LLM-Summarized Interest Distillation for Multimedia Recommendation

Meng Jian[1] , **Zhuoyang Xia**[1] , **Haolun Fan**[1,2] and **Lifang Wu**[1*]
[1]Beijing University of Technology
[2]Inner Mongolia Transportation Design and Research Institute CO., LTD
jianmeng648@163.com, xiazhuoyang2024@163.com, xingyue52077@outlook.com, lfwu@bjut.edu.cn

Abstract

Recommendation system technology faces the fundamental challenge of data sparsity. Although incorporating auxiliary textual data works to enrich interactions, lengthy and unstructured item descriptions often fail to depict user interests, as they primarily describe item attributes. Furthermore, the sparse interactions provide incomplete interest patterns, leading to biased recommendations. To address these issues, this work proposes an LLM-summarized interest distillation (SID) model to enrich interactions by interest summarization and retrieval augmentation. Interest summarization leverages large language model to improve user interest descriptions and filters out interest-irrelevant details, generating compact, user-centric interest texts. Retrieval augmentation utilizes semantic neighbors as multiple teachers to adaptively transfer their multimodal collaborative knowledge to augment behaviorally sparse student users, thereby enriching interest learning for recommendation. It enhances interactions both qualitatively, through improved interest descriptions, and quantitatively, through expanded semantic neighbors. Extensive experiments show the superior performance of SID over the state-of-the-art models, demonstrating its effectiveness of interest summarization and retrieval augmentation to improve recommendations.

1 Introduction

Due to e-commerce and various online services, the rapid growth of digital multimedia data increases the need for personalized services. Identifying user-preferred content from vast data effectively is a crucial topic in recent years. Collaborative filtering, at the core of recommendation systems, predicts user preferences by mining historical user-item interactions. However, data sparsity, particularly with long-tail items [Yin *et al.*, 2012] and users with limited activity records [Idrissi and Zellou, 2020], poses a significant challenge in modeling interests based solely on interactions. Aiming at enhancing recommendations, recent studies have incorporated auxiliary textual data to enrich user and item representations. Despite their enhancements, inadequate exploration of semantic signals also limits the understanding of users' personalized interests.

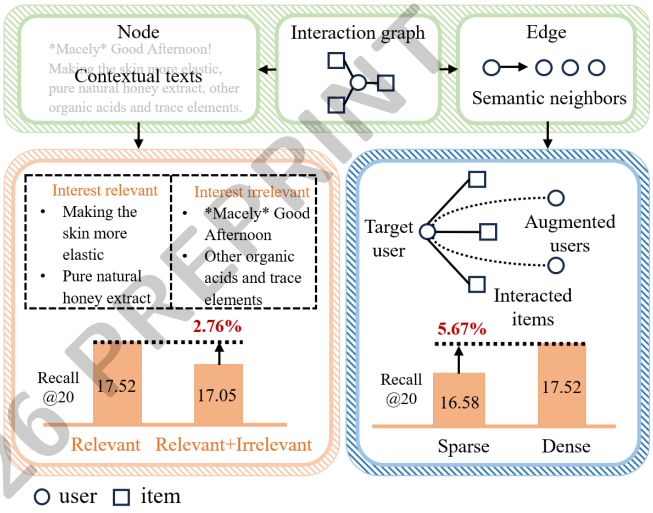

Figure 1: The motivation of data enhancement with contextual texts and semantic neighbors.

Pioneering studies [Wang *et al.*, 2024] leverage contextual signals from interactions by encoding textual data of items and aggregating them to represent user interests. However, empirical findings indicate that lengthy and unstructured text sequences struggle to depict user interests as they primarily describe items' attributes rather than user interests. On this point, we conduct an empirical study in Figure 1 illustrating that interest-irrelevant descriptions of items lead interest modeling to deviate from downstream recommendation objectives, thereby hindering recommendation performance. This phenomenon inspires a systematic manner to remove redundant data and extract interest-relevant signals when depicting users' interests through item description texts, thereby enhancing the description of interactions. Additionally, the structural neighbors of the sparse interaction graph characterize incomplete interest patterns, leading to biased recommendations. Most recent works have turned to data augmentation [Yu *et al.*, 2024; Wu *et al.*, 2024] for alleviating the data sparsity issue by incorporating task-relevant semantic neigh-

*Corresponding author.

bors. As the empirical study in Figure 1, data augmentation broadens the data landscape and better infers the interests of sparse users, enhancing the recommendation performance. This result suggests exploring an effective data augmentation method to provide more collaborative signals and enhance interest modeling for recommendations.

For interest modeling, this work investigates data enhancement with contextual texts and semantic neighbors of interactions. We propose a LLM-summarized interest distillation (SID) model to enrich interactions in interest description quality and collaborative neighbors quantity for multimedia recommendation. It employs LLMs' understanding and reasoning ability to summarize users' interests from their interactions for depicting users' interests in compacted texts. With this dedicated interest representation, semantically similar neighbors are retrieved, which serve as teachers to transfer multimodal collaborative knowledge to behavioral sparse users. Both interest summarization and retrieval augmentation work to enrich interest modeling for promoting recommendations. The primary contributions of this work are summarized as follows.

- We propose an LLM-summarized interest distillation (SID) model by interest summarization and retrieval augmentation to enhance interactions in both interest descriptions and collaborative neighbors.
- LLMs' understanding and reasoning capabilities are utilized to summarize users' interests from interaction contextual texts, removing interest irrelevant descriptions, thereby improving interest modeling in texts.
- Instead of directly augmenting graph structure, the retrieval-augmented neighbors work as multiple teachers to transfer their multimodal knowledge to the student node, adaptively augmenting collaborative signals for recommendation.

2 Related Work

2.1 Multimedia Recommendation

The multimedia recommendation system integrates contextual information of behavioral interactions for recommendation. Representative VBPR [He and McAuley, 2016] integrates visual information into embeddings of matrix factorization [Rendle *et al.*, 2009] to jointly represent user interests. LGMRec [Guo *et al.*, 2024] learns multimodal embeddings in a local graph and combines the global embeddings from a hypergraph to comprehensively represent interests. SMORE [Ong and Khong, 2025] integrates multimodal features in the frequency domain and performs multimodal graph learning to represent interests. Besides modality integration, recent works optimize modality alignment to fuse multiple modalities in interest modeling. BM3 [Zhou *et al.*, 2023] optimizes interaction graph reconstruction with inter- and intra-modality alignment to aggregate contextual contents into recommender learning. DiffMM [Jiang *et al.*, 2024] injects modality-aware knowledge into graph learning of interactions and highlights modality consistency for multimodal interest representation. Existing multimedia models usually focus on embedding and aligning modalities to represent interests. However, they ignore to explore the quality of modality data for learning interests and recommenders. It motivates the extraction of interest-relevant signals from behavior contextual modalities to improve their effectiveness in aiding behavioral data for interest learning.

