

Full length article

Dual-space feature representation learning network for multimodal recommender systems

Yuzhuo Dang ^a, Wanyu Chen ^b, Zhiqiang Pan ^a, Xin Zhang ^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
Modality-aligned
Modality-unique
Cross-modal item-item graph
Attentional feature fusion

ABSTRACT

Multimodal recommendation has garnered great attention for its ability to utilize both user-item interaction behaviors and multimodal data about items, enabling more accurate modeling of user preferences. Although existing methods enhance representation learning of items through modality alignment and auxiliary item-item graph learning, they face two critical drawbacks: (1) Aligning the semantic discrepancies between modalities often leads to the suppression of modality-unique features; (2) Relying solely on uni-modal item-item graphs fails to capture inter-modal interactions between items. To address these issues, we propose a novel Dual-space feature Representation learning Network (DRNet) for multimodal recommender systems. Specifically, to preserve comprehensive item features, we design a dual-space disentanglement module that maps the raw modal features of items into two independent spaces, and introduces dedicated losses for each space to help the model learn accurate modality-aligned and modality-unique features. Furthermore, to bridge the gap in inter-modal mining, we develop the Cross-modal item-item Graph Construction (CGC) strategies. By recognizing the heterogeneity of the two spaces, we utilize edge union operations in the aligned space and establish cross-modal node connections in the unique space, thereby achieving the simultaneous mining of intra-modal and inter-modal interactions. Finally, a Multi-head Attentional Feature Fusion (MAFF) mechanism is employed to adaptively integrate the dual-space feature representations. Extensive experiments on four publicly available datasets demonstrate that DRNet achieves the state-of-the-art performance compared to baselines. The code and data of DRNet are available at <https://github.com/Yuzhuo-Dang/DRNet>.

1. Introduction

Recommender systems serve as a crucial tool to mitigate the information overload by providing information filtering and personalized service across diverse platforms, including online marketplaces, social networking sites, and short video services [1–3]. Conventional recommender systems generally rely on a unitary type of data, such as user behaviors, to analyze user's interests and provide personalized recommendations [4,5]. However, with the emergence of multimodal data on on-line platforms, users increasingly perform comprehensive evaluations that integrate diverse modalities (e.g., image, text, and audio) during their decision-making processes. As a result, researchers shift their focus to advancing Multimodal Recommender Systems (MMRS), which integrate users' behavior data with items' multimodal features to learn the comprehensive user preferences [6–9]. In addition, the

processing and integration of multimodal data provides an effective solution to the issues of data sparsity and cold-start scenarios that are pervasive in conventional recommender systems, thus boosting their generalization capability and robustness [10–12].

The mainstream methods of MMRS typically leverage pre-trained models for multimodal feature extraction, subsequently integrating these features with historical interaction signals to capture the multi-granularity user preferences [13–15]. Specifically, early approaches primarily focus on fusing the interaction embeddings and multimodal features through summation, concatenation, or attention mechanism [16–18]. Recently, research focus shift towards capturing higher-order relationships by employing Graph Neural Network (GNN) [19] on the user-item bipartite graph and item-item affinity graphs [20,21]. A representative work, LATTICE [20], constructs uni-modal item-item

* Corresponding author.

E-mail addresses: dangyuzhuo@nudt.edu.cn (Y. Dang), wanyuchen@nudt.edu.cn (W. Chen), panzhiqiang@nudt.edu.cn (Z. Pan), zhangxin16@nudt.edu.cn (X. Zhang), duanyuxiao19@nudt.edu.cn (Y. Duan), caifei08@nudt.edu.cn (F. Cai), chenhonghui@nudt.edu.cn (H. Chen).<https://doi.org/10.1016/j.aei.2026.104462>

Received 2 November 2025; Received in revised form 14 January 2026; Accepted 12 February 2026

Available online 27 February 2026

1474-0346/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

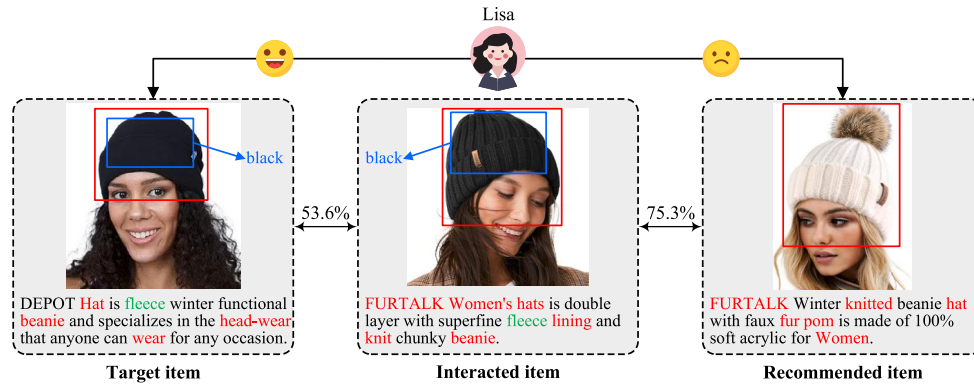


Fig. 1. A clothing recommendation case from the Amazon platform shows the impact of modality-unique features. The red boxes/text denote the modality-aligned contents, while the blue and green parts represent the modality-unique features. The center, left, and right images correspond to Lisa's interacted item, her intended purchase, and the recommendation generated by the DiffCL model, respectively. The scores between images indicate the cosine similarity calculated based on the multimodal features learned by DiffCL.

affinity graphs by assessing feature similarity metrics between items, thereby uncovering the latent structural relationships in each modality. Furthermore, several studies integrate Self-Supervised Learning (SSL) [22,23] into this field, aligning and refining item representations by exploiting inherent supervised signals between modal features, as well as between these features and interaction signals [24,25].

Despite these notable achievements, existing methods still face the challenge of adequately mining multimodal features, which is mainly reflected in the following two aspects:

(1) Multimodal alignment suppresses modality-unique features.

Existing methods primarily focus on aligning the semantic discrepancies between modalities to generate the fused multimodal features [26, 27]. However, this process acts as a double-edged sword, while it effectively captures the modality-aligned features (i.e., shared semantic commonalities across modalities, such as categories or brands), it leads to the suppression of modality-unique features (i.e., non-overlapping attributes specific to a single modality, such as visual contours or functional descriptions in text). Disregarding these modality-unique features can lead to biased user preference modeling, ultimately degrading the recommendation accuracy. As illustrated in Fig. 1, the cosine similarity between the interacted and recommended items reaches 75.3% when utilizing the modality-aligned features learned by the baseline DiffCL [28]. However, Lisa's latent intentions are “fleece” in the textual modality and “black” in the visual modality. These modality-unique features are discarded in the multimodal alignment. Consequently, the cosine similarity between the target and interacted items drops to 53.6%, leading to a recommendation failure for the target item.

(2) Lack of mining for inter-modal interactions between items.

To model the latent semantic relationships among items, some studies typically construct the uni-modal item-item affinity graph utilizing either low-dimensional item modal features learned from the modality-aware layer [20] or raw item modal features extracted via the pre-trained models [21,24]. Based on this, the item node information is propagated and aggregated through Graph Convolutional Network (GCN), thereby uncovering the semantic relationships hidden in the item modal features. However, such methods are limited to learning the intra-modal interactions within the uni-modal graph structures and cannot capture the inter-modal correlations among items. We argue that a comprehensive graph learning for item modal features should not be limited to mining the intra-modal interactions in the uni-modal item-item graphs, but should additionally construct the cross-modal item-item graphs to simultaneously capture both intra-modal and inter-modal interactions. Therefore, how to construct the effective cross-modal item-item graphs starting from the existing uni-modal item-item graphs is a key issue worthy of exploration.

To address the aforementioned challenges, we propose a novel Dual-space feature Representation learning Network (DRNet) for multimodal recommender systems. First, to mitigate the suppression of modality-unique features, we design a dual-space disentanglement module. Unlike existing methods that pursue a unified alignment, we disentangle the raw item modal features extracted from the pre-trained models into modality-aligned and modality-unique spaces, thereby both shared and specific features are effectively captured without mutual interference. Specifically, we design an alignment loss to capture shared semantics in the modality-aligned space, alongside a uniqueness loss to preserve modality-specific details in the modality-unique space that is orthogonal to modality-aligned features. Second, to bridge the gap in complex inter-modal interactions, we design the Cross-modal item-item Graph Construction (CGC) strategies. Given that the differences of feature distributions between the two disentangled spaces lead to divergent interaction logic in the uni-modal item-item graphs, we devise dedicated CGC strategies specifically tailored to each space. In detail, in the modality-aligned space, we utilize the edge union operations to aggregate item interactions across different modalities, while in the modality-unique space, we establish cross-modal node connections to mine the correlations between distinct modalities. After that, we performed graph convolution operations on the user-item interaction graph to derive the high-order collaborative signals, leveraging the learned item modal features and randomly initialized user representations. Finally, we design a Multi-head Attentional Feature Fusion (MAFF) mechanism that dynamically and adaptively integrates the modality-aligned and modality-unique item features or user representations through learnable attention weights. Extensive experiments conducted on four publicly available datasets demonstrate the superiority of DRNet above the baselines in terms of Recall@K and NDCG@K.

In summary, the major contributions of this work are summarized as follows:

- We emphasize the critical role of modality-unique features in learning feature representations for multimodal recommender systems. To this end, we design a dual-space disentanglement module to effectively capture both the modality-aligned and modality-unique features that influence user preferences.
- To accommodate the heterogeneity of feature distributions in the dual space, we devise dedicated CGC strategies for each of them, and the constructed cross-modal item-item graphs can simultaneously mine intro-modal and inter-modal interactions among the items.
- We design a MAFF mechanism that introduces an attention mechanism into the representation fusion process, which can adaptively integrates the modality-aligned and modality-unique item features or user presentations.

- We conduct extensive experiments on four publicly available datasets to validate the effectiveness of our proposed DRNet. The experimental results demonstrate that DRNet achieves the state-of-the-art performance in multimodal recommendation.

The remainder of this work is structured as follows. First, Section 2 reviews the related work on GCN-based and multimodal recommender systems. Then, Section 3 presents the notations, task definition, and the process of constructing uni-modal item–item graphs. Next, Section 4 provides a comprehensive introduction to our proposed DRNet. After that, Section 5 describes the datasets, evaluation metrics, baselines, and experimental settings, and Section 6 contains an in-depth analysis and discussion of the experimental results. Finally, we conclude this work and outline the directions for future research in Section 7.

2. Related work

In this section, we first review the representative methods of GCN-based recommender systems, followed by an introduction to the related work on multimodal recommender systems.

2.1. GCN-based recommender system

Graph Convolutional Network (GCN) demonstrates obvious advantages in various fields, including traffic prediction, drug development, and recommender systems, due to its powerful ability to model non-Euclidean spatial data [29,30]. In the scenario of recommendation, each observed user–item interaction pair constitutes an edge in the graph [31–33]. GCN is applied to this structured data makes it possible to obtain the high-order representations of users and items by propagating and aggregating features among neighborhood nodes.

The early research primarily focuses on integrating GCN into collaborative filtering architectures. For instance, Wang et al. [34] explicitly propagate user preferences through the multi-hop connections in the user–item bipartite interaction graph. However, its nonlinear transformation module results in low training efficiency. To overcome this limitation, He et al. [35] eliminate both feature transformation and nonlinear activation modules from the vanilla GCN, thus reducing the training cost and maintaining competitive performance. However, an increase in the number of graph convolution layers results in the over-smoothed representations of nodes, leading to the model losing its discriminative power [36,37]. In response, Wu et al. [38] propose SGL, which employs SSL to mitigate the node representation degradation, and Jin et al. [39] introduce a generalized framework that simultaneously reduces the inter-dimensional feature correlations while implicitly promoting the feature independence. In addition, knowledge graph-enhanced recommender systems, such as KGIN [40], which incorporates external knowledge into the recommendation, thereby mitigates data sparsity issues and improving the interpretability of recommendation results. In modeling complex interactions, the traditional GCN is limited by its local perception, making it challenging to capture long-range dependencies effectively [41,42]. Inspired by the transformer architecture [43], graph transformers are gradually emerging. Among them, Ying et al. [41] propose Graphormer, which is the first to apply transformer to GCN. It realizes full graph information propagation through the self-attention mechanism, but the computational complexity increases squarely with the number of nodes. To improve the training efficiency, Xing et al. [44] introduce a hierarchical graph transformer enhanced by collaborative training mechanisms, which partitions the complete graph into a series of sub-graphs and captures the local information within sub-graphs and the global information between sub-graphs.

The aforementioned works indicate that GCN possesses unique advantages in processing the user–item interaction graph by effectively modeling topological structures and extracting richer interaction signals, thereby enhancing the recommendation accuracy.

2.2. Multimodal recommender system

With the rapid development of multimodal representation learning technologies, recommendation based on multimodal features have gradually become a research hotspot [13,24,45]. Early methods, such as VBPR [6], validates the value of modal data for modeling the fine-grained user preferences by integrating visual features with interaction signals. Subsequently, inspired by GCN, Wei et al. [13] propose the multimodal graph convolution that employs structured neighborhood propagation to refine modal representations. In addition, Wei et al. [46] introduce an adaptive edge-weighting mechanism in the interaction graph learning to suppress the noise in implicit feedback.

However, these methods only focus on processing multimodal features on the interaction graph while neglecting the potential relationships between items or users, which could provide additional valuable insights for recommendations. Therefore, for the inter-item modeling, Zhang et al. [20] propose LATTICE, which constructs item–item affinity graphs by computing the feature similarity between items, then employs graph convolution to mine latent semantic information between items. As an optimization, Zhou and Shen [21] argue the inefficiency of dynamically constructing item–item graphs in LATTICE and propose constructing frozen item–item graphs prior to training. In addition, Guo et al. [47] build multimodal hyper-edges based on the item attributes to model global structural dependencies. On the other hand, for the inter-user modeling, Wang et al. [48] propose DualGNN, which constructs the user–user relation graph by analyzing user co-occurrence patterns, then learns modal user preferences through a modality-level attention mechanism.

Recent work has introduced SSL to explore the complex relationship between interaction signals and multimodal features. For example, Tao et al. [49] employ data augmentation techniques to generate the multi-view user–item interaction graphs and use contrastive learning to align cross-view representations of nodes. Furthermore, to avoid the complex modeling of auxiliary graphs, Zhou et al. [50] employ a simple dropout method to generate potential modal features and design three contrastive learning losses to jointly optimize user and item representations. Nonetheless, Li et al. [25] employ generative SSL models to address the insufficient modeling accuracy and visibility bias stemming from incomplete modal data. In addition, some studies point out that modal features often contain inherent noise unrelated to user preferences, which can degrade the quality of recommendations. Therefore, Yu et al. [24] utilize the user behaviors as guidance to eliminate the redundant text descriptions and image backgrounds from the raw modal features. Moreover, Kieu et al. [51] leverage Vision–Language Models for fine-grained representation learning, and propose an adaptive information-aware fusion mechanism to balance complementary and redundant signals. Recently, Ping et al. [52] recognize the behavioral complexities in multimodal interactions and propose a factor-level disentanglement framework for item multimodal features.

Although the aforementioned multimodal recommender systems have achieved significant progress, they still face limitations in modality-unique feature extraction and cross-modal graph structure learning. To overcome these limitations, we propose a dual-space representation learning to disentangle raw modal features into modality-aligned and modality-unique features. In addition, we design the construction of cross-modal item–item graphs to comprehensively explore intra-modal and inter-modal interactions among multimodal features.

