MLLMRec: A Preference Reasoning Paradigm with Graph Refinement for Multimodal Recommendation

Yuzhuo Dang
College of Systems Engineering, National University of Defense Technology
Changsha, China
dangyuzhuo@nudt.edu.cn

Xin Zhang
College of Systems Engineering, National University of Defense Technology
Changsha, China
zhangxin16@nudt.edu.cn

Zhiqiang Pan
College of Systems Engineering, National University of Defense Technology
Changsha, China
panzhiqiang@nudt.edu.cn

Yuxiao Duan
College of Systems Engineering, National University of Defense Technology
Changsha, China
duanyuxiao19@nudt.edu.cn

Wanyu Chen
College of Electronic Countermeasures, National University of Defense Technology
Hefei, China
wanyuchen@nudt.edu.cn

Fei Cai*
College of Systems Engineering, National University of Defense Technology
Changsha, China
caifei08@nudt.edu.cn

Honghui Chen*
College of Systems Engineering, National University of Defense Technology
Changsha, China
chenhonghui@nudt.edu.cn

arXiv:2508.15304v3 [cs.IR] 10 Sep 2026

ABSTRACT

Multimodal recommendation combines the user historical behaviors with the modal features of items to capture the tangible user preferences, presenting superior performance compared to the conventional ID-based recommender systems. However, existing methods still encounter two key problems in the representation learning of users and items, respectively: (1) the initialization of multimodal user representations is either agnostic to historical behaviors or contaminated by irrelevant modal noise, and (2) the widely used *K*NN-based item-item graph contains noisy edges with low similarities and lacks audience co-occurrence relationships. To address such issues, we propose MLLMRec, a novel preference reasoning paradigm with graph refinement for multimodal recommendation. Specifically, on the one hand, the item images are first converted into high-quality semantic descriptions using a multimodal large language model (MLLM), thereby bridging the semantic gap between visual and textual modalities. Then, we construct a behavioral description list for each user and feed it into the MLLM to reason about the purified user preference profiles that contain the latent interaction intents. The reasoned profiles and the multimodal descriptions of items, together with their ID embeddings, are propagated over the user-item interaction graph to absorb the high-order collaborative signals. On the other hand, we develop the threshold-controlled denoising and topology-aware enhancement strategies to refine the suboptimal item-item graph, which are applied to both the multimodal and ID item representations to improve the accuracy of item representation learning. Extensive experiments on three publicly available datasets demonstrate that MLLMRec achieves the state-of-the-art performance. The source code is provided at https://github.com/Yuzhuo-Dang/MLLMRec.git.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Multimodal Recommendation, User Preference Reasoning, Item-Item Graph Refinement, Multimodal Large Language Model

ACM Reference Format:
Yuzhuo Dang, Xin Zhang, Zhiqiang Pan, Yuxiao Duan, Wanyu Chen, Fei Cai, and Honghui Chen. 2018. MLLMRec: A Preference Reasoning Paradigm with Graph Refinement for Multimodal Recommendation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

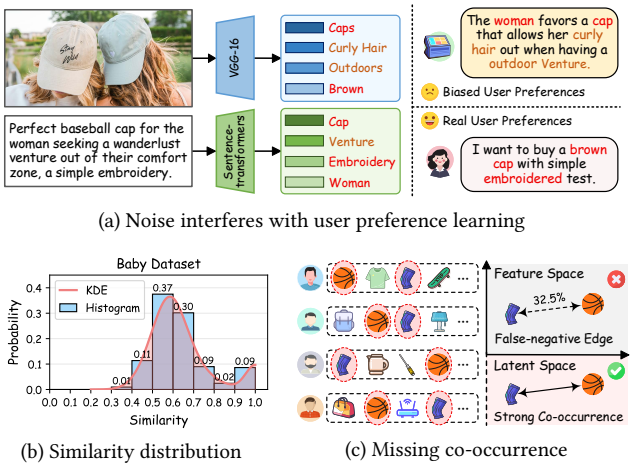
1 INTRODUCTION

With the exponential growth of data scale and the diversification of content forms on online platforms, multimodal recommender systems (MMRS) have emerged as a pivotal technology for mitigating the information overload [21, 30]. By integrating the various item

modalities (e.g., text, image, and audio) with the historical user-item interactions, MMRS facilitate a more nuanced understanding of user preferences, thereby delivering highly personalized services in increasingly complex information environments [15, 27].

The evolution of MMRS has progressed through several distinct paradigms. Pioneering research, exemplified by VBPR [10], focuses on simple feature fusion by augmenting the ID embeddings with the visual features. Subsequently, to capture the high-order collaborative signals, graph-based methods like MMGCN [33] and GRCN [32] introduce graph convolutional networks (GCNs) to propagate the multimodal information across the user-item interaction graph. Recognizing that the item-side correlations are often underutilized, some studies shift their focus toward mining the latent structural patterns among items. For instance, LATTICE [40] constructs the tunable item-item graphs by computing the cosine similarities between items based on the raw and projected modal features. Further, FREEDOM [41] argues that the frozen item-item graphs constructed using only the raw features yield superior performance. Most recently, the burgeoning field of large language models (LLMs) open new frontiers for MMRS [1, 4, 29]. For example, LLMRec [31] leverages the vast world knowledge of LLMs to generate the supplementary attributes for users and items. Moreover, DOGE [20] employs multimodal large language models (MLLMs) to address the issue of granularity inconsistency across modalities, achieving the cross-modal relation mining.