2.2 LLM-Based Recommendation

Large language models have gained widespread recognition due to their textual understanding capability with extensive world knowledge. A growing body of studies [Zhang *et al.*, 2024; Sheng *et al.*, 2025] has leveraged LLM to encode users and items, extracting semantic features to depict the interaction graph in recommender learning. For example, NoteLLM [Zhang *et al.*, 2024] utilizes LLM to encode items into a special token, which is fed into a graph contrastive learning for representing and recommending items. LRD [Yang *et al.*, 2024] follows discrete variational autoencoder to encode items with LLM and extracts latent relations for reconstructing items and modeling users. AlphaRec [Sheng *et al.*, 2025] encodes items via LLM to initialize graph nodes and performs graph propagation to capture user interests with both semantic and structural signals. Beyond the LLMs' understanding capability, recent research [Tsai *et al.*, 2024; Ren *et al.*, 2024] has also witnessed their reasoning advantage to mine interests implicitly hidden in users' interactions. Representative RLMRec [Ren *et al.*, 2024] harnesses LLM to generate detailed user and item profiles for aiding behavioral signals. It effectively addresses the potential noise and task-irrelevant information from items' descriptions. The advanced reasoning and summarization abilities of large language models, surpassing their encoding proficiency, position them as a powerful tool for investigating the contextual texts of interest learning.

2.3 Contrastive Learning-Based Recommendation

Contrastive learning encodes the data structure by discerning samples, where the inherent self-supervised signals effectively address the data sparsity in personalized recommendation tasks. Representative SGL [Wu *et al.*, 2021] augments graph views through random perturbation and adopts contrastive learning to ensure view consistency for embedding learning. CLCRec [Wei *et al.*, 2021] uses contrastive learning to maximize mutual information between user-item and item-item pairs to reformulate representation learning. SimGCL [Yu *et al.*, 2022] adds random noises on node embeddings to form contrastive views for contrastive learning, demonstrating InfoNCE loss is primary to align learning views. LightGCL [Cai *et al.*, 2023] contrasts global and local collaborative views, achieving unconstrained structural refinement in embedding learning. LightCCF [Zhang *et al.*, 2025] builds contrastive learning to bring users closer to their positive items and separate them from other users' positive pairs to improve neighborhood aggregation. Since InfoNCE loss possesses excellent neighborhood aggregation capability in contrastive learning, it is desired to transfer knowledge from multimodal neighbors for interest learning.

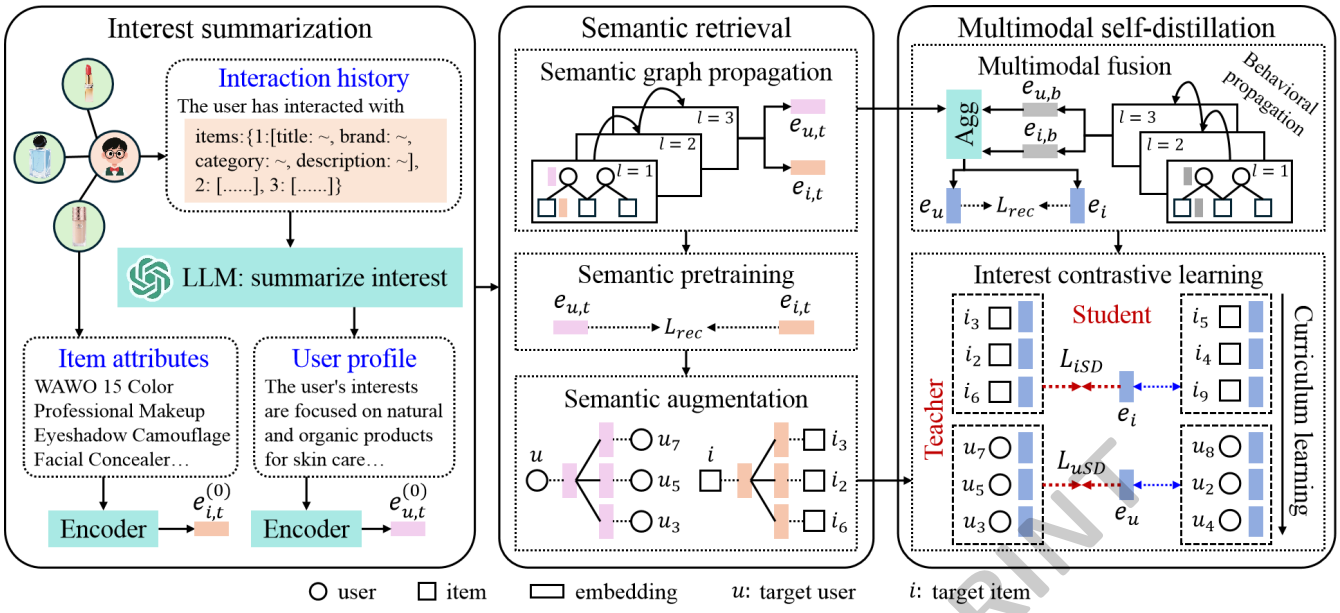

Figure 2: The framework of the proposed SID model, including three modules: (a) Interest summarization module adopts LLM to describe users' interests. (b) Semantic retrieval module filters semantic similar neighbors to serve as teachers. (c) Multimodal self-distillation module transfer the multimodal knowledge from teachers to the target student for interest modeling.

3 Problem Definition

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items with $|\mathcal{U}| = M$ and $|\mathcal{I}| = N$. For user $u \in \mathcal{U}$, $\mathcal{I}_u$ is the set of items that user u has interacted with, $\mathcal{I}_u \subseteq \mathcal{I}$. The historical behaviors are recorded in a binary interaction matrix $\mathcal{A} \in \{0,1\}^{M \times N}$, where $\mathcal{A}_{ui} = 1$ if $i \in \mathcal{I}_u$, otherwise $\mathcal{A}_{ui} = 0$. Besides, item i is semantically depicted with textual attributes of title, brand, category, and description, aggregated in $\mathcal{P}_i$. The target of this work is to learn a recommender model $f(\cdot)$ for predicting user-item interaction score $\hat{y}_{ui}$ as

$$\hat{y}_{ui} = f(u, i | \mathcal{A}_{ui}, \mathcal{P}_i, \Theta), \tag{1}$$

where Θ is the set of trainable parameters. The all-ranking protocol sorts the items by descending $\hat{y}_{ui}$ to recommend high-score ones for users.

