ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230631

Enhancing Dyadic Relations with Homogeneous Graphs for Multimodal Recommendation

Hongyu Zhou[a], Xin Zhou[b], Lingzi Zhang[a] and Zhiqi Shen[a;*]

[a]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[b]Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, Singapore

Abstract. User-item interaction data in recommender systems is a form of dyadic relation, reflecting user preferences for specific items. To generate accurate recommendations, it is crucial to learn representations for both users and items. Recent multimodal recommendation models achieve higher accuracy by incorporating multimodal features, such as images and text descriptions. However, our experimental findings reveal that current multimodality fusion methods employed in state-of-the-art models may adversely affect recommendation performance without compromising model architectures. Moreover, these models seldom investigate internal relations between item-item and user-user interactions. In light of these findings, we propose a model that enhances the dyadic relations by learning Dual RepresentAtions of both users and items via constructing homogeneous Graphs for multimOdal recommeNdation. We name our model as DRAGON. Specifically, DRAGON constructs user-user graphs based on commonly interacted items and item-item graphs derived from item multimodal features. Graph learning on both the user-item heterogeneous and homogeneous graphs is used to obtain dual representations of users and items. To capture information from each modality, DRAGON employs an effective fusion method, attentive concatenation. Extensive experiments on three public datasets and eight baselines show that DRAGON can outperform the strongest baseline by 21.41% on average. Our code is available at https://github.com/hongyurain/DRAGON.

1 Introduction

As society evolves, recommender systems have become indispensable tools to assist users in finding products and services tailored to their preferences. Previous work [1, 17, 7, 26] have examined historical user-item interactions, which can be regarded as a form of dyadic relation, to capture user preferences. However, these methods exhibit suboptimal performance due to the sparse nature of interactions between users and items in real-world datasets.

To alleviate the data sparsity problem, recent multimodal recommender systems that utilize multimodal information (*e.g.*, item descriptive texts, product images) to enhance recommendation performance have gained considerable attention. A line of research [11, 6] integrates multimodal features as supplementary information to improve latent item representations within the classic collaborative filtering framework. Inspired by the success of graph neural networks (GNNs) in recommendation [20, 7], recent studies have focused on modeling user-item interactions as a bipartite graph and integrating multimodal information with graph structure. For instance, MMGCN [24] constructs a user-item bipartite graph for each modality to obtain modal-specific representations for better understanding user preferences. GRCN [23] introduces a graph refine layer capable of identifying noisy edges and eliminating false-positive edges to clarify the structure of the user-item interaction graph. DualGNN [19] and LATTICE [25] incorporate either user-user or item-item relations into the user-item interactions, achieving state-of-the-art recommendation performance. Although these models demonstrate effective recommendation accuracy, we posit that high-order relations on both sides of dyadic relations can be explored simultaneously to fully address the data sparsity issues. Inspired by the dual representation learning mechanism [27], we enhance the representation learning of users and items by incorporating their dual representations to capture both the inter- and intra-relations between users and items.

Furthermore, we experimentally reveal that these methods fail to effectively fuse the modality features. Specifically, we conduct an ablation study of multimodal features on two competitive multimodal models, DualGNN [19] and LATTICE [25]. The results presented in Table 1 show that the performance of these models fed with a single modality, especially textual features, outperforms that with both modalities. This finding poses a meaningful question: *How can we effectively fuse the multimodal information for recommendation?*

Table 1. Performance of DualGNN [19] and LATTICE [25] utilizing features in different modalities. R and N denote evaluation metrics Recall and NDCG. T and V denote textual and visual information.

Dataset	Metric	DualGNN			LATTICE		
		V&T	T	V	V&T	T	V
Baby	R@10	0.0448	**0.0612**	0.0511	**0.0547**	0.0546	0.0492
	R@20	0.0716	**0.0943**	0.0830	0.0850	**0.0874**	0.0781
	N@10	0.0240	**0.0331**	0.0278	**0.0292**	0.0287	0.0265
	N@20	0.0309	**0.0417**	0.0360	0.0370	**0.0371**	0.0339
Sports	R@10	0.0568	**0.0697**	0.0615	0.0620	**0.0625**	0.0572
	R@20	0.0859	**0.1060**	0.0926	0.0953	**0.0971**	0.0887
	N@10	0.0310	**0.0379**	0.0335	0.0335	**0.0336**	0.0312
	N@20	0.0385	**0.0473**	0.0415	0.0421	**0.0425**	0.0393
Clothing	R@10	0.0454	**0.0524**	0.0420	0.0492	**0.0521**	0.0408
	R@20	0.0683	**0.0798**	0.0636	0.0733	**0.0749**	0.0614
	N@10	0.0241	**0.0281**	0.0229	0.0268	**0.0290**	0.0221
	N@20	0.0299	**0.0351**	0.0283	0.0330	**0.0348**	0.0273

To address this question, we investigate the performance of various modality fusion methods, including **Max-pooling**, **Mean-pooling**, **Attentively Sum**, and **Attentively Concatenation**. Our experiments indicate that the late-fusion approach, **Attentively Concatenation**, which directly concatenates the textual and visual features as the multimodal representation, achieves the best performance.

* Corresponding Author. Email: zqshen@ntu.edu.sg.

In light of these findings, we propose a framework that learns Dual RepresentAtions of both users and items via constructing homogeneous Graphs for multimOdal recommeNdation (DRAGON). Specifically, DRAGON constructs a heterogeneous user-item bipartite graph for each modality to learn the modality-specific representations. It then employs the direct attentive Concatenation fusion method to better exploit the learned modality-specific information. To learn dual representations, we construct two homogeneous graphs based on the user co-occurrence and the item semantic features to capture the user preference from neighboring users and the latent item content semantic from neighboring items. Finally, DRAGON leverages the learned dual representations of users and items to make recommendations. Extensive experiments are conducted on three public datasets to show the effectiveness of our proposed method.