Despite these remarkable achievements, existing MMRS still exhibit the following two challenges in the representation learning at both the user and item sides: (1) **Inaccurate initialization of user representations.** Current approaches typically initialize the user representations through either random initialization or aggregating the modal features of the historically interacted items. However, both paradigms are flawed. Specifically, the former method is inherently behavior-agnostic, neglecting to leverage the rich preference signals in the interactions and thus exacerbating the cold-start issue. The latter approach is often noise-contaminated, as it fails to filter out the modal features encoded by pre-trained models, leading to the indiscriminate aggregation of irrelevant features (e.g., the brown-marked noise in Figure 1(a)). More critically, these methods are incapable of capturing the latent interaction intents hidden in the user decision-making process. Consequently, they are forced to rely on the sophisticated subsequent behavioral modeling to compensate for the informational deficiencies in the initial representations. (2) **Structural noise in the item-item graph topology.** The widely used KNN sparsification strategy [2] for the item-item graph construction is limited by its fixed number of neighbors. This mechanism compels items lacking strongly similar neighbors to associate with low-similarity ones, inevitably introducing false-positive edges. As the evidence reveals in Figure 1(b), approximately 12% of edges in the Baby dataset link items with a similarity score below 0.5. In addition, the KNN strategy depends exclusively on the similarity of modal features while ignoring the latent audience co-occurrence correlations, which are the vital collaborative signals in the field of recommender systems [19]. Consequently, items that are semantically distant but behaviorally correlated remain disconnected (e.g., the basketball and genouillere in Figure 1(c)), resulting in false-negative edges. Both types of structural noise disrupt the



(a) Noise interferes with user preference learning

(b) Similarity distribution

(c) Missing co-occurrence

Figure 1: Illustrations of (a) Noise interferes with user preference learning, (b) Distribution of similarity on the Baby dataset, and (c) Missing co-occurrence.

efficacy of the subsequent graph convolution, ultimately leading to semantic drift in the item representation learning.

To address such limitations, we propose a novel preference reasoning paradigm with graph refinement (MLLMRec) for multimodal recommendation. Specifically, to ensure high-quality initialization of user representations, we first utilize the MLLM to convert the item images into semantic descriptions, thereby facilitating the cross-modal alignment. Then, these visual descriptions are integrated with the textual metadata to form the comprehensive multimodal descriptions of items. Distinct from conventional feature extraction, we embed these item-level semantic descriptions into the historical interaction sequences to construct the natural language descriptions of the user behaviors. Building upon this, we leverage the advanced cognitive priors of the MLLM to reason over these descriptions, distilling latent interaction intents while effectively filtering modality-inherent noise to yield the behavior-aware and robust initial user preference profiles. These profiles, together with the multimodal descriptions of items and their ID embeddings, are then propagated over the user-item interaction graph to absorb the high-order collaborative signals. In addition, to rectify the structural noise in the item-item graph, we design two refinement strategies. A threshold-controlled denoising strategy is implemented to prune the spurious edges with low similarity, while a topology-aware enhancement strategy is developed to recover the latent audience co-occurrence correlations. Together, these strategies construct a more reliable item-item graph topology, providing a robust structural foundation for the subsequent graph convolution to capture high-order features from both the multimodal and ID item representations with greater precision. Finally, the multimodal and ID representations are concatenated to form the final user and item embeddings, and an InfoNCE contrastive loss is employed to align the two views. Extensive experiments on three datasets including Baby, Sports, and Clothing validate the effectiveness and robustness of MLLMRec.

Our main contributions can be summarized as follows:

- We propose an innovative multimodal recommendation paradigm that utilizes the MLLM to reason about the user preference profiles, thereby distilling the latent interaction intents and filtering out the modality-inherent noise. Moreover, these reasoned representations are propagated together with the ID embeddings over the user-item graph and aligned via an InfoNCE contrastive loss, yielding robust final user and item embeddings.
- We develop two plug-and-play graph refinement strategies, comprising threshold-controlled denoising and topology-aware enhancement, to eliminate the low-similarity semantic edges and recover the latent co-occurrence links, respectively, thereby constructing the more reliable item-item graph.
- Extensive experiments conducted on three publicly available datasets demonstrate that MLLMRec achieves the state-of-the-art performance on most evaluation metrics, and the ablation studies verify the effectiveness of each module.