4 Methodology

This section details the proposed LLM-summarized interest distillation (SID) for multimedia recommendation. As illustrated in Figure 2, SID involves three core modules of interest summarization, semantic retrieval, and multimodal self-distillation.

4.1 Interest Summarization

In multimedia scenarios, items' textual attributes that users have interacted with carry far richer semantic collaborative signals than implicit behavioral interaction data. They can directly describe user interests through textual semantics, helping alleviate data sparsity issues. However, items' attribute texts inevitably introduce data noise and interest-irrelevant information, which can interfere with interest modeling and increase the risk of language model hallucination. For describing user interests with texts, this work investigates prompt engineering, which emphasizes the goal of interest modeling, assigns the model a specific role-playing identity, and imposes clear formatting constraints. It leverages LLM's understanding and reasoning capabilities to filter out interest-irrelevant content, which guarantees the quality of semantic texts in describing user interests.

In the proposed SID, LLM (Llama3-8B) works with prompt engineering, including a system prompt Q_s and a user prompt Q_u, as illustrated in Figure 3. Q_s defines the LLM's role as a precision assistant for JSON summaries, specifies the focus of personal interests only, describes item attributes by title, brand, category, and description, and outlines the input format as a list of JSON strings. Besides, Q_u provides a user-specific interaction history $\mathcal{P}_i, i = 1, 2, ..., n$ in JSON format, specifies output requirements with 100–200 words, and includes a reference user interest profile for guidance. The prompt Q_s and Q_u guide LLM to summarize the user's interests and generate a comprehensive user profile that adheres to the specified JSON format represented in Equation (2).

$$\mathcal{P}_u = LLM(Q_s, Q_u), \tag{2}$$

where $\mathcal{P}_u$ is the generated user profile to describe user interests in texts. This profile comprehensively describes user interests by capturing both explicit preferences from interaction history and implicit preferences inferred from contextual and attribute-based contents. As Figure 2, the textual user profile $\mathcal{P}_u$ and item attributes $\mathcal{P}_i$ are encoded with a standard language model to embed users and items as Equation (3).

$$\mathbf{e}^{(0)}_{u,t} = proj(Enc(\mathcal{P}_u)), \quad \mathbf{e}^{(0)}_{i,t} = proj(Enc(\mathcal{P}_i)), \tag{3}$$

where $\mathbf{e}^{(0)}_{u,t}$ and $\mathbf{e}^{(0)}_{i,t}$ are the initialized semantic embeddings of user u and item i, the semantic projector $proj(\cdot)$ performs

System Prompt

You are a precision assistant generating structured JSON summaries of user interests. The summary must focus on **personal interest of the user**.
I will provide you with information about item that the user has interacted as well as the attributes of this item (title, brand, categories and description).
Here is the instructions:
Each interacted item will be described in JSON format, with the **following attributes**:
{**"title"**: "the title of the item which is interacted by the user",
"brand": "the brand of the item which is interacted by the user", (if there is no brand, I will set this value to " "),
"category": "several tags describing the item which is interacted by the user" (if there is no categories, I will set this value to " "),
"description": "the detailed description of the item which is interacted by the user"}
The information I will give you: **PURCHASSED ITEM**: a list of **JSON strings** describing the interaction history of the user with the attributes of items interacted by the user.

User Prompt

Given the interaction history of user and the **attributes** of the item in this interaction history.
{"The user has interacted with items": {
"0": {**"title"**: "WAWO 15 Color Professionl Makeup Eyeshadow Camouflage Facial Concealer Neutral Palette", **"brand"**: "COKA", **"category"**: "Beauty Makeup Face Concealers & Neutralizers", **"description"**: "An extensive range of 15 multiple vibrant long wear concealer colour...."},
"11650": {"title": "Vitamin C Serum for Face 20% - With Vegan Hyaluronic Acid…", "brand": "Amara Organics", "category": "Beauty Skin Care Face Creams & Moisturizers Fluids & Lotions Fluids", "description": "Drop A Decade From Your FaceEnjoy radiant…"},}}
Requirements:
The value should be a summary of the user's interests, no less than 100 words and no more than 200 words.
Please provide your answer in **JSON format**. Your response must be a valid JSON object with **a single key** that is the number "1" in string format, and the value is **a concise summary of the user's interests** based on their interaction history in string format, following this structure: **"1": "A summarization of the personal interest of this user and briefly explain your reasoning for the summarization."**
Do not include any text outside the JSON structure

User Profile

The user appears to have a strong interest in iPhone 4 and 4S accessories, particularly cases and screen protectors. He has interacted with numerous products that offer protection and style for their device, such as body armor defender cases, armband pouches, and bling diamond finishing screen protectors. Additionally, he has shown interest in home button stickers and customizing their phone's appearance. This suggests a focus on functionality and aesthetics when it comes to their iPhone accessories. The user may be looking for ways to **personalize and safeguard their device** while also **maintaining its style and performance**.

Figure 3: An example of system prompt, user prompt, and the generated user profile in interest summarization.

two linear projections with a $LeakyReLU(\cdot)$ nonlinear activation, and $Enc(\cdot)$ represents the textual encoder Qwen2.5-3B in which the last hidden state of the last token are taken as the encoder output.

4.2 Semantic Retrieval

To extract semantic collaborative signals, the proposed SID propagates the textual embeddings of users and items on the interaction graph by graph convolution as:

$$\mathbf{e}_{u,t}^{(l)} = \sum_{i \in \mathcal{N}_u} \frac{\mathbf{e}_{i,t}^{(l-1)}}{\sqrt{|\mathcal{N}_u|\,|\mathcal{N}_i|}}, \quad \mathbf{e}_{i,t}^{(l)} = \sum_{u \in \mathcal{N}_i} \frac{\mathbf{e}_{u,t}^{(l-1)}}{\sqrt{|\mathcal{N}_i|\,|\mathcal{N}_u|}}, \tag{4}$$

where $\mathbf{e}_{u,t}^{(l)}$ and $\mathbf{e}_{i,t}^{(l)}$ represent the learned semantic embeddings of user u and item i at the lth propagation layer, $l = 1, 2, ..., L$, $\mathcal{N}_u$ and $\mathcal{N}_i$ denote the first-hop neighboring node set of user u and item i in the interaction graph. The multi-order semantic collaborative signals are then aggregated by integrating the layerwise semantic embeddings as:

