DOGE: LLMs-Enhanced Hyper-Knowledge Graph Recommender for Multimodal Recommendation

Fanshen Meng, Zhenhua Meng, Ru Jin, Rongheng Lin*, Budan Wu

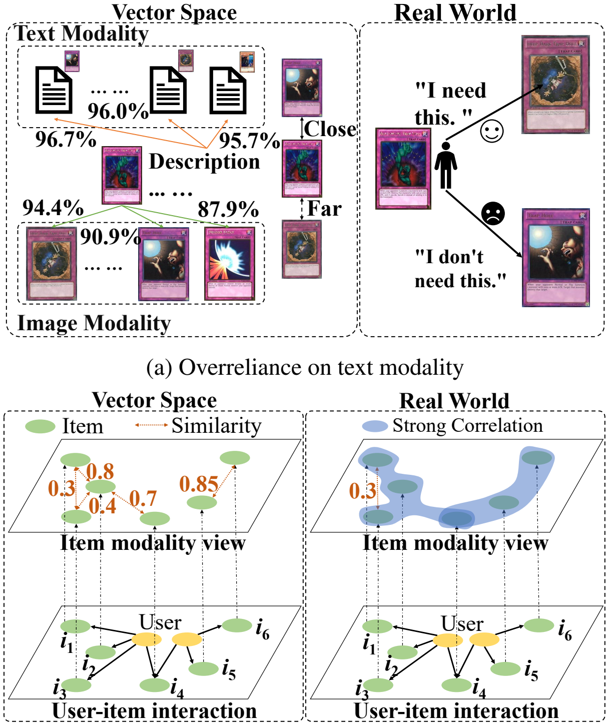
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
{mengfanshen, zhmeng, rjin, rhlin, wubudan}@bupt.edu.cn

Abstract

In recent years, there has been a burgeoning interest in multimodal recommender systems within the recommendation systems domain. These systems aim to understand user preferences by leveraging both user interaction data and multimodal information associated with items. This approach frequently results in superior recommendation accuracy compared to traditional models that rely solely on user-item interactions. Despite the advancements of these methods, there is a relatively low utilization of image features in propagating item-item characteristics, an overreliance on text feature similarity, and a frequent neglect of the deep relationships between items, users, and modalities. In response to these challenges, we introduce a novel model termed LLMs-Enhance**D** Hyper-Kn**O**wledge **G**raph R**E**commender for Multimodal Recommendation (**DOGE**). DOGE utilizes large language models (LLMs) to understand image information under the guidance of text information, generating cross-modal features that effectively enhance the relationship between text and image modalities. Subsequently, DOGE constructs a Hyper-Knowledge Graph (HKG) using user-item interaction information and modality features enhanced by large language models. This graph encompasses a wide range of item-item and user-user binary relations and hyper-relations, effectively expanding the feature propagation mechanisms and mitigating the overreliance on text modality. By learning on heterogeneous user-item graphs and homogeneous item-item, user-user graphs, DOGE enhances potential effective propagation between item features and user features, acquiring more effective feature representations of users and items. Comprehensive experimentation across three public real-world datasets illustrates that DOGE attains state-of-the-art (SOTA) performance, exhibiting a 7.2% improvement over the strongest baseline.

Introduction

As multimedia content rapidly expands, recommender systems have become essential for assisting users in discovering personalized products and value chain partners. Leveraging multimodal information, multimodal recommender systems (MRSs) have garnered considerable attention for effectively mitigating the data sparsity issues inherent in real-world datasets.



(a) Overreliance on text modality

(b) Strong correlation among items

Figure 1: Illustrations of (a) Overreliance on text modality, and (b) Strong correlation among items.

Early MRSs (Liu, Wu, and Wang 2017; Wei et al. 2021) fuse multimodal features with item ID embeddings. Some also apply attention mechanisms to model user preferences (Chen et al. 2017, 2019; Liu et al. 2019). However, these methods only capture low-order user-item interactions. The capabilities of graph neural networks (GNNs) in modeling high-order semantics have propelled graph-based recommendation systems to the forefront of multimodal recommendation research. For instance, MMGCN (Wei et al. 2019) disaggregates multimodal features into multiple spatial views and aggregates information within each space using graph convolutional networks (GCNs) to better model user preferences. MGCN (Yu et al. 2023) introduces a

*Corresponding author

behavior-aware fuser to further model the importance of different modal views of multimodal item features. Moreover, researchers have complemented user-item relations by constructing auxiliary graph structures. For example, DualGNN (Wang et al. 2021) builds a homogeneity graph of users to improve recommendation performance. Furthermore, DRAGON (Zhou et al. 2023b) integrates user graph and the frozen item graph (Zhou and Shen 2023), and learns simultaneously on both the homogeneous graph and the heterogeneous user-item graph to obtain dual embeddings of users and items, achieving optimal performance. In recent years, due to the powerful semantic understanding capabilities of large language models (LLMs), some researchers have attempted to use LLMs to enhance side information. For example, LLMRec(Wei et al. 2024) leverages LLMs to generate user profiles, thereby enhancing user representation, and has made significant progress.

