Subspace-Aware Graph Construction and Contrastive Alignment for Multimodal Recommendation with Large Language Models

Haodong Li[1,2], Lianyong Qi[1,2*], Weiming Liu[3], Fan Wang[4], Chong Li[1,2], Shengye Pang[5], Wenwen Gong[6*], Yanwei Xu[7], Xiaoxiao Chi[8], Yang Zhang[9], Xiaokang Zhou[10,11]

[1]College of Computer Science and Technology, China University of Petroleum (East China), China
[2]Shandong Key Laboratory of Intelligent Oil and Gas Industrial Software, China
[3]ByteDance Inc., Singapore
[4]College of Computer Science and Technology, Zhejiang University, China
[5]School of Computer Engineering and Science, Shanghai University, China
[6]College of Information and Electrical Engineering, China Agricultural University, China
[7]School of Computer Science, Peking University, China
[8]School of Computing, Macquarie University, Australia
[9]Anuradha and Vikas Sinha Department of Data Science, University of North Texas, USA
[10]Faculty of Business and Data Science, Kansai University, Japan
[11]RIKEN Center for Advanced Intelligence Project, Japan
lhd_hfut@163.com, lianyongqi@upc.edu.cn, lwming95@gmail.com, fanwang97@zju.edu.cn, z23070165@s.upc.edu.cn, pangsy@shu.edu.cn, wenweng@cau.edu.cn, yanwei.xu@pku.edu.cn, xiaoxiao.chi@students.mq.edu.au, yang.zhang@unt.edu, zhou@kansai-u.ac.jp

Abstract

Multimedia content offers additional context for recommender systems to better understand user interests. Existing studies on multimodal recommendation primarily focus on constructing item-item semantic graphs. However, most of these methods capture only shallow semantic structures based on feature similarity and struggle to model more complex or cross-entity semantic relationships (e.g., user-item). Moreover, in these methods, collaborative signals often dominate and suppress semantic knowledge, which limits its role in representation learning. To address these issues, we propose SCALE, a novel framework that combines Subspace-aware graph Construction and contrastive Alignment for multimodaL recommendation with large languagE models. Specifically, we first use large language models and encoders to extract user and item features. Following the subspace clustering assumption, we apply the Orthogonal Matching Pursuit algorithm to mine complex semantic structures within the item-item, user-user, and user-item spaces, and integrate them into a unified semantic graph. We then perform graph convolution on both the semantic and interaction graphs, and aggregate the results for recommendation. Furthermore, contrastive losses are employed to enhance semantic fusion and alignment. Extensive experiments on five real-world datasets demonstrate that SCALE significantly outperforms state-of-the-art multimodal recommendation models, highlighting its effectiveness in modeling complex relationships and integrating semantic knowledge with collaborative signals.

*Corresponding authors

Introduction

The explosive growth of data in the Internet era has resulted in severe information overload, thereby prompting the development of personalized recommendation systems. With the increasing prevalence of multimedia content, multimodal recommendation systems (MRS) have attracted growing attention in recent years. By integrating heterogeneous information sources (e.g., text, images, audio), MRS can better capture user preferences and item characteristics, thus enabling more accurate and diverse recommendations.

Semantic knowledge from multimodal sources can effectively complement collaborative signals in user–item interactions, producing more expressive user and item representations (Zhou et al. 2023a; Bai et al. 2024; Liu et al. 2025). However, a significant semantic gap remains, as multimodal content is often diverse, high-dimensional, and noisy, which sharply contrasts with the structured, sparse, and high-quality nature of interaction records. Consequently, MRS face two key challenges: (1) **Challenge 1**: How to effectively extract comprehensive semantic knowledge from multimodal content? (2) **Challenge 2**: How to appropriately integrate this semantic knowledge with collaborative signals?

Inspired by graph neural networks (Kipf and Welling 2017; Hu et al. 2019), early collaborative filtering (CF) models represent user–item interactions as a bipartite graph (Wang et al. 2019; He et al. 2020), while some studies further construct item–item semantic graphs based on multimodal features (Zhang et al. 2021; Zhou and Shen 2023). Recently, approaches leverage large language models (LLMs) (Ren et al. 2024), social graphs (Wei et al. 2024), and user co-occurrence matrices (Wang et al. 2023) to model user–user relationships. However, most of these methods

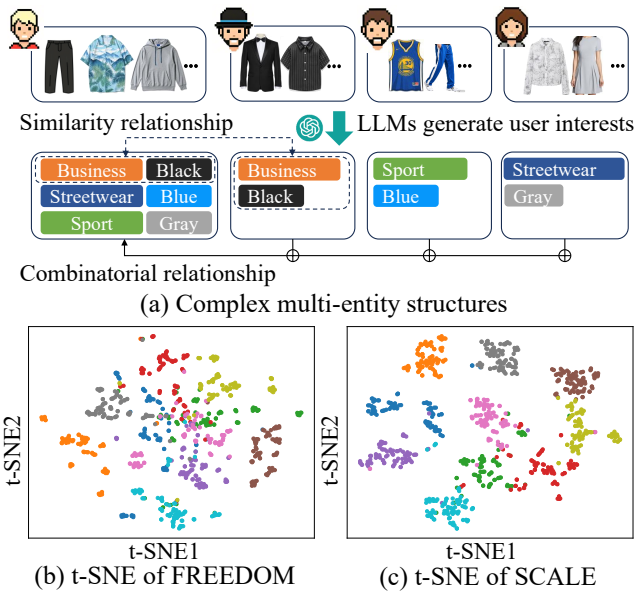


Figure 1: (a) Semantic relationships on user interests. (b)-(c) The t-SNE visualization of item representations from FREEDOM and SCALE, where each color denotes a distinct item category. For instance, green points correspond to baby toys.

build semantic graphs from multimodal features by selecting a fixed number of neighbors based on feature similarity, which primarily captures shallow, pairwise similarities within entities (i.e., user–user and item–item). This strategy may fails to capture complex multi-entity structures, such as the combinatorial user-interest patterns illustrated in Figure 1(a). Furthermore, cross-entity semantic relations (i.e., user–item) are often overlooked. Therefore, traditional MRS cannot effectively address **Challenge 1**.