$$\mathbf{e}_{u,t} = \frac{1}{L+1} \sum_{l=0}^{L} \mathbf{e}_{u,t}^{(l)}, \quad \mathbf{e}_{i,t} = \frac{1}{L+1} \sum_{l=0}^{L} \mathbf{e}_{i,t}^{(l)}, \tag{5}$$

where $e_{u,t}$ and $e_{i,t}$ are the aggregated semantic embeddings of user u and item i. Here, semantic pretraining is performed with the recommendation task in Section 4.4 to optimize the semantic projector $proj(\cdot)$ in Equation (3), which guarantees the quality of semantic embeddings in the subsequent retrieval augmentation. As Figure 2, homogeneous similar users and items are retrieved by evaluating interest consistency along the semantic embeddings $e_{u,t}$ and $e_{i,t}$ as:

$$\begin{aligned} g_u &= top_k(u' | \cos(\mathbf{e}_{u,t}, \mathbf{e}_{u',t}), u' \in |\mathcal{U} \backslash u|), \\ g_i &= top_k(i' | \cos(\mathbf{e}_{i,t}, \mathbf{e}_{i',t}), i' \in |\mathcal{I} \backslash i|), \end{aligned} \tag{6}$$

where g_u and g_i are the set of users/items retrieved for user u and item i, $\cos(\cdot)$ measures the semantic cosine similarity, and $top_k(\cdot)$ returns top-k users u' for u and top-k items i' for i. This semantic retrieval measures pairwise semantic similarity between the target user/item and all the other users/items and selects the top-k neighbors by descending the candidates. The retrieved users/items act to augment neighboring collaborative signals for the sparse target node from the semantic view. It augments the sparse interaction data in collaborative neighbors quantitatively.

4.3 Multimodal Self-Distillation

In addition to the semantic collaboration, the behavioral collaborative signals are propagated on the interaction graph to learn behavior embeddings for users and items. As Figure 2, the behavior propagation performs the same graph convolution in Equations (4) and (5), where the behavior embeddings $\mathbf{e}_{u,b}^{(0)}$ and $\mathbf{e}_{i,b}^{(0)}$ are randomly initialized as learnable parameters in the proposed SID. The derived behavior embeddings $\mathbf{e}_{u,b}$ and $\mathbf{e}_{i,b}$ aggregate multi-order behavioral collaborative signals. SID simply fuses the semantic and behavior embeddings as:

$$\mathbf{e}_u = \mathbf{e}_{u,b} + \mathbf{e}_{u,t}, \quad \mathbf{e}_i = \mathbf{e}_{i,b} + \mathbf{e}_{i,t} \tag{7}$$

where $\mathbf{e}_u$ and $\mathbf{e}_i$ are the fused multimodal embeddings of user u and item i. In the multimodal view, the retrieved neighbors g_u and g_i are treated as multiple teachers to augment the sparse interactions. Beyond directly augmenting the graph structure [Fan *et al.*, 2023], we deploy an adaptive knowledge transfer manner by multimodal self-distillation to augment the interactions. The self-distillation is implemented with InfoNCE contrastive learning [Yuan *et al.*, 2025] to transfer knowledge from teachers to the target user/item in Equation (8).

$$\begin{aligned} \mathcal{L}_{uSD}(u) &= -\log \frac{\exp\left(s(\mathbf{e}_u, \mathbf{e}_{g_u})/\tau_{sd}\right)}{\sum_{u' \in \mathcal{B}_u} \exp\left(s(\mathbf{e}_u, \mathbf{e}_{g_{u'}})/\tau_{sd}\right)}, \\ \mathcal{L}_{iSD}(i) &= -\log \frac{\exp\left(s(\mathbf{e}_i, \mathbf{e}_{g_i})/\tau_{sd}\right)}{\sum_{i' \in \mathcal{B}_i} \exp\left(s(\mathbf{e}_i, \mathbf{e}_{g_{i'}})/\tau_{sd}\right)}, \end{aligned} \tag{8}$$

where $s(\mathbf{e}_u, \mathbf{e}_{g_u}) = \sum_{j \in g_u} cos(\mathbf{e}_u, \mathbf{e}_j)$ and $s(\mathbf{e}_i, \mathbf{e}_{g_i}) = \sum_{j \in g_i} cos(\mathbf{e}_i, \mathbf{e}_j)$ aggregate the cosine similarity between $\mathbf{e}_{u/i}$ and $\mathbf{e}_j$, τ_{sd} is the temperature coefficient, $\mathcal{B}_u$ and $\mathcal{B}_i$ are respectively the sets of users and items in a batch. In optimizing $\mathcal{L}_{uSD}(u)$ and $\mathcal{L}_{iSD}(i)$, the gradient flow through the teacher embeddings is halted, keeping them frozen. It ensures

the stability of the teachers in guiding interest learning. By minimizing the embedding disparity, the multimodal collaborative knowledge is aggregated from multiple teachers and injected into the target interest embedding for augmenting the sparse learning.

4.4 Recommendation and Optimization

As Figure 2, the learned multimodal embeddings $\mathbf{e}_u$ and $\mathbf{e}_i$ are utilized to match user interests. The pairwise matching is implemented with the inner product to predict the interaction score as Equation (9).

$$\hat{y}_{ui} = \mathbf{e}_u \cdot \mathbf{e}_i. \tag{9}$$

To optimize the recommendation task, a adaptive weighted Bayesian personalized ranking loss $\mathcal{L}_{rec}$ is applied in SID to encourage positive item ranked above negative ones as Equation (10).

$$\begin{aligned} \mathcal{L}_{rec} &= -\sum_{(u,i,j)\in O} \log\left[\omega_{u,i} \cdot \sigma(\hat{y}_{u,i} - \hat{y}_{u,j})\right], \\ \omega_{u,i} &= \frac{e^{\hat{y}_{u,i}/\tau_p}}{e^{\hat{y}_{u,i}/\tau_p} + \sum_{(u,j)\in O^-} e^{\hat{y}_{u,j}/\tau_p}}, \end{aligned} \tag{10}$$

where $O = \{(u,i,j) \mid (u,i) \in O^+, (u,j) \in O^-\}$, O^+ denotes the positive instance set, and O^- is the negative instance set, $\sigma(\cdot)$ is a sigmoid function, $\omega_{u,i}$ is the adaptive weight to measure the discrepancy between the target positive $u-i$ and negative instances for user u in a batch, and τ_p is the temperature coefficient. Particularly, in the semantic retrieval, the recommendation loss $\mathcal{L}_{rec}$ is performed to pretrain the semantic embeddings $\mathbf{e}_{u,t}$ and $\mathbf{e}_{i,t}$ to ensure the quality of retrieval augmentation. Besides, the recommendation and the auxiliary self-distillation tasks are jointly optimized with a curriculum learning strategy [Wang *et al.*, 2021], which gradually introduces distillation from a single side to dual sides. Following the sequential modeling of items and users in the text, we first perform self-distillation on the item side. Subsequently, we introduce the more complex user-side distillation, thereby reducing the optimization difficulty during the model training phase. The curriculum learning on multimodal embeddings is represented in Equation (11).

$$\mathcal{L}_{main} = \begin{cases} \mathcal{L}_{rec} + \beta(\mathcal{L}_{iSD}), & t \le T, \\ \mathcal{L}_{rec} + \beta(\mathcal{L}_{iSD} + \mathcal{L}_{uSD}), & t > T \end{cases} \tag{11}$$

where β is the weight to control the contribution of self-distillation, t represents the number of training epochs, and T denotes the threshold of training epochs to control the training complexity, converting from a single side to dual sides.