Despite these significant achievements, current MRSs still face some limitations: (1) The significance of multimodal side information is uneven. FREEDOM's (Zhou and Shen 2023) hyper-parameter sensitivity study demonstrated that with the increase in modality weight (the weight of the image modality) in frozen item graph, the Recall@20 indicator shows an overall downward trend. This indicates that item features mainly learn along the side of higher text similarity in the item homogeneous graph. As illustrated in Figure 1(a), items that are significantly more similar in the visual modality may be farther apart in the vector space due to slight differences in the text modality. This shows that the information from the image modality has not been effectively mined, and the relationship between text and image pairs has not been adequately explored. Such over-reliance may result in a decline in recommendation performance due to missing or contaminated text features. (2) Relying solely on the propagation of item similarity adjacency in the vector space of item features poses challenges in fully addressing users' diverse needs. The propagation process of the homogeneous graph depends on the modality similarity of items. However, in the real world, even if the modality similarity between item groups is not high, there may still be strong associations between different item groups, and these associations can be hyper-relational, as shown in Figure 1(b). To address these issues, we propose an LLMs-Enhance**D** Hyper-Kn**O**wledge **G**raph R**E**commender for Multimodal Recommendation (**DOGE**). DOGE introduces a method of constructing a LLMs-enhanced Hyper-Knowledge Graph (HKG), effectively improving recommendation performance in the multimodal domain. Specifically:

- Modality Enhancement Method Based on LLMs: We propose a method to enhance modality relationships using a multimodal LLM. First, we input images as prompts to the multimodal LLM along with the corresponding text information. Then, we treat the LLM as a virtual knowledge base, utilizing it to understand image features and output image cues to strengthen the connections between modalities.
- Construction of a HKG for Recommendation: By deeply exploring the multi-dimensional relationships between users and items, we construct a HKG for recommendation. This graph extends multi-dimensional hyper-relations based on the user-item interactions and modality views, effectively meeting the diverse needs of recommendations.
- Our method's efficacy is demonstrated through tests conducted on three real-world datasets available to the public. Compared to the strongest baseline models, our method shows an average improvement of 7.2%.

Related Work

Multi-Modal Recommendation

Compared to traditional collaborative filtering models (He and McAuley 2016; Mao et al. 2021a; Papadakis et al. 2022), multimodal recommendation systems alleviate data sparsity issues present in real datasets by leveraging abundant multimodal information and user behaviors, effectively enhancing recommendation performance and garnering considerable attention (Wu et al. 2022a; Zhou et al. 2023a). Early multimodal recommendation systems typically integrate multimodal information into ordinary collaborative filtering frameworks. For example, VBPR is the first model to incorporate visual information (He and McAuley 2016). However, these methods often struggle to capture high-order features which are highly beneficial for enhancing recommendation effectiveness. As GNNs evolve rapidly, researchers have attempted to apply GNNs to recommendation systems (Kipf and Welling 2016; Hu et al. 2019; Berg, Kipf, and Welling 2017; Ying et al. 2018; Wang et al. 2019). GCNs (Wu et al. 2022b; Mao et al. 2021b; Chen et al. 2020) have garnered significant attention among GNNs for their capability to capture high-order semantic information of users and items (Zhang et al. 2022; Wei et al. 2023). LightGCN (He et al. 2020) retains only the neighborhood aggregation part of GCN for collaborative filtering, simplifying the structure to better suit recommendation scenarios. MGCN proposes a behavior-aware fuser to adaptively learn the importance of different modal features, comprehensively modeling user preferences. Recently, some studies employ auxiliary graphs to enhance user or item feature representations (Sun et al. 2019; Li et al. 2022; Sun et al. 2020). For the user-item graph, FREEDOM introduces an edge pruning technique to remove noise caused by unintended interactions. DRAGON learns dual representations of items in both heterogeneous and homogeneous graphs and employs attention-concatenation fusion to combine multimodal features, achieving optimal performance. To perceive comprehensive user interests, LGMRec (Guo et al. 2024) creates local graph embeddings and global hypergraph embeddings to learn local topological relations and global dependencies.

Recommendation System Enhanced by LLMs

With the rapid development of LLMs, an increasing number of LLM-powered recommendation systems have been proposed. Benefiting from the strong semantic understanding capabilities of LLMs, these methods effectively address

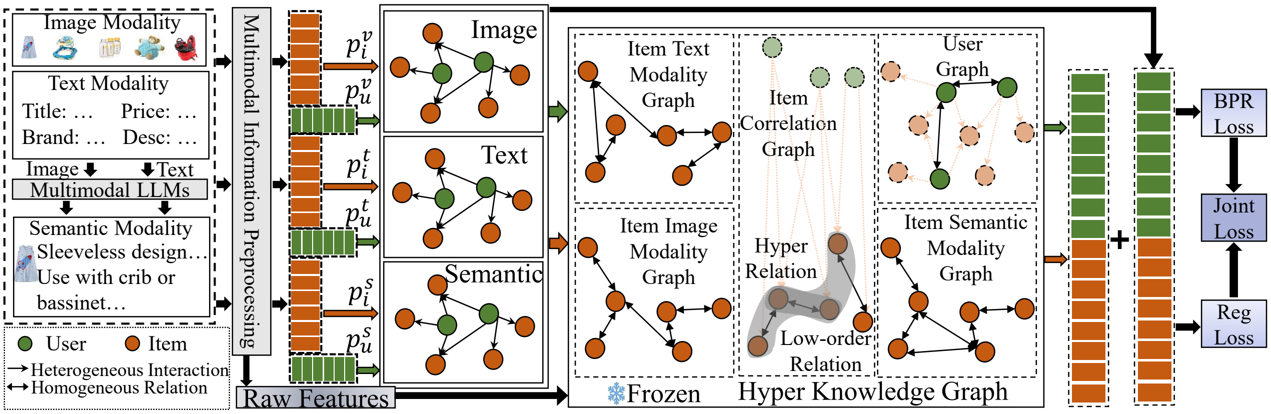


Figure 2: The overall framework of DOGE. Multimodal features represent different modal item information input into DOGE. User embeddings are randomly generated. The heterogeneous graph is used for feature propagation between users and items. The frozen Hyper Knowledge Graph is used for multi-dimensional propagation of user-user and item-item graphs.