In CF, items that are closer to a user in the representation space are more likely to be recommended (Rendle et al. 2009; Liu et al. 2021, 2022a). To achieve this, graph-based recommenders improve the proximity of potential interacting entities via neighborhood aggregation (Mao et al. 2021; Liu et al. 2022b). Beyond modeling interactions, MRS need to incorporate semantic knowledge into representation learning. Therefore, FREEDOM (Zhou and Shen 2023) performs graph convolution on item–item semantic graphs to bring semantically related entities closer, as shown in Figure 1(b). However, this implicit aggregation struggles to effectively integrate collaborative signals with semantic information, leading to overlaps among semantically unrelated nodes (e.g., green, pink, and red nodes in Figure 1(b)). As a result, existing methods are still insufficient for fully addressing **Challenge 2**.

To address the above challenges, we propose SCALE, a novel framework that combines subspace-aware graph construction and contrastive alignment for multimodal recommendation with large language models. For **Challenge 1**, we construct a semantic graph that captures comprehensive semantic relationships between users and items. First, we extract features from textual metadata using LLMs. Following the subspace clustering assumption (Elhamifar and Vidal 2013), we adopt the Orthogonal Matching Pursuit (OMP) algorithm (Mallat and Zhang 1993) to mine complex semantic structures within item-item, user-user, and user-item spaces. Moreover, we build feature similarity graphs and combine them with subspace-aware graphs to better capture comprehensive relationships. For **Challenge 2**, we independently model semantic knowledge and collaborative signals through graph convolution, then align and integrate their embeddings to obtain unified representations. Furthermore, we propose a relation alignment loss that pulls related entities closer and pushes unrelated ones apart in the representation space, thereby enhancing the influence of semantic information on the final representations, as illustrated in Figure 1(c).

Our contributions are summarized as follows:

- We propose a semantic graph construction method that captures comprehensive semantic relationships, including combinatorial structures among multiple entities.
- We propose SCALE, a novel multimodal recommendation framework that effectively extracts and aligns semantic and collaborative signals.
- Extensive experiments on multiple real-world datasets demonstrate the effectiveness and superiority of SCALE in multimodal recommendation.

Related Work

Multimodal Recommendation. The expansion of multimedia content facilitates refined user interest modeling in MRS (Zhou et al. 2023a). Early work typically incorporates multimodal content as side information in CF models (He and McAuley 2016b). Later studies apply graph convolutions and capture item-item relationships (Kim et al. 2022; Liu et al. 2022c; Zhou and Shen 2023). In parallel, other approaches explore user–user interest relationships by co-occurrence matrices (Wang et al. 2023) or social graphs (Wei et al. 2024). However, inconsistencies between social relationships and user interests limit the effectiveness of social graphs in recommendation tasks (Xiao et al. 2023). More importantly, these methods primarily capture intra-entity relationships (e.g., user–user, item–item), while neglecting cross-entity semantic relations (user–item).

Large Language Models for Recommendation. The success of LLMs across diverse domains has accelerated their adoption in recommendation research (Wu et al. 2024). Some approaches employ LLMs as recommenders to directly generate items of interest (Bao et al. 2023; Xu et al. 2024; Tennenholtz et al. 2024). However, the high computational cost of these methods poses a significant barrier to their widespread application in MRS (Zhou et al. 2025; Li et al. 2025). To address this, recent methods leverage LLMs to enhance user and item semantic representations in recommendation models (Liu et al. 2024; Xi et al. 2024; Zhang et al. 2024; Ren et al. 2024). However, while LLMs enable these methods to effectively extract multimodal information from raw data, they still face challenges in balancing and disentangling collaborative signals and semantic knowledge.

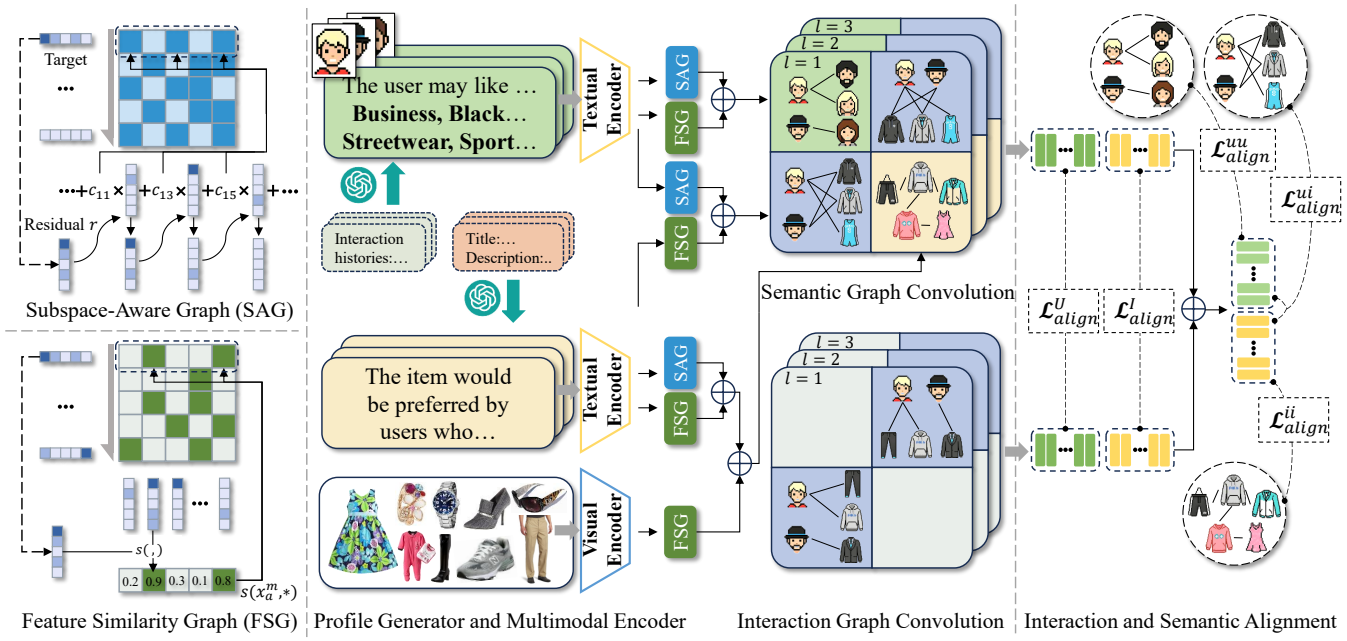


Figure 2: The overall framework of SCALE. LLMs and encoders extract features to build a semantic graph. Convolutions on semantic and interaction graphs are combined for recommendation, guided by contrastive losses.