3. Preliminaries

Prior to presenting our model, we first introduce the notation used in this work and define the multimodal recommendation task. Subsequently, we review the conventional method for constructing a uni-modal item–item affinity graph in previous work. This is the preliminary knowledge and basic process for building cross-modal item–item graphs.

Table 1

Main notations used in DRNet.

Notations	Descriptions
$\mathcal{U} = \{u_1, u_2, \dots, u_{ \mathcal{U} }\}$	a set of users
$\mathcal{I} = \{i_1, i_2, \dots, i_{ \mathcal{I} }\}$	a set of items
$\mathbf{R}, \mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$	the matrix and graph of historical user-item interactions
$\mathcal{M} = \{v, t\}$	a set of modalities includes visual and textual modalities
$\mathbf{E}_I^m$	the raw modal features extracted from pre-trained models
$\mathcal{D}^c \mid c \in \{ma, mu\}$	the modality-aligned and modality-unique spaces
$\mathbf{E}_{I, \mathcal{I}}^{m, ma}, \mathbf{E}_{I, \mathcal{I}}^{m, mu}$	the mapped modal features of items in $\mathcal{D}^{ma}$ and $\mathcal{D}^{mu}$
$\mathcal{G}_I^{m, ma}, \mathcal{G}_I^{m, mu}$	the uni-modal item-item affinity graphs in $\mathcal{D}^{ma}$ and $\mathcal{D}^{mu}$
$\mathbf{S}^{m, ma}, \mathbf{S}^{m, mu}$	the adjacency matrices of the graphs $\mathcal{G}_I^{m, ma}$ and $\mathcal{G}_I^{m, mu}$
n_{um}	the top- n value for constructing $\mathcal{G}_I^{m, ma}$ and $\mathcal{G}_I^{m, mu}$
$\mathcal{G}_{I, \mathcal{I}}^{ma}, \mathcal{G}_{I, \mathcal{I}}^{mu}$	the cross-modal item-item graphs in $\mathcal{D}^{ma}$ and $\mathcal{D}^{mu}$
$\mathbf{S}^{ma}, \mathbf{S}^{mu}$	the adjacency matrices of the graphs $\mathcal{G}_{I, \mathcal{I}}^{ma}$ and $\mathcal{G}_{I, \mathcal{I}}^{mu}$
α	the textual importance score for constructing $\mathcal{G}_{I, \mathcal{I}}^{ma}$
n_{cm}	the top- n value for constructing $\mathcal{G}_{I, \mathcal{I}}^{mu}$
$\mathbf{X}_I^{ma}, \{\mathbf{X}_I^{v, mu}, \mathbf{X}_I^{t, mu}\}$	the enhanced modal features of items in $\mathcal{D}^{ma}$ and $\mathcal{D}^{mu}$
$\mathbf{X}_{\mathcal{U}}^{ma}, \{\mathbf{X}_{\mathcal{U}}^{v, mu}, \mathbf{X}_{\mathcal{U}}^{t, mu}\}$	the randomly initialized user representations in $\mathcal{D}^{ma}$ and $\mathcal{D}^{mu}$
$\mathbf{L}_{li}$	the layers of convolution in the item-item graph learning
$\mathbf{H}_{I, \mathcal{I}}^{ma}, \{\mathbf{H}_{I, \mathcal{I}}^{v, mu}, \mathbf{H}_{I, \mathcal{I}}^{t, mu}\}$	the high-order modal features of items in $\mathcal{D}^{ma}$ and $\mathcal{D}^{mu}$
$\mathbf{H}_{\mathcal{U}}^{ma}, \{\mathbf{H}_{\mathcal{U}}^{v, mu}, \mathbf{H}_{\mathcal{U}}^{t, mu}\}$	the high-order modal user representations in $\mathcal{D}^{ma}$ and $\mathcal{D}^{mu}$
$\mathbf{L}_{ui}$	the layers of convolution in the user-item graph learning
F	the number of attention heads in MAFF
$\mathbf{H}_{\mathcal{U}}, \mathbf{H}_I$	the global multimodal representations of users and items
d_m, d_{ma}, d_{mu}, d	the dimensions of $\mathbf{E}_I^m, \mathcal{D}^{ma}, \mathcal{D}^{mu}$, and global representations
$\hat{y}_{ui}$	the preference score of user u for item i
K	the number of items recommended to each user
$\mathcal{L}, \mathcal{L}_{bpr}, \mathcal{L}_{ma}, \mathcal{L}_{mu}$	the joint, BPR, alignment, and uniqueness loss
λ_1, λ_2	the trade-off hyper-parameters for weighting $\mathcal{L}_{ma}$ and $\mathcal{L}_{mu}$

3.1. Notations and task definition

Given a set of users $\mathcal{U}$ and a set of item $\mathcal{I}$, the user-item historical interactions are encoded as a Boolean matrix $\mathbf{R} = [r_{u,i}]$, where $r_{u,i} = 1$ indicates an observed interaction between user u and item i , otherwise $r_{u,i} = 0$. The interaction matrix $\mathbf{R}$ can be constructed as a bipartite graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{\mathcal{U} \cup \mathcal{I}\}$ represents the node set, and $\mathcal{E} = \{(u, i) \mid r_{u,i} = 1\}$ denotes the edge set. In our work, we concentrate on dealing with both visual and textual modalities, i.e., $m \in \mathcal{M} = \{v, t\}$. The raw modal features of item i are denoted as $\mathbf{E}_I^m \in \mathbb{R}^{|\mathcal{I}| \times d_m}$, where each row corresponds to a d_m -dimensional feature vector extracted from the pre-trained model for the modality m . In addition, we establish multimodal user representations $\mathbf{X}_{\mathcal{U}} = \{\mathbf{X}_{\mathcal{U}}^{ma}, \mathbf{X}_{\mathcal{U}}^{v, mu}, \mathbf{X}_{\mathcal{U}}^{t, mu}\}$ in both the modality-aligned and modality-unique spaces through random initialization. For clarity, Table 1 summarizes the main notations used in our proposed DRNet.

Task Definition. Based on the aforementioned notations, we formulate the multimodal recommendation task as: top- $K(\hat{y}_{uI}) = \mathcal{F}(\mathcal{G}, \mathbf{E}_I^m, \mathbf{X}_{\mathcal{U}})$. Here, the top- K items are recommended to user u based on the predicted preference scores $\hat{y}_{uI}$. The function $\mathcal{F}(\cdot)$ is the feature representation learning process, which consists of four components in this work: dual-space disentanglement, cross-modal item-item graph learning, user-item interaction graph learning, as well as fusion and prediction.

3.2. Uni-modal item-item graph construction

Previous studies have shown that constructing uni-modal item-item affinity graphs and applying GCN on them can effectively capture latent semantic relationship signals between items, thereby enhancing the modal feature representation of items [20,24]. Specifically, LAT-TICE [20] first assesses the semantic relationships between items by computing the cosine similarity of modal features as follows:

$$s_{a,b}^m = \frac{(\mathbf{e}_a^m)^T \mathbf{e}_b^m}{\|\mathbf{e}_a^m\| \cdot \|\mathbf{e}_b^m\|}, \quad (1)$$

where $\mathbf{S}^m = [s_{a,b}^m]_{1 \leq a,b \leq |\mathcal{I}|} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$ denotes the similarity score matrix for modality m , $\langle i_a, i_b \rangle$ is a pair of items, $\mathbf{e}_i^m$ represents the modal

features of item i , $\|\cdot\|$ indicates the computation of vector modulo, and $(\cdot)^T$ denotes the operation of matrix transpose. Furthermore, the top- n sparsification strategy [53] is employed to eliminate the unreliable edges with low confidence scores, thereby reduce both noise and computational complexity. For each item, only the n edges with highest similarity scores are retained, which is formulated as:

$$\tilde{s}_{a,b}^m = \begin{cases} 1, & s_{a,b}^m \in \text{top-}n(\mathbf{S}_{a,:}^m), \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where n is an adjustable hyper-parameter, and vector $\mathbf{S}_{a,:}^m \in \mathbb{R}^{1 \times |\mathcal{I}|}$ is the a th row of $\mathbf{S}^m$, denoting the set of similarity scores between item i_a and all items, thus $\tilde{\mathbf{S}}^m = [\tilde{s}_{a,b}^m] \in \{0, 1\}^{|\mathcal{I}| \times |\mathcal{I}|}$ is a sparse Boolean matrix that serves as the uni-modal item-item graph's adjacency matrix. As a result, we obtain the uni-modal item-item affinity graph $\mathcal{G}_I^m = (\mathcal{I}, \mathcal{E}^m)$, where $\mathcal{I}$ is the set of nodes denoting items and $\mathcal{E}^m = \{\langle i_a, i_b \rangle \mid \tilde{s}_{a,b}^m = 1\}$ denotes the edge set. In summary, the combination of Eq. (1) and Eq. (2) forms the Uni-modal item-item Graph Constructor $\mathcal{G}_I^m = \text{UGC}(\mathbf{E}_I^m, n)$.

The aforementioned uni-modal item-item graphs only can capture intra-modal semantic relationships, neglecting to explore inter-modal interactions. To address this limitation, we design Cross-modal item-item Graph Construction (CGC) strategies in this work, which enables the integration of both intra-modal and inter-modal relationships.

4. Approach

In this section, we first introduce the overall framework of DRNet in Section 4.1, followed by a detailed description of its four key modules, including dual-space disentanglement in Section 4.2, cross-modal item-item graph learning in Section 4.3, user-item interaction graph learning in Section 4.4, as well as fusion and prediction in Section 4.5. Finally, we present the optimization process of DRNet and summarize the algorithm in Section 4.6.

4.1. Overview

As illustrated in Fig. 2, the overall framework of our proposed DRNet consists of four components: (i) **Dual-Space Disentanglement**, which maps the raw modal features extracted from the pre-trained models into two independent spaces to capture the modality-aligned and modality-unique features through the designed loss functions; (ii) **Cross-Modal Item-Item Graph Learning**, which constructs the cross-modal item-item graphs based on the uni-modal item-item affinity graphs and performs convolution operations on the cross-modal item-item graphs to obtain the enhanced modal features of items; (iii) **User-Item Interaction Graph Learning**, which performs convolution operations on the user-item graph using the item modal features and randomly initialized user representations, thereby obtaining the high-order representations; (iv) **Fusion and Prediction**, which employs the designed MAFF mechanism to integrate the modal features of items from the dual spaces as well as the user representations, ultimately generating top- K recommendations based on the calculated user preference scores.

In the model optimization phase, we combine the designed alignment loss, uniqueness loss, and BPR loss to jointly train the parameters in DRNet, thereby achieving more accurate recommendations. The detailed descriptions of each components are provided in the subsequent sections.

4.2. Dual-space disentanglement

Since existing methods for multimodal feature learning primarily focus on mining the aligned features across modalities while neglecting the unique details of the specific modality, we design a dual-space disentanglement module that maps the raw modal features into the modality-aligned and modality-unique features. This module enables DRNet to effectively capture both aligned and unique information

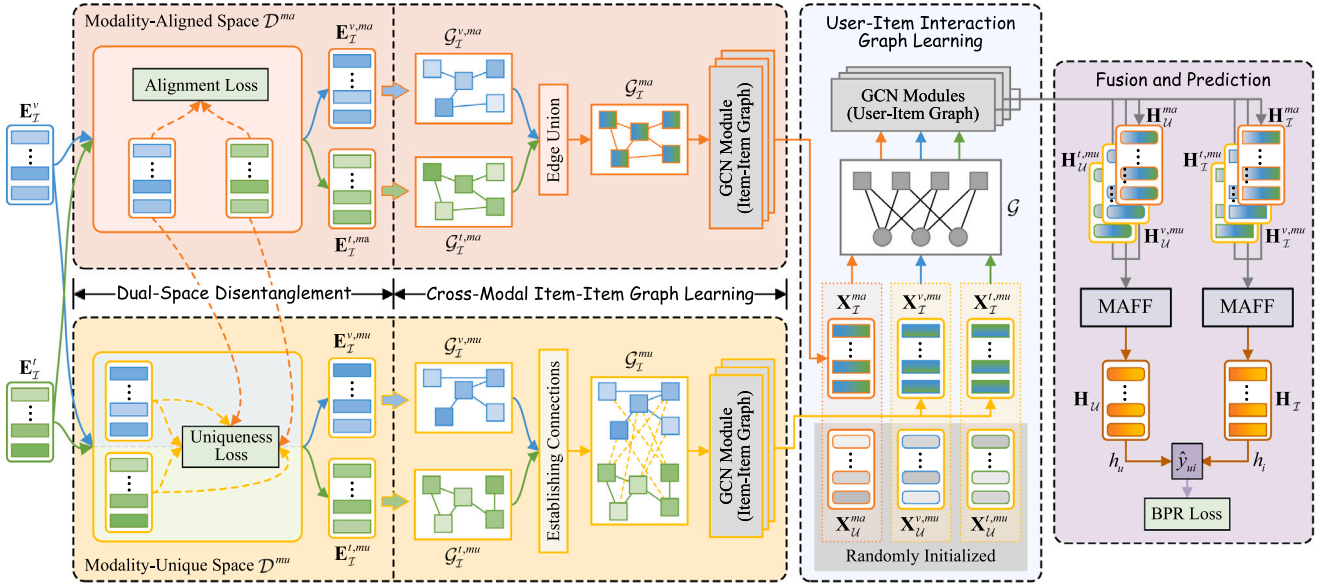


Fig. 2. The overall framework of DRNet. The user-item interaction graph $\mathcal{G}$, raw item modal features $\mathbf{E}_I^m$, and randomly initialized user representations $\mathbf{X}_U^c$ as input, and user preference scores $\hat{y}_{ui}$ as output.

across modalities. Specifically, we utilize the Multilayer Perceptron (MLP) to map the raw modal features $\mathbf{E}_I^m$ extracted from the pre-trained models [54,55] into two distinct feature spaces $\mathcal{D}^c | c \in \{ma, mu\}$.

The first component is the **modality-aligned space** $\mathcal{D}^{ma} \in \mathbb{R}^{d_{ma}}$, which captures the aligned features across modalities. Here is a description of the feature mapping process:

$$\mathbf{E}_I^{m,ma} = \sigma_1(\mathbf{E}_I^m \mathbf{W}^{m,ma} + \mathbf{b}^{m,ma}), \quad (3)$$

where $\mathbf{E}_I^{m,ma} \in \mathbb{R}^{|I| \times d_{ma}}$ denotes the modality-aligned features of items, $\mathbf{W}^{m,ma} \in \mathbb{R}^{d_m \times d_{ma}}$ and $\mathbf{b}^{m,ma} \in \mathbb{R}^{d_{ma}}$ are the transformation matrix and bias vector of the modality-aligned MLP, respectively, and $\sigma_1(\cdot)$ denotes the LeakyReLU activation function. This spatial constraint serves to minimize the modal discrepancies, thereby improving the alignment between heterogeneous modal features. For this purpose, we employ the Noise Contrastive Estimation Loss (InfoNCE) [56] as the alignment loss $\mathcal{L}_{ma}$ and minimize it to enhance the feature similarity between different modalities for identical item, which is defined as follows:

$$\mathcal{L}_{ma} = - \sum_{i \in I} \log \frac{\exp(\cos(\mathbf{e}_i^{v,ma}, \mathbf{e}_i^{t,ma})/\tau)}{\sum_{i' \in I} \exp(\cos(\mathbf{e}_i^{v,ma}, \mathbf{e}_{i'}^{t,ma})/\tau)}, \quad (4)$$

where $\mathbf{e}_i^{m,ma}$ is the row vector of $\mathbf{E}_I^{m,ma}$ representing the modality-aligned features of item i , which contains visual and textual features. In addition, $\exp(\cdot)$ denotes the natural exponential function, and τ is the temperature parameter used to modulate the smoothness of similarity distribution.