various challenges. LLM4RS (Dai et al. 2023) applies ChatGPT directly to recommendation systems, evaluating the usability of large language models within such systems. LLMRank (Hou et al. 2024) assists LLMs in making recommendations by designing prompts and guiding strategies. When multiple generators retrieve ranking candidates, zero-shot LLMs challenge traditional recommendation models effectively. Additionally, some researchers leverage information generated by LLMs to enhance existing recommendation models. For instance, OpenGraph (Xia, Kao, and Huang 2024) introduces a data augmentation mechanism by incorporating LLMs to mitigate data sparsity issues. LLMRec (Wei et al. 2024) improves recommendation systems through three LLM-based enhancement strategies, including enhancing user-item interactions, user node information, and item node information. It also develops a denoising data mechanism to uphold the accuracy and reliability of augmented data. To address potential noise and bias from implicit feedback, the RLMRec (Ren et al. 2024) framework enhances representation learning through LLMs. It combines auxiliary text signals and cross-view alignment to improve the representation quality and robustness of recommendation systems.

Methodology

Problem Formulation

Assuming a given collection of users $\mathcal{U}$, containing n users $u \in \mathcal{U}$, and a collection of items $\mathcal{I}$, containing q items $i \in \mathcal{I}$. Each item i has multiple modalities $m \in \mathcal{M} = \{v, t\}$, where v represents the visual modality (i.e., images), and t represents the text modality (i.e., natural language text). We use $p_i^m \in \mathbb{R}^{d \times |\mathcal{I}|}$ and $r_i^m \in \mathbb{R}^{d_m \times |\mathcal{I}|}$ to denote the embedding and raw features of item i, and $p_u^m \in \mathbb{R}^{d \times |\mathcal{U}|}$ to represent the preference features of user u in modality m, where d and d_m denote the embedding dimension and raw features dimension, respectively. The historical behaviors of users can be represented as $\mathcal{R} \in [0, 5]^{|\mathcal{U}| \times |\mathcal{I}|}$, where each $\mathcal{R}_{u,i} = 0$ indicates no interaction between the user and item, otherwise it represents the satisfaction level of the user for that item. Thus, the interaction data can be represented as an interaction graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N} = \{\mathcal{U} \cup \mathcal{I}\}$ represents the node set, and historical interactions $\mathcal{R}_{u,i} \neq 0$ are considered as the set of edges $\mathcal{E}$ in the graph. In this paper, an additional semantic modality s generated by LLMs is also introduced, denoted as $m' \in \mathcal{M}' = \{v, t, s\}$.

Semantic Relationship Enhanced Feature Generation

To maximize the use of text and image features and reduce the potential risk of depending too heavily on text modality during homogeneous graph propagation, we propose semantic relationship enhanced features. We ingeniously design a prompt to describe the visual and textual features of the products to a multimodal LLM (Liu et al. 2024; Bavishi et al. 2023). Specifically, the product's image is first divided into multiple image patches, each representing a part of the overall information. These patches are projected onto the front part of the prompt to provide information about the products in the image. Then, semantic information is generated by the LLM:

$$s = SentenceTransformer(LLM(image, text)). \quad (1)$$

The prompt is designed as "[$\{image\}$: *Could you please introduce the characteristics of the* $\{$*task name*$\}$ *products in the image? And explain their potential uses and what items can be used together. The item title is* $\{$*item title*$\}$.]", where $\{\cdot\}$ denote different contents injected based on the task. The text information is combined with the image information by the prompt. Then, the low-information-density raw visual features of the items in the image are transformed into high-information-density semantic features through multi-layer transformers guided by the text information. These semantic features contain both textual and visual information, serving as an additional layer of information that significantly reinforces the interaction between textual and visual modalities.

Constructing Hyper-Knowledge Graph

To better enhance the relationships between items, we construct a frozen graph that enhances item relationships. We partly adopt the approach of FREEDOM (Zhou and Shen 2023), which calculates the similarity between the original features of each item i in each modality $m \in \mathcal{M}$ and constructs adjacency relationships with the top-K similar nodes, forming a modality-aware item-item similarity score graph $\mathcal{G}_m = \{\mathcal{I}, \mathcal{E}_m\}$, where $\mathcal{E}_m = \{c^m_{i,i'} | i, i' \in \mathcal{I}\}$. Specifically, for each pair of items (i, i'), we use cosine similarity to calculate the similarity score $c^m_{i,i'}$ between items i and i' in modality m. Similarly, for the semantic modality generated by the LLM, we also construct a semantic similarity graph $\mathcal{G}_s$ as:

$$c^m_{i,i'} = \frac{(r^m_i)^\top r^m_{i'}}{\|r^m_i\| \, \|r^m_{i'}\|}, \qquad c^s_{i,i'} = \frac{(r^s_i)^\top r^s_{i'}}{\|r^s_i\| \, \|r^s_{i'}\|}, \tag{2}$$

where $c^m_{i,i'}$ and $c^s_{i,i'}$ can be represented by $c^{m'}_{i,i'}$, which is seen as the weight of edges in a fully connected graph. To sparsify this graph, we retain the top-K similar items. Specifically, for each modality in m', if i' is among the top-K similar items, we retain the relationship between i and i' to preserve the foundational structure of the most related items:

$$c^{m'}_{i,i'} = \begin{cases} 1, & \text{if } c^{m'}_{i,i'} \in \text{top-K}(c^{m'}_i), \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

To capture the semantic relationships among frequently co-occurring items, we construct two homogeneous graphs: an item co-occurrence graph $\mathcal{G}_i = \{\mathcal{I}, \mathcal{E}_i\}$ and a hyper-relational graph $\mathcal{G}_{hi} = \{\mathcal{I}, \mathcal{E}_{hi}\}$. We begin by defining an item-item co-occurrence matrix A^{co}, in which each element $a^{co}_{i,i'}$ represents the frequency with which items i and i' have been jointly interacted with by the same user. To alleviate the influence of noise and retain salient relational information, we employ a top-K sparsification strategy. Specifically, for each item i, we identify the top-K items most frequently co-occurring with it, based on its corresponding row a^{co}_i:

$$\tilde{a}^{co}_{i,i'} = \begin{cases} 1, & \text{if } a^{co}_{i,i'} \in \text{top-K}(a^{co}_i), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

As a result, the edge set of the co-occurrence graph is given by $\mathcal{E}_i = \{(i, i') \mid i, i' \in \mathcal{I}, \tilde{a}^{co}_{i,i'} > 0\}$.