Methodology

In this section, we elaborate on each component of SCALE. The overall framework is illustrated in Figure 2.

Preliminaries

Let $\mathcal{U}$ and $\mathcal{I}$ denote the user and item sets, with $|\mathcal{U}| = M$ and $|\mathcal{I}| = N$, respectively. The user-item interaction matrix is denoted as $\mathbf{R} \in \mathbb{R}^{M \times N}$, where $\mathbf{R}_{u,i} = 1$ if user $u \in \mathcal{U}$ has interacted with item $i \in \mathcal{I}$, and $\mathbf{R}_{u,i} = 0$ otherwise. Let $\mathcal{M}$ denote the set of modalities. For each modality $m \in \mathcal{M}$, we represent the feature vector of user u or item i as $\boldsymbol{x}_u^m, \boldsymbol{x}_i^m \in \mathbb{R}^{d_m}$, where d_m denotes the dimensionality for modality m. In this work, we consider two settings: one where $\mathcal{M}$ includes both visual (v) and textual (t) modalities, and another where only the textual modality is available.

Profile Generator and Multimodal Encoder

Leveraging the strong ability of LLMs to comprehend and summarize textual content (Ren et al. 2024; Ma, Ren, and Huang 2024), we construct informative item profiles:

$$F_i = LLMs(P_I, T_i), \tag{1}$$

where P_I is the predefined prompt and T_i denotes the textual metadata of item i (e.g., title and description). The prompt defines both format and instruction, enabling LLMs to generate profiles that highlight the item's most appealing characteristics. Beyond item profiles, we employ LLMs to construct user profiles based on interaction histories:

$$F_u = LLMs(P_U, \{\mathcal{I}_i : i \in \mathcal{N}_u\}), \tag{2}$$

where P_U is the predefined user prompt, $\mathcal{N}_u$ denotes the set of items that user u has interacted with, and $\mathcal{I}_i$ comprises the user reviews and metadata of item i. Like P_I, P_U is designed to guide LLMs in inferring user interests.

Moreover, we employ a text encoder (Izacard et al. 2022) to transform the user profile F_u and item profile F_i into representations $\boldsymbol{x}_u^t, \boldsymbol{x}_i^t \in \mathbb{R}^{d_t}$, and a visual encoder (He and McAuley 2016a) to convert the item image into $\boldsymbol{x}_i^v \in \mathbb{R}^{d_v}$.

Algorithm 1: Orthogonal Matching Pursuit algorithm

Require: Target signal $\boldsymbol{x} \in \mathbb{R}^d$, dictionary $\mathbf{D} \in \mathbb{R}^{d \times k_s}$, sparsity level k_o
Ensure: Sparse coefficient vector $\boldsymbol{c} \in \mathbb{R}^{k_s}$
1: Initialize residual $\boldsymbol{r} \leftarrow \boldsymbol{x}$, support set $S \leftarrow \emptyset$, $\boldsymbol{x} \leftarrow 0$
2: **for** $t = 1$ to k_o **do**
3: Select index $j^* = \arg\max_j |\langle \mathbf{D}_j, \boldsymbol{r} \rangle|$
4: Update support set: $S \leftarrow S \cup \{j^*\}$
5: Solve least squares: $\boldsymbol{c}_S = \arg\min_z \|\boldsymbol{x} - \mathbf{D}_S z\|_2$
6: Update residual: $\boldsymbol{r} \leftarrow \boldsymbol{x} - \mathbf{D}_S \boldsymbol{c}_S$
7: **end for**
8: Set $\boldsymbol{c}_j \leftarrow 1$ for $j \in S$, others remain 0
9: **return** $\boldsymbol{c}$

Semantic Graph Construction

Subspace-Aware Graph Construction. Subspace clustering assumes that data from multiple classes lie in a union of low-dimensional subspaces (Elhamifar and Vidal 2013). For example, facial images of the same person under different lighting conditions often reside in the same subspace. Each data point can thus be represented as a linear or affine combination of other points within that subspace. In MRS, we exploit this self-expressiveness property to construct a user-

user semantic graph. Specifically, we assume that users with similar interests lie in a local low-dimensional subspace. Consequently, each user's representation can be sparsely reconstructed from those of its neighbors:

$$\min \|\mathbf{c}_u\|_1 \quad \text{s.t.} \quad \boldsymbol{x}_u^t = \mathbf{D}_u \boldsymbol{c}_u, \tag{3}$$

where $\boldsymbol{x}_u^t$ denotes the textual feature of user u. To reduce computational cost, we construct the dictionary $\mathbf{D}_u \in \mathbb{R}^{d_t \times k_s}$ using the representations of the top-k_s users most similar to user u. The OMP algorithm is then employed to compute the sparse coefficient vector $\boldsymbol{c}_u \in \mathbb{R}^{k_s}$. OMP aims to approximate a target vector as a sparse combination of dictionary atoms (Mallat and Zhang 1993). At each iteration, the algorithm selects the atom most correlated with the current residual and updates the residual accordingly to progressively minimize the approximation error, as shown in Algorithm 1. In SCALE, we leverage this property to iteratively identify the user interest most aligned with the current residual, thereby uncovering representations that exhibit combinatorial relationships with the target, as shown in the subspace-aware graph (SAG) in Figure 2. Subsequently, we construct the user-user graph $\mathbf{C}^U \in \mathbb{R}^{M \times M}$ by concatenating the sparse coefficient vectors of all users. Specifically, $\mathbf{C}^U_{u,v} = 1$ if the coefficient corresponding to user v in $\boldsymbol{c}_u$ is non-zero; otherwise, $C^U_{u,v} = 0$.

Furthermore, items with similar attributes tend to reside in the same underlying subspace and can be linearly represented by one another. Therefore, we treat item features as both dictionary atoms and reconstruction targets, and apply the OMP algorithm to reconstruct each item, resulting in the item-item graph $\mathbf{C}^I \in \mathbb{R}^{N \times N}$. Similarly, we reconstruct user features from item features to build the user-item graph $\mathbf{C}^{UI} \in \mathbb{R}^{M \times N}$, offering complementary inter-entity semantics in addition to the user-item interactions.