2 Related Work

2.1 Multimodal Recommendation

Collaborative filtering (CF) based models are widely employed [8, 20, 33, 32] in recommender systems. These CF-based methods leverage historical interactions between users and items to predict user preferences. However, they often suffer from data sparsity issues, as user-item interactions are typically limited in real-world datasets.

To mitigate this problem, massive multimodal content information has been utilized to improve recommendation performance. For instance, VBPR [6], the first model to consider visual information, leverages the visual features from a pre-trained Convolutional Neural Network (CNN) to augment the matrix factorization by incorporating the visual features with ID embeddings. Inspired by the fact that humans process the modality information with varying attention, VECF [4] models user attention on different regions of images and reviews to better capture user preferences. Recently, Graph Neural Networks (GNNs) have gained increased attention in the context of multimodal-based recommender systems. MMGCN [24] enhances the quality of learned user and item representations by constructing a modality-specific user-item bipartite graph and adapting the message-passing mechanism of GNNs. Building on MMGCN, GRCN [23] introduces a graph refine layer capable of identifyling noisy edges and eliminating false-positive edges to refine the structure of the user-item interaction graph. DualGNN[27] incorporates an attention mechanism to capture user preferences across different modalities, while constructing a user-user graph to learn the user preference from neighboring users. LATTICE [25] builds an item-item graph for each modality, combining them to form a latent modality-fused item graph. Through graph convolution operations, items can share information from highly linked affinities within the graph to enhance their representations. Authors in [30] demonstrate that learning the item-item graph in LATTICE yields negligible improvement in recommendation performance, and freezing the graph is more beneficial for recommendations. Self-supervised learning techniques have also proven to be effective in multimodal recommendation systems. SLMRec [18] integrates self-supervised learning tasks into GNNs to uncover latent patterns from multi-modalities, thereby learning powerful representations. BM3 [34] bootstraps latent representations of both ID embeddings and multimodal features using a contrastive view generator and designs three contrastive objective functions to optimize representations for effective and efficient recommendations. For an in-depth exploration of multimodal recommender systems, we recommend consulting the comprehensive survey conducted by [29].

2.2 Multimodal Fusion

Identifying a fused multi-modal representation that is complementary and comprehensible can significantly enhance performance. Technically speaking, multi-modal fusion integrates information from different modalities to create a multimodal representation applicable to various tasks, such as link prediction [13, 12] and node classification [28], etc. It can be categorized into early fusion, late fusion and hybrid fusion [2]. Early fusion incorporates extracted features at the beginning, while late fusion integrates information after each modality has completed its decision-making process (*e.g.*, classification or regression). Hybrid fusion combines the two methods. For example, ACNet [9] employs an early fusion method based on the attention mechanism. Regarding late fusion, NMCL [22] utilizes the cooperative networks for each modality to perform feature augmentation with the attention mechanism, followed by late fusion applied to predictions from various modalities. CELFT [21] designs a hybrid fusion method, combining both early and late fusion, to overcome the limitations of single fusion methods.

Multimodal recommendation models [24, 25, 27] typically apply mean-pooling, attentive sum or max-pooling. Our experimental results suggest that those models use a single modality representation leads to better performance than utilizing the multimodal representation learned from these fusion methods. It indicates that sum and max pooling methods may result in information loss when performing multimodal fusion. To address this issue, we adopt attentive concatenation fusion without reducing the embedding dimension, which has proven more effective in combining information from different modalities for recommendations.

3 Methodology

This section provides a detailed explanation of our proposed model DRAGON. Fig. 1 presents the overall architecture of DRAGON, which consists of four main components: (1) Graph learning on a modality-specific heterogeneous graph to obtain uni-modal representations; (2) A Multimodal representation learning module that captures user preferences across different modalities and extracts complementary information from each modality; (3) Graph learning on homogeneous graphs to capture the co-occurrence relations between users and the semantic relations between items; (4) A predictor module that ranks candidate items based on scores calculated from final user

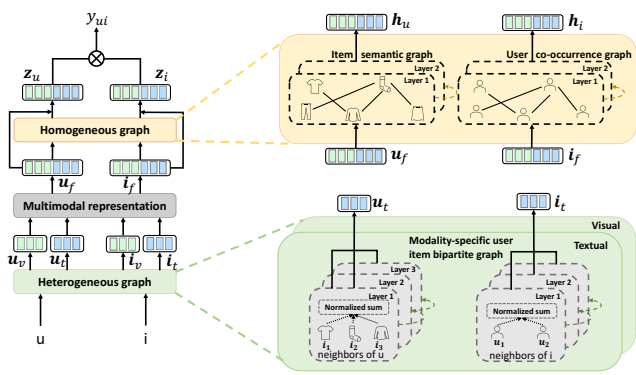


Figure 1. An overview of our proposed DRAGON.

3.1 Preliminary

Given a set of N users $u \in \mathcal{U}$, a set of M items $i \in \mathcal{I}$. We model the dyadic relations of user interactions as a user-item bipartite graph

$\mathcal{G} = \{\mathcal{U}, \mathcal{I}, \mathcal{E}\}$, where we regard the historical interactions as the set of edges in the graph denoted by $\mathcal{E}$. Besides the user-item interactions, each item is associated with multimodal content information $m \in \{v, t\}$, where v and t represent the visual and textual features respectively. We denote the modality feature for an item i as $\boldsymbol{x}_i^m \in \mathbb{R}^{d_m}$, where d_m denotes the feature dimension of modality m. In this paper, we only consider the visual and textual modalities denoted by v and t. However, the proposed framework can be easily extended to scenarios involving more than two modalities.