5 Experiments

5.1 Experiment Settings

Datasets

To verify the effectiveness of SID in multimedia recommendation, experiments are performed on three public datasets [McAuley *et al.*, 2015], including Beauty, Toys, and Cellphones. These datasets provide user-item interaction records and textual descriptions for each item. The key statistical indicators of the datasets are summarized in Table 1. Each dataset is split into training and test sets at an $8:2$ ratio. User-item interaction pairs recorded in the interaction history are defined as positive pairs, while user-item pairs without observed interactions in the history are treated as negative pairs. For negative sampling [Robinson *et al.*, 2021], 256 negative pairs are randomly selected for each positive pair to support self-distillation and the downstream recommendation optimization.

Dataset	#Users	#Items	#Interactions	Density
Beauty	22,363	12,101	198,502	0.073%
Toys	19,412	11,924	167,597	0.072%
Cellphones	27,879	10,429	194,439	0.067%

Table 1: Statistics of Beauty, Toys, and Cellphones datasets.

Evaluation Metrics

To evaluate the performance, the all-ranking protocol is employed on the test set. For each user, all the items are ranked based on their predicted interaction scores, and the top-N unobserved items are selected to form the user's recommendation list, where N is the length of the recommendation list, which is set to 20 by default. For quantitative evaluation, three widely used metrics are adopted, including Recall@N, NDCG@N, and Precision@N, where higher values of these metrics indicate better performance. To ensure the reliability and statistical significance of the experimental results, each experiment is repeated 10 times, and the average value of them is reported in the analysis.

Parameters

The semantic projector $proj(\cdot)$ and behavior embeddings $\{\mathbf{e}^{(0)}_{u,b}, \mathbf{e}^{(0)}_{i,b}\}$ are initialized using the Xavier normal distribution. The textual encoder derives a 2048-dimensional textual representation, and the semantic and behavior embeddings are set to 64 dimensions. The graph convolution is performed with 3 layers for semantic and behavior propagation. The learning rate is set to 3e-4, and the Adam optimizer [Zhang, 2018] is applied to train the model. The hyperparameters are tuned with empirical experiments. By default, on the datasets, the temperature coefficient for recommendation loss τ_p is set as 0.2, τ_{sd} in self-distillation loss is set by $\{0.2, 0.2, 0.3\}$, the strength β of self-distillation is set 0.2, and the size k of the retrieved neighbors is set as 10. To ensure experimental fairness, the original parameter settings of all baseline models are retained, and the encoder is unified to be consistent with Qwen2.5-3B as in this work. All experiments are conducted in an environment with PyTorch 2.2.0 and Python 3.8.19.

5.2 Performance Comparison

We compare the proposed SID with representative baselines, including graph-based models (MFBPR [Koren *et al.*, 2009], NGCF [Wang *et al.*, 2019], and LightGCN [He *et al.*, 2020]), contrastive learning-based models (SGL [Wu *et al.*, 2021], SimGCL [Yu *et al.*, 2022]), multimedia models (BPR_T [He and McAuley, 2016], BM3_T [Zhou *et al.*, 2023], SMORE_T [Ong and Khong, 2025]), and LLM-based models (RLMRecCon [Ren *et al.*, 2024], RLM-Gen [Ren *et al.*, 2024], Al-

	Beauty			Toys			Cellphones		
	Recall	NDCG	Precision	Recall	NDCG	Precision	Recall	NDCG	Precision
MFBPR	0.0992	0.0502	0.0080	0.0880	0.0467	0.0060	0.1041	0.0501	0.0062
NGCF	0.1144	0.0547	0.0091	0.1066	0.0543	0.0073	0.1378	0.0645	0.0081
LightGCN	0.1373	0.0709	0.0111	0.1295	0.0673	0.0088	0.1577	0.0756	0.0092
SGL	0.1483	0.0769	0.0121	0.1387	0.0716	0.0096	0.1649	0.0804	0.0097
SimGCL	0.1530	0.0785	0.0124	0.1437	0.0736	0.0100	0.1744	0.0850	0.0102
BPR-T	0.1177	0.0569	0.0089	0.1145	0.0567	0.0075	0.1113	0.0519	0.0066
BM3-T	0.1146	0.0557	0.0088	0.1018	0.0488	0.0067	0.1454	0.0680	0.0083
SMORE-T	0.1496	0.0723	0.0112	0.1486	0.0700	0.0099	0.1513	0.0730	0.0087
RLMRec-Con	0.1469	0.0726	0.0116	0.1305	0.0634	0.0089	0.1544	0.0737	0.0091
RLMRec-Gen	0.1285	0.0637	0.0105	0.1180	0.0587	0.0081	0.1492	0.0719	0.0087
AlphaRec	0.1588	0.0755	0.0119	0.1520	0.0721	0.0101	0.1697	0.0763	0.0098
SID	**0.1752**	**0.0887**	**0.0137**	**0.1729**	**0.0870**	**0.0116**	**0.1828**	**0.0867**	**0.0107**
Improv.	10.33%	12.99%	10.48%	13.75%	18.21%	14.85%	4.82%	2.00%	4.90%

Table 2: Top-20 recommendation performance by Recall, NDCG, and Precision on Beauty, toys, and cellphones datasets.