Feature Similarity Graph Construction. Beyond the aforementioned subspace-aware graphs, which models complex multi-entity semantics, we introduce feature similarity graphs (FSG) to capture shallow pairwise relations using kNN sparsification (Chen, Fang, and Saad 2009). Specifically, the similarity score between entity $a \in A$ and entity $b \in B$ is computed via the cosine similarity of their modality-specific features $\boldsymbol{x}_a^m$ and $\boldsymbol{x}_b^m$:

$$s(\boldsymbol{x}_a^m, \boldsymbol{x}_b^m) = \frac{(\boldsymbol{x}_a^m)^{\mathrm{T}} \boldsymbol{x}_b^m}{\|\boldsymbol{x}_a^m\| \cdot \|\boldsymbol{x}_b^m\|}, \tag{4}$$

For each entity $a \in A$, we retain the top-k_f similar neighbors, resulting in a sparse adjacency matrix $\mathbf{G}^m$ defined as:

$$\mathbf{G}^m_{a,b} = \begin{cases} 1, & \text{if } \boldsymbol{x}_b^m \in \text{top-}k_f(\boldsymbol{s}(\boldsymbol{x}_a^m, *)) \\ 0, & \text{otherwise} \end{cases}, a \in A, b \in B, \tag{5}$$

where $\boldsymbol{s}(\boldsymbol{x}_a^m, *)$ denotes a vector of similarity scores between entity a and all entities in B, as illustrated in the FSG in Figure 2. By varying A, B, and the modality m, we construct the following graphs: (1) $\mathbf{G}_U$, the user-user similarity graph ($A = B = \mathcal{U}$, $m = t$); (2) $\mathbf{G}_I^v$, the item-item visual graph ($A = B = \mathcal{I}$, $m = v$); (3) $\mathbf{G}_I^t$, the item-item textual graph ($A = B = \mathcal{I}$, $m = t$); and (4) $\mathbf{G}_{UI}$, the user-item similarity graph ($A = \mathcal{U}$, $B = \mathcal{I}$, $m = t$). We then integrate these with the subspace-aware graphs to form a semantic graph that captures comprehensive semantic relations:

$$\hat{\mathbf{G}} = \begin{bmatrix} \mathbf{C}^U & \mathbf{C}^{UI} \\ (\mathbf{C}^{UI})^{\mathrm{T}} & \mathbf{C}^I \end{bmatrix} \oplus \begin{bmatrix} \mathbf{G}_U & \mathbf{G}_{UI} \\ \mathbf{G}_{UI}^{\mathrm{T}} & \mathbf{G}_I^v \oplus \mathbf{G}_I^t \end{bmatrix}, \tag{6}$$

where $\oplus$ denotes an element-wise logical intersection: $(\mathbf{C} \oplus \mathbf{G})_{i,j} = 1$ if both $\mathbf{C}_{i,j} \neq 0$ and $\mathbf{G}_{i,j} \neq 0$, and 0 otherwise.

Interaction and Semantic Graph Convolution

To capture high-order interactions and semantics, we separately perform graph convolutions (He et al. 2020) on the interaction and semantic graphs. Specifically, after applying edge pruning (Zhou and Shen 2023) on the interaction graph $\tilde{\mathbf{G}} = \begin{bmatrix} 0 & \mathbf{R} \\ \mathbf{R}^{\mathrm{T}} & 0 \end{bmatrix}$, we conduct graph convolution as follows:

$$\tilde{\mathbf{H}}^{(l+1)} = \tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{G}} \tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{H}}^{(l)}, \quad l \in [0, L_I), \tag{7}$$

where $\tilde{\mathbf{D}} \in \mathbb{R}^{(M+N) \times (M+N)}$ is the degree matrix of $\tilde{\mathbf{G}}$, and $\tilde{\mathbf{D}}_{i,i}$ indicates the number of nonzero elements in the i-th row. $\tilde{\mathbf{H}}^{(0)} = \mathbf{Z} \in \mathbb{R}^{(M+N) \times d}$ denotes a randomly initialized ID embedding matrix, with the first M rows for users and the remaining N for items. We average the outputs of all L_I graph convolution layers to obtain the interaction-based representation: $\tilde{\mathbf{H}} = \frac{1}{L_I+1} \sum_{l=0}^{L_I} \tilde{\mathbf{H}}^{(l)}$. Similarly, we perform L_S layers of graph convolutions on the semantic graph $\hat{\mathbf{G}}$, and define the semantic-based representation as the output of the final layer: $\hat{\mathbf{H}} = \hat{\mathbf{H}}^{(L_S)} \in \mathbb{R}^{(M+N) \times d}$.

Interaction and Semantic Alignment Module

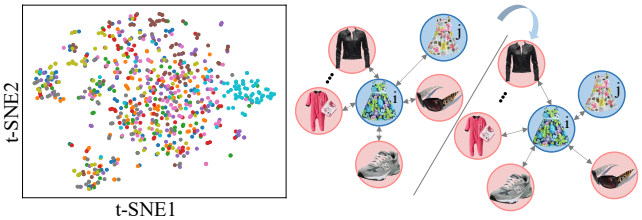
Semantic Alignment. To bridge the semantic gap between interaction and multimodal data, we preserve ID embeddings and independently perform graph convolutions on interaction and semantic graphs, thereby mitigating the noise and representational entanglement that may result from directly replacing ID embeddings with multimodal features. Moreover, we introduce a semantic alignment loss to facilitate the integration of collaborative and multimodal information. The user-side semantic alignment loss is defined as:

$$\mathcal{L}^U_{\text{align}} = -\sum_{u=1}^{M} \log \frac{\exp(s(\tilde{\boldsymbol{h}}_u, \hat{\boldsymbol{h}}_u)/\tau_s)}{\sum_{u'=1}^{M} \exp(s(\tilde{\boldsymbol{h}}_u, \hat{\boldsymbol{h}}_{u'})/\tau_s)}, \tag{8}$$

where τ_s is a temperature parameter, and $\tilde{\boldsymbol{h}}_u, \hat{\boldsymbol{h}}_u \in \mathbb{R}^d$ denote the u-th row of the interaction- and semantic-based representations $\tilde{\mathbf{H}}$ and $\hat{\mathbf{H}}$, respectively. The item-side semantic alignment loss $\mathcal{L}^I_{\text{align}}$ is defined similarly. To alleviate overfitting and reduce computational cost, we apply the contrastive loss only to a sampled subset of users and items during training. The overall semantic alignment loss is defined as:

$$\mathcal{L}_{\text{align}_s} = \mathcal{L}^U_{\text{align}} + \mathcal{L}^I_{\text{align}}. \tag{9}$$

The final representation is defined as $\mathbf{H} = \tilde{\mathbf{H}} + \hat{\mathbf{H}}$, where $\mathbf{H}_U \in \mathbb{R}^{M \times d}$ and $\mathbf{H}_I \in \mathbb{R}^{N \times d}$ denote the first M and the last N rows of $\mathbf{H}$, respectively. Let $\boldsymbol{h}_u$ and $\boldsymbol{h}_i$ represent the u-th row of $\mathbf{H}_U$ and the i-th row of $\mathbf{H}_I$.