The second component is the **modality-unique space** $\mathcal{D}^{mu} \in \mathbb{R}^{d_{mu}}$, which captures the unique features across modalities. The feature mapping procedure operates according to the following formulation:

$$\mathbf{E}_I^{m,mu} = \sigma_1(\mathbf{E}_I^m \mathbf{W}^{m,mu} + \mathbf{b}^{m,mu}), \quad (5)$$

where $\mathbf{E}_I^{m,mu} \in \mathbb{R}^{|I| \times d_{mu}}$ denotes the modality-unique features of items, $\mathbf{W}^{m,mu} \in \mathbb{R}^{d_m \times d_{mu}}$ and $\mathbf{b}^{m,mu} \in \mathbb{R}^{d_{mu}}$ are the transformation matrix and bias vector of the modality-unique MLP, respectively. This spatial constraint is to maximize the inter-modal discrepancy and ensure the separation of modality-unique features from their modality-aligned features. Specifically, to maximize the modality gap in space $\mathcal{D}^{mu}$, we enforce a soft orthogonality constraint on different modal features of the same item, thereby penalizing redundant representations across modalities. Furthermore, to ensure feature disentanglement, this

constraint is extended to intra-space features, ensuring that the two spaces capture complementary rather than redundant information. In summary, the uniqueness loss $\mathcal{L}_{mu}$ is expressed as:

$$\mathcal{L}_{mu} = \|\mathbf{e}_i^{v,mu} \mathbf{e}_i^{t,mu}\|_F^2 + \|(\mathbf{e}_i^{t,ma})^T \mathbf{e}_i^{t,mu}\|_F^2 + \|(\mathbf{e}_i^{v,ma})^T \mathbf{e}_i^{v,mu}\|_F^2, \quad (6)$$

where $\mathbf{e}_i^{m,mu}$ is the row vector of $\mathbf{E}_I^{m,mu}$, corresponding to the modality-unique features of item i , and $\|\cdot\|_F^2$ denotes the squared Frobenius norm which is the method for achieving the soft orthogonality constraint [57].

4.3. Cross-modal item-item graph learning

Based on the mapped modality-aligned and modality-unique features $\mathbf{E}_I^{m,c} = \{\mathbf{E}_I^{m,ma}, \mathbf{E}_I^{m,mu}\}$ in the dual spaces, we employ the uni-modal item-item graph constructor $\text{UGC}(\mathbf{E}_I^m, n)$ introduced in Section 3.2 to construct the corresponding uni-modal item-item graphs in each space, which is defined as follows:

$$\mathcal{G}_I^{m,c} = \text{UGC}(\mathbf{E}_I^{m,c}, n_{um}), \quad (7)$$

where $\mathcal{G}_I^{m,c}$ denotes the uni-modal item-item affinity graphs in modality-aligned and modality-unique spaces, n_{um} is a hyper-parameter that determines the top- n value, and the output $\mathcal{G}_I^{m,c}$ represents the uni-modal item-item graphs, which include $\{\{\mathcal{G}_I^{v,ma}, \mathcal{G}_I^{t,ma}\}, \{\mathcal{G}_I^{v,mu}, \mathcal{G}_I^{t,mu}\}\}$. However, these graphs only contains intra-modal semantic relationships. To further obtain the inter-modal interaction information, we construct cross-modal item-item graphs building upon the uni-modal graphs. Given the obvious differences in feature distributions between the dual spaces, we design distinct strategies to achieve the Cross-modal item-item Graph Construction (CGC).

4.3.1. Cross-modal graph construction and learning in $\mathcal{D}^{ma}$

In the modality-aligned space $\mathcal{D}^{ma}$, the representations of the same nodes across different modalities maintain semantic consistency under the constraint of Eq. (4). Consequently, the uni-modal graph structures are considered to be topologically similar. However, affected by data noise or collection gaps, visual features may miss certain correlations that textual features can capture, and vice versa. To leverage the complementary associations identical node pairs across modalities, we perform an edge union operation to combine the uni-modal graphs. Specifically, the adjacency matrix of the cross-modal item-item graph

is generated through an aggregation of the adjacency matrices of the uni-modal graphs, which is mathematically expressed as:

$$s_{a,b}^{ma} = \tilde{s}_{a,b}^{v,ma} \vee \tilde{s}_{a,b}^{t,ma} = \begin{cases} 1, & \tilde{s}_{a,b}^{v,ma} = 1 \text{ or } \tilde{s}_{a,b}^{t,ma} = 1, \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where $\mathbf{S}^{ma} = [s_{a,b}^{ma}] \in \{0,1\}^{|I| \times |I|}$ represents the adjacency matrices of the cross-modal item-item graph in $\mathcal{D}^{ma}$, $\tilde{\mathbf{S}}^{v,ma}$ and $\tilde{\mathbf{S}}^{t,ma}$ are the adjacency matrices of the uni-modal item-item graphs $\mathcal{G}_I^{v,ma}$ and $\mathcal{G}_I^{t,ma}$, respectively. The symbol $\vee$ denotes the Logical OR operation between the two different elements. Based on the above, we can obtain the cross-modal item-item graph $\mathcal{G}_I^{ma} = \{I, \mathcal{E}^{ma}\}$ in $\mathcal{D}^{ma}$, where I is the set of node denoting items and $\mathcal{E}^{ma} = \{\langle i_a, i_b \rangle | s_{a,b}^{ma} = 1\}$. The node features are the fusion of visual and textual features:

$$\mathbf{E}_I^{ma} = (1 - \alpha)\mathbf{E}_I^{v,ma} \oplus \alpha\mathbf{E}_I^{t,ma}, \quad (9)$$

where $\mathbf{E}_I^{ma} \in \mathbb{R}^{|I| \times 2d_{ma}}$ denotes the node features in graph $\mathcal{G}_I^{ma}$, α is a weighting factor that balances the contributions of visual versus textual modalities in feature fusion, and $\oplus$ represents the concatenation operation. The above operations constitute the modality-aligned CGC ma , i.e., $\mathcal{G}_I^{ma} = \text{CGC}^{ma}(\mathbf{E}^{m,ma}, n_{um}, \alpha)$.

Subsequently, we employ LightGCN [35], which is a lightweight GCN model, to perform message propagation and aggregation on graph $\mathcal{G}_I^{ma}$, thereby capturing the features of neighboring nodes. In detail, the convolutional operation at the l -th layer is mathematically expressed as:

$$\mathbf{E}_I^{ma(l)} = \left((\mathbf{N}^{ma})^{-\frac{1}{2}} \mathbf{S}^{ma} (\mathbf{N}^{ma})^{-\frac{1}{2}} \right) \mathbf{E}_I^{ma(l-1)}, \quad \text{and} \quad \mathbf{E}_I^{ma(0)} = \mathbf{E}_I^{ma}, \quad (10)$$

where $\mathbf{E}_I^{ma(l)}$ denotes the output of l -th layer. $\mathbf{N}^{ma}$ is a diagonal matrix computed from $\mathbf{S}^{ma}$, and $\mathbf{N}_{aa}^{ma} = \sum s_{a,:}^{ma}$. Generally, previous works combine the results of each layer to form the final output, ensuring comprehensively retention of information [35,48]. However, due to the introduction of the top- n strategy, graph $\mathcal{G}_I^{ma}$ contains self-loops, which allows the features of the previous layer to be preserved during the convolution process. Therefore, to prevent the feature redundancy, we employ the last layer's output as the enhanced multimodal features:

$$\mathbf{X}_I^{ma} = \mathbf{E}_I^{ma(L_{ii})}, \quad (11)$$

where $\mathbf{X}_I^{ma} \in \mathbb{R}^{|I| \times 2d_{ma}}$ denotes the enhanced multimodal features of items in $\mathcal{D}^{ma}$, L_{ii} is a hyper-parameter that controls the layers of convolution in the cross-modal item-item graph learning.

4.3.2. Cross-modal graph construction and learning in $\mathcal{D}^{mu}$

In the modality-unique space $\mathcal{D}^{mu}$, the representations of different modalities contain unique features and exhibit orthogonal relationships under the constraint of Eq. (6). Therefore, the graph structures of distinct modalities should be treated as independent views, and constructing a cross-modal item-item graph requires establishing connections between them. Specifically, we leverage the cosine similarity (i.e., Eq. (1)) to calculate the connection confidence scores $\mathbf{S}^{c,mu}$ between inter-modal nodes:

$$s_{a,b}^{c,mu} = \frac{(\mathbf{e}_{i_a}^{v,mu})^T \mathbf{e}_{i_b}^{t,mu}}{\|\mathbf{e}_{i_a}^{v,mu}\| \cdot \|\mathbf{e}_{i_b}^{t,mu}\|}, \quad (12)$$

and then employ the top- n strategy (i.e., Eq. (2)) to obtain the adjacency matrix $\tilde{\mathbf{S}}^{c,mu} = [\tilde{s}_{a,b}^{c,mu}] \in \{0,1\}^{|I| \times |I|}$ between uni-modal graphs:

$$\tilde{s}_{a,b}^{c,mu} = \begin{cases} 1, & s_{a,b}^{c,mu} \in \text{top-}n_{cm}(s_{a,:}^{c,mu}), \\ 0, & \text{otherwise,} \end{cases} \quad (13)$$

where n_{cm} is a hyper-parameter that control the top- n value. Subsequently, we concatenate $\tilde{\mathbf{S}}^{c,mu}$ with the adjacency matrices of the uni-modal graphs to build the united adjacency matrix $\mathbf{S}^{mu} \in \mathbb{R}^{2|I| \times 2|I|}$ of the cross-modal item-item graph in the modality-unique space as follows:

$$\mathbf{S}^{mu} = \begin{bmatrix} \tilde{\mathbf{S}}^{v,mu} & \tilde{\mathbf{S}}^{c,mu} \\ (\tilde{\mathbf{S}}^{c,mu})^T & \tilde{\mathbf{S}}^{t,mu} \end{bmatrix}, \quad (14)$$

where $\tilde{\mathbf{S}}^{v,mu}$ and $\tilde{\mathbf{S}}^{t,mu}$ is the adjacency matrices of the visual and textual item-item graphs $\mathcal{G}_I^{v,mu}$ and $\mathcal{G}_I^{t,mu}$, respectively. Then, we can construct the cross-modal item-item graph $\mathcal{G}_I^{mu} = \{I^{mu}, \mathcal{E}^{mu}\}$ in $\mathcal{D}^{mu}$, where the node set $I^{mu} = I^{v,mu} \cup I^{t,mu}$, and $\mathcal{E}^{mu} = \{\langle i_a, i_b \rangle | s_{a,b}^{mu} = 1\}$. In a word, these operations constitute the modality-unique CGC mu , i.e., $\mathcal{G}_I^{mu} = \text{CGC}^{mu}(\mathbf{E}^{m,mu}, n_{um}, n_{cm})$.

Subsequently, similar to Eq. (10), we obtain the graph convolutional output on the united graph $\mathcal{G}_I^{mu}$, which is mathematically expressed as:

$$\mathbf{E}_I^{mu(l)} = \left((\mathbf{N}^{mu})^{-\frac{1}{2}} \mathbf{S}^{mu} (\mathbf{N}^{mu})^{-\frac{1}{2}} \right) \mathbf{E}_I^{mu(l-1)}, \quad \text{and} \quad \mathbf{E}_I^{mu(0)} = \begin{bmatrix} \mathbf{E}_I^{v,mu} \\ \mathbf{E}_I^{t,mu} \end{bmatrix}, \quad (15)$$

where $\mathbf{E}_I^{mu(0)} \in \mathbb{R}^{2|I| \times d_{mu}}$ denotes the initial matrix at the 0-th layer. $\mathbf{E}_I^{mu(l)}$ denotes the output of l -th layer. $\mathbf{N}^{mu}$ is a diagonal matrix computed from $\mathbf{S}^{mu}$, and $\mathbf{N}_{aa}^{mu} = \sum s_{a,:}^{mu}$. Consistent with the operation in the modality-aligned space, we take the output of layer L_{ii} as the result:

$$\mathbf{X}_I^{mu} = \begin{bmatrix} \mathbf{X}_I^{v,mu} \\ \mathbf{X}_I^{t,mu} \end{bmatrix} = \mathbf{E}_I^{mu(L_{ii})}, \quad (16)$$

where $\mathbf{X}_I^{v,mu} \in \mathbb{R}^{|I| \times d_{mu}}$ and $\mathbf{X}_I^{t,mu} \in \mathbb{R}^{|I| \times d_{mu}}$ denotes the enhanced visual and textual features of items in $\mathcal{D}^{mu}$, respectively.

4.4. User-item interaction graph learning

To extract the high-order collaborative signals from the user-item bipartite interaction graph $\mathcal{G}$, we utilize GCN modules to propagate and aggregate the node representations. Specifically, we first initialize three embedding matrices $\mathbf{X}_U^{ma} \in \mathbb{R}^{|U| \times 2d_{ma}}$, and $\mathbf{X}_U^{v,mu}, \mathbf{X}_U^{t,mu} \in \mathbb{R}^{|U| \times d_{mu}}$ as user representations. Then, we combine the above user embeddings with the corresponding enhanced item multimodal features to construct the node feature matrices for the user-item graphs $\mathcal{G}_{U,I}^{ma}, \mathcal{G}_{U,I}^{v,mu}$ and $\mathcal{G}_{U,I}^{t,mu}$. Subsequently, three GCNs are employed in parallel to process the above three graphs. Taking the first graph $\mathcal{G}_{U,I}^{ma}$ as an example, the specific graph convolution operations are as follows:

$$\mathbf{X}_{U,I}^{ma(l)} = \left((\mathbf{D})^{-\frac{1}{2}} \mathbf{M} (\mathbf{D})^{-\frac{1}{2}} \right) \mathbf{X}_{U,I}^{ma(l-1)}, \quad (17)$$

where

$$\mathbf{X}_{U,I}^{ma(0)} = \begin{bmatrix} \mathbf{X}_U^{ma} \\ \mathbf{X}_I^{ma} \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{R} \\ \mathbf{R}^T & \mathbf{0} \end{bmatrix},$$

and $\mathbf{D}$ is a diagonal matrix computed from $\mathbf{M}$, and $\mathbf{D}_{aa} = \sum \mathbf{M}_{a,:}$. $\mathbf{X}_I^{ma}$ is the enhanced multimodal feature obtained from Eq. (11), and $\mathbf{X}_{U,I}^{mu(0)} \in \mathbb{R}^{(|U|+|I|) \times 2d_{ma}}$ denote the concatenated feature matrix of both user and item representations. Subsequently, the output of each convolutional layer are summed to obtain high-order representations, which is mathematically expressed as:

$$\mathbf{H}_{U,I}^{ma} = \begin{bmatrix} \mathbf{H}_U^{ma} \\ \mathbf{H}_I^{ma} \end{bmatrix} = \sum_{l=0}^{L_{ui}} \mathbf{X}_{U,I}^{ma(l)}, \quad (18)$$

where $\mathbf{H}_U^{ma} \in \mathbb{R}^{|U| \times 2d_{ma}}$ and $\mathbf{H}_I^{ma} \in \mathbb{R}^{|I| \times 2d_{ma}}$ denotes the high-order representations of users and items in $\mathcal{D}^{ma}$, respectively. L_{ui} is a hyper-parameter that controls the layers of convolution in the user-item interaction graph learning.