For the hyper-relational graph, we consider all items interacted with by a single user as forming a hyperedge. The set of hyperedges is denoted as $\mathcal{E}_{he} = \{e^{he}_1, e^{he}_2, \ldots, e^{he}_k\}$ containing k hyperedges, where each hyperedge e^{he} includes h items. In this work, we focus only on hyperedges where $h > 2$. For each subset of $\mathcal{E}_{he}$, we construct its power set:

$$\mathcal{H} = \{(e^p, w(e^p)) \mid e^p \in \mathcal{P}(e^{he}), \forall e^{he} \in \mathcal{E}_{he}\}, \tag{5}$$

where $\mathcal{P}(\cdot)$ represents the power set operation, w represents the frequency of e^p. We limit the subset size and frequency in the power set to avoid overly complex computations. To simplify the propagation process, we define a hypergraph matrix $\mathcal{H}^* \in \{0,1\}^{|\mathcal{I}| \times |\mathcal{H}|}$. If item i belongs to the j-th hyperedge e^p_j in $\mathcal{H}$, we set its element h^*_{ij} to 1. We directly obtain the adjacency matrix of the hyperedges as: $A^{hi} = \mathcal{H}^* \cdot \mathcal{H}^{*\top}$. Next, we sparsify this matrix as follows:

$$a^{hi}_{i,i'} = \begin{cases} 1, & \text{if } a^{hi}_{i,i'} \in \text{top-K}(a^{hi}_i) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Finally, we define the hyper-relational edge set as $\mathcal{E}_{hi} = \{(i, i') \mid i, i' \in \mathcal{I}, a^{hi}_{i,i'} > 0\}$.

As the item-item relationships are enhanced, the original two item relationship graphs are expanded into five types of item-item relationships $r \in R = \{v, t, s, i, hi\}$. To alleviate the problem of gradient explosion, we perform matrix normalization on the adjacency matrix $\mathcal{A}_r$ of each graph $\mathcal{G}_r$ as: $\widetilde{\mathcal{A}}_r = D^{-\frac{1}{2}} \mathcal{A}_r D^{-\frac{1}{2}}$, where D is the diagonal matrix of $\mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$. Based on the multi-item relationship graph obtained using the above method, we aggregate the propagation paths of each relationship graph through weighted summation to construct the overall item relationship adjacency matrix as: $\hat{\mathcal{A}_R} = \sum_{r \in R} \alpha_r \widetilde{\mathcal{A}}_r$.

We construct a user similarity graph $\mathcal{G}_u = \{\mathcal{U}, \mathcal{E}_u\}$, where $\mathcal{E}_u = \{c_{u,u'} | u, u' \in \mathcal{U}\}$ represents the edges between users u and u', and the edge values denote user similarity. The similarity is calculated as follows:

$$c_{u,u'} = \sum_{k=1}^{\mathcal{N}_{u,u'}} \frac{1}{1 + \left| r^k_u - r^k_{u'} \right|}, \tag{7}$$

where $\mathcal{N}_{u,u'}$ represents the number of items that users u and u' interact with together, and r^k_u represents the rating of user u for the k-th item. For each user, we retain the top-K users with the highest similarity $c_{u,u'}$:

$$c_{u,u'} = \begin{cases} c_{u,u'}, & \text{if } c_{u,u'} \in \text{top-K}(c_u), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Propagating on Homogeneous and Heterogeneous Graphs

To capture representations of users and items across multiple modalities in heterogeneous graphs, we adopt the representation learning approach of LightGCN (He et al. 2020) to train our user-item graph $\mathcal{G}$ as:

$$\begin{aligned} \left(p^{m'}_u\right)^{(k+1)} &= \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \left(p^{m'}_i\right)^{(k)}, \\ \left(p^{m'}_i\right)^{(k+1)} &= \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i|}\sqrt{|\mathcal{N}_u|}} \left(p^{m'}_u\right)^{(k)}, \end{aligned} \tag{9}$$

where $p^{m'}_u$ is randomly initialized, and $p^{m'}_i$ represents a multimodal feature based on the pre-trained model embedding. $\mathcal{N}_u$ and $\mathcal{N}_i$ represent the 1-hop neighbors of u and i respectively. After K layers of data propagation, the user and item feature representations are obtained as:

$$p^{m'}_u = \sum_{k=0}^{K} \left(p^{m'}_u\right)^{(k)}, \quad p^{m'}_i = \sum_{k=0}^{K} \left(p^{m'}_i\right)^{(k)}. \tag{10}$$

We adopt attention-based concatenation (Zhou et al. 2023b) to integrate the features learned from the heterogeneous graph:

$$u_{rep} = (W_{Att})^\top \odot [p^v_u : p^s_u : p^t_u], i_{rep} = [p^v_i : p^s_i : p^t_i], \tag{11}$$

where W_{Att} is a learnable parameter, $\odot$ represents element-wise multiplication.