3.2 Dual Representation Learning

Learning the representations of users and items is critical for the recommendation system. All representation learning-based techniques assume the existence of a common representation containing consistent knowledge of different views of items [27]. Distinct item views contain specific discriminant information in addition to consistent knowledge about this item. We construct the heterogeneous and homogeneous graphs together to learn dual representations of both user and item, capturing both internal associations and relationships between users and items.

3.2.1 Heterogeneous Graph

To learn modality-specific user and item representations, we construct a user-item graph for each modality, which is denoted as $\mathcal{G}_m$. Following MMGCN [24], we maintain the same graph structure $\mathcal{G}$ for different $\mathcal{G}_m$, but only retain the node features associated with a specific modality m. We adopt LightGCN [7] to encode $\mathcal{G}_m$. As shown in [7], LightGCN simplifies graph convolutional operations by excluding the feature transformation and nonlinear activation modules to improve recommendation performance while easing the model optimization process. Specifically, the user and item representations at the $(l+1)$-th graph convolution layer of $\mathcal{G}_m$ are derived as follows:

$$\boldsymbol{u}_m^{(l+1)} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \boldsymbol{i}_m^{(l)}, \quad \boldsymbol{i}_m^{(l+1)} = \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \boldsymbol{u}_m^{(l)}. \tag{1}$$

where $\mathcal{N}_u$ and $\mathcal{N}_i$ are the set of first hop neighbors of u and i in $\mathcal{G}_m$. $\boldsymbol{u}_m^{(0)}$ is randomly initialized and $\boldsymbol{i}_m^{(0)}$ is initialized with $\boldsymbol{x}_i^m$. The symmetric normalization $\frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}}$ is used to normalize the modality features learned from each layer which avoids the increase of scale when performing the graph convolutional operations.

After L layers of data propagation, we combine the representations from every GCN layer using element-wise summation to derive the modality-specific representations for users and items. Formally,

$$\boldsymbol{u}_m = \sum_{l=0}^{L} \boldsymbol{u}_m^{(l)}, \qquad \boldsymbol{i}_m = \sum_{l=0}^{L} \boldsymbol{i}_m^{(l)}. \tag{2}$$

In such cases, the historical interactions and modality information have been encoded into the final single-modal representations of users and items. These operations are applied to each modality by propagating on the modality-specific user-item bipartite graphs to learn the representations for each modality.

3.2.2 Homogeneous Graph

In addition to employing the heterogeneous graph, which encodes the dyadic relation between users and items, we argue that recommendation performance can be further enhanced by modeling the internal relations between users or items. For the two homogeneous graphs, we pre-establish and freeze them to maintain the initial co-occurrence relation and semantic meaning.

User Co-occurrence Graph. Based on the assumption that users who have interacted with similar items typically have similar preferences, we argue that the user's preference pattern is hidden inside the co-occurrence items and we construct a homogeneous user co-occurrence graph to learn the internal relations. However, the numbers of co-occurrence items between users span a broad range. Generally, the user will have a high number of commonly interacted items with a small group of users but few items with other users. We only consider those with more commonly interacted items with the user to capture similar preferences. To explicitly model the item co-occurrence patterns of users, we construct a homogeneous user co-occurrence graph $\tilde{\mathcal{G}} = \{\mathcal{U}, \mathcal{P}_u\}$, where $\mathcal{P}_u = \{e_{u,u'} | u, u' \in \mathcal{U}\}$ denotes the edges between user nodes in $\tilde{\mathcal{G}}$ and $e_{u,u'}$ record the number of items that commonly interacted with u and u'.

For every user $u \in \mathcal{U}$, we retain its top-k users with the highest number of commonly interacted items. Specifically, we keep the edge weight $e_{u,u'}$ if u' belongs to the top-k users. Otherwise, the edge weight is 0.

$$e_{u,u'} = \begin{cases} e_{u,u'} & \text{if } e_{u,u'} \in \text{top-k}(e_u), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Upon establishing the user co-occurrence graph, we incorporate the attention mechanism during graph propagation. The weight used for aggregating neighboring nodes for a user is computed using the softmax function to maximize the effect of neighboring users with a higher number of commonly interacted items. The representation of u learned from $\tilde{\mathcal{G}}$ at layer $l+1$ is denoted as $\boldsymbol{h}_u^{(l+1)}$, which is derived as follows:

$$\boldsymbol{h}_u^{(l+1)} = \sum_{u' \in \mathcal{N}_u} \frac{exp(e_{u,u'})}{\sum_{\hat{u} \in \mathcal{N}_u} exp(e_{u,\hat{u}})} \boldsymbol{h}_{u'}^{(l)}. \tag{4}$$

where $e_{u,u'}$ indicates the number of common interacted items between u and u' and $\mathcal{N}_u$ denotes the neighbors of user u in $\tilde{\mathcal{G}}$. In this case, the representation of each user can be enhanced based on neighbors in the co-occurrence graph.

We experimented with alternative methods to construct the user graph, such as averaging the features of neighbor items to represent the user or using contrastive loss to minimize the difference between neighboring users. However, these methods did not yield satisfactory performance. Ultimately, we found that the simple but efficient co-occurrence method worked best. In the future, there may be potential to develop more effective approaches for constructing the user graph.