The other two GCNs also follow the operations described in Eqs. (17) and (18), resulting in the following high-order features:

$$\mathbf{H}_{U,I}^{v,mu} = \begin{bmatrix} \mathbf{H}_U^{v,mu} \\ \mathbf{H}_I^{v,mu} \end{bmatrix}, \quad \mathbf{H}_{U,I}^{t,mu} = \begin{bmatrix} \mathbf{H}_U^{t,mu} \\ \mathbf{H}_I^{t,mu} \end{bmatrix}, \quad (19)$$

where $\mathbf{H}_U^{v,mu} \in \mathbb{R}^{(|U|+|I|) \times d_{mu}}$ and $\mathbf{H}_I^{t,mu} \in \mathbb{R}^{(|U|+|I|) \times d_{mu}}$ denote the high-order representations of the visual and textual modalities in $\mathcal{D}^{mu}$, respectively.

4.5. Fusion and prediction

4.5.1. Multi-head attentional feature fusion

Motivated by the efficacy of attention mechanisms in multimodal feature fusion [33,58,59], we design a Multi-head Attentional Feature Fusion (MAFF) mechanism. This module is specifically engineered to adaptively integrate modality-aligned and modality-unique representations for both users and items, building upon established multi-head attention architectures [17,60]. Taking the fusion of user representations as an example, we first concatenate the three types of representations obtained from Eqs. (18) and (19) as follows:

$$\tilde{\mathbf{H}}_U = \mathbf{H}_U^{ma} \oplus \mathbf{H}_U^{v,mu} \oplus \mathbf{H}_U^{t,mu}, \quad (20)$$

where $\tilde{\mathbf{H}}_U \in \mathbb{R}^{|U| \times 2(d_{ma} + d_{mu})}$ denotes the composite user feature matrix, and $\oplus$ represents the concatenation operator. Then, we use a feature learning module based on the multi-head attention mechanism to capture the inter-representation relationships and dynamically adjust their relative importance, which is formulated as:

$$\text{head}_f = \text{SOFTMAX} \left(\frac{(\tilde{\mathbf{H}}_U \mathbf{W}_h^Q)(\tilde{\mathbf{H}}_U \mathbf{W}_h^K)^T}{\sqrt{d/F}} \right) (\tilde{\mathbf{H}}_U \mathbf{W}_h^V), \quad (21)$$

$$\mathbf{H}_U = (\text{head}_1 \oplus \text{head}_2 \oplus \dots \oplus \text{head}_F) \mathbf{W}^O,$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{(2d_{ma} + 2d_{mu}) \times d/F}$ denote the transformation matrices, and $\mathbf{W}^O \in \mathbb{R}^{d \times d}$ is the weight matrix for linear projection. d denotes the dimension of the global representations, and F is the number of attention heads. The output of f th self-attention module is called head_f , and the concatenated outputs from all modules are transformed through the linear layer to generate the global representation $\mathbf{H}_U$ of users. Similarly, we can obtain the global multimodal features $\mathbf{H}_I$ of items by performing the Eqs. (20) and (21), i.e., $\mathbf{H}_I = \text{MAFF}(\mathbf{H}_I^{ma}, \mathbf{H}_I^{v,mu}, \mathbf{H}_I^{t,mu})$.

4.5.2. Prediction and recommendation

For deriving the items that may be attractive to the user, we employ the inner product operation to calculate the user-item preference scores as follows:

$$\hat{y}_{ui} = \mathbf{h}_u \cdot (\mathbf{h}_i)^T, \quad (22)$$

where $\mathbf{h}_u$ and $\mathbf{h}_i$ denote the row vectors of $\mathbf{H}_U$ and $\mathbf{H}_I$, respectively, and $\hat{y}_{ui}$ represents the preference score between them. Since a higher score reflects a stronger predicted user preference, we rank the items by descending scores and select the most relevant top- K items as recommendations to the user.

4.6. Joint optimization

In this work, we leverage the implicit feedback between users and items (e.g., clicks, interactions, or purchases) as preference indicators. And, we utilize the Bayesian Personalized Ranking (BPR) loss [61], based on the assumption that user-interacted items are stronger indicators of preference than non-interacted ones, to assist in training DRNet. This loss is formulated as follows:

$$\mathcal{L}_{bpr} = - \sum_{(u,i_a,i_b) \in \mathcal{T}} \log \left(\sigma \left(\hat{y}_{ui_a} - \hat{y}_{ui_b} \right) \right), \quad (23)$$

where $\mathcal{T} = \{(u, i_a, i_b) | r_{u,i_a} = 1, r_{u,i_b} = 0\}$ is the set of sample triplets, $r_{u,i}$ is an element of the interaction matrix $\mathbf{R}$, and i_b is an item randomly selected from those not interacted with user u .

Finally, we jointly minimize the alignment loss $\mathcal{L}_{ma}$ (from Eq. (4)), uniqueness loss $\mathcal{L}_{mu}$ (from Eq. (6)), and BPR loss $\mathcal{L}_{bpr}$ to optimize the parameters in DRNet. The joint loss $\mathcal{L}$ is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{bpr} + \lambda_1 \mathcal{L}_{ma} + \lambda_2 \mathcal{L}_{mu}, \quad (24)$$

where λ_1 and λ_2 are hyper-parameters balancing the contributions of the alignment loss and uniqueness loss, respectively. To explicitly illustrate the complete training procedure of DRNet, we provide a detailed description in Algorithm 1.

Algorithm 1 Training process of DRNet.

Input: The historical user-item interaction graph $\mathcal{G}$, the raw item modal features $\mathbf{E}_I^m$.

Output: The preference score $\hat{y}_{ui}$ between users and items.

- 1: Initialize: Xavier initialized $\mathbf{X}_U^{ma}, \mathbf{X}_U^{v,mu}, \mathbf{X}_U^{t,mu}$.
- 2: **for** epoch in range(epochs) **do**
- // Dual-Space Disentanglement
- 3: $\mathbf{E}_U^{m,ma} \leftarrow \text{MLP}(\mathbf{E}_U^m)$; // Modality-Aligned Space
- 4: $\mathbf{E}_U^{m,mu} \leftarrow \text{MLP}(\mathbf{E}_U^m)$; // Modality-Unique Space
- // Cross-Modal Item-Item Graph Learning
- 5: $\mathcal{G}_I^{ma} = \{I, \mathcal{E}^{ma}\} \leftarrow \text{CGC}^{ma}(\mathbf{E}_I^{m,ma}, n_{um}, \alpha)$;
- 6: $\mathcal{G}_I^{mu} = \{I, \mathcal{E}^{mu}\} \leftarrow \text{CGC}^{mu}(\mathbf{E}_I^{m,mu}, n_{um}, n_{em})$;
- 7: $\mathbf{X}_I^{ma} \leftarrow \text{Item-ItemGCN}(\mathcal{G}_I^{ma}, \mathbf{E}_I^{m,ma})$;
- 8: $\mathbf{X}_I^{mu} \leftarrow \text{Item-ItemGCN}(\mathcal{G}_I^{mu}, \mathbf{E}_I^{m,mu})$;
- // User-Item Interaction Graph Learning
- 9: $\mathbf{H}_U^{ma}, \mathbf{H}_I^{ma} \leftarrow \text{User-ItemGCN}(\mathcal{G}, \mathbf{X}_U^{ma}, \mathbf{X}_I^{ma})$;
- 10: $\mathbf{H}_U^{v,mu}, \mathbf{H}_I^{v,mu} \leftarrow \text{User-ItemGCN}(\mathcal{G}, \mathbf{X}_U^{v,mu}, \mathbf{X}_I^{v,mu})$;
- 11: $\mathbf{H}_U^{t,mu}, \mathbf{H}_I^{t,mu} \leftarrow \text{User-ItemGCN}(\mathcal{G}, \mathbf{X}_U^{t,mu}, \mathbf{X}_I^{t,mu})$;
- // Fusion and Prediction
- 12: $\mathbf{H}_U \leftarrow \text{MAFF}(\mathbf{H}_U^{ma}, \mathbf{H}_U^{v,mu}, \mathbf{H}_U^{t,mu})$;
- 13: $\mathbf{H}_I \leftarrow \text{MAFF}(\mathbf{H}_I^{ma}, \mathbf{H}_I^{v,mu}, \mathbf{H}_I^{t,mu})$;
- 14: $\hat{y}_{ui} = \mathbf{h}_u \cdot \mathbf{h}_i$;
- // Joint Optimization
- 15: $\mathcal{L} = \mathcal{L}_{bpr} + \lambda_1 \mathcal{L}_{ma} + \lambda_2 \mathcal{L}_{mu}$;
- 16: Optimize the MLPs, $\mathbf{X}_U^{ma}, \mathbf{X}_U^{v,mu}, \mathbf{X}_U^{t,mu}$, and MAFFs with back-propagation;
- 17: **end for**
- 18: **return** $\hat{y}_{ui}$.

5. Experiments

We propose the following six research questions to guide the experiments for validating the effectiveness of our proposed DRNet.

- **RQ1:** Does DRNet outperform the state-of-the-art baselines in the multimodal recommendation task? (See Section 6.1.1)
- **RQ2:** How efficient is DRNet compared with baselines in terms of training time and memory costs? (See Section 6.1.2)
- **RQ3:** How do the components in DRNet contribute to the improvement of recommendation performance? (See Section 6.2)
- **RQ4:** Do the dimensions of the modality-aligned and modality-unique spaces affect model performance? (See Section 6.3)
- **RQ5:** What are the advantages of having two distinct CGCs designed for the dual spaces? (See Section 6.4)
- **RQ6:** What is the impact of hyper-parameters optimized through grid search on DRNet? (See Section 6.5)
- **RQ7:** Does DRNet effectively extract the modality-aligned and modality-unique features of items? (See Section 6.6)
- **RQ8:** To what extent can DRNet effectively capture the modality-unique features that are suppressed by multimodal alignment? (See Section 6.7)

5.1. Datasets and evaluation metrics

We conduct extensive experiments on four publicly available datasets to demonstrate the effectiveness of DRNet. The statistics of the datasets are shown in Table 2.

Among the datasets listed in Table 2, the first three are sourced from the Amazon platform,¹ which are (i) Baby, (ii) Sports and Outdoors, as well as (iii) Clothing, Shoes, and Jewelry. These datasets are abbreviated as *Baby*, *Sports*, and *Clothing* for conciseness. In addition, to test the robustness of DRNet on data from different platforms, we select *MicroLens*,² which is a micro-video recommendation dataset. All

¹ <http://jmcauley.ucsd.edu/data/amazon>

² <https://recsys.westlake.edu.cn/>

Table 2
Statistics of the experimental datasets.

Datasets	#Users	#Items	#Interactions	Density
<i>Baby</i>	19,445	7050	160,792	0.117%
<i>Sports</i>	35,598	18,357	296,337	0.045%
<i>Clothing</i>	39,387	23,033	278,677	0.031%
<i>MicroLens</i>	98,129	17,228	705,174	0.004%

the aforementioned datasets contain both visual and textual modalities of items. Specifically, for the three Amazon datasets, the images of items serve as the visual modality, while the titles and text descriptions are considered as the textual modality. For *MicroLens*, the covers and titles of videos are regarded as the visual and textual modalities, respectively. Following the data processing from previous works [21,24,47], the 5-core filtering is employed, retaining only users and items with more than five interactions. Furthermore, we adopt an 8:1:1 ratio to randomly split the historical interactions in each dataset into training, validation, and test sets.

Evaluation Metrics. In this work, we employ Recall@ K (top- K Recall rate) and NGCG@ K (top- K Normalized Discounted Cumulative Gain) metrics, which are commonly used in recommendation and information retrieval to evaluate model performance and abbreviated as R@ K and N@ K , respectively. Specifically, R@ K quantifies the fraction of items in the recommendation list that the user is really interested in, while N@ K considers both the ranking and position of recommended items to assess the quality of the recommendation list. Experimental results report the outputs of $K = 10$ and $K = 20$, which are calculated as the average recommendation scores in the test set.

5.2. Baselines

To demonstrate DRNet's superiority in multimodal recommendation, we evaluate it against the following three categories of existing recommendation methods:

(i) General recommenders:

- **BPR** [61] is a personalized ranking algorithm based on Bayesian posterior optimization that learns a recommender by maximizing the likelihood function of the user's preference ranking for scenarios with implicit feedback data.
- **VBPR** [6] is one of the early models that integrated uni-modal data into interaction-based recommender systems, which integrates visual features of items with the BPR to further model the visual-based user preferences.

(ii) GCN-based recommenders:

- **SGL** [38] generates multi-view interaction graphs through node dropping, edge dropping, and random walk. It then maximizes the representation consistency of identical nodes across different augmented views to mitigate the adverse effects of high-degree nodes.
- **LightGCN** [35] modifies the GCN variant by simplifying both the feature transformation and nonlinear activation modules, thereby improving computational efficiency while preserving representation learning accuracy for users and items.

(iii) Multimodal recommenders:

- **MMGCN** [13] integrates the visual, audio, and textual modal data of items with the user-item interactions, and it learns user preferences under different modalities through the modality-specific interaction graph learning.
- **GRCN** [46] proposes a soft denoising strategy that dynamically mitigates the false positive edges of user-item interactions based on the multimodal information of items, thus purifying the user-item interaction graph.

Table 3
Hyper-parameter values for four datasets.

Datasets	d_{ma}, d_{mu}, d	τ	L_{ii}	L_{ui}	F	α	n_{um}	n_{em}	λ_1	λ_2
<i>Baby</i>	64	0.2	1	2	5	0.8	10	4	1e-01	1e-03
<i>Sports</i>						0.8	5	3	1	1e-03
<i>Clothing</i>						0.8	5	2	1e-02	1e-03
<i>MicroLens</i>						0.6	5	3	1e-04	1e-01

- **LATTICE** [20] constructs modality-aware item-item affinity graphs and performs GCN on it to improve the multimodal feature representations of items, thereby uncovering latent semantic information within the modal content.
- **SLMRec** [49] incorporates SSL into the multimodal recommendation via contrastive learning, where it generate multiple views of the observed interaction graph through data augmentation techniques such as feature dropout and feature masking.
- **BM3** [50] is a lightweight and efficient model that generates contrastive views through data augmentation and aligns inter-modal and intra-modal features to integrate multimodal information.
- **MGCN** [24] mitigates the noisy signals in multimodal features with the assistance of interaction information and designs a behavior-aware fusion module that adaptively learns the importance of different modalities.
- **LGMRec** [47] models the user preference by jointly utilizing global and local graphs, where the former captures dependencies in hyper-graph structures while the latter focuses on modality-related interaction information.
- **AD-DRL** [9] disentangles the modal features of items based on the high-level and low-level attribute categories and designs a hierarchical attribute-aware representation learning to separate both intra-modal and inter-modal features.
- **DSL-MRS** [62] proposes diffusion-based structured learning to reduce the information loss in uni-modal item-item graphs cause by graph sparsification, and utilizes graph attention networks to capture complex higher-order item interactions.
- **DiffCL** [28] introduces the diffusion-enhanced contrastive learning that mitigates data noise through generative views and combines the cross-modal alignment mechanism guided by ID embeddings to improve semantic consistency.

