press/v9/glorot10a.html.

[57] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR (Poster), 2015, URL http://arxiv.org/abs/1412.6980.

[58] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

[59] G.R. Terrell, D.W. Scott, Variable kernel density estimation, Ann. Statist. (1992) 1236–1265.

[60] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, in: NeurIPS, 2023, URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[61] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H. Zheng, M. Sun, RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback, in: CVPR, IEEE, 2024, pp. 13807–13816, http://dx.doi.org/10.1109/CVPR52733.2024.01310.

[62] C. Yang, F. Zhu, J. Han, S. Hu, Invariant meets specific: A scalable harmful memes detection framework, in: ACM Multimedia, ACM, 2023, pp. 4788–4797, http://dx.doi.org/10.1145/3581783.3611761.

[63] Q. Qin, Y. Huo, L. Huang, J. Dai, H. Zhang, W. Zhang, Deep neighborhood-preserving hashing with quadratic spherical mutual information for cross-modal retrieval, IEEE Trans. Multimed. 26 (2024) 6361–6374, http://dx.doi.org/10.1109/TMM.2023.3349075.

[64] Z. Qiu, P. Yang, C. Xiao, S. Wang, X. Xiao, J. Qin, C. Liu, T. Wang, B. Lei, 3D multimodal fusion network with disease-induced joint learning for early alzheimer's disease diagnosis, IEEE Trans. Med. Imag. 43 (9) (2024) 3161–3175, http://dx.doi.org/10.1109/TMI.2024.3386937.