Item Semantic Graph. Multimodal features offer rich and valuable content information about items, but previous studies [23, 19] neglect the significant underlying semantic relations of item features. Inspired by [25], we argue that item features are objective and we could establish the modality-specific homogeneous item graphs based on raw features to learn the internal relations between items. Specifically, we construct the modality-aware item semantic graph $\hat{\mathcal{G}}_m = \{\mathcal{I}, \mathcal{P}_m^i\}$ for each modality m, where $\mathcal{P}_m^i = \{e_{i,i'}^m | i, i' \in \mathcal{I}\}$ denotes the edges between item nodes in $\hat{\mathcal{G}}_m$. For an edge $e_{i,i'}$, its weight is calculated by the cosine similarity between original modality features of $\boldsymbol{x}_i^m$ and $\boldsymbol{x}_{i'}^m$:

$$e_{i,i'}^m = \frac{(\boldsymbol{x}_i^m)^\top \boldsymbol{x}_{i'}^m}{\|\boldsymbol{x}_i^m\| \|\boldsymbol{x}_{i'}^m\|}. \tag{5}$$

The derived $\hat{\mathcal{G}}_m$ is a fully connected graph where edge weights are calculated based on the similarity scores of modality features of

connected nodes. Next, we make the graph sparse by retaining the top-k similar items of every item. As the $\hat{\mathcal{G}}_m$ is a weighted graph, we convert it into an unweighted graph that captures the fundamental relation structure of the most related items [3]. Formally,

$$e^m_{i,i'} = \begin{cases} 1 & \text{if } e^m_{i,i'} \in \text{top-k}(e^m_i), \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Since we get one item semantic graph for each modality, we combine them by performing weighted summation based on the importance score α_m that indicates the contribution of each modality and the summation is 1. Formally, $\hat{\mathcal{G}} = \{\mathcal{I}, \mathcal{P}_i\}$, where $\mathcal{P}_i = \{e_{i,i'}|i, i' \in \mathcal{I}\}$ and $e_{i,i'} = \sum_{m \in M} \alpha_m e^m_{i,i'}$.

LATTICE [25] designs an item-item graph structure similar to ours but the item-item structure learning of LATTICE is proven to be dispensable as disclosed by [30]. We decide to freeze it during training so that it could reach a better performance.

After establishing the item semantic graph, we apply the graph convolution operation on it to capture the item-item relationship:

$$\boldsymbol{h}^{(l+1)}_i = \sum_{i' \in \mathcal{N}_i} e_{i,i'} \boldsymbol{h}^{(l)}_{i'}. \tag{7}$$

where $\mathcal{N}_i$ denotes the neighbors of item i in $\mathcal{G}^i$. Both $\boldsymbol{h}^{(0)}_u$ and $\boldsymbol{h}^{(0)}_i$ are initialized with their fused representations $\boldsymbol{u}_f$ and $\boldsymbol{i}_f$, which are introduced in the following section.

3.3 Multimodal Fusion

A crucial factor influencing multimodal recommendation accuracy is multimodal fusion. As mentioned in Section 1, some previous multimodal recommendation models utilizing single-modal information outperform those using multimodal information without changing the model structure. We hypothesize their fusion methods may fail to capture modality-specific characteristics and even corrupt the learned single-modality representation. Our aim is to learn multimodal representations that can capture complementary information not contained within a single modality. To fuse the single modal features derived from modality-specific user-item graphs, we apply the **Attentive Concatenation** for user multimodal embeddings, which capture user preferences across different modalities, and direct **Concatenation** for item multimodal embeddings. The attention weight α for users is initialized to 0.5 learned during the training to capture important scores of different modalities. Formally,

$$\boldsymbol{u}_f = \alpha \boldsymbol{u}_v \parallel (1-\alpha)\boldsymbol{u}_t, \qquad \boldsymbol{i}_f = \boldsymbol{i}_v \parallel \boldsymbol{i}_t. \tag{8}$$

where $\parallel$ denotes the concatenation operation. By performing attentive concatenation, we assume the single modality representations carry the richest information for each modality and this operation can capture the intact complementary information from each modality. The attention weight α will help measure the importance of each modality influencing user preferences. Our approach involves utilizing attentive concatenation for user representations while employing concatenation for items. This is because the importance of the different modalities in the item representation has already been captured through the combination of item semantic graphs mentioned earlier. To validate the effectiveness of our approach, we conducted an ablation study comparing different modality fusion methods, and the results demonstrate the superiority of our method in terms of performance.

Regularization is employed to prevent overfitting when the model is too complex and starts fitting the noise in the data rather than the underlying patterns. Modality weight α has been added as the regularization penalty term to the loss function, which helps reduce model complexity by encouraging smaller values for the model parameters and preventing them from taking on large values that may cause overfitting. By adding a penalty term to the loss function, regularization encourages the model to learn simpler representations of the data that are more likely to generalize to new, unseen data.

3.4 Integration with Dual Representations

We integrate representations of users and items learned from heterogeneous (*i.e.*, Modality-specific User-Item Graphs) and homogeneous (*i.e.*, User Co-occurrence Graph & Item Semantic Graph) graphs to form their dual representations, ensuring that interactions between users and items, as well as their internal relations, are effectively captured. We perform element-wise summation on the outputs learned from the three graphs to generate the dual representations for u and i:

$$\boldsymbol{z}_u = \boldsymbol{u}_f + \boldsymbol{h}^{L_u}_u, \qquad \boldsymbol{z}_i = \boldsymbol{i}_f + \boldsymbol{h}^{L_i}_i. \tag{9}$$

where $\boldsymbol{z}_u$ and $\boldsymbol{z}_i$ denote the final representations of user u and item i. L_u and L_i denote the number of GCN layers for the user co-occurrence graph and item semantic graph respectively.