5.3. Experimental settings

In this work, we implement all baselines and DRNet using the MMRec toolbox [63]. To ensure fair comparisons, we adopt the pre-extracted item visual and textual features, as well as the experimental settings in the MMRec. Specifically, the raw visual and textual features are extracted using the pre-trained VGG16 [54] and Sentence2Vec [55] models, with dimensions of $d_v = 4096$ and $d_t = 384$, respectively. We conduct experiments under the PyTorch framework and run them on an NVIDIA RTX4090 GPU with 24 GB memory. We initialize the model parameters via the Xavier [64] and train the models with the Adam optimizer [65]. All models employ a batch size of 2048 and learning rate of 0.001 during training. Furthermore, the training process incorporates early stopping based on R@20 performance, terminating if no improvement occurs for 20 consecutive epochs, with a maximum training duration of 1000 epochs. In DRNet, the dimensions of the modality-aligned space, modality-unique space, and global representation are set to 64, i.e., $d_{ma} = d_{mu} = d = 64$. In addition, we configure the number of attention heads F in MAFF to 5, and the temperature τ in alignment loss to 0.2 as in the previous work [47]. Following Yu et al. [24], we set the layers L_{ii} and L_{ui} of convolution in the cross-modal item-item graph learning and user-item interaction graph learning to 1 and 2, respectively. Moreover, we leverage the grid search to find the optimal values of remaining hyper-parameters. Specifically, the top- n value n_{um} for constructing the uni-modal item-item graphs

Table 4

Overall performance of baselines and DRNet on the four datasets. The upper and lower tables present the results of R@K and N@K, respectively. The optimal and suboptimal results are highlighted in **bold** and underline, respectively. Improv. indicates the percentage improvement in metrics for DRNet compared to the optimal baseline. A triangle (Δ) denotes that the improvement is statistically significant based on the t -test (p -value < 0.05).

Models	<i>Baby</i>		<i>Sports</i>		<i>Clothing</i>		<i>MicroLens</i>	
	R@10	R@20	R@10	R@20	R@10	R@20	R@10	R@20
BPR	0.0357	0.0575	0.0432	0.0653	0.0187	0.0279	0.0528	0.0818
VBPR	0.0423	0.0663	0.0558	0.0856	0.0280	0.0414	0.0563	0.0870
SGL	0.0461	0.0727	0.0548	0.0832	0.0327	0.0508	0.0638	0.0982
LightGCN	0.0479	0.0754	0.0569	0.0864	0.0340	0.0526	0.0660	0.1015
MMGCN	0.0378	0.0615	0.0370	0.0605	0.0197	0.0328	0.0416	0.0689
GRCN	0.0532	0.0824	0.0559	0.0877	0.0424	0.0650	0.0652	0.1009
LATTICE	0.0545	0.0849	0.0618	0.0950	0.0465	0.0709	0.0694	0.1056
SLMRec	0.0538	0.0809	0.0660	0.0988	0.0441	0.0658	<u>0.0751</u>	<u>0.1132</u>
BM3	0.0563	0.0881	0.0654	0.0978	0.0420	0.0618	0.0684	0.1049
MGCN	0.0616	0.0961	0.0726	0.1102	0.0638	0.0941	0.0743	0.1128
LGMRec	<u>0.0641</u>	<u>0.0995</u>	0.0693	0.1056	0.0557	0.0837	0.0702	0.1086
AD-DRL	0.0596	0.0942	0.0732	<u>0.1106</u>	0.0642	0.0947	0.0735	0.1112
DSL-MRS	0.0618	0.0970	0.0725	0.1086	0.0621	0.0916	0.0724	0.1098
DiffCL	0.0639	0.0985	<u>0.0735</u>	0.1098	<u>0.0645</u>	<u>0.0952</u>	0.0748	0.1125
DRNet	0.0652Δ	0.1015Δ	0.0778Δ	0.1149Δ	0.0678Δ	0.0996Δ	0.0783Δ	0.1175Δ
Improv.	1.72%	2.01%	5.85%	3.89%	5.12%	4.62%	4.26%	3.80%

Models	<i>Baby</i>		<i>Sports</i>		<i>Clothing</i>		<i>MicroLens</i>	
	N@10	N@20	N@10	N@20	N@10	N@20	N@10	N@20
BPR	0.0192	0.0249	0.0241	0.0298	0.0103	0.0126	0.0273	0.0347
VBPR	0.0223	0.0284	0.0307	0.0384	0.0159	0.0193	0.0293	0.0372
SGL	0.0247	0.0316	0.0299	0.0373	0.0180	0.0226	0.0333	0.0419
LightGCN	0.0257	0.0328	0.0311	0.0387	0.0188	0.0236	0.0345	0.0432
MMGCN	0.0200	0.0261	0.0193	0.0254	0.0101	0.0135	0.0205	0.0276
GRCN	0.0282	0.0358	0.0306	0.0389	0.0225	0.0283	0.0341	0.0437
LATTICE	0.0289	0.0368	0.0334	0.0419	0.0258	0.0312	0.0361	0.0455
SLMRec	0.0284	0.0357	0.0359	0.0441	0.0238	0.0293	<u>0.0396</u>	<u>0.0495</u>
BM3	0.0298	0.0382	0.0351	0.0434	0.0229	0.0282	0.0358	0.0450
MGCN	0.0337	0.0426	0.0392	0.0491	0.0343	0.0422	0.0386	0.0488
LGMRec	<u>0.0342</u>	<u>0.0431</u>	0.0373	0.0467	0.0303	0.0374	0.0366	0.0464
AD-DRL	0.0332	0.0418	0.0398	<u>0.0501</u>	0.0346	0.0428	0.0379	0.0477
DSL-MRS	0.0338	0.0425	0.0389	0.0488	0.0337	0.0415	0.0371	0.0469
DiffCL	0.0340	0.0429	<u>0.0405</u>	0.0499	<u>0.0348</u>	<u>0.0431</u>	0.0392	0.0491
DRNet	0.0351Δ	0.0444Δ	0.0427Δ	0.0523Δ	0.0373Δ	0.0455Δ	0.0414Δ	0.0515Δ
Improv.	2.63%	3.02%	5.43%	4.39%	7.18%	5.57%	4.55%	4.04%

is tuned in {5, 10, 15, 20}, the textual importance score α in CGC^{ma} is adjusted in {0.5, 0.6, 0.7, 0.8, 0.9}, and the top- n value n_{cm} in CGC^{mu} is tuned in {1, 2, 3, 4, 5}. The weights λ_1 and λ_2 of the alignment loss and uniqueness loss are both chosen from {1, 1e-01, 1e-02, 1e-03, 1e-04}. Table 3 presents the optimal hyper-parameter configurations of DRNet on four datasets.

6. Results and discussion

In this section, we provide the experimental results in tables or figures, along with in-depth discussions to answer the six research questions.

6.1. Overall performance

6.1.1. Effectiveness (RQ1)

To demonstrate DRNet outperforms the state-of-the-art baselines, we conduct experiments on four datasets. Table 4 presents the overall performance of baselines and DRNet on the R@K and N@K metrics. From the experimental results, we have the following findings:

(1) **Incorporating modal data yields obvious performance gains for interaction-based recommender systems.** For instance, VBPR improves the R@20 and N@20 metrics on the *Baby* dataset by 15.3% and 14.1%, respectively, by linearly integrating visual modal features

on top of BPR. LATTICE, MGCN, and LGMRec, which employ LightGCN as their graph convolution backbone, further improve recommendation performance by effectively modeling multimodal features through item-item graph learning, multi-view mining, and hyper-graph learning, respectively. In addition, AD-DRL improving performance by capturing fine-grained multimodal features by the attribute-aware disentangled representation learning. More recently, DSL-MRS and DiffCL use diffusion models to respectively supplement modality semantics and generate denoised contrastive views, thereby further improving the precision of representation learning. However, MMGCN exhibits inferior performance to LightGCN, due to its simplistic multimodal feature fusion and failure to model user's modality preference heterogeneity.

(2) **Representation learning can be further improved by employing GCN to aggregate and propagate features on graphs.** Specifically, SGL captures the high-order user-item relationships by preforming GCN on the user-item interaction graph, thus outperforming BPR. LightGCN simplifies the vanilla GCN, making it a lightweight yet effective approach for recommendations. In the models of multimodal graph learning, GRGN shows superior performance on the *Baby* datasets, attributable to its effective graph purification mechanism that is more effective in denoising on the dense datasets. SLMRec utilizes additional self-supervised signals to make it more adaptable to sparse datasets (i.e., *MicroLens* dataset). Moreover, LATTICE, MGCN and DSL-MRS achieve obviously performance improvements through the item-item graph learning. LGMRec exhibits sub-optimal performance on the *Baby*

Table 5
Efficiency comparison of DRNet against baselines.

Datasets	Metrics	MMGCN	GRCN	LATTICE	SLMRec	BM3	MGCN	LGMRec	AD-DRL	DSL-MRS	DiffCL	DRNet
<i>Baby</i>	Time (s/epoch)	1.85	1.28	0.53	0.97	0.59	0.98	1.25	0.87	1.28	1.62	1.21
	Memory (GB)	1.92	1.98	3.31	1.16	1.21	1.35	1.02	1.25	1.32	1.87	1.48
<i>Sports</i>	Time (s/epoch)	5.68	3.57	2.77	2.27	1.72	2.67	3.04	2.05	3.12	5.27	3.17
	Memory (GB)	3.13	3.54	17.53	1.96	2.55	2.62	2.39	2.52	2.79	4.24	3.06
<i>Clothing</i>	Time (s/epoch)	5.93	3.10	20.19	2.05	1.66	2.54	2.98	1.96	3.02	5.98	3.28
	Memory (GB)	3.25	3.59	22.65	2.22	2.93	3.28	2.81	3.09	3.22	4.66	3.62
<i>MicroLens</i>	Time (s/epoch)	51.98	16.76	4.51	7.32	2.73	6.21	11.25	5.22	7.10	13.37	7.54
	Memory (GB)	12.06	6.03	14.42	2.08	2.16	2.46	3.37	2.34	3.21	5.16	2.79

dataset, as it performs hyper-graph learning to capture the structural dependencies.

(3) **Our proposed DRNet obtains statistically significant performance gains over all baselines across four datasets.** Specifically, we perform five independent experiments with varied random seeds and statistically analyze the results using the t -test. The results show that the performance gains of DRNet over the optimal baseline is statistically significant (p -value < 0.05). For instance, DRNet achieves improvements of 2.01%, 3.89%, 4.62%, and 3.80% in the term of R@20 on four datasets compared to the optimal baseline. This performance advantage stems from the ability of DRNet to effectively extract both modality-aligned and modality-unique features, as well as to mine the intra-modal and inter-modal interactions between items. In addition, we find that DRNet achieves greater improvement on the ranking-sensitive N@K metric compared to the R@K, suggesting its ability to generate recommendation lists with more accurate item rankings aligned with user preferences.

(4) **DRNet exhibits strong robustness and can adapt to data scenarios of different scales and densities.** As shown in Table 4, the optimal baseline on four datasets are LGMRec (on the *Baby* dataset), DiffCL (on the *Sports* and *Clothing* datasets), and SLMRec (on the *MicroLens* dataset). We attribute this difference in performance distribution to the fact that LGMRec captures global dependencies through hyper-graph learning but its performance degrades with increasing data scale due to noise sensitivity, while DiffCL, equipped with a diffusion-based contrastive learning denoising framework, demonstrates superior performance on larger and more complex datasets. In addition, SLMRec compensates the lack of collaborative signals in the sparse dataset through SSL, thereby outperforming other models on the sparse *MicroLens* dataset. In contrast, DRNet achieves state-of-the-art performance on data scenarios of different scales and densities, due to its adequate feature mining and learning as well as adaptive feature fusion mechanism.

6.1.2. Efficiency (RQ2)

To further analyze the operational efficiency of our model, we record the average training time per epoch and GPU memory costs for both DRNet and the baselines. Table 5 presents the results for all models running on a 24 GB NVIDIA RTX4090 GPU.

From Table 5, we can observe that DRNet has similar time and space costs compared to the current baselines. On the *Baby* dataset, DRNet achieves an approximately 25% reduction in training time compared to the state-of-the-art model DiffCL, while requiring less memory. This efficiency is mainly attributed to the fact that DRNet circumvents the time-consuming diffusion model involving multi-step iterative inference, replacing it with lightweight multilayer perceptrons and efficient attention mechanisms. Furthermore, DRNet maintains stable resource consumption even on the large-scale dataset *MicroLens*. This is due to our employed top- n sparsification strategy [53], which effectively constrains the density of the uni-modal item-item graphs. By limiting the number of neighbors per node, this strategy prevents the computational complexity from escalating quadratically or exponentially as the number of item nodes increases.

Balancing effectiveness and efficiency, DRNet achieves the best results across all datasets in terms of Recall@20 and NDCG@20, while incurring small computational costs. These results demonstrate that DRNet has great practical significance and strong systemic robustness for engineering applications.

6.2. Ablation study (RQ3)

To evaluate the contribution of each component in DRNet, we construct the following model variants: (1) $w/o D^{ma}$, which disregards the modality-aligned features and relies solely on the modality-unique features for recommendations. (2) $w/o D^{mu}$, which disregards the modality-unique features and relies solely on the modality-aligned features for recommendations. (3) w/o CGL, which removes the cross-modal item-item graph learning module, consequently failing to mine the intra-modal and inter-modal interactions between items. (4) w/o CGC, (5) w/o CGC^{ma}, (6) w/o CGC^{mu}, which abandon constructing cross-modal item-item graphs in dual spaces, modality-aligned space, and modality-unique space, respectively, and only capture the intra-modal interactions in the uni-modal item-item graphs. (7) w/o IGL, which eliminates the user-item interaction graph learning module, thus ignoring to mine the high-order collaborative signals. (8) w/o MAFF, which replaces the multi-head attentional feature fusion mechanism by concatenation to fuse the features from different spaces.

Table 6 shows the experimental results, it can be observed that all components in DRNet contribute obviously to the model performance improvement, collectively driving DRNet to achieve state-of-the-art performance. Additionally, we observe that DRNet $w/o D^{mu}$ exhibits a more obvious performance drop than DRNet $w/o D^{ma}$ across four datasets, indicating that the modality-unique features learned by DRNet are more important than the modality-aligned features. This also validates our proposal that extracting modality-aligned features is necessary. Furthermore, we notice that the contributions of different components vary across datasets. For example, w/o CGL has the largest performance decrease on the *Clothing* dataset, with R@20 and N@20 metrics decreasing by 27.9% and 29.7%, respectively, and w/o IGL exhibits the most obvious performance decrease on the *Baby* dataset, with R@20 and N@20 metrics decreasing by 24.3% and 23.6%, respectively. We attribute this to the fact that the *Clothing* dataset has a larger set of items, which contains richer semantic association information between items, and the *Baby* dataset has a higher interaction density, which contains richer signals of user-item interactions. Moreover, the experimental results on the modal variants of the CGC reveals that the CGC strategies designed for the modality-aligned and modality-unique space both can effectively improve the model performance, with CGC^{mu} contributing more to performance improvement. To further analyze the roles and advantages of our designed two CGC strategies, we design a cross-contrast factorial experiment in Section 6.4. In summary, the components in DRNet work collectively to build a comprehensive multi-modal recommender system, and removing any component will destroy the integrity of DRNet, which leads to performance degradation.

Table 6

Performance of DRNet variants with different components removed on four datasets. “w/o” denotes the variant without the corresponding component.