In the homogeneous user-user graph, we utilize an attention method to aggregate the neighboring nodes of users and obtain f_u (Wang et al. 2021). For the homogeneous item-item graph, we employ an item feature propagation layer for feature propagation as:

$$f_i^{(k+1)} = \sum_{i' \in \mathcal{N}_i} \hat{\mathcal{A}}_R f_{i'}^k, \tag{12}$$

where $\mathcal{N}_i$ is the set of neighbors of item i, R represents the 5 types of item-item relationships mentioned above, and f_u^0, f_i^0 represent the fused results of users u_{rep} and items i_{rep}.

Then, we add the results propagated from the user-user and item-item homogeneous graphs to the results aggregated from the heterogeneous graph, obtaining the final user representation $u_{rep}^{'}$ and item representation $i_{rep}^{'}$:

$$u_{rep}^{'} = u_{rep} + f_u, \qquad i_{rep}^{'} = i_{rep} + f_i, \tag{13}$$

where f_u and f_i are the final results output from the user and item homogeneous graphs respectively.

Optimization

The model parameters are optimized using the Bayesian Personalized Ranking (BPR) loss (Rendle et al. 2012):

$$\mathcal{L} = \sum_{u,i^+,i^- \in D} \left(-\ln \sigma(z_{u,i^+} - z_{u,i^-})\right) + \beta \left\|\Theta\right\|_2^2, \tag{14}$$

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

Experiment Detail

Dataset

To evaluate the effectiveness of our proposed model, we perform extensive experiments on real-world Amazon datasets (McAuley et al. 2015). For a deeper analysis of the model's performance in handling larger datasets with sparser data, we select the Baby, Home and Kitchen, and Electronics datasets which we refer to as Baby, Kitchen and Electronics. For each dataset, we utilize the 5-core setting to ensure that each item or user is associated with at least 5 interactions. Following the data generation method in MMRec (Zhou 2023), We utilize pre-trained sentence-transformers (Reimers and Gurevych 2019) to convert text and semantic features into model-usable vectors, with a dimension of 384, and utilize the published original visual features with a dimension of 4096. The statistics of the data are summarized in the Table 1.

Evaluation Metrics

Following prior settings, we divide historical interactions into training, validation, and test sets using an 8:1:1 ratio. To assess top-K recommendation performance, we apply two standard evaluation metrics: R@k (Recall) and N@k (Normalized Discounted Cumulative Gain), where k is calculated for both 10 and 20, reporting the average results for users.

Dataset	Users	Items	Interactions	Sparsity
Baby	19445	7050	160792	99.88%
Kitchen	66519	28237	551682	99.97%
Electronics	192403	63001	1689188	99.98%

Table 1: Statistical overview of the datasets.

Baselines

To evaluate the effectiveness of our proposed model, we compare DOGE with other baselines. The first category is the most popular GCN-based method, and the second category comprises 12 multimodal recommendation models, including the current SOTA models.

i) General Model:

- LightGCN (He et al. 2020), a popular GCN-based method, simplifies GCN modules for recommendation.

ii) Multimedia Models:

- VBPR (He and McAuley 2016), a classic multimodal recommendation method, integrates visual and ID embeddings of each item for representation.
- MMGCN (Wei et al. 2019) fuses representations from multiple modalities of items for recommendation.
- DualGNN (Wang et al. 2021) build a user graph from the user-item interaction to enhance user representations.
- GRCN (Wei et al. 2020) enhances previous models by filtering out false-positive interactions.
- LATTICE (Zhang et al. 2021) establishes item-item graphs to enhance item representations.
- SLMRec (Tao et al. 2022) proposes three data augmentation methods and introduces self-supervised learning into multimodal recommendation.
- BM3 (Zhou et al. 2023c) perturbs representations through a dropout mechanism without requiring randomly sampled negative examples.
- MICRO (Zhang et al. 2022) extends LATTICE by learning item-item graphs from multimodal features.
- FREEDOM (Zhou and Shen 2023) denoises the user-item interaction, and makes the modality graph freezing.
- MGCN (Yu et al. 2023) purifies item modal features and adaptively learns the importance of different modalities.
- DRAGON (Zhou et al. 2023b) learns dual representations of users and items, and integrates multimodal features using attention concatenation.
- LGMRec (Guo et al. 2024) combines local graph processing and global hypergraph processing to improve recommendation accuracy and robustness.

Parameter Settings

We implement our proposed model using PyTorch within the MMRec (Zhou 2023) framework, setting the user and item embedding dimensions for all models to 64. Model parameters are initialized with the Xavier method (Glorot and Bengio 2010), employed the Adam optimizer (Kingma