(a) t-SNE of LightGCN (b) Aligned Item Representations

Figure 3: (a) The t-SNE visualization of different item category representations from LightGCN. (b) Comparison of item representations before and after relation alignment.

Relation Alignment. As shown in Figure 3(a), collaborative filtering models (e.g., LightGCN (He et al. 2020)) capture user–item interactions but struggle to incorporate semantic understanding. Although FREEDOM (Zhou and Shen 2023) leverages item-item semantic graphs to enhance semantic discrimination, its effect may be disturbed by dominant collaborative signals (see Figure 1(b)). To better integrate semantic knowledge into representation learning, we introduce a relation alignment loss, defined for item-item relations as:

$$\mathcal{L}_{\text{align}}^{ii} = -\sum_{(i,j)\in\mathcal{P}} \log \frac{\exp(s_{i,j}/\tau_r)}{\exp(s_{i,j}/\tau_r) + \sum_{v\in\mathcal{V}_i}\exp(s_{i,v}/\tau_r)}, \tag{10}$$

where $s_{i,j} \triangleq s(\boldsymbol{h}_i, \boldsymbol{h}_j)$, τ_r is a temperature parameter, and $\mathcal{P} = \{(i,j) \mid \mathbf{C}^I_{i,j} = 1\}$ denotes the set of positive item pairs in $\mathbf{C}^I$. $\mathcal{V}_i$ represents a randomly sampled set of negative items that satisfy $\mathbf{C}^I_{i,v} = 0$. Figure 3 (b) illustrates how $\mathcal{L}^{ii}_{\text{align}}$ influences item representation learning, explicitly pulling semantically related entities closer in the representation space. The relation alignment losses $\mathcal{L}^{uu}_{\text{align}}$ and $\mathcal{L}^{ui}_{\text{align}}$ are defined in a similar manner to $\mathcal{L}^{ii}_{\text{align}}$, with samples drawn from $\mathbf{C}^U$ and $\mathbf{C}^{UI}$, respectively. The overall relation alignment loss is formulated as:

$$\mathcal{L}_{\text{align}_r} = \mathcal{L}^{uu}_{\text{align}} + \mathcal{L}^{ii}_{\text{align}} + \mathcal{L}^{ui}_{\text{align}}. \tag{11}$$

Optimization

We optimize the model with the Bayesian Personalized Ranking (BPR) loss (Rendle et al. 2009), defined as:

$$\mathcal{L}_{\text{bpr}} = -\sum_{(u,i,j)\in\mathcal{D}} \log\left(\sigma(y_{u,i} - y_{u,j})\right), \tag{12}$$

where $\sigma(\cdot)$ is the Sigmoid function, and $y_{u,i} = \boldsymbol{h}_u^{\mathrm{T}}\boldsymbol{h}_i$ represents the predicted preference score of user u for item i. $\mathcal{D}$ denotes the training set, which consists of triplets (u,i,j). (u,i) is a positive pair satisfying $\mathbf{R}_{u,i} = 1$, and (u,j) is a negative pair with $\mathbf{R}_{u,j} = 0$, where item j is randomly sampled from the non-interacted items of user u. The BPR loss promotes higher predicted scores for positive user–item pairs compared to negative ones. The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{bpr}} + \lambda_s\mathcal{L}_{\text{align}_s} + \lambda_r\mathcal{L}_{\text{align}_r}, \tag{13}$$

where λ_s and λ_r denote the weights of the semantic alignment loss and relation alignment loss.

Experiment

In this section, we conduct comprehensive experiments on several real-world datasets to answer the following questions: (1) **RQ1**: How does SCALE perform in comparison with state-of-the-art multimodal recommendation models? (2) **RQ2**: How does SCALE perform compared to other LLM-enhanced recommendation models? (3) **RQ3**: How do different components of SCALE contribute to the overall performance? (4) **RQ4**: How does SCALE perform under different hyperparameter settings?

Experimental Settings

Datasets. We conduct experiments on five widely used real-world datasets, including four Amazon (McAuley et al. 2015) subsets (Baby, Sports, Clothing, and Book) and Yelp. On Baby, Sports, and Clothing, we compare the performance of SCALE with other multimodal recommendation models. These datasets provide both textual descriptions and images for each item. We employ a pre-trained encoder (Izacard et al. 2022) to extract textual representations and adopt 4096-dimensional image features (He and McAuley 2016a) as visual inputs. Moreover, we evaluate SCALE against LLM-enhanced recommendation models on Book and Yelp.
Baselines. We compare our proposed SCALE with representative SOTA models, including traditional recommendation models (BPR (Rendle et al. 2009), LightGCN (He et al. 2020), and LayerGCN (Zhou et al. 2023b)), multimodal recommendation models (VBPR (He and McAuley 2016b), LATTICE (Zhang et al. 2021), SLMRec (Tao et al. 2023), BM3 (Zhou et al. 2023c), FREEDOM (Zhou and Shen 2023), MGCN (Yu et al. 2023), LGMRec (Guo et al. 2024), DiffMM (Jiang et al. 2024), and MENTOR (Xu et al. 2025)), and LLM-enhanced recommendation models (KAR (Xi et al. 2024) and RLMRec (Ren et al. 2024)).
Evaluation Protocols. Following (Zhang et al. 2021; Zhou and Shen 2023), we randomly split the interaction data into training, validation, and testing sets with a ratio of 8:1:1. We adopt two widely used ranking metrics, Recall@H (R@H) and NDCG@H (N@H), and evaluate model performance on the testing set with $H = 10$ and 20.
Implementation Details. We employ Llama-3.2-3B to generate user and item profiles. The dimensions of the textual and visual features, d_t and d_v, are 768 and 4096, respectively. We set the batch size to 2048 and the learning rate to 1e−3, using the Adam optimizer (Kingma and Ba 2015). User and item ID embeddings are initialized with the Xavier method (Glorot and Bengio 2010), with the embedding dimension d set to 64. The loss weights λ_s and λ_r are set to 1e−2 and 5e−2, respectively. The negative sample size $|\mathcal{V}|$ in the relation alignment loss is set to 1. The temperature parameters τ_s and τ_r are set to 0.6 and 0.2, respectively. The sparsity level k_o is selected from $\{2, 5, 10, 12, 15\}$. k_s and k_f are set to 800 and 5, respectively. The edge pruning ratio ρ is set to 0.6. The number of graph convolution layers for the interaction graph (L_I) and the semantic graph (L_S) is selected from 1 to 5. The optimal hyperparameters are determined via grid search. SCALE is implemented based on the MMRec framework (Zhou et al. 2023a) and adopts an early stopping strategy based on Recall@20 on the validation set.