Model variants	Baby		Sports		Clothing		MicroLens	
	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20
DRNet	0.1015	0.0444	0.1149	0.0523	0.0996	0.0455	0.1175	0.0515
w/o D^{ma}	0.0988	0.0437	0.1110	0.0502	0.0932	0.0424	0.1149	0.0500
w/o D^{mu}	0.0945	0.0419	0.1101	0.0492	0.0922	0.0422	0.1131	0.0490
w/o CGL	0.0955	0.0418	0.1080	0.0485	0.0718	0.0320	0.1139	0.0496
w/o CGC	0.0967	0.0425	0.1121	0.0506	0.0926	0.0412	0.1146	0.0501
w/o CGC ^{ma}	0.0991	0.0435	0.1129	0.0511	0.0952	0.0432	0.1159	0.0508
w/o CGC ^{mu}	0.0986	0.0429	0.1124	0.0509	0.0933	0.0428	0.1149	0.0502
w/o IGL	0.0768	0.0339	0.0859	0.0379	0.0854	0.0384	0.1078	0.0458
w/o MAFF	0.0986	0.0432	0.1126	0.0518	0.0937	0.0432	0.1153	0.0504

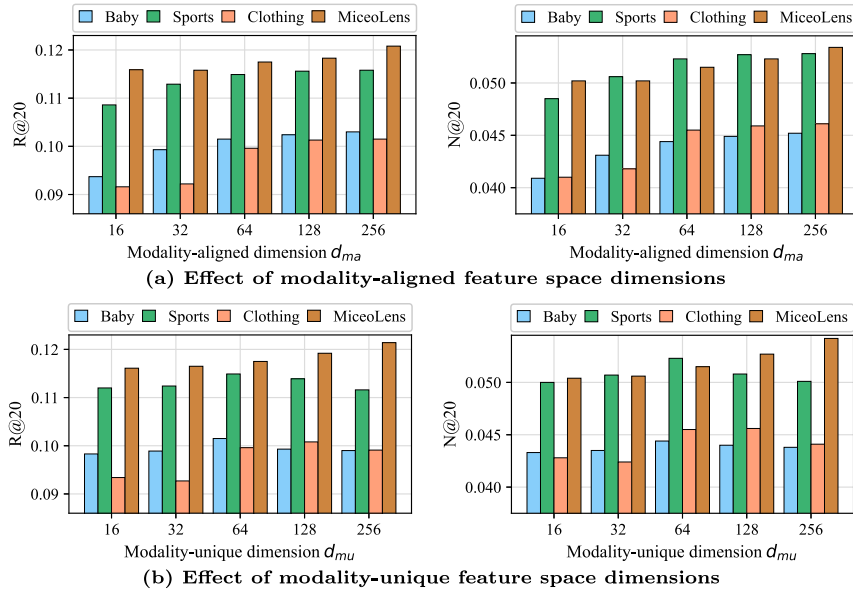


Fig. 3. Performance of DRNet with the values of d_{ma} and d_{mu} in {16, 32, 64, 128, 256} on four datasets.

6.3. Effect of feature space dimensions (RQ4)

For preventing inflated performance resulting from excessively high feature dimensions, we follow the previous methods [20,47,49] to set the dimensions of both the modality-aligned space D^{ma} and modality-unique space D^{mu} to 64, i.e., $d_{ma} = d_{mu} = 64$. In this section, we adjust the dimension of one space in {16, 32, 64, 128, 256} while keeping the other space dimension at 64 to study the impact of feature space dimensions on model performance. The experimental results are shown in Fig. 3.

As shown in Fig. 3(a), for the modality-aligned space D^{ma} , increasing the value of d_{ma} obviously improves the model performance across four datasets. This improvement stems from the richness of aligned features, where higher dimensions better preserve semantic information, thereby improving the representation of aligned features. However, as shown in Fig. 3(b) for modality-unique space D^{mu} , excessively high d_{mu} values lead to performance degradation on the *Baby*, *Sports*, and *Clothing* datasets. This occurs because modality-unique features primarily contain minor details, and lower-dimensional representations are sufficient to capture this information. In contrast, higher dimensions may introduce redundant features that interfere with model learning and potentially lead to over-fitting. Notably, the performance of DRNet continuously improves on the *MicroLens* dataset as the dimension of modality-unique features increases. This phenomenon may be attributed to the fact that *MicroLens*, as a micro-video dataset, contains rich visual dynamic information and complex textual descriptions,

requiring a higher dimensions in D^{mu} for adequate representations. Furthermore, we observe that on the *Clothing* dataset, increasing the value of d_{mu} results in a more obvious performance improvement, whereas the improvement is relatively modest on the *Baby* dataset. This discrepancy may be related to the varying complexity and diversity of modality-unique features across different datasets. The above experimental results show that selecting appropriate values for d_{ma} and d_{mu} based on the complexity and modality richness of the dataset is crucial in practical applications.

6.4. Cross-contrast factorial experiment (RQ5)

Considering the differences in feature distributions between D^{ma} and D^{mu} , we design dedicated Cross-modal item-item Graph Construction (CGC) strategies for each space. Specifically, we perform the edge union operation in D^{ma} , while treating the item-item graphs of each modality as independent views in D^{mu} and establishing connections between them. To analyze the effectiveness of CGC^{ma} and CGC^{mu} and their adaptability to respective spaces, we conduct a cross-contrast experiment to test the performance of DRNet under all combinations of the dual spaces and two strategies. The experimental results are shown in Table 7.

From Table 7, we observe that DRNet reaches the optimal performance across all four datasets when CGC^{ma} is used in D^{ma} and CGC^{mu} is used in D^{mu} , which reflects the suitability of our dedicated CGC strategies for each space. Additionally, all the remaining combinations

Table 7

Performance of DRNet with the two different CGCs on four datasets. The optimal and the suboptimal results are highlighted in **bold** and underline, respectively.

D^{ma}	D^{mu}	Baby		Sports		Clothing		MicroLens	
		R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20
CGC ^{ma}	CGC ^{mu}	0.1015	0.0444	0.1149	0.0523	0.0996	0.0455	0.1175	0.0515
CGC ^{ma}	CGC ^{ma}	0.0971	0.0427	0.1123	0.0507	0.0929	0.0419	0.1148	0.0502
CGC ^{mu}	CGC ^{mu}	<u>0.1001</u>	<u>0.0439</u>	<u>0.1136</u>	<u>0.0519</u>	<u>0.0975</u>	<u>0.0448</u>	<u>0.1164</u>	<u>0.0509</u>
CGC ^{mu}	CGC ^{ma}	0.0962	0.0421	0.1116	0.0503	0.0918	0.0416	0.1142	0.0498

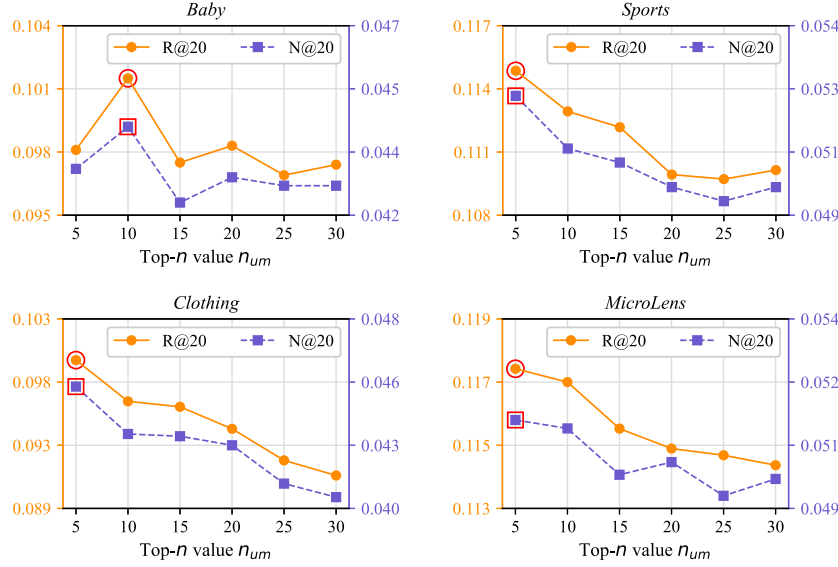


Fig. 4. Performance of DRNet with the values of n_{um} in $\{5, 10, 15, 20, 25, 30\}$ on four datasets. The optimal point is highlighted in red.

result in performance degradation, with the orthogonal combination (i.e., using CGC^{mu} in D^{ma} and CGC^{ma} in D^{mu}) exhibiting the most severe decline. These performance degradation is due to the fact that applying CGC^{mu} in D^{ma} fails to fully utilize the inter-modal interactions, while using CGC^{ma} in D^{mu} introduces noisy interactions and destroys the orthogonal relationships between modalities. Furthermore, we find that when both spaces adopt CGC^{mu}, DRNet exhibits superior overall performance compared to using the CGC^{ma} uniformly, while achieving suboptimal results across all four datasets. We attribute this phenomenon to that applying CGC^{mu} in D^{ma} only leads to insufficient mining of inter-modal interaction information, unlike employing CGC^{ma} in D^{mu} introduces destructive noisy interactions. In summary, these experimental results not only validate the advantages of our dedicated CGCs designed for each space, but also reveal the robustness of the CGC^{mu} strategy.

6.5. Sensitivity analysis (RQ6)

In this section, we test the performance of DRNet under different hyper-parameters that optimized through grid search, including the top-n value n_{um} in Section 6.5.1, the textual importance score α in Section 6.5.2, the top-n value n_{cm} in Section 6.5.3, and the loss weights λ_1 and λ_2 in Section 6.5.4. In each experiment, only the corresponding hyper-parameter is adjusted, while other hyper-parameters remain at their optimal values.

6.5.1. Impact of the top-n value n_{um}

The top-n value n_{um} in Eq. (7) controls the number of edges in constructing uni-modal item-item graphs, which influences the degree of the intra-modal semantic relationships. To investigate its impact on model performance, we adjust the value of n_{um} in the range $\{5, 10, 15, 20, 25, 30\}$.

As shown in Fig. 4, the recommendation performance of DRNet exhibits different trends across four datasets as n_{um} changes, and the optimal values of n_{um} varies. Specifically, on the *Baby* dataset, as n_{um} increases, the performance of DRNet first rises sharply and reaches its peak at $n_{um} = 10$, while further increase of n_{um} leads to a fluctuating decrease on model performance. In addition, on the *Sports*, *Clothing*, and *MicroLens* datasets, DRNet achieves optimal performance at $n_{um} = 5$, while the performance gradually decreases with increasing values of n_{um} . Overall, DRNet tends to achieve optimal performance at smaller n_{um} values. This may be due to the fact that a larger n_{um} values introduce edges between items with low relevance, thereby introducing irrelevant noise signals and reducing the accuracy of multimodal feature representations of items. This finding is consistent with that in the previous research [20], which highlights that choosing different number of neighbors has a obvious impact on the quality of the item-item graph.

6.5.2. Impact of the textual importance score α

The textual importance score α is utilized in CGC^{ma} to control the weight of textual features when fused with visual features in D^{ma} , thus affecting the multimodal representations of nodes in the cross-modal item-item graph C_T^{ma} . For exploring the relative contributions of visual and textual modalities in D^{ma} , we tune the value of α in a step of 0.1 between 0.1 and 0.9.

Fig. 5 illustrates the performance of DRNet across four datasets as the textual importance score α varies. We observe that on the *Baby*, *Sports*, and *Clothing* datasets, the recommendation performance initially improves as α increases, peaking at $\alpha = 0.8$, and then declines. This reflects that textual modality performs a more important role than visual modality in D^{ma} across these three datasets. In addition, the two evaluation metrics, R@20 and N@20, show the most obvious increase on the *Sports* dataset, indicating that there is a large gap between the visual and textual features, and that the correct choice of textual

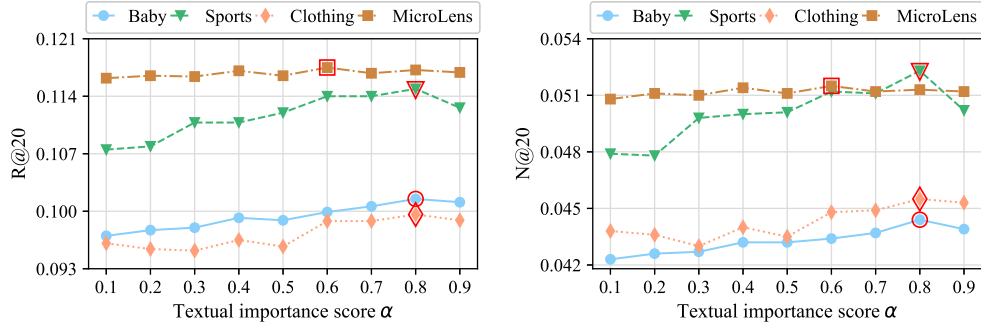


Fig. 5. Performance of DRNet with values of α in the range of 0.1 to 0.9 on four datasets. The optimal point is highlighted in red.

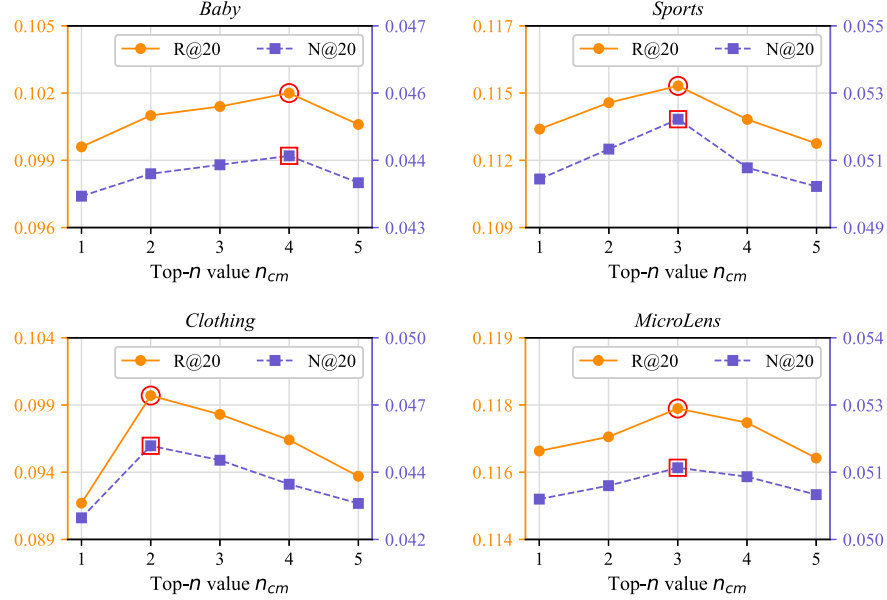


Fig. 6. Performance of DRNet with the values of n_{cm} in $\{1, 2, 3, 4, 5\}$ on four datasets. The optimal point is highlighted in red.

modality weight has a great impact on performance improvement. In contrast, DRNet exhibits strong robustness to variations in α on the *MicroLens* dataset. Specifically, the R@20 metric fluctuates slightly between 0.116 and 0.118, and the N@20 metric fluctuates slightly between 0.050 and 0.052, both of which reach their maximum values at $\alpha = 0.6$. This stability suggests that in the *MicroLens* dataset, there is a high degree of consistency between visual and textual modal features in D^{ma} , consequently adjusting the weight of textual modality has a moderate impact on the model performance.

6.5.3. Impact of the top-n value n_{cm}

As defined in CGC^{mu} , the parameter n_{cm} controls the connection density between uni-modal item-item graphs in D^{mu} . In other words, n_{cm} determines the number of cross-modal edges in the modality-unique item-item graph $G_{I'}^{mu}$. To study the impact of n_{cm} on the model performance, we perform experiments with its value varying in $\{1, 2, 3, 4, 5\}$.

The experimental results are shown in Fig. 6, and we can observe that on the all four datasets, the performance of DRNet shows an increasing and then decreasing trend with the increase of n_{cm} . Taking the *Baby* dataset as an example, when n_{cm} increases from 1 to 4, the R@20 metric rises from 0.0991 to 0.1015 and the N@20 metric rises from 0.0435 to 0.0444. This phenomenon validates the effectiveness of capturing inter-modal semantic relationships by constructing edges between item-item graphs of different modalities in D^{mu} . However, when n_{cm} exceeds a certain threshold (e.g., $n_{cm} = 4$ on the *Baby* dataset), the model performance begins to decline. This suggests that excessive

cross-modal connections may introduce conflicting features between modalities, thereby interfering with multimodal representation learning. This trend is consistent across datasets, but the peaks of n_{cm} vary. Specifically, the optimal value is 2 for the *Sports* and *MicroLens* datasets, while it is 3 for the *Clothing* dataset. This reflects the differences in inter-modal interactions among items in different datasets. Overall, a smaller n_{cm} makes it difficult for DRNet to fully explore inter-modal interactions, while a larger n_{cm} may lead to performance degradation due to the conflicts between modalities. Therefore, it is essential to select a suitable n_{cm} value to balance inter-modal interactions and modal conflicts for different datasets.