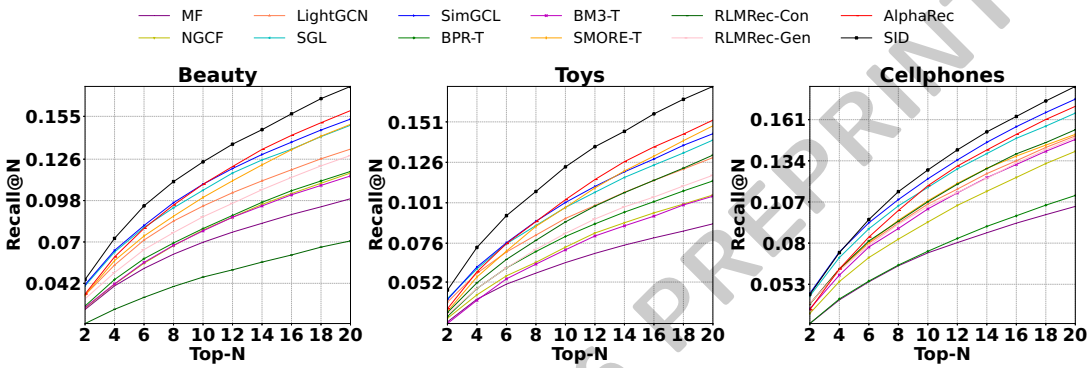

Figure 4: Performance comparison of top-N recommendations by Recall on Beauty, Toys, and Cellphones datasets.

phaRec [Sheng *et al.*, 2025]). For fair comparison, we remove or institute the visual modality with the textual one in multimedia models, retaining the behavioral and the textual modalities in comparison. The performance of SID and the baselines is compared by Recall@N in Figure 4. The performance by NDCG and Precision exhibits a similar trend; therefore, their figures are omitted. Table 2 presents the top-20 recommendation performance comparing SID and the baselines. In the table, the suboptimal performance is underlined, and the optimal performance is shown in bold.

It can be observed that SID outperforms the baselines consistently on the datasets. By Recall@20, SID yields relative increases of 10.33%, 13.75%, and 4.82%, respectively. The improvements verify the effectiveness of interest summarization and retrieval augmentation in SID to enhance multimodal interest learning for recommendation. SGL and SimGCL outperform MFBPR, NGCF, and LightGCN. It confirms that contrastive learning can effectively enhance graph learning for recommendation. Meanwhile, the superiority of SID over SGL and SimGCL indicates that retrieval augmented knowledge works more significant than self-discrimination in contrastive learning for the downstream recommendation task. The representative multimedia SMORE_T achieves better performance than LightGCN on the Beauty and Toys datasets, but slightly lower performance on the Cellphones dataset, which implies the external semantic information

	Beauty			Toys			Cellphones		
	Recall	NDCG	Precision	Recall	NDCG	Precision	Recall	NDCG	Precision
w/o reasoning	0.1705	0.0876	0.0134	0.1714	0.0868	0.0105	0.1800	0.0862	0.0104
w/o text	0.1373	0.0709	0.0111	0.1295	0.0673	0.0088	0.1577	0.0756	0.0092
w/o RA	0.1670	0.0850	0.0132	0.1616	0.0826	0.0109	0.1789	0.0859	0.0104
w/o SD	0.1444	0.0734	0.0115	0.1347	0.0691	0.0090	0.1404	0.0672	0.0082
w/o RA+SD	0.1658	0.0849	0.0130	0.1608	0.0824	0.0108	0.1781	0.0858	0.0104
w/o uSD	0.1730	0.0874	0.0135	0.1715	0.0866	0.0114	0.1807	0.0865	0.0105
w/o iSD	0.1702	0.0873	0.0134	0.1639	0.0840	0.0111	0.1804	0.0864	0.0105
w/o curriculum	0.1724	0.0878	0.0135	0.1696	0.0859	0.0114	0.1824	0.0866	0.0106
SID	**0.1752**	**0.0887**	**0.0137**	**0.1729**	**0.0870**	**0.0116**	**0.1828**	**0.0867**	**0.0107**

Table 3: Ablation performance of top-20 recommendations. RA: retrieval augmentation, SD: self-distillation.

works actively but requires an elaborate design for interest modeling. Additionally, SMORE_T yields lower performance than SimGCL, which underscores the significance of contrastive learning for recommendation. Representative LLM model, AlphaRec, exhibits comparable performance to SimGCL. The performance validates the potential of LLM in depicting interests with semantic information to support the downstream recommendation task.

5.3 Ablation Study

We conduct experiments to verify the role of interest summarization and retrieval augmentation in SID in Table 3. The key operations are removed from SID successively to derive variants compared with SID. SID w/o reasoning removes LLM reasoning and directly encodes items' textual attributes to embed user interest in semantics. SID w/o text removes the textual modality and retains only the behavioral modal-

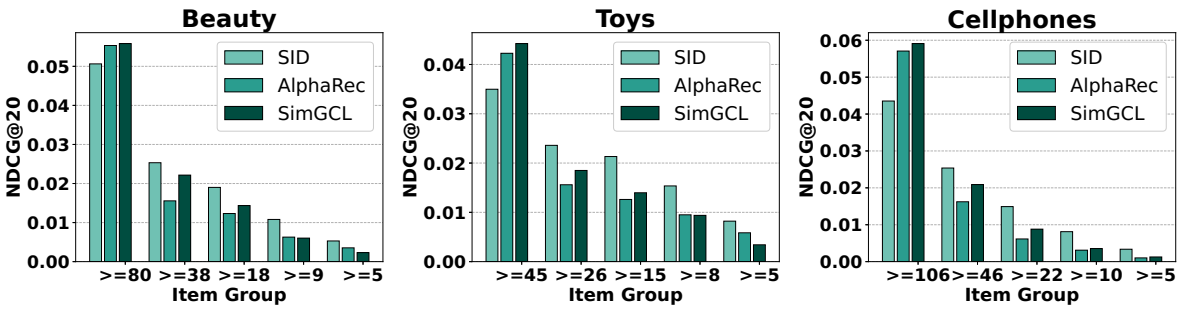

Figure 5: Performance comparison on item groups of varying popularity by NDCG@20. The x-axis is the lower bound of interaction size in each item group.