3.5 Optimization

To optimize the parameters of DRAGON for the recommendation task, we leverage the Bayesian Personalized Ranking (BPR) loss [17], which aims to score higher for the positive item than the negative one. We construct a triplet set $\mathcal{R}$ that includes the triplet (u, i, j) for each user u with the positive item i and a randomly sampled negative item j that has no interactions with u. The loss function $\mathcal{L}_{rec}$ is defined as follows,

$$\begin{aligned} &\mathcal{R} = \{(u,i,j)|(u,i) \in \mathcal{E}, (u,j) \notin \mathcal{E}\}, \\ &\mathcal{L}_{rec} = \sum_{(u,i,j) \in \mathcal{R}} -\ln \sigma(y_{ui} - y_{uj}) + \lambda \|\boldsymbol{\Theta}\|_2. \end{aligned} \tag{10}$$

$y_{ui} = \boldsymbol{z}^\top_u \boldsymbol{z}_i$ calculates the inner product of $\boldsymbol{z}_u$ and $\boldsymbol{z}_i$, λ is the L_2 regularization weight, and $\boldsymbol{\Theta}$ denotes model parameters.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

We conduct experiments on three categories Baby, Sports and Clothing from the Amazon dataset [14], which contains product descriptions and images as textual and visual features. We follow previous work [6, 7, 25] to use implicit feedback that tracks users' preferences by monitoring performed actions like review here but not considers the exact ratings. Moreover, We retain the 5-core setting for users and items, ensuring that each user or item is associated with at least 5 interactions. We use the pre-trained sentence-transformers [16] to extract text features with a dimension equal to 384 and follow [25] to use the published 4096-dimensional visual features. The dataset statistics are summarised in Table 2. Sparse data is a common issue in recommendation systems that arises from the fact that the majority of users typically only interact with a small subset of items, resulting in a sparse user-item interaction matrix. Data sparsity is calculated by dividing the number of interactions by the product of the number of items and users.

Table 2. Statistics of the datasets.

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

4.1.2 Baselines

To evaluate the performance of our proposed model DRAGON, we compare it with traditional recommendation models that only utilize u-i interaction and multimodal models that reach the current best performance. We conform to the settings and hyperparameter search methods of the baseline papers, ensuring fair comparison.
1) General recommendation models:

- **BPR** [17] optimizes the user and item representations utilizing the matrix factorization method.
- **LightGCN** [7] simplifies the Graph Convolution Network by discarding the feature transformation and nonlinear activation modules.

2) Multimodal recommendation models:

- **VBPR** [6] integrates visual features into item representations. For a fair comparison, we combine text and vision features to learn item representations.
- **DualGNN** [19] proposes using representations learned from modality-specific graphs and fusing the representations of neighbors in the user correlation graph.
- **GRCN** [23] locates and removes the false-positive edges in the graph. It then learns representations of items and users by conducting information propagation and aggregation in the refined graph.
- **LATTICE** [25] introduces an item-item graph on each modality and obtains the latent item semantic graph by aggregating information from all modalities.
- **SLMRec** [18] uses self-supervised learning techniques that supplement the supervised tasks to uncover the hidden signals from the data itself with contrastive loss.
- **BM3** [34] uses self-supervised learning techniques that bootstrap latent representations of both ID embeddings and multimodal features. It designs a multi-modal contrastive loss to optimize the objective functions without negative samples.

4.1.3 Evaluation Metrics

We follow the settings as previous models [25, 30] to randomly split the historical interactions with the ratio of 8:1:1 as train, valid and test sets. Moreover, we adopt the widely used metrics Recall@K and NDCG@K (denoted by R@K and N@K) to evaluate the top-K recommendation performance. We empirically set $K = 10$ and 20. For each metric, we compute the performance of each user in the testing data and report the average performance over all users.

4.1.4 Implementation Details

We implement our proposed model by PyTorch [15] and embed the users and items with a dimensional size of 64 for all models. We use the Xavier method [5] to initialize the embedding parameters, utilize Adam [10] as the optimizer, and fix the mini-batch size to 2048. All models are evaluated on a Tesla V100 32GB GPU card. The optimal hyper-parameters are determined via grid searches on the validation set: we do a grid search on the learning rate in {1e-1, 1e-2, 1e-3, 1e-4, 1e-5,1e-6}, regularization weight in {1e-1, 1e-2, 1e-3, 1e-4, 1e-5} and importance score α_m for image weight in the item semantic graph in {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9}. We fix the number of GCN layers in the heterogeneous graph and homogeneous graph with $L = 2$ and $L_u = L_i = 1$, respectively. we consider k of top-k in the user co-occurrence graph as empirical value setting to 40 and setting to 10 in the item semantic graph for baby, 5 for sports and clothing. We set the maximum number of epochs to 1000 and adopt the early stopping strategy. That is, the model terminates when R@20 on the validation set does not increase for 20 successive epochs. To ensure a fair comparison, all baseline models as well as our proposed model have been integrated into the unified multimodal recommendation framework, MMRec [31].

4.2 Performance Comparison

As depicted in Table 3, we compare the recommendation performance of the state-of-the-art methods with our proposed model, leading to the following observations:

- All GCN-based methods outperform traditional MF-based recommendation models (*i.e.*, BPR and VBPR), demonstrating the effectiveness of modeling the historical interactions using a graph with graph convolutional operations.
- Across all evaluation metrics, including Recall and NDCG, DRAGON surpasses all baseline models on every dataset. For example, in terms of R@20, DRAGON improves upon the strongest baseline on the datasets Baby, Sports, and Clothing by 15.63%, 15.05%, and 33.56% respectively. This improvement is attributed to the dual representations and the multimodal fusion method. Learning dual representations from the heterogeneous and homogeneous graphs captures both the historical interactions and the internal relations among each set of dyadic objects (*i.e.*, users or items). A homogeneous graph aids in learning relevant characteristics from the neighbors, while the fusion method enhances the multimodal representation by acquiring complementary information from each single modality.
- Multimodal recommendation models outperform general recommendation models. GRCN, LATTICE, SLMRec, and BM3 are multimodal models that outperform all general methods. VBPR, which builds upon the BPR framework by introducing modality information, outperforms BPR on all datasets.
  However, some multimodal models rely heavily on the representativeness of multimodal characteristics of items, resulting in inconsistent performance across various datasets. For example, DualGNN is built upon LightGCN and outperforms it on the Clothing dataset but is less effective on Baby and Sports. It is possible that multimodal features are more critical for revealing item characteristics in the clothing dataset but are less informative in the other two datasets, in which DualGNN underperforms LightGCN.