6.5.4. Impact of the loss weights λ_1 and λ_2

In Eq. (24), λ_1 and λ_2 control the weights of the alignment loss and uniqueness loss, respectively, thereby affecting the direction of parameter optimization. To find the most appropriate values of λ_1 and λ_2 for each datasets, we perform a grid search in the range $\{1, 1e-01, 1e-02, 1e-03, 1e-04\}$.

The experimental results are shown in Fig. 7. Based on these heatmaps, we can clearly observe the variation trends in the model performance across different datasets when tuning the values of λ_1 and λ_2 . Specifically, on the *Baby* dataset, the R@20 and N@20 metrics peak at different combinations of parameters, with R@20 reaches its maximum value of 0.1015 when $\lambda_1 = 1e-01$ and $\lambda_2 = 1e-03$, while N@20 achieves its maximum value of 0.0445 when $\lambda_1 = 1e-02$ or $1e-03$ and $\lambda_2 = 1e-03$. However, across the remaining three datasets, the peaks

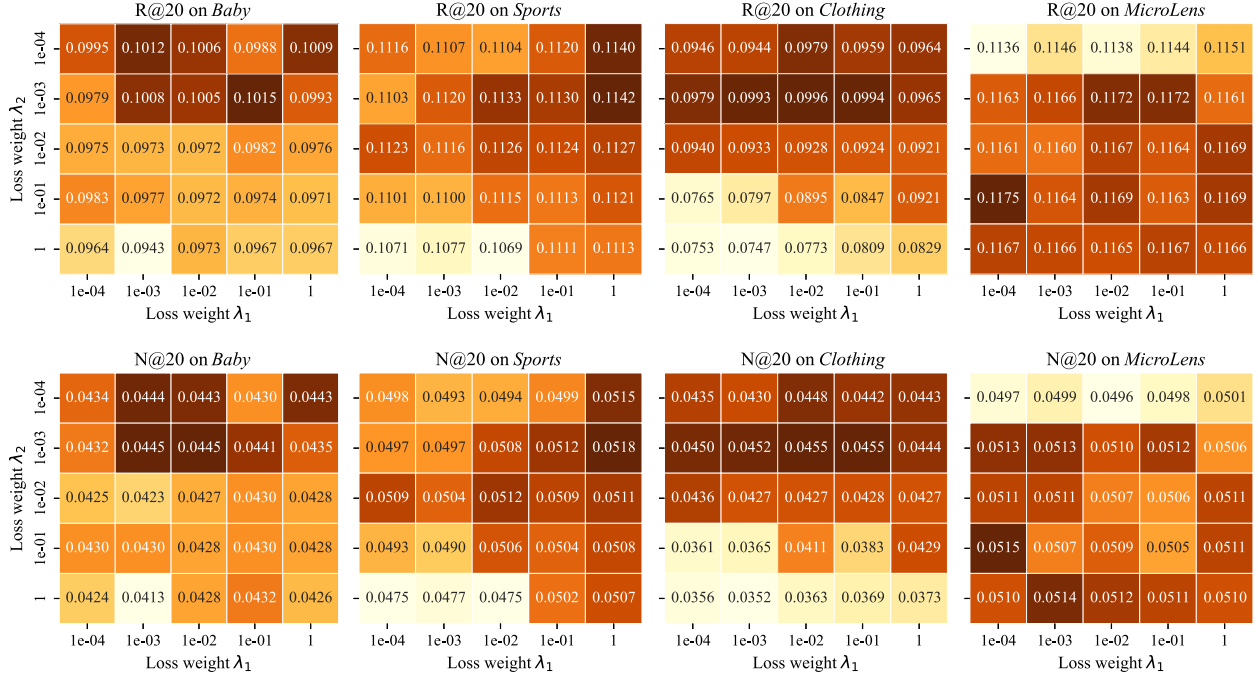


Fig. 7. Performance of DRNet with the values of λ_1 and λ_2 in {1, 1e-01, 1e-02, 1e-03, 1e-04} on four datasets. Darker colors represent the larger values, while lighter hues correspond to the smaller values.

of R@20 and N@20 occur under the same parameter combinations of their respective, i.e., on the *Sports* dataset when $\lambda_1 = 1$ and $\lambda_2 = 1e-03$, on the *Clothing* dataset when $\lambda_1 = 1e-02$ and $\lambda_2 = 1e-03$, and on the *MicroLens* dataset when $\lambda_1 = 1e-04$ and $\lambda_2 = 1e-01$. Notably, on the *Baby*, *Sports*, and *Clothing* datasets, the optimal and suboptimal results are clustered around the larger λ_1 and smaller λ_2 values, indicating that it is more important to learn accurate modality-aligned features than to learn modality-unique features on these datasets. However, on the *MicroLens* dataset, the smaller λ_1 and larger λ_2 values lead to the optimal performance, reflecting that learning accurate modality-unique features is more valuable on this dataset. Furthermore, through the horizontal and vertical comparative analysis on each heatmaps, we find that the variation of λ_2 has a greater impact on model performance than λ_1 , which suggests that selecting an appropriate weight for the uniqueness loss is more important for improving the performance of DRNet in practical applications.

6.6. Visualization and statistical analysis (RQ7)

To evaluate the effectiveness and distinctness of DRNet in extracting the modality-aligned and modality-unique features, we conduct both qualitative and quantitative analyses of the learned features across four datasets. Specifically, on one hand, we randomly select 500 items and use t-SNE [66] to map the visual and textual features from the modality-aligned and modality-unique spaces into a 2D plane for visualization. On the other hand, we introduce euclidean distance, silhouette coefficient [67], and Calinski-Harabasz index [68] as statistical metrics to quantitatively evaluate the effect of representation learning.

6.6.1. Visualization analysis

Although the nonlinear t-SNE algorithm inherently distorts the original directional relationships between features during dimensionality reduction, it effectively preserves their relative distances, thereby accurately reflecting the overall distribution patterns of the feature space. The visualization results are shown in Fig. 8. First, we observe the distribution of visual and textual features within each space. Specifically, in the modality-aligned space, the visual and textual features

exhibit highly similar distributions with obvious overlap. This demonstrates that the designed alignment loss enables DRNet to effectively capture the modality-consistent semantic information. Conversely, in the modality-unique space, the visual and textual features are separated and distributed in different regions. This phenomenon verifies that the designed uniqueness loss ensures the independence of features across modalities and allows DRNet to effectively extract modality-unique details. Second, we investigate the inter-space distribution of each modality. It is observed that both visual and textual features are separated between the two spaces with minimal overlap. This intuitively indicates that the modality-aligned and modality-unique features learned by DRNet encode distinct semantic information, facilitated by the uniqueness loss. Notably, the above feature separation phenomenon is particularly obvious on the *MicroLens* dataset. We attribute this to the rich semantic features inherent in entertainment-oriented micro-video contexts, which enable DRNet to learn specialized attributes across different representation spaces.

In addition, we find that the modality-unique visual features exhibit higher clustering density compared to their textual counterparts. A plausible explanation is that the visual features extracted by the pre-trained model VGG16 [54] are 4096-dimensional, containing high-level semantics, and after being reduced in dimension using MLP, they converge to a relatively compact space. In contrast, the textual features extracted by the pre-trained model Sentence2Vec [55] are 384-dimensional, containing more specific and discrete keyword and attribute descriptions, and after being reduced in dimension using MLP, they still have higher diversity and variability. Such disparity is absent in the modality-aligned space due to the semantic synchronization imposed by the alignment constraint.

6.6.2. Statistical analysis

Table 8 presents the results of three statistical metrics. The higher values across these metrics indicate greater inter-cluster distances, suggesting a higher degree of separation between the two categories of features. We calculate the intra-space and inter-space statistical results to assess the accuracy of representation learning within each space and the semantic distinction between the two spaces, respectively.

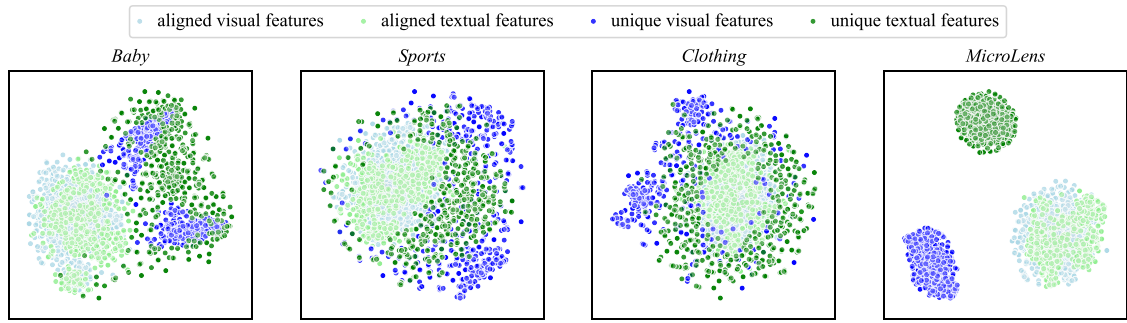


Fig. 8. The distribution of modality-aligned and modality-unique features on four datasets. The blue and green points denote visual and textual features, respectively. Specifically, the muted tones indicate modality-aligned features ($E_I^{m,ma}$ in Eq. (3)), whereas the saturated colors represent modality-unique features ($E_I^{m,mu}$ in Eq. (5)).

Table 8

Quantitative results of feature distributions. The upper table presents the distribution statistics between visual and textual features within the modality-aligned and modality-unique spaces, respectively. The lower table shows the statistical results comparing modality-aligned and modality-unique distributions for both visual and textual features independently.

Intra-space statistical results						
Datasets	Euclidean Distance		Silhouette Coefficient		Calinski–Harabasz Index	
	D^{ma}	D^{mu}	D^{ma}	D^{mu}	D^{ma}	D^{mu}
<i>Baby</i>	3.685	65.013	0.007	0.294	161.398	1357.445
<i>Sports</i>	2.228	26.337	0.005	0.152	102.858	843.338
<i>Clothing</i>	7.032	38.472	0.002	0.127	17.115	574.854
<i>MicroLens</i>	0.171	0.926	0.002	0.870	0.084	913 962.462
Inter-space statistical results						
Datasets	Euclidean Distance		Silhouette Coefficient		Calinski–Harabasz Index	
	visual	textual	visual	textual	visual	textual
<i>Baby</i>	46.213	64.736	0.184	0.317	655.961	1556.915
<i>Sports</i>	25.018	18.042	0.297	0.218	698.273	551.634
<i>Clothing</i>	32.350	25.686	0.109	0.124	481.636	607.964
<i>MicroLens</i>	1.048	0.886	0.848	0.851	694 743.703	685 496.142

From the intra-space evaluation, it is evident that the values for euclidean distance, silhouette coefficient, and Calinski–Harabasz index in the modality-aligned space are lower than those in the modality-unique space. This indicates that the visual and textual features of the same item are indeed closer in the modality-aligned space, whereas they are further apart in the modality-specific space and there is a greater separation of categories. These findings verify the accuracy of representation learning guided by our designed loss functions in these two spaces. Regarding the inter-space evaluation, the statistical results for both visual and textual features in the two spaces are relatively high, reaching a magnitude comparable to the intra-space statistics in the modality-unique space. This indicates that the uniqueness loss defined in Eq. (6) effectively ensures that the modality-aligned and modality-unique features learned by DRNet obviously semantic distinction, successfully achieving the disentanglement of these two feature categories.

6.7. Case study (RQ8)

To intuitively demonstrate the actual performance of DRNet in capturing comprehensive user preferences, this section provides a comparative analysis of typical recommendations from the *Baby* and *Clothing* datasets. To ensure the generalizability of our findings, we focus on the users with only a single interaction in the test set. This subset covers 17,859 users in the *Baby* dataset (approximately 91.8% of the total 19,445) and 37,937 users in the *Clothing* dataset (approximately 96.3% of the total 39,387). We compare the recommended items of DRNet against the state-of-the-art baseline, DiffCL. Table 9 presents typical success cases alongside a failure case that reveals the limitations of DRNet.

Advantage analysis. In the *Baby* dataset case, although the items interacted with by user “A3NG17DQWSV6DY” vary in category, they exhibit a distinct tropical rainforest color scheme visually and emphasize health protection textually. Due to its reliance on modality alignment mechanism, DiffCL overly emphasize the coarse-grained shared semantics of “baby furniture”. Consequently, fine-grained modality-specific features, such as color palettes and technical functionalities, are suppressed as noise during alignment process, leading to the recommendation of mass-market best-selling products. In contrast, DRNet successfully retain these highly distinctive preference signals through its specially designed modality-unique space, making the recommendation results highly consistent with the user’s historical preferences in terms of aesthetic intent and functional requirements. A similar phenomenon is observed in the *Clothing* dataset for user “A33X5I19RL9CCZ”, where DiffCL capture only the modality-aligned semantics of “women’s sleepshirt”, but loses the non-skirted visual attributes and the textual details regarding floral patterns. DRNet not only can capture the modality-aligned features, but also can uncover the modality-unique attributes that are overlooked by the baseline. This demonstrates the superiority of its dual-space disentanglement in handling complementary multimodal information.

Error analysis. Although DRNet performs well in most scenarios, a typical failure case in the *Clothing* dataset reveals challenges posed the quality of raw modal data. This case shows that due to the complex visual content (e.g., backgrounds and multiple entities) or exaggerated textual descriptions can cause the modality-unique space to incorrectly retain noise features irrelevant to actual user preferences. Even more seriously, such feature biases are propagated and accumulated through cross-modal item–item graph learning, ultimately resulting in semantically drifted items appearing in the recommendation list.

Table 9

Successful and failed cases of DRNet. In cases (a) and (b), the items interacted with by users exhibit prominent modality-unique features, with DRNet recommends the right answer. In case (c), the multimodal data is contaminated with noise, the alignment-only baseline DiffCL succeeds, whereas DRNet fails.