Dataset	Baby				Kitchen				Electronics			
Metrics	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
LightGCN(SIGIR'20)	0.0479	0.0754	0.0257	0.0328	0.0315	0.0452	0.0173	0.0209	0.0363	0.054	0.0204	0.025
VBPR(AAAI'16)	0.0423	0.0663	0.0223	0.0284	0.0248	0.0367	0.014	0.017	0.0293	0.0458	0.0159	0.0202
MMGCN(MM'19)	0.0378	0.0615	0.02	0.0261	0.0172	0.0284	0.0086	0.0115	0.0213	0.0343	0.0112	0.0146
DualGNN(TMM'21)	0.0448	0.0716	0.024	0.0309	0.0299	0.0434	0.0165	0.02	0.0363	0.0541	0.0202	0.0248
GRCN(MM'20)	0.0539	0.0833	0.0288	0.0363	0.0349	0.0505	0.0195	0.0235	0.0349	0.0529	0.0195	0.0241
LATTICE(MM'21)	0.0547	0.085	0.0292	0.037	—	—	—	—	—	—	—	—
SLMRec(TMM'22)	0.054	0.081	0.0285	0.0357	0.0352	0.0515	0.0196	0.0238	0.0443	0.0651	0.0249	0.0303
BM3(WWW'23)	0.0564	0.0883	0.0301	0.0383	0.0312	0.0462	0.0173	0.0212	0.0437	0.0648	0.0247	0.0302
MICRO(TKDE'22)	0.0584	0.0929	0.0318	0.0407	—	—	—	—	—	—	—	—
FREEDOM(MM'23)	0.0627	0.0992	0.033	0.0424	0.04	0.0584	0.0225	0.0273	0.0396	0.0601	0.022	0.0273
MGCN(MM'23)	0.062	0.0964	0.0339	0.0427	0.0405	0.0593	0.023	0.0279	0.0433	0.0639	0.0242	0.0295
DRAGON(ECAI'23)	0.0662	0.1021	0.0345	0.0435	0.045	0.0655	0.0251	0.0304	0.045	0.0678	0.025	0.0309
LGMRec(AAAI'24)	0.0644	0.1002	0.0349	0.044	0.0425	0.0628	0.0239	0.0292	0.0440	0.0665	0.0244	0.0303
DOGE†(Our method)	**0.0715**	**0.1098**	**0.0390**	**0.0486**	**0.0463**	**0.0672**	**0.026**	**0.0315**	**0.0471**	**0.0701**	**0.0265**	**0.0325**
DOGE(Our method)	**0.0719**	**0.11**	**0.0391**	**0.0489**	**0.0465**	**0.0677**	**0.0264**	**0.0319**	**0.0481**	**0.0718**	**0.027**	**0.0331**
improv.	8.61%	7.71%	12.03%	11.13%	3.33%	3.35%	5.17%	4.93%	6.88%	5.89%	8%	7.11%

Table 2: Performance of DOGE and baseline models on three datasets, with the best results indicated in bold, second best results indicated with underline, and "improv."indicating the percentage improvement of DOGE over the best baseline model. "—"indicates the model cannot be trained on a single GeForce RTX 4090 24GB. † indicates that, like the baseline model, DOGE do not use rating information to construct the user graph but relies solely on implicit feedback data.

and Ba 2014), and set the batch size to 2048. We construct a hyperparameter grid for learning rate and regularization weight, with values ranging from {1e-1, 1e-2, 1e-3, 1e4} for both, resulting in 16 parameter combinations. Additionally, when the learning rate and regularization weight are determined to be optimal, we conduct a grid search for the importance score $\alpha_{r'}$ of non-text weights in the item-item graph, varying from 0.1 to 0.9 in increments of 0.1, where $r^{'} \in R^{'} = R - \{t\}$. We fix the layers of our heterogeneous graph at 2 and the layers of our homogeneous graph at 1. The top-K is set to 40 for the user graph, and for the item graph $\mathcal{G}_r$, K is set to 10. We establish 1000 epochs as the upper limit for training, employing early stopping after 20 epochs, driven by the R@20 measure. Model training is conducted using an RTX 4090 GPU equipped with 24GB of memory. We follow all settings and hyperparameter strategies from the baseline papers, striving to ensure that baseline models performed at their strongest.

Experiment Result

Overall Performance

We compare the currently SOTA methods with our proposed model, and the results are shown in Table 2. The following conclusions are drawn: (1) The introduction of additional effective information can significantly improve recommendation performance. Models such as GRCN, LATTICE, and the current SOTA multimodal methods show significant improvements over the most popular general model, LightGCN. However, these methods tend to overlook the semantic information expressed in the image modality. We introduce semantic modality features generated by LLMs by projecting image information into prompts and combining them with the text modality of products. This effectively enhances the text-visual modality relationship, enabling the effective expression of multimodal information, ultimately resulting in the best outcomes. (2) The HKG enables effective feature propagation between nodes. Compared to other models, DualGNN and DRAGON, which construct user auxiliary graphs, have shown superior performance. Therefore, we believe that mining and utilizing node relationships in homogeneous graphs can effectively enhance model performance. Based on this idea, we construct a hyper-knowledge graph. The results show that DOGE performs excellently on small, medium, and large datasets. Our method improves by an average of 9.9% over LGMRec and by an average of 7.2% over the optimal baseline method DRAGON on all datasets.

Ablation Study

We evaluate the contribution of each component to improving recommendation performance. Specifically, we decouple the DOGE model and conduct experiments by sequentially removing the semantic modality features and HKG components. We compare these variants with the strongest baseline, DRAGON, and set up two variants:

- DOGE$_{\text{w/o s}}$: We remove the semantic modality enhancement, and the item modality features consist of textual and visual parts.
- DOGE$_{\text{w/o HKG}}$: We remove the HKG. Items propagate based on the direction of similarity between text and image modalities, with a text-to-image weight ratio of 9:1.

Figure 3 records the R@20 and N@20 of these variants on three datasets, indicating that all components significantly

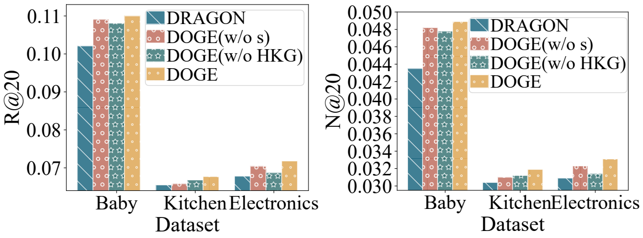


Figure 3: Performance evaluation of various DOGE model variants.