Additionally, we evaluate the scalability of DRAGON on a larger dataset Electronic of Amazon dataset with around 1.7M interactions, 200K users and 63K items. LATTICE [25] consumes more memory than the other baselines which could not be handled by the 32GB GPU card. SLMRec [18] needs to find the optimal parameter by grid search on more than 200 parameter sets which takes a long training time. DRAGON outperforms these baselines on the large graph. Although SLMRec is the strongest baseline on Electronic, DRAGON achieves an improvement of 4.76% in terms of R@20 compared to SLMRec.

Table 3. Performance of baselines in terms of Recall and NDCG. Best results are in **boldface** and the second best is underlined. "%Imp" denotes the relative improvement of DRAGON over the best baseline.

Dataset	Metric	General Model		Multimodal Model							
		BPR	LightGCN	VBPR	DualGNN	GRCN	SLMRec	LATTICE	BM3	DRAGON	%Imp
Baby	R@10	0.0357	0.0479	0.0423	0.0448	0.0539	0.0529	0.0547	0.0564	**0.0662**	17.38%
	R@20	0.0575	0.0754	0.0663	0.0716	0.0833	0.0775	0.0850	0.0883	**0.1021**	15.63%
	N@10	0.0192	0.0257	0.0223	0.0240	0.0288	0.0290	0.0292	0.0301	**0.0345**	14.62%
	N@20	0.0249	0.0328	0.0284	0.0309	0.0363	0.0353	0.0370	0.0383	**0.0435**	13.58%
Sports	R@10	0.0432	0.0569	0.0558	0.0568	0.0598	0.0663	0.0620	0.0656	**0.0752**	13.42%
	R@20	0.0653	0.0864	0.0856	0.0859	0.0915	0.0990	0.0953	0.0980	**0.1139**	15.05%
	N@10	0.0241	0.0311	0.0307	0.0310	0.0332	0.0365	0.0335	0.0355	**0.0413**	13.15%
	N@20	0.0298	0.0387	0.0384	0.0385	0.0414	0.0450	0.0421	0.0438	**0.0512**	13.78%
Clothing	R@10	0.0206	0.0361	0.0281	0.0454	0.0424	0.0442	0.0492	0.0422	**0.0671**	36.38%
	R@20	0.0303	0.0544	0.0415	0.0683	0.0662	0.0659	0.0733	0.0621	**0.0979**	33.56%
	N@10	0.0114	0.0197	0.0158	0.0241	0.0223	0.0241	0.0268	0.0231	**0.0365**	36.19%
	N@20	0.0138	0.0243	0.0192	0.0299	0.0283	0.0296	0.0330	0.0281	**0.0443**	34.24%

4.3 Ablation Study

In this section, we conduct exhaustive experiments to examine the behaviors of our proposed model under various settings.

4.3.1 Effect of different components of DRAGON

We devise the following variants of DRAGON based on the homogeneous graphs employed and compare with the strongest baselines (LATTICE, SLMRec and BM3) to investigate the contribution of different components of DRAGON:

- DRAGON_{UI} omits the homogeneous graphs and relies solely on the heterogeneous graph.
- DRAGON_{UU} incrementally incorporates the user co-occurrence graph into DRAGON_{UI}. This variant captures the relations between users, signifying that only users have dual representations.
- DRAGON_{II} incrementally integrates the item semantic graph into DRAGON_{UI}. This variant captures the relations between items, indicating that only items have dual representations.

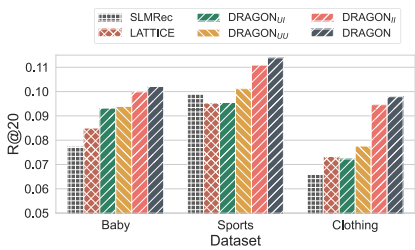


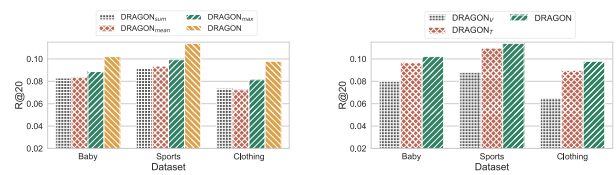
Figure 2. Effect of different components of DRAGON.

Fig. 2 presents our comparison results in terms of Recall@20.

The following observations demonstrate that all components of DRAGON contribute to its performance: (1) DRAGON_{UI}, which exclusively employs our multimodal fusion, achieves comparable or even higher accuracy than the strong baselines. This finding highlights the efficiency of the attentive concatenation method. (2) $\text{DRAGON}_{UI} < \text{DRAGON}_{UU} < \text{DRAGON}_{II}$ illustrates the benefits of incorporating homogeneous graphs. However, we observe that utilizing the user co-occurrence graph does not significantly improve the results for the Baby dataset. As Table 2 indicates, the Baby dataset is less sparse than the other two datasets, suggesting that user relation co-occurrence patterns in denser datasets are primarily captured by the user-item interaction graph. Consequently, the user co-occurrence graph contributes less in this case. (3) $\text{DRAGON}_{UU} < \text{DRAGON}_{II}$ shows that the primary performance improvement results from the addition of an item semantic graph. The discrepancy in performance can be attributed to our approach of constructing a user co-occurrence graph based on user-item interactions, which emphasizes user relations already contained in interactions. In contrast, the item semantic graph is constructed using raw features, enabling the capture of additional modality similarity relations beyond those relations present in the user-item interactions. (4) DRAGON achieves the best performance, demonstrating the effectiveness of integrating the multimodal fusion method and dual representation learning.