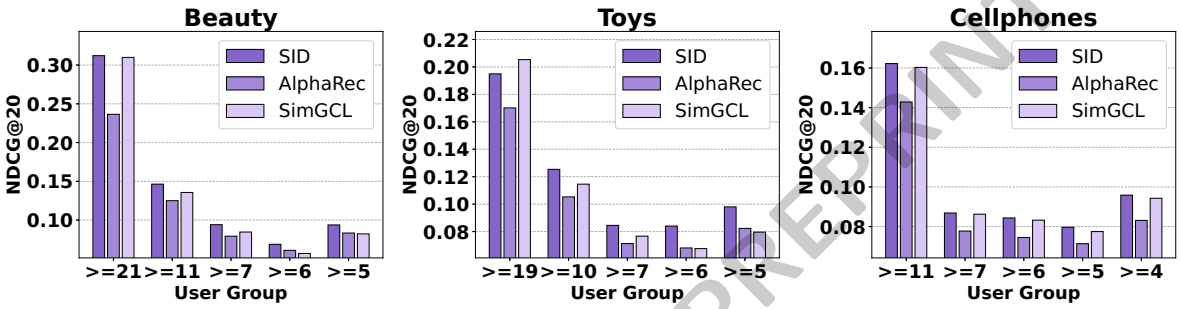

Figure 6: Performance comparison on user groups of varying interaction sparsity by NDCG@20. The x-axis signifies the lower limit of interaction size in each user group.

ity for interest modeling. SID w/o RA removes retrieval augmentation and perturbs the target user/item to augment data for contrastive learning. SID w/o SD removes self-distillation and uses the retrieved users/items to augment the graph structure for recommendation. SID w/o RA+SD removes retrieval augmentation and self-distillation, preserving single task $\mathcal{L}_{rec}$ to train the model. SID w/o uSD and w/o iSD removes the self-distillation on the user/item side and merely performs knowledge transfer on the item/user side. SID w/o curriculum removes the curriculum learning strategy and implements knowledge transfer immediately on both dual sides. The performance of SID w/o reasoning decreases compared to SID, which indicates that LLM can filter out interest-irrelevant signals and improve the quality of texts in describing interests. SID w/o text drops the performance sharply, proving that the semantic embedding encoded by LLM acts crucially to capture interest signals in the downstream retrieval and distillation. The performance of SID w/o RA, w/o SD, and w/o RA+SD degrades apparently, confirming that retrieval augmentation contributes to enhancing the interactions with semantic neighbors. SID w/o SD exhibits a more notable decline than that w/o RA. These trends highlight that both semantic retrieval and self-distillation play important roles in multimodal interest learning, and the multimodal distillation is relatively more critical. The performance degradation by SID w/o uSD and w/o iSD verifies the effectiveness of dual-side knowledge transfer. SID w/o curriculum decreases the performance, illustrating that knowledge transfer on the user side is more complex than the item side, and their joint learning would disturb each other, hindering the performance. It can be concluded that interest summarization and retrieval augmentation are active in enriching the sparse interactions for recommendation.

5.4 Long-Tail Analysis

We conduct experiments on items of varying popularity to verify the ability of SID in recommending long-tail items. The performance of SID is compared with representative baselines, SimGCL and AlphaRec, which are provided in Figure 5. In each dataset, items are sorted by their interaction size in descending order and divided into 5 groups, where each group contains approximately the same number of interactions. For instance, on the Beauty dataset, the minimum number of interactions for items in each group is 80, 38, 18, 9, and 5, respectively. As the popularity decreases, SID degrades less than SimGCL and AlphaRec. In groups with either moderate or extremely sparse interactions, SID consistently achieves superior performance, which indicates that SID recommends more long-tail items. This characteristic makes item recommendations more balanced and effectively mitigates overemphasizing popular items over long-tail ones. Overall, SID is effective and competitive in recommending long-tail items.

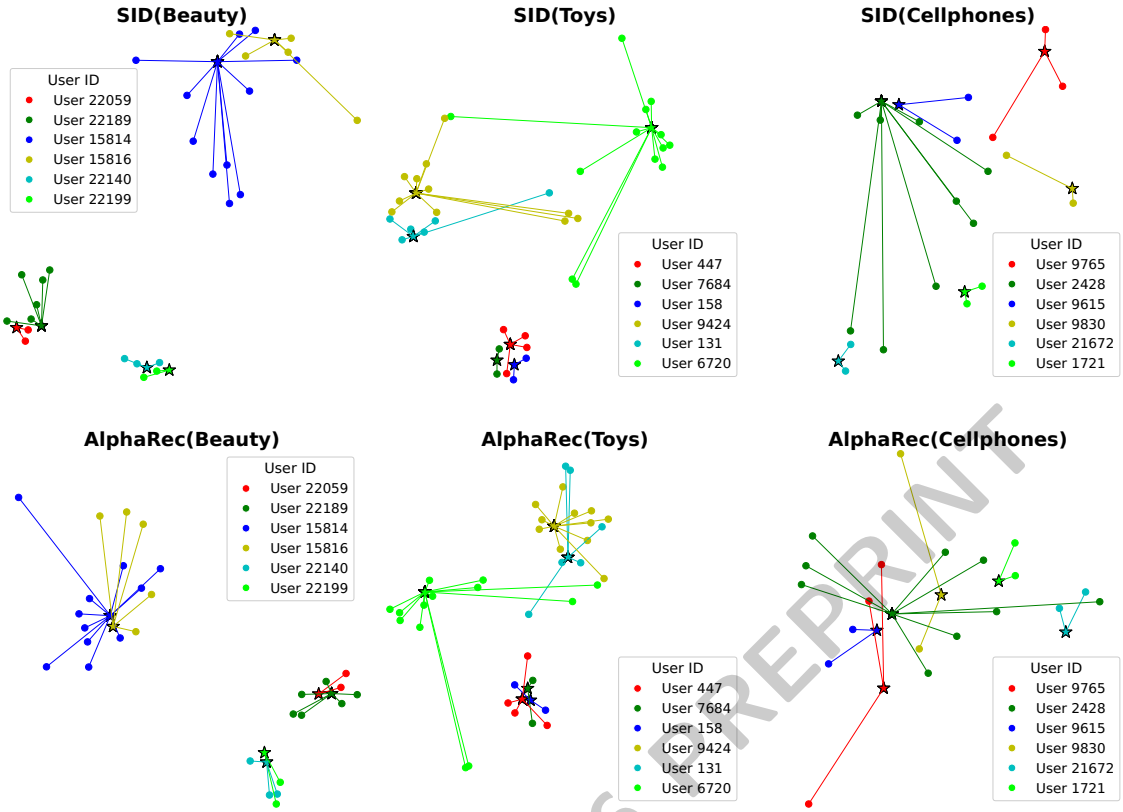

Figure 7: Embedding distribution by SID (1st row) and AlphaRec (2nd row), where stars represent users, and circles of the same color indicate the items interacted with by the corresponding user.