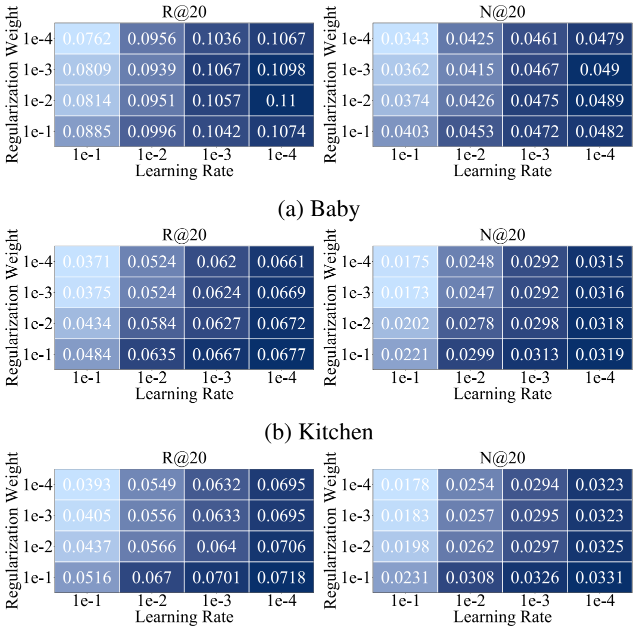
contribute to the final model. (1) In all datasets, the removal of either the semantic modality or HKG led to a drop in model performance. Thus, we believe both modules positively contribute to enhancing the model's performance. (2) The removal of HKG has the greatest impact on Electronics. This is because Electronics has a higher sparsity, and the new paths brought by item relationship enhancement can guide effective propagation of item features, increasing the affinity of features for semantically similar items and thereby improving recommendation effectiveness more significantly. (3) The fact that DOGE$_{\text{w/o s}}$ performs the worst among the variants indicates that the effect of semantic modality enhancement is most pronounced in the Kitchen dataset. We believe this is because the product features in the Kitchen dataset have lower data quality in the text modality, and the image modality struggles to effectively represent product features. Therefore, the addition of semantic modalities effectively enhances recommendation effectiveness.

Effect of Learning Rate and Regularization Weight

We set the regularization weights and learning rates for DOGE within the range of {1e-1, 1e-2, 1e-3, 1e-4}. Figure 4(a), Figure4(b), and Figure4(c) respectively show the model's performance under different learning rate and regularization configurations on the three datasets measured by R@20 and N@20. We observe that higher learning rates increase the model's sensitivity to regularization weights, whereas lower learning rates reduce this sensitivity, whereas when the learning rate is reduced, the sensitivity decreases. The model performs strongly across all datasets when the learning rate is set to 1e-4, and under the same learning rate, the performance variations among different regularization weights are relatively modest. The experimental results across different parameter configurations are smooth, confirming the strong recommendation performance of DOGE is not influenced by random processes during training.

Effect of Different Modal Proportions

To investigate the impact of different modalities on recommendation performance, we adjust the proportions of different modalities under the same conditions and observed their R@20 and N@20. The experimental results are shown in the Figure 5. In the figure, $\alpha_{r'}$ represents the sum of the proportion of non-text modalities in the propagation process of the homogeneous graph. It is observed that as $\alpha_{r'}$ increases, the two evaluation metrics in all three datasets initially increase and then decrease, reaching optimal performance at $\alpha_{r'} = \{0.4, 0.6\}$, which effectively highlights the increased contribution of the visual and semantic modality. Furthermore, during the decline, no precipitous drop occurred, and the model's performance remained very stable. This indicates that we effectively reduced the dependence of recommendation performance on text modality on the item homogeneous graph.



R@20

Regularization Weight	1e-1	1e-2	1e-3	1e-4
1e-4	0.0762	0.0956	0.1036	0.1067
1e-3	0.0809	0.0939	0.1067	0.1098
1e-2	0.0814	0.0951	0.1057	0.11
1e-1	0.0885	0.0996	0.1042	0.1074

N@20

Regularization Weight	1e-1	1e-2	1e-3	1e-4
1e-4	0.0343	0.0425	0.0461	0.0479
1e-3	0.0362	0.0415	0.0467	0.049
1e-2	0.0374	0.0426	0.0475	0.0489
1e-1	0.0403	0.0453	0.0472	0.0482

Learning Rate

(a) Baby

R@20

Regularization Weight	1e-1	1e-2	1e-3	1e-4
1e-4	0.0371	0.0524	0.062	0.0661
1e-3	0.0375	0.0524	0.0624	0.0669
1e-2	0.0434	0.0584	0.0627	0.0672
1e-1	0.0484	0.0635	0.0667	0.0677

N@20

Regularization Weight	1e-1	1e-2	1e-3	1e-4
1e-4	0.0175	0.0248	0.0292	0.0315
1e-3	0.0173	0.0247	0.0292	0.0316
1e-2	0.0202	0.0278	0.0298	0.0318
1e-1	0.0221	0.0299	0.0313	0.0319

Learning Rate

(b) Kitchen

R@20

Regularization Weight	1e-1	1e-2	1e-3	1e-4
1e-4	0.0393	0.0549	0.0632	0.0695
1e-3	0.0405	0.0556	0.0633	0.0695
1e-2	0.0437	0.0566	0.064	0.0706
1e-1	0.0516	0.067	0.0701	0.0718

N@20

Regularization Weight	1e-1	1e-2	1e-3	1e-4
1e-4	0.0178	0.0254	0.0294	0.0323
1e-3	0.0183	0.0257	0.0295	0.0323
1e-2	0.0198	0.0262	0.0297	0.0325
1e-1	0.0231	0.0308	0.0326	0.0331

Learning Rate

(c) Electronics

Figure 4: The performance of DOGE across different learning rates and regularization weights on the Baby, Kitchen, and Electronics datasets.