Model	Baby				Sport				Clothing			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BPR	0.0379	0.0607	0.0202	0.0261	0.0452	0.0690	0.0252	0.0314	0.0211	0.0315	0.0118	0.0144
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0361	0.0544	0.0197	0.0243
LayerGCN	0.0529	0.0820	0.0281	0.0355	0.0594	0.0916	0.0323	0.0406	0.0371	0.0566	0.0200	0.0247
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
LATTICE	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
SLMRec	0.0529	0.0775	0.0290	0.0353	0.0663	0.0990	0.0365	0.0450	0.0452	0.0675	0.0247	0.0303
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
FREEDOM	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0629	0.0941	0.0341	0.0420
MGCN	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0641	0.0945	0.0347	0.0428
LGMRec	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0555	0.0828	0.0302	0.0371
DiffMM	0.0623	0.0975	0.0328	0.0411	0.0671	0.1017	0.0377	0.0458	0.0522	0.0791	0.0288	0.0354
MENTOR	0.0678	0.1048	0.0362	0.0450	0.0763	0.1139	0.0409	0.0511	0.0668	0.0989	0.0360	0.0441
SCALE	**0.0965**	**0.1341**	**0.0550**	**0.0646**	**0.0874**	**0.1257**	**0.0494**	**0.0593**	**0.0713**	**0.1065**	**0.0393**	**0.0482**

Table 1: Performance comparison with baselines on three datasets. The best results are bolded and the second best results are underlined. Improv. indicates the improvement of SCALE on second best results.

Backbone	Model	Book		Yelp	
		R@20	N@20	R@20	N@20
LightGCN	KAR	0.1416	0.0863	0.1194	0.0756
	RLMRec	0.1483	0.0903	0.1230	0.0776
SGL	KAR	0.1372	0.0875	0.1208	0.0761
	RLMRec	0.1537	0.0947	0.1263	0.0798
LightGCN	SCALE	**0.1624**	**0.1004**	**0.1280**	**0.0801**

Table 2: Performance comparison with LLM-enhanced CF models on the Book and Yelp datasets.

Performance Comparison

Multimodal Recommendation Models (RQ1). Table 1 reports a comparison between SCALE and other recommendation models on multiple datasets. We observe that:

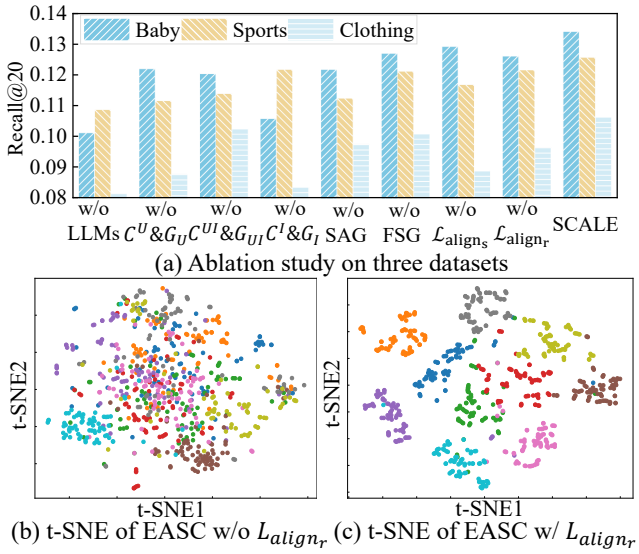
(1) SCALE consistently outperforms all baselines by a large margin across all metrics on the Baby, Sports, and Clothing datasets. Specifically, it achieves improvements ranging from 7.68% to 27.96% on Recall@20 compared to the strongest baselines, demonstrating the effectiveness of our proposed method. We attribute this superior performance to three key factors: accurate LLM-generated profiles, comprehensive subspace-aware graph construction, and effective contrastive alignment.

(2) SCALE achieves greater performance improvement on the Baby dataset compared to the other datasets. This may be attributed to the higher semantic density of the Baby dataset. In contrast, datasets such as Clothing suffer from a high rate of missing descriptions (93.83%) and severe textual redundancy (94.10%). SCALE relies heavily on multimodal features, especially textual ones, for both graph construction and relation alignment. Thus, richer textual input contributes to more informative LLM-generated profiles and better model performance.

LLM-enhanced CF Models (RQ2). To further evaluate the effectiveness of SCALE, we compare it with several LLM-enhanced CF models, with results presented in Table 2. For a fair comparison, all models utilize the same LLM-generated content and textual representations. The results show that SCALE consistently outperforms the other methods across all evaluation metrics. We attribute this superiority to the following reasons. KAR (Xi et al. 2024) directly incorporates textual features into user interest learning, while RLMRec (Ren et al. 2024) encodes user and item profiles derived from LLMs and aligns them with ID-based representations. However, due to the semantic gap between textual and ID-based representations, such fusion may introduce irrelevant semantic information into the final representations, thereby reducing their clarity and weakening the collaborative signals. In contrast, SCALE aligns ID-based representations through contrastive views or graph structures, without directly incorporating textual features at any stage. This design ensures that semantic knowledge and collaborative signals remain distinct in representations, thereby preserving their respective contributions.

Ablation Study (RQ3)

In this section, we evaluate the contribution of each component of SCALE to the overall performance. Figure 4(a) presents the results on the Baby dataset when each component is ablated individually. We observe that all components contribute positively to the model's effectiveness, especially the LLMs module. Specifically, removing the semantic or relation alignment losses leads to performance drops, underscoring the importance of aligning semantic knowledge with collaborative signals. Similarly, removing the subspace-aware graph or feature similarity graph degrades performance, confirming their complementary roles in capturing Comprehensive semantic relations. Moreover, excluding inter-entity (e.g., user-item) or intra-entity (e.g., user-user, item-item) relations from the semantic graph also results in performance degradation, highlighting the necessity of multi-type relation modeling in SCALE.