4.3.2 Effect of different modality fusion method

We identify modality fusion issue and utilize the direct concatenation with attention in our proposed model. We compare its performance with the fusion methods mentioned in [19][25]. We replace the fusion of DRAGON with weighted sum (denoted as DRAGON_{sum}), mean (denoted as DRAGON_{mean}) and weighted max (denoted as DRAGON_{max}) to demonstrate the superiority of concatenation fusion used in DRAGON. Fig. 3(a) shows comparison results in terms of Recall@20. Clearly, our fusion method outperforms others as it can more effectively capture complementary information of each modality and attend information from all modalities for recommendation.



(a) Comparison of different fusion methods (b) Comparison of different modality

Figure 3. Ablation study on fusion methods and feature modalities.

4.3.3 Effect of single modality vs. multi-modalities

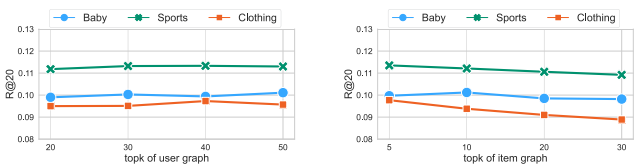
In the Introduction, we reveal performance of previous multimodal models might degrade under multimodal settings. Hence, we compare performance of DRAGON under uni-modal and multimodal settings. DRAGON_V, DRAGON_T, and DRAGON denote the models that utilize visual, textual, and both modalities respectively. Fig. 3(b) shows our comparison results in terms of Recall@20. We observe that: (1) Models with different single modality information have different performances. Textual modality performs better than visual modality. (2) DRAGON with multimodal information outperforms those utilizing single modality, demonstrating that fusing different modalities of information can improve performance of DRAGON. Thus, fusion method in our proposed model is indispensable.

4.4 Hyper-parameter Sensitivity Study

4.4.1 Effect of top-k in homogeneous graph

We conduct an ablation study to investigate the influence of the top-k values on the user co-occurrence graph. The user co-occurrence graph is established based on the shared items between users, and we only retain the top-k users with the highest number of commonly interacted items. We search for the top-k from a set of {20, 30, 40, 50}. Fig. 4(a) displays the comparison results in terms of Recall@20. We find that the performance is not highly sensitive to the choice of the top-k value. This may be because the performance is not heavily influenced by the user co-occurrence graph as shown in Fig. 2. Another reason is that closely linked users predominantly share similar user behavior information, thus increasing the top-k value has a limited impact on overall performance.

Similarly, we investigate the influence of the top-k values on the item semantic graph. The edge weights are the cosine similarity scores of modality features of connected nodes. We retain the top-k similar items for each item. The top-k value is searched in {5,10,20,30}. The comparison result, shown in Fig. 4(b) in terms of Recall@20, reveals that DRAGON achieves the best performance with small k values. Baby dataset reaches the best performance with $k = 10$ and the other two datasets reach the best with $k = 5$. Large k values lead to performance drop in all the datasets.



(a) Top-k value of user graph (b) Top-k value of item graph

Figure 4. Effects of top-k in constructing homogeneous graphs.

4.4.2 Effect of latent dimensionality

The embedding dimensionality d of our model is searched from {2^4, 2^5, 2^6, 2^7,2^8}. Fig. 5(a) indicates that our model's performance consistently improves with larger latent dimension sizes. However, increasing the dimension size also requires increased computational resources. Therefore, we use a moderate size for the latent dimension to balance performance and resource consumption. In our experiments, we set the embedding size at a fixed value of 64, which not only attains a performance comparable to larger sizes but also necessitates fewer computational resources.

4.4.3 Effect of importance score α_m

As presented in section 3.2.2, the importance score α_m is used to indicate the contribution of each modality when integrating the modality-specific item semantic graphs. We examine α_m from 0.1 to 0.9 with an increment of 0.1 to control the visual modality importance of items, with text modality importance equal to $1 - \alpha_m$. Fig. 5(b) shows the comparison result in terms of Recall@20. We observe that increasing α_m results in a performance drop across all datasets, indicating that the textual modality is more informative than the visual modality in constructing the item semantic graph. Moreover, The three datasets achieve the highest performance with the α_m among {0.1,0.2,0.3}, indicating that the importance of visual modality varies across different datasets.

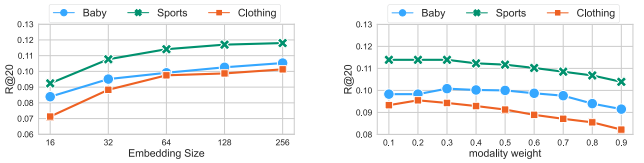
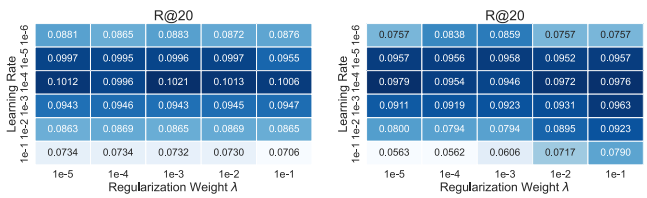


Figure 5. Performance changing of DRAGON *w.r.t* embedding size and the ratio of modality.

4.4.4 Effect of learning rate and regularization weight

To determine the optimal learning rate and regularization weight for DRAGON, we conduct a sensitivity analysis using Recall@20 as performance metrics. We explore learning rates within the range {1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1} and regularization weights λ in the set {1e-5, 1e-4, 1e-3, 1e-2, 1e-1}. Fig. 6(a) and Fig. 6(b) display the performance of DRAGON under various combinations of learning rates and regularization weights on the Baby and Clothing datasets. Based on these results, we make the following observations: (1) A learning rate of 1e-4 achieves the best performance across all regularization weights. Excluding the extremely large and small learning rates, our model demonstrates strong performance when changing learning rates among the set {1e-3, 1e-4, 1e-5}. This further highlights the efficiency and stability of our model. (2) Performance is less sensitive to the regularization weight compared to the learning rate. Nevertheless, optimizing the regularization weight parameter set in DRAGON also makes little contributes to improved performance.