5.5 Sparsity Analysis

We conduct experiments on users of varying interaction sparsity to validate the ability of SID to deal with the sparsity issue. The performance comparison on varying sparsity is given in Figure 6. In each dataset, users are partitioned into 5 groups based on interaction density, ranking from dense to sparse. Within each group, the number of user interactions is approximately uniform. For instance, in the Beauty dataset, users were categorized into groups with the lower bound of interactions 21, 11, 7, 6, and 5, respectively. SID outperforms AlphaRec and SimGCL on most user groups. As interactions become increasingly sparse, the degradation of SID is relatively small, illustrating a more pronounced advantage. In summary, SID is competitive in addressing the sparse interaction issue.

5.6 Visualization

To show the discriminative ability, we further visualize the embedding distribution by t-SNE to compare SID with AlphaRec in Figure 7. The results reveal notable differences between SID and AlphaRec. By SID, users and their interacted items exhibit more compactness. There is minimal intermixing between clusters of users and their interactions. In comparison, AlphaRec displays a far more disordered distribution where the clusters are in a mixed, chaotic state. It validates the effectiveness of SID in capturing users' discriminative personality in interest modeling.

6 Conclusion

We have proposed an LLM-summarized interest distillation (SID) model to enhance users' sparse interactions with high-quality interest descriptions and more collaborative neighbors for multimedia recommendation. SID leverages LLMs' understanding and reasoning capabilities to depict user interest with semantic texts, which effectively reduces interest-irrelevant descriptions from their interactions. With these summarized texts, we retrieve semantic neighbors to transfer their multimodal knowledge to augment the target user for promoting recommendations. Extensive experiments on real-world datasets demonstrate that SID significantly outperforms representative baselines, verifying the ability of interest summarization and retrieval augmentation in enriching the sparse interactions. Ablation studies further validate the necessity of the compacted interest description and augmented knowledge distillation. The long-tail and sparsity analysis confirm that SID can effectively alleviate the sparse interaction issue. In future work, we will explore interest summarization ability to further promote interest learning with semantic augmentation.

Acknowledgements

This work was supported by the Beijing Natural Science Foundation under Grant NO. L241053, S&T Program of Hebei under Grant No.202621502030001, and the National Natural Science Foundation of China under Grant NO. 62176011.

References

[Cai *et al.*, 2023] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. LightGCL: Simple yet effective graph contrastive learning for recommendation. In *Proceedings of the International Conference on Learning Representation*, 2023.

[Fan *et al.*, 2023] Ziwei Fan, Ke Xu, Zhang Dong, Hao Peng, Jiawei Zhang, and Philip S Yu. Graph collaborative signals denoising and augmentation for recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2037–2041, 2023.

[Guo *et al.*, 2024] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. LGMRec: local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8454–8462, 2024.

[He and McAuley, 2016] Ruining He and Julian McAuley. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, 2016.

[He *et al.*, 2020] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.

[Idrissi and Zellou, 2020] Nouhaila Idrissi and Ahmed Zellou. A systematic literature review of sparsity issues in recommender systems. *Social Network Analysis and Mining*, 10(1), 2020.

[Jiang *et al.*, 2024] Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. DiffMM: Multimodal diffusion model for recommendation. In *Proceedings of the ACM International Conference on Multimedia*, pages 7591–7599, 2024.

[Koren *et al.*, 2009] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[McAuley *et al.*, 2015] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[Ong and Khong, 2025] Rongqing Kenneth Ong and Andy WH Khong. Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 773–781, 2025.

[Ren *et al.*, 2024] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Representation learning with large language models for recommendation. In *Proceedings of the ACM web conference*, pages 3464–3475, 2024.

[Rendle *et al.*, 2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 452–461, 2009.

[Robinson *et al.*, 2021] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *Proceedings of the International Conference on Learning Representations*, 2021.

[Sheng *et al.*, 2025] Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. Language representations can be what recommenders need: Findings and potentials. In *Proceedings of the International Conference on Learning Representations*, 2025.

[Tsai *et al.*, 2024] Alicia Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, Taibai Xu, Zemin Zhang, Lichan Hong, Ed H Chi, and Xinyang Yi. Leveraging llm reasoning enhances personalized recommender systems. In *Findings of the Association for Computational Linguistics*, pages 13176–13188, 2024.

[Wang *et al.*, 2019] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 165–174, New York, NY, USA, 2019.

[Wang *et al.*, 2021] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Wang *et al.*, 2024] Yuhao Wang, Yichao Wang, Zichuan Fu, Xiangyang Li, Wanyu Wang, Yuyang Ye, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. LLM4MSR: An LLM-enhanced paradigm for multi-scenario recommendation. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2472–2481, 2024.

[Wei *et al.*, 2021] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. Contrastive learning for cold-start recommendation. In *Proceedings of the ACM international conference on multimedia*, pages 5382–5390, 2021.

[Wu *et al.*, 2021] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval*, pages 726–735, 2021.

[Wu *et al.*, 2024] Junda Wu, Cheng Chun Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, and Julian Mcauley. CoRAL:collaborative retrieval-augmented large language models improve long-tail recommendation. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

[Yang *et al.*, 2024] Shenghao Yang, Weizhi Ma, Peijie Sun, Qingyao Ai, Yiqun Liu, Mingchen Cai, and Min Zhang. Sequential recommendation with latent relations based on large language model. In *Proceedings of the International ACM SIGIR conference on research and development in information retrieval*, pages 335–344, 2024.

[Yin *et al.*, 2012] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. Challenging the long tail recommendation. *Proceedings of the VLDB Endowment*, 5(9):896–907, 2012.

[Yu *et al.*, 2022] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1294–1303, 2022.

[Yu *et al.*, 2024] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Jundong Li, and Zi Huang. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Automatic Control*, 36(1):21, 2024.

[Yuan *et al.*, 2025] Meng Yuan, Zhao Zhang, Wei Chen, Chu Zhao, Tong Cai, Deqing Wang, Rui Liu, and Fuzhen Zhuang. Hek-cl: Hierarchical enhanced knowledge-aware contrastive learning for recommendation. *ACM Transactions on Information Systems*, 2025.

[Zhang *et al.*, 2024] Chao Zhang, Shiwei Wu, Haoxin Zhang, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. NoteLLM: A retrievable large language model for note recommendation. In *Proceedings of the ACM Web Conference*, pages 170–179, 2024.

[Zhang *et al.*, 2025] Yu Zhang, Yiwen Zhang, Yi Zhang, Lei Sang, and Yun Yang. Unveiling contrastive learning's capability of neighborhood aggregation for collaborative filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

[Zhang, 2018] Zijun Zhang. Improved adam optimizer for deep neural networks. In *Proceedings of the IEEE/ACM International Symposium on Quality of Service*, 2018.

[Zhou *et al.*, 2023] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multimodal recommendation. In *Proceedings of the ACM web conference 2023*, pages 845–854, 2023.