(a) Ablation study on three datasets

(b) t-SNE of EASC w/o L_{align_r} (c) t-SNE of EASC w/ L_{align_r}

Figure 4: (a) The effect of each component. (b)-(c) The t-SNE visualization of SCALE without/with $\mathcal{L}_{\text{align}_r}$.

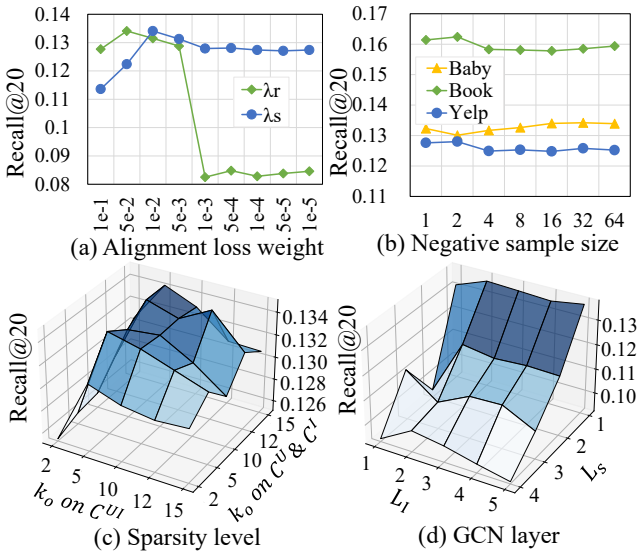


Figure 5: SCALE performance across hyperparameters.

To validate the effectiveness of the relation alignment loss $\mathcal{L}_{\text{align}_r}$, we visualize the distribution of different item category representations learned by SCALE. Figures 4(b) and 4(c) present the t-SNE visualization without and with $\mathcal{L}_{\text{align}_r}$, respectively. As shown in Figure 4(b), items within the same category remain relatively cohesive even without relation alignment, which can be attributed to semantic graph convolution, as it aggregates representations of semantically related items. Figure 4(c) shows that applying the relation alignment loss leads to clearer inter-group separation. This suggests that the mechanism not only brings semantically related items closer but also enhances the distinction between unrelated ones, thereby effectively incorporating semantic information into the final representations.

Sensitivity Analysis (RQ4)

Alignment Loss Weight. We vary λ_s and λ_r from $1\text{e}{-5}$ to 0.1 and report the Recall@20 results on the Baby dataset in Figure 5(a). We observe that SCALE benefits from appropriate loss weights, while performance degrades when either λ_s or λ_r is too small or too large. These findings highlight the importance of balancing semantic knowledge and collaborative signals, as overemphasizing either can degrade recommendation performance.

Negative Sample Size. Figure 5(b) illustrates the performance of SCALE on three datasets with varying negative sample sizes $|\mathcal{V}|$ in the relation alignment loss. We observe that the performance of SCALE is insensitive to $|\mathcal{V}|$, while more negative samples greatly increase memory cost and training difficulty. This is likely because randomly sampled negatives are not guaranteed to be semantically or interactively unrelated to the target, thereby diminishing the effectiveness of negative sampling in optimization.

Sparsity Level. Figure 5(c) shows the results on the Baby dataset when varying the sparsity level k_o across different semantic graphs ($\mathbf{C}^U$, $\mathbf{C}^I$, and $\mathbf{C}^{UI}$). We observe that SCALE achieves its best performance when the sparsity level of each graph is set to 10. We hypothesize that a small k_o may fail to capture certain meaningful combinatorial relationships, thereby hindering the model from learning comprehensive representations. In contrast, an overly large k_o may bring in irrelevant entities into the OMP-based graph, resulting in increased noise and decreased performance.

Number of Graph Convolutional Layer. As illustrated in Figure 5(d), we evaluate the performance of SCALE with different combinations of L_I and L_S on the Baby dataset. We observe a decline in performance as L_S increases, indicating that modeling higher-order relations on the semantic graph is ineffective. This may be because even a few irrelevant entities are gradually propagated through successive layers of graph convolution, leading to an amplified impact of noise. Moreover, SCALE performs best when L_I is set to 2 or 5, indicating that multi-hop message passing effectively enhances recommendation diversity and personalization.

Conclusion

In this paper, we propose a novel multimodal recommendation model SCALE, which can capture comprehensive semantic relationships and effectively integrate semantic and collaborative information. Specifically, we apply the OMP algorithm to mine complex semantic structures within user-user, item-item, and user-item spaces, and integrate them into a unified semantic graph. We then perform graph convolution on both interaction and semantic graphs, and introduce a semantic alignment loss to encourage consistency between their outputs. Furthermore, we propose a relation alignment loss that explicitly pulls semantically related entities closer in the representation space, thereby reinforcing the role of semantic knowledge in representation learning. Extensive experiments on five real-world datasets validate the effectiveness of our proposed method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62572486, No. 62506364), Natural Science Foundation of Shandong Province (No. ZR2023MF007).

References

Bai, H.; Wu, L.; Hou, M.; Cai, M.; He, Z.; Zhou, Y.; Hong, R.; and Wang, M. 2024. Multimodality Invariant Learning for Multimedia-Based New Item Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, 677–686. ACM.

Bao, K.; Zhang, J.; Zhang, Y.; Wang, W.; Feng, F.; and He, X. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, 1007–1014.

Chen, J.; Fang, H.; and Saad, Y. 2009. Fast Approximate *k*NN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *Journal of Machine Learning Research*, 10: 1989–2012.

Elhamifar, E.; and Vidal, R. 2013. Sparse Subspace Clustering: Algorithm, Theory, and Applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11): 2765–2781.

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*, volume 9 of *JMLR Proceedings*, 249–256. JMLR.org.

Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; and Ruan, B. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *AAAI*, 8454–8462. AAAI Press.

He, R.; and McAuley, J. J. 2016a. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, 507–517. ACM.

He, R.; and McAuley, J. J. 2016b. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, 144–150.

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, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, 639–648.