(a) Baby (b) Clothing

Figure 6. Sensitivity analyses on the DRAGON hyper-parameters.

5 Conclusion

In this paper, we aim to solve the modality fusion issue and learn better representations for dyadic-related users and items. Therefore, we develop a novel model, named DRAGON, to learn the dual representations of users and items by constructing homogeneous and heterogeneous graphs. In particular, we first construct the modality-specific user-item bipartite graph to learn the modality features. After getting the representations of each modality, we utilize the late concatenation fusion method to learn the multimodal features. Then, we construct the user co-occurrence graph to capture the co-occurrence relations between users and the item semantic graph to capture the semantic relations between items. Therefore, we learn both inter- and intra-relations of the dyadic-related users and items. Finally, we conduct extensive experiments on three datasets to demonstrate the effectiveness of our proposed model. Our experimental finding on multimodal fusion could shed light on the design of future multimodal recommender systems.

6 Acknowledgements

This work was supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore.

References

[1] Gediminas Adomavicius and Alexander Tuzhilin, 'Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions', *IEEE transactions on knowledge and data engineering*, **17**(6), 734–749, (2005).

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, 'Multimodal machine learning: A survey and taxonomy', *IEEE transactions on pattern analysis and machine intelligence*, **41**(2), 423–443, (2018).

[3] Jie Chen, Haw-ren Fang, and Yousef Saad, 'Fast approximate knn graph construction for high dimensional data via recursive lanczos bisection.', *Journal of Machine Learning Research*, **10**(9), 1989–2012, (2009).

[4] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha, 'Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation', in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 765–774, (2019).

[5] Xavier Glorot and Yoshua Bengio, 'Understanding the difficulty of training deep feedforward neural networks', in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, (2010).

[6] Ruining He and Julian McAuley, 'Vbpr: visual bayesian personalized ranking from implicit feedback', in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 144–150, (2016).

[7] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng 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*, pp. 639–648, (2020).

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua, 'Neural collaborative filtering', in *Proceedings of the 26th international conference on world wide web*, pp. 173–182, (2017).

[9] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang, 'Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation', in *IEEE International Conference on Image Processing*, pp. 1440–1444, (2019).

[10] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[11] Qiang Liu, Shu Wu, and Liang Wang, 'Deepstyle: Learning user preferences for visual recommendation', in *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pp. 841–844, (2017).

[12] Qidong Liu, Enguang Yao, Chaoyue Liu, Xin Zhou, Yafei Li, and Mingliang Xu, 'M2gcn: multi-modal graph convolutional network for modeling polypharmacy side effects', *Applied Intelligence*, 1–12, (2022).

[13] Wenxuan Liu, Hao Duan, Zeng Li, Jingdong Liu, Hong Huo, and Tao Fang, 'Entity representation learning with multimodal neighbors for link prediction in knowledge graph', in *2021 7th International Conference on Computer and Communications*, pp. 1628–1634, (2021).

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

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., 'Pytorch: An imperative style, high-performance deep learning library', *Advances in neural information processing systems*, **32**, (2019).

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

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

[18] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua, 'Self-supervised learning for multimedia recommendation', *IEEE Transactions on Multimedia*, (2022).

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

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

[21] Yifan Wang, Xing Xu, Wei Yu, Ruicong Xu, Zuo Cao, and Heng Tao Shen, 'Combine early and late fusion together: A hybrid fusion framework for image-text matching', in *2021 IEEE International Conference on Multimedia and Expo*, pp. 1–6, (2021).

[22] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen, 'Neural multimodal cooperative learning toward micro-video understanding', *IEEE Transactions on Image Processing*, **29**, 1–14, (2019).

[23] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua, 'Graph-refined convolutional network for multimedia recommendation with implicit feedback', in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3541–3549, (2020).

[24] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua, 'Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video', in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1437–1445, (2019).

[25] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang, 'Mining latent structures for multimedia recommendation', in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3872–3880, (2021).

[26] Lingzi Zhang, Yong Liu, Xin Zhou, Chunyan Miao, Guoxin Wang, and Haihong Tang, 'Diffusion-based graph contrastive learning for recommendation with implicit feedback', in *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part II*, pp. 232–247. Springer, (2022).

[27] Wei Zhang, Zhaohong Deng, Kup-Sze Choi, Jun Wang, and Shitong Wang, 'Dual representation learning for one-step clustering of multi-view data', *arXiv preprint arXiv:2208.14450*, (2022).

[28] Zufan Zhang, Xieliang Li, and Chenquan Gan, 'Multimodality fusion for node classification in d2d communications', *IEEE Access*, **6**, 63748–63756, (2018).

[29] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen, 'A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions', *arXiv preprint arXiv:2302.04473*, (2023).

[30] Xin Zhou, 'A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation', *arXiv preprint arXiv:2211.06924*, (2022).

[31] Xin Zhou, 'Mmrec: Simplifying multimodal recommendation', *arXiv preprint arXiv:2302.03497*, (2023).

[32] Xin Zhou, Donghui Lin, Yong Liu, and Chunyan Miao, 'Layer-refined graph convolutional networks for recommendation', in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pp. 1247–1259. IEEE, (2023).

[33] Xin Zhou, Aixin Sun, Yong Liu, Jie Zhang, and Chunyan Miao, 'Selfcf: A simple framework for self-supervised collaborative filtering', *ACM Transactions on Recommender Systems*, **1**(2), 1–25, (2023).

[34] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang, 'Bootstrap latent representations for multi-modal recommendation', in *Proceedings of the ACM Web Conference 2023*, p. 845–854, (2023).