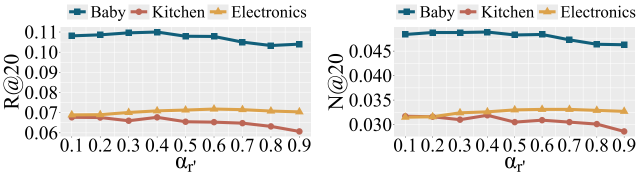


Figure 5: Relationship between $\alpha_{r'}$ ratio and DOGE performance in item homogeneous graphs.

Conclusion

We propose a DOGE recommender for multimodal recommendation. Specifically, we leverage a LLM to understand the image and text information of items, generating a semantic modality. The semantic modality strengthens the relationship between the image modality and the text modality. Then, we construct a HKG to explore the connections between nodes to address the issue of over-reliance on text modality in item homogeneous graphs present in many multimodal recommendation models. Finally, we utilize three datasets from Amazon to conduct extensive experiments.

Acknowledgments

Support for this project was provided by the National Key Research and Development Program of China (No.2021YFB3300700) and BUPT Excellent Ph.D. Students Foundation (CX20241001).

References

Bavishi, R.; Elsen, E.; Hawthorne, C.; Nye, M.; Odena, A.; Somani, A.; and Taşırlar, S. 2023. Introducing our Multimodal Models.

Berg, R. v. d.; Kipf, T. N.; and Welling, M. 2017. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*.

Chen, J.; Zhang, H.; He, X.; Nie, L.; Liu, W.; and Chua, T.-S. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, 335–344.

Chen, L.; Wu, L.; Hong, R.; Zhang, K.; and Wang, M. 2020. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 27–34.

Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; and Zha, H. 2019. 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*, 765–774.

Dai, S.; Shao, N.; Zhao, H.; Yu, W.; Si, Z.; Xu, C.; Sun, Z.; Zhang, X.; and Xu, J. 2023. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 1126–1132.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; and Ruan, B. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8454–8462.

He, R.; and McAuley, J. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 639–648.

Hou, Y.; Zhang, J.; Lin, Z.; Lu, H.; Xie, R.; McAuley, J.; and Zhao, W. X. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, 364–381. Springer.

Hu, F.; Zhu, Y.; Wu, S.; Wang, L.; and Tan, T. 2019. Hierarchical graph convolutional networks for semi-supervised node classification. *arXiv preprint arXiv:1902.06667*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, G.; Guo, Z.; Li, J.; and Wang, C. 2022. MDGCF: Multi-dependency graph collaborative filtering with neighborhood-and homogeneous-level dependencies. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 1094–1103.

Liu, F.; Cheng, Z.; Sun, C.; Wang, Y.; Nie, L.; and Kankanhalli, M. 2019. User diverse preference modeling by multimodal attentive metric learning. In *Proceedings of the 27th ACM international conference on multimedia*, 1526–1534.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, Q.; Wu, S.; and Wang, L. 2017. Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, 841–844.

Mao, K.; Zhu, J.; Wang, J.; Dai, Q.; Dong, Z.; Xiao, X.; and He, X. 2021a. SimpleX: A simple and strong baseline for collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1243–1252.

Mao, K.; Zhu, J.; Xiao, X.; Lu, B.; Wang, Z.; and He, X. 2021b. UltraGCN: ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 1253–1262.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43–52.

Papadakis, H.; Papagrigoriou, A.; Panagiotakis, C.; Kosmas, E.; and Fragopoulou, P. 2022. Collaborative filtering recommender systems taxonomy. *Knowledge and Information Systems*, 64(1): 35–74.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ren, X.; Wei, W.; Xia, L.; Su, L.; Cheng, S.; Wang, J.; Yin, D.; and Huang, C. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024*, 3464–3475.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Sun, J.; Zhang, Y.; Guo, W.; Guo, H.; Tang, R.; He, X.; Ma, C.; and Coates, M. 2020. Neighbor interaction aware graph convolution networks for recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 1289–1298.

Sun, J.; Zhang, Y.; Ma, C.; Coates, M.; Guo, H.; Tang, R.; and He, X. 2019. Multi-graph convolution collaborative filtering. In *2019 IEEE international conference on data mining (ICDM)*, 1306–1311. IEEE.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T.-S. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*.

Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25: 1074–1084.

Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T.-S. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, 165–174.

Wei, W.; Huang, C.; Xia, L.; and Zhang, C. 2023. Multimodal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, 790–800.

Wei, W.; Ren, X.; Tang, J.; Wang, Q.; Su, L.; Cheng, S.; Wang, J.; Yin, D.; and Huang, C. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 806–815.

Wei, Y.; Wang, X.; Li, Q.; Nie, L.; Li, Y.; Li, X.; and Chua, T.-S. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5382–5390.

Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T.-S. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*, 3541–3549.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, 1437–1445.

Wu, L.; He, X.; Wang, X.; Zhang, K.; and Wang, M. 2022a. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4425–4445.

Wu, S.; Sun, F.; Zhang, W.; Xie, X.; and Cui, B. 2022b. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5): 1–37.

Xia, L.; Kao, B.; and Huang, C. 2024. Opengraph: Towards open graph foundation models. *arXiv preprint arXiv:2403.01121*.

Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 974–983.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B.-K. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6576–6585.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, 3872–3880.

Zhang, J.; Zhu, Y.; Liu, Q.; Zhang, M.; Wu, S.; and Wang, L. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Zhou, H.; Zhou, X.; Zeng, Z.; Zhang, L.; and Shen, Z. 2023a. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473*.

Zhou, H.; Zhou, X.; Zhang, L.; and Shen, Z. 2023b. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. *arXiv preprint arXiv:2301.12097*.

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

Zhou, X.; and Shen, Z. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 935–943.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023c. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, 845–854.