Hu, F.; Zhu, Y.; Wu, S.; Wang, L.; and Tan, T. 2019. Hierarchical Graph Convolutional Networks for Semi-supervised Node Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 4532–4539. ijcai.org.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Trans. Mach. Learn. Res.*, 2022.

Jiang, Y.; Xia, L.; Wei, W.; Luo, D.; Lin, K.; and Huang, C. 2024. DiffMM: Multi-Modal Diffusion Model for Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 7591–7599. ACM.

Kim, T.; Lee, Y.; Shin, K.; and Kim, S. 2022. MARIO: Modality-Aware Attention and Modality-Preserving Decoders for Multimedia Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, 993–1002. ACM.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, oulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Li, H.; Qi, L.; Liu, W.; Xu, X.; Dou, W.; Cao, Y.; Zhang, X.; Beheshti, A.; and Zhou, X. 2025. Balancing User-Item Structure and Interaction with Large Language Models and Optimal Transport for Multimedia Recommendation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, 3027–3035. ijcai.org.

Liu, Q.; Wu, X.; Zhao, X.; Wang, Y.; Zhang, Z.; Tian, F.; and Zheng, Y. 2024. Large Language Models Enhanced Sequential Recommendation for Long-tail User and Item. *CoRR*, abs/2405.20646.

Liu, W.; Chen, C.; Xu, J.; Liao, X.; Wang, F.; Zheng, X.; Fu, Z.; Pei, R.; and Wang, J. 2025. Joint Similarity Item Exploration and Overlapped User Guidance for Multi-Modal Cross-Domain Recommendation. In *WWW*, 2882–2893. ACM.

Liu, W.; Su, J.; Chen, C.; and Zheng, X. 2021. Leveraging distribution alignment via stein path for cross-domain cold-start recommendation. *Advances in Neural Information Processing Systems*, 34: 19223–19234.

Liu, W.; Zheng, X.; Hu, M.; and Chen, C. 2022a. Collaborative Filtering with Attribution Alignment for Review-based Non-overlapped Cross Domain Recommendation. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 1181–1190. ACM.

Liu, W.; Zheng, X.; Hu, M.; and Chen, C. 2022b. Collaborative filtering with attribution alignment for review-based non-overlapped cross domain recommendation. In *Proceedings of the ACM web conference 2022*, 1181–1190.

Liu, W.; Zheng, X.; Su, J.; Hu, M.; Tan, Y.; and Chen, C. 2022c. Exploiting Variational Domain-Invariant User Em-

bedding for Partially Overlapped Cross Domain Recommendation. In *SIGIR*, 312–321. ACM.

Ma, Q.; Ren, X.; and Huang, C. 2024. XRec: Large Language Models for Explainable Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, 391–402.

Mallat, S.; and Zhang, Z. 1993. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12): 3397–3415.

Mao, K.; Zhu, J.; Xiao, X.; Lu, B.; Wang, Z.; and He, X. 2021. UltraGCN: Ultra Simplification of Graph Convolutional Networks for Recommendation. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, 1253–1262. ACM.

McAuley, J. 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, Santiago, Chile, August 9-13, 2015*, 43–52. ACM.

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, WWW 2024, Singapore, May 13-17, 2024*, 3464–3475. ACM.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, 452–461.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T. 2023. Self-Supervised Learning for Multimedia Recommendation. *IEEE Trans. Multim.*, 25: 5107–5116.

Tennenholtz, G.; Chow, Y.; Hsu, C.; Jeong, J.; Shani, L.; Tulepbergenov, A.; Ramachandran, D.; Mladenov, M.; and Boutilier, C. 2024. Demystifying Embedding Spaces using Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2023. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Trans. Multim.*, 25: 1074–1084.

Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, rance, July 21-25, 2019*, 165–174. ACM.

Wei, C.; Hu, C.; Wang, C.; and Huang, S. 2024. Time-Aware Multibehavior Contrastive Learning for Social Recommendation. *IEEE Trans. Ind. Informatics*, 20(4): 6424–6435.

Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Gu, H.; Shen, T.; Qin, C.; Zhu, C.; Zhu, H.; Liu, Q.; Xiong, H.; and Chen, E. 2024. A survey on large language models for recommendation. *World Wide Web (WWW)*, 27(5): 60.

Xi, Y.; Liu, W.; Lin, J.; Cai, X.; Zhu, H.; Zhu, J.; Chen, B.; Tang, R.; Zhang, W.; and Yu, Y. 2024. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, 12–22. ACM.

Xiao, X.; Dai, H.; Dong, Q.; Niu, S.; Liu, Y.; and Liu, P. 2023. Incorporating Social-Aware User Preference for Video Recommendation. In *WISE*, volume 14306 of *Lecture Notes in Computer Science*, 544–558. Springer.

Xu, J.; Chen, Z.; Yang, S.; Li, J.; Wang, H.; and Ngai, E. C. H. 2025. MENTOR: Multi-level Self-supervised Learning for Multimodal Recommendation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 12908–12917. AAAI Press.

Xu, L.; Zhang, J.; Li, B.; Wang, J.; Cai, M.; Zhao, W. X.; and Wen, J. 2024. Prompting Large Language Models for Recommender Systems: A Comprehensive Framework and Empirical Analysis. *CoRR*, abs/2401.04997.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B. 2023. Multi-View Graph Convolutional Network for Multimedia Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 6576–6585.

Zhang, A.; Chen, Y.; Sheng, L.; Wang, X.; and Chua, T. 2024. On Generative Agents in Recommendation. In *SIGIR*, 1807–1817. ACM.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining Latent Structures for Multimedia Recommendation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 3872–3880. ACM.

Zhou, H.; Zhou, X.; Zeng, Z.; Zhang, L.; and Shen, Z. 2023a. A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions. *CoRR*, abs/2302.04473.

Zhou, P.; Liu, C.; Ren, J.; Zhou, X.; Xie, Y.; Cao, M.; Rao, Z.; Huang, Y.; Chong, D.; Liu, J.; Kim, J. B.; Wang, S.; Wong, R. C.; and Kim, S. 2025. When Large Vision Language Models Meet Multimodal Sequential Recommendation: An Empirical Study. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, 275–292. ACM.

Zhou, X.; Lin, D.; Liu, Y.; and Miao, C. 2023b. Layer-refined Graph Convolutional Networks for Recommendation. In *39th IEEE International Conference on Data Engineering, ICDE 2023*, 1247–1259.

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, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 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, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, 845–854.