Case	(a) From the <i>Baby</i> dataset	(b) From the <i>Clothing</i> dataset	(c) From the <i>Clothing</i> dataset
User	ID: A3NG17DQWSV6DY	ID: A33X5119RL9CCZ	ID: A98NPCU9ELO0Y
Items in Training	 ID: B0045JKUU2 Title: Fisher-Price Rainforest Healthy Care High Chair	 ID: B00417BEZQ Title: Dreams and Company Women's Plus Size, long tag-free sleepshirt	 ID: B00CL4RL78 Title: Hollywood Star Fashion Women's Drape Pockets Light Weight Flyaway Cardigan Shawl
	 ID: B001NJC5QY Title: Summer Infant 3-Stage Superseat Highchair, Green	 ID: B005TTQOEI Title: Only Necessities Women's Plus Size Print Lounger	 ID: B005AKKYUQ Title: The Saress Beach Cover-Up-Short-Choose from a Variety Of Stylish Colors
	 ID: B0015CA3A4 Title: Tiny Love Tummy Time Fun Activity Mat, Frog	 ID: B0041762P8 Title: Only Necessities Women's Plus Size Long Lounger with Bib & Border Print	 ID: B006ZUA9DC Title: Ever Pretty Double V-neck Floral Printed Satin Short NWT Party Dress
DiffCL	 ID: B000MAER3Y Title: Chicco KeyFit 30 Infant Car Seat, Midori 	 ID: B0007WFGWS Title: National Crinkle Cotton Dress 	 ID: B005TDLRRS Title: Allegra K Lady Horizontal Striped Scoop Neck Racer Back Design Casual Tank Dress 
DRNet	 ID: B000I2WB6G Title: Fisher-Price Rainforest Bouncer 	 ID: B00008695M Title: Shadowline Short Gown with Flower Petal Embroidery and Flutter Sleeves 	 ID: B009H6NPBE Title: Women's Emma Chiffon Floral Multicolor Fashion Scarf 

7. Conclusions and future work

This work focuses on the recommendation in multimodal scenarios and presents a dual-space feature representation learning network, designated as DRNet. In contrast to existing methods, DRNet incorporates more comprehensive feature extraction and semantic mining modules, which obviously enhances the representation learning capability in multimodal recommendation. Specifically, DRNet maps the pre-extracted modal features of items into dual feature representation spaces to decouple them into the modality-aligned and modality-unique features. Moreover, we design loss functions for each space to help DRNet capture the cross-modal consistency information and modality-unique details that influence user preferences. To further explore multimodal semantic relationships, we extend existing uni-modal item-item graph learning by using edge union operations and establishing cross-modal connections in the modality-aligned and modality-unique spaces, respectively. These strategies enable the intra-modal and inter-modal interactions between items to be simultaneously mined during graph convolution operations while effectively promoting the flow of semantic information across modalities. In addition, we design a Multi-head Attentional Feature Fusion (MAFF) mechanism, leveraging the advantages of attention mechanisms to dynamically integrate item features or user representations from the dual spaces. Finally, through extensive experiments on four publicly available datasets, we validate the significant advantages of DRNet over baselines and demonstrate the effectiveness of each component in DRNet.

Despite its effectiveness, DRNet has several limitations. First, this work relies on the quality of features extracted by the pre-trained models, which restricts its adaptability. When raw modal data contains high noise or suffers from low quality, the dual-space disentanglement module may fail to capture effective semantic information. Furthermore, the cross-modal item-item graph learning may amplify these noisy signals, potentially degrading recommendation performance. Second, the current graph construction is limited to the item

side due to the absence of explicit user modality profiles in existing datasets, thereby precluding the exploration of latent semantic relationships among users. In the future, we aim to employ Multimodal Large Language Models (MLLMs) [69,70] to directly extract more accurate modality-aligned and modality-unique features. Moreover, we plan to leverage MLLMs to generate user profiles, which will facilitate the construction of user-side modal graphs and the uncovering of multimodal user-user dependencies.

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. **Xin Zhang:** Writing – review & editing, Validation, Investigation. **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] P. Covington, J. Adams, E. Sargin, Deep neural networks for YouTube recommendations, in: Proceedings of the 10th ACM Conference on Recommender Systems, RecSys'16, ACM, 2016, pp. 191–198.
- [2] N. Khan, Z. Ma, A. Ullah, K. Polat, Similarity attributed knowledge graph embedding enhancement for item recommendation, *Inform. Sci.* 613 (2022) 69–95.
- [3] S. Wan, G. Wang, Z. Ming, Y. Yan, A.B. Nellippallil, J.K. Allen, F. Mistree, Knowledge graph-based representation and recommendation for surrogate modeling method, *Adv. Eng. Inform.* 62 (2024) 102706.
- [4] 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: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'21, ACM, 2021, pp. 1813–1817.
- [5] T. Hasan, R.C. Bunescu, Topic-level Bayesian surprise and serendipity for recommender systems, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys'23, ACM, 2023, pp. 933–939.
- [6] R. He, J.J. McAuley, VBPR: visual Bayesian personalized ranking from implicit feedback, in: AAAI. Proceedings of the 30th AAAI Conference on Artificial Intelligence, AAAI'16, 2016, pp. 144–150.
- [7] J. Magalhães, T. Chua, T. Mei, A.F. Smeaton, The next generation multimodal conversational search and recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, MM'21, ACM, 2021, pp. 953–954.
- [8] T. Zang, Y. Zhu, R. Zhang, C. Wang, K. Wang, J. Yu, Contrastive multi-view interest learning for cross-domain sequential recommendation, *ACM Trans. Inf. Syst.* 42 (3) (2024) 76:1–76:30.
- [9] Z. Li, F. Liu, Y. Wei, Z. Cheng, L. Nie, M.S. Kankanhalli, Attribute-driven disentangled representation learning for multimodal recommendation, in: Proceedings of the 32nd ACM International Conference on Multimedia, MM'24, ACM, 2024, pp. 9660–9669.
- [10] X. Pan, Y. Chen, C. Tian, Z. Lin, J. Wang, H. Hu, W.X. Zhao, Multimodal meta-learning for cold-start sequential recommendation, in: Proceedings of the 31st ACM International Conference on Information and Knowledge Management, CIKM'22, ACM, 2022, pp. 3421–3430.
- [11] Y. Xu, L. Zhu, Z. Cheng, J. Li, Z. Zhang, H. Zhang, Multi-modal discrete collaborative filtering for efficient cold-start recommendation, *IEEE Trans. Knowl. Data Eng.* 35 (1) (2023) 741–755.
- [12] Y. Dang, Z. Pan, X. Zhang, W. Chen, F. Cai, H. Chen, Discrepancy learning guided hierarchical fusion network for multi-modal recommendation, *Knowl.-Based Syst.* 317 (2025) 113496.
- [13] 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: Proceedings of the 27th ACM International Conference on Multimedia, MM'19, ACM, 2019, pp. 1437–1445.
- [14] P. Müller, M. Dietz, D. Schiller, D. Thomas, G. Zhang, P. Gebhard, E. André, A. Bulling, MultiMediate: Multi-modal group behaviour analysis for artificial mediation, in: Proceedings of the 29th ACM International Conference on Multimedia, MM'21, ACM, 2021, pp. 4878–4882.
- [15] K. Liu, F. Xue, D. Guo, L. Wu, S. Li, R. Hong, MEGCF: multimodal entity graph collaborative filtering for personalized recommendation, *ACM Trans. Inf. Syst.* 41 (2) (2023) 30:1–30:27.
- [16] Q. Xu, F. Shen, L. Liu, H.T. Shen, GraphCAR: Content-aware multimedia recommendation with graph autoencoder, in: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'18, ACM, 2018, pp. 981–984.
- [17] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, M.S. Kankanhalli, User diverse preference modeling by multimodal attentive metric learning, in: Proceedings of the 27th ACM International Conference on Multimedia, MM'19, ACM, 2019, pp. 1526–1534.
- [18] W. Wei, C. Huang, L. Xia, C. Zhang, Multi-modal self-supervised learning for recommendation, in: Proceedings of the 26th International Conference on World Wide Web, WWW'23, ACM, 2023, pp. 790–800.
- [19] R. van den Berg, T.N. Kipf, M. Welling, Graph convolutional matrix completion, 2017, *CoRR* abs/1706.02263.
- [20] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, L. Wang, Mining latent structures for multimedia recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, MM'21, ACM, 2021, pp. 3872–3880.
- [21] X. Zhou, Z. Shen, A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, MM'23, ACM, 2023, pp. 935–943.
- [22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, *Adv. Neural Inf. Process. Syst.* 33 (2020) 18661–18673.
- [23] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, Z. Huang, Self-supervised learning for recommender systems: A survey, *IEEE Trans. Knowl. Data Eng.* 36 (1) (2024) 335–355.
- [24] P. Yu, Z. Tan, G. Lu, B. Bao, Multi-view graph convolutional network for multimedia recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, MM'23, ACM, 2023, pp. 6576–6585.
- [25] J. Li, S. Wang, Q. Zhang, S. Yu, F. Chen, Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations, in: Proceedings of the 28th International Conference on World Wide Web, WWW'25, ACM, 2025, pp. 2787–2798.
- [26] W. Yang, Z. Fang, T. Zhang, S. Wu, C. Lu, Modal-aware bias constrained contrastive learning for multimodal recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, MM'23, ACM, 2023, pp. 6369–6378.
- [27] 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.
- [28] Q. Song, J. Hu, L. Xiao, B. Sun, X. Gao, S. Li, DiffCL: A diffusion-based contrastive learning framework with semantic alignment for multimodal recommendations, *IEEE Trans. Neural Networks Learn. Syst.* 36 (10) (2025) 18587–18597.
- [29] H. Huang, Y. Song, Y. Wu, J. Shi, X. Xie, H. Jin, Multitask representation learning with multiview graph convolutional networks, *IEEE Trans. Neural Networks Learn. Syst.* 33 (3) (2022) 983–995.
- [30] F. Mostofi, O.B. Tokdemir, V. Togan, A decision-support productive resource recommendation system for enhanced construction project management, *Adv. Eng. Inform.* 62 (2024) 102793.
- [31] Z. Pan, F. Cai, W. Chen, C. Chen, H. Chen, Collaborative graph learning for session-based recommendation, *ACM Trans. Inf. Syst.* 40 (4) (2022) 72:1–72:26.
- [32] S. Wu, F. Sun, W. Zhang, X. Xie, B. Cui, Graph neural networks in recommender systems: A survey, *ACM Comput. Surv.* 55 (5) (2023) 97:1–97:37.
- [33] X. Zhang, R. Yan, Y. Fan, J. Zhang, H. Yuan, C. Wei, Cross-domain attention transfer network for recommendation, *Adv. Eng. Inform.* 68 (2025) 103667.
- [34] X. Wang, X. He, M. Wang, F. Feng, T. Chua, Neural graph collaborative filtering, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, ACM, 2019, pp. 165–174.
- [35] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, LightGCN: Simplifying and powering graph convolution network for recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'20, ACM, 2020, pp. 639–648.
- [36] Q. Li, Z. Han, X. Wu, Deeper insights into graph convolutional networks for semi-supervised learning, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI'18, AAAI Press, 2018, pp. 3538–3545.
- [37] J.H. Giraldo, K. Skianis, T. Bouwmans, F.D. Malliaros, On the trade-off between over-smoothing and over-squashing in deep graph neural networks, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM'23, ACM, 2023, pp. 566–576.
- [38] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, X. Xie, Self-supervised graph learning for recommendation, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'21, ACM, 2021, pp. 726–735.
- [39] W. Jin, X. Liu, Y. Ma, C.C. Aggarwal, J. Tang, Feature overcorrelation in deep graph neural networks: A new perspective, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD'22, ACM, 2022, pp. 709–719.
- [40] X. Wang, T. Huang, D. Wang, Y. Yuan, Z. Liu, X. He, T. Chua, Learning intents behind interactions with knowledge graph for recommendation, in: Proceedings of the 24th International Conference on World Wide Web, WWW'21, ACM / IW3C2, 2021, pp. 878–887.
- [41] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, T.Y. Liu, Do transformers really perform badly for graph representation? *Adv. Neural Inf. Process. Syst.* 34 (2021) 28877–28888.
- [42] H. Pei, X. Liu, X. Huang, M. Wu, Z. Wen, F. Zhao, A personalized recommendation method under the cloud platform based on users' long-term preferences and instant interests, *Adv. Eng. Inform.* 54 (2022) 101763.
- [43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [44] Y. Xing, X. Wang, Y. Li, H. Huang, C. Shi, Less is more: on the over-globalizing problem in graph transformers, 2024, *arXiv preprint arXiv:2405.01102*.
- [45] H. Liu, Y. Wei, F. Liu, W. Wang, L. Nie, T. Chua, Dynamic multimodal fusion via meta-learning towards micro-video recommendation, *ACM Trans. Inf. Syst.* 42 (2) (2024) 47:1–47:26.
- [46] Y. Wei, X. Wang, L. Nie, X. He, T. Chua, Graph-refined convolutional network for multimedia recommendation with implicit feedback, in: Proceedings of the 28th ACM International Conference on Multimedia, MM'20, ACM, 2020, pp. 3541–3549.
- [47] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, B. Ruan, LGMRec: Local and global graph learning for multimodal recommendation, in: Proceedings of the 38th AAAI Conference on Artificial Intelligence, AAAI'24, AAAI Press, 2024, pp. 8454–8462.
- [48] 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.

- [49] 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.
- [50] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, F. Jiang, Bootstrap latent representations for multi-modal recommendation, in: *Proceedings of the 26th International Conference on World Wide Web, WWW'23*, ACM, 2023, pp. 845–854.
- [51] H. Kieu, M. Xu, T.T. Huynh, D.D. Le, Enhancing multimodal recommendations with vision-language models and information-aware fusion, 2025, *CoRR* abs/2511.02113.
- [52] Y. Ping, S. Wang, Z. Yang, Y. Dong, R. Jia, MRFFD: multimodal recommender based on feature fusion and decoupling, *Neural Comput. Appl.* (2025) 1–22.
- [53] J. Chen, H. Fang, Y. Saad, Fast approximate kNN graph construction for high dimensional data via recursive lanczos bisection, *J. Mach. Learn. Res.* 10 (2009) 1989–2012.
- [54] J. Ni, J. Li, J.J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'19*, Association for Computational Linguistics, 2019, pp. 188–197.
- [55] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'19*, Association for Computational Linguistics, 2019, pp. 3980–3990.
- [56] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: *Proceedings of the 13th Artificial Intelligence and Statistics, AISTATS'10*, in: *JMLR Proceedings*, vol. 9, JMLR.org, 2010, pp. 297–304.
- [57] D. Hazarika, R. Zimmermann, S. Poria, MISA: modality-invariant and -specific representations for multimodal sentiment analysis, in: *Proceedings of the 28th ACM International Conference on Multimedia, MM'20*, ACM, 2020, pp. 1122–1131.
- [58] Y.H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, *ACL'19*, in: *Findings of the 57th Association for Computational Linguistics*, vol. 2019, 2019, p. 6558.
- [59] R. Lin, H. Hu, Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis, *IEEE Trans. Multimed.* 26 (2023) 2740–2755.
- [60] Y. Kim, T. Kim, W. Shin, S. Kim, MONET: modality-embracing graph convolutional network and target-aware attention for multimedia recommendation, in: *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM'24*, ACM, 2024, pp. 332–340.
- [61] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in: *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI'09*, AUAI Press, 2009, pp. 452–461.
- [62] N. Khan, D.S. Sisodia, Exploiting diffusion-based structured learning for item interactions representations in multimodal recommender systems, *Inf. Process. Manage.* 62 (3) (2025) 104075.
- [63] H. Zhou, X. Zhou, Z. Zeng, L. Zhang, Z. Shen, A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions, 2023, *CoRR* abs/2302.04473.
- [64] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, AISTATS'10*, in: *JMLR Proceedings*, vol. 9, JMLR.org, 2010, pp. 249–256.
- [65] D.P. Kingma, Adam: A method for stochastic optimization, 2014, *arXiv preprint arXiv:1412.6980*.
- [66] L. van der Maaten, Accelerating t-SNE using tree-based algorithms, *J. Mach. Learn. Res.* 15 (1) (2014) 3221–3245.
- [67] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.* 20 (1987) 53–65.
- [68] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, *Comm. Statist. Theory Methods* 3 (1) (1974) 1–27.
- [69] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, R. Cucchiara, The revolution of multimodal large language models: A survey, in: *Findings of the 62nd Association for Computational Linguistics, ACL'24*, Association for Computational Linguistics, 2024, pp. 13590–13618.
- [70] J. Jiang, C. Ma, X. Song, H. Zhang, J. Luo, Corvid: Improving multimodal large language models towards chain-of-thought reasoning, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 3034–3046.