2 RELATED WORKS

2.1 Multimodal Recommendation

Unlike conventional recommender systems that rely solely on the user-item interactions, MMRS leverage the rich item-side modalities to facilitate a more comprehensive understanding of user preferences [3, 36]. This field has evolved through several key paradigms. Early efforts, such as VBPR [10], introduce the visual features as auxiliary signals to bolster the collaborative filtering (CF). With the advent of GCNs, researchers shift toward the high-order representation learning. For instance, MMGCN [33] performs message passing on the modality-specific user-item interaction graphs. To further capture the latent item-side correlations, subsequent works like LATTICE [40] and FREEDOM [41] construct the item-item semantic graphs to perform GCNs. However, these methods ignore the different impacts of various modalities on the user decisions. To address this, MGCN [38] incorporates the behavior-aware modal weights, while GUME [16] explicitly models user interests through the extended modality interactions. More recently, self-supervised learning (SSL) has been introduced into MMRS to mitigate the data sparsity and the modal noise. For example, MMSSL [30] employs data augmentation and contrastive learning to mine the inter-modal dependencies. Moreover, MENTOR [35] argues that the excessive modality alignment may lead to the loss of interaction information, proposing a multi-level alignment strategy instead. ModalSync [18] further integrates the supervised and self-supervised paradigms, achieving the synergistic optimization of the behavior modeling and multimodal learning. However, these methods primarily focus on the item feature extraction while resorting to the shallow and incorrect initializations for the user representations. Furthermore, the constructed item-item graph is noisy and incomplete, which impede the learning of robust item representations.

2.2 LLMs-Enhanced Recommendation

With the superior capabilities demonstrated by LLMs in natural language processing tasks, researchers have increasingly explored their potential applications in the recommender systems [17, 28, 39]. Early explorations, such as P5 [6], directly employs a series of prompts to convert various recommendation metadata into a generative sequence-to-sequence framework. Chat-Rec [5] feeds the user behavior sequences into Chat-GPT to generate the personalized recommendations. Subsequently, to align the vast world knowledge of LLMs with the recommendation task, TALLRec [1] uses LoRA [12] for the parameter-efficient fine-tuning of LLaMA, thereby improving its applicability to recommendation. Moreover, ED^2 [37] proposes a dual dynamic indexing mechanism that enables the synergistic fusion between semantic knowledge and historical interactions. Moreover, to overcome the inherent knowledge limitations of LLMs, RecMind [29] integrates the database querying and web searching to improve the reasoning accuracy, while introducing the self-inspiring learning for better planning. Furthermore, CIKGRec [13] designs a cross-domain contrastive learning framework that aligns the auxiliary information domain with the recommendation task. Recently, several studies pivot toward MLLMs to handle richer content formats. For instance, DOGE [20] employs MLLMs to translate the item images into the descriptive text, effectively unifying disparate the modal granularities. Meanwhile, NEGGEN [14] leverages the generative prowess of MLLMs to synthesize the modality balanced negative samples, thereby providing the robust supervisory signals. However, the existing MLLMs-based methods primarily treat MLLMs as the sophisticated feature translators or data augmentors. They overlook the potential of MLLMs to perform preference reasoning by synthesizing the multimodal contents and the interaction patterns.

3 TASK FORMULATION

Let $\mathcal{U} = \{u\}$ and $\mathcal{I} = \{i\}$ denote the sets of users and items, respectively. Their observed historical interactions are represented by a Boolean matrix $\mathbf{R} = [r_{u,i}] \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $r_{u,i} = 1$ indicates an implicit feedback between user u and item i, and $r_{u,i} = 0$ otherwise. Beyond the collaborative backbone, each item $i \in \mathcal{I}$ is associated with the multimodal attributes $\mathcal{M} = \{v, t\}$, where v and t denote the visual and textual domains, respectively [16, 41]. In addition, we represent the collection of item images as $\mathcal{D}^v = \{i^v\}$ and the corresponding textual descriptions as $\mathcal{D}^t = \{i^t\}$.

Given the interaction matrix $\mathbf{R}$ and the multimodal item metadata $\{\mathcal{D}^v, \mathcal{D}^t\}$, the objective of our work is to learn a predictive function $\mathcal{F} : (\mathbf{R}, \mathcal{D}^v, \mathcal{D}^t) \rightarrow \hat{y}_{ui}$, where $\hat{y}_{ui} \in [0, 1]$ denotes the predicted preference score of user u toward item i. The framework $\mathcal{F}$ aims to capture the synergistic effects between collaborative signals and multimodal semantics. Finally, items are ranked in descending order of their predicted scores $\hat{y}_{ui}$ to generate a personalized top-K recommendation list for each user.

4 METHODOLOGY

Figure 2 illustrates the overall framework of our proposed MLLMRec, which synergizes the reasoning stages, the collaborative representation learning, and the refined item-item graph learning for multimodal recommendation. In the subsequent sections, we elaborate on the design of each component.

4.1 Reasoning Strategy

To bridge the semantic gap across heterogeneous modalities and homogenize the data granularities, we first convert the item images into semantic descriptions. Subsequently, we construct the personalized prompts for each user to guide the MLLM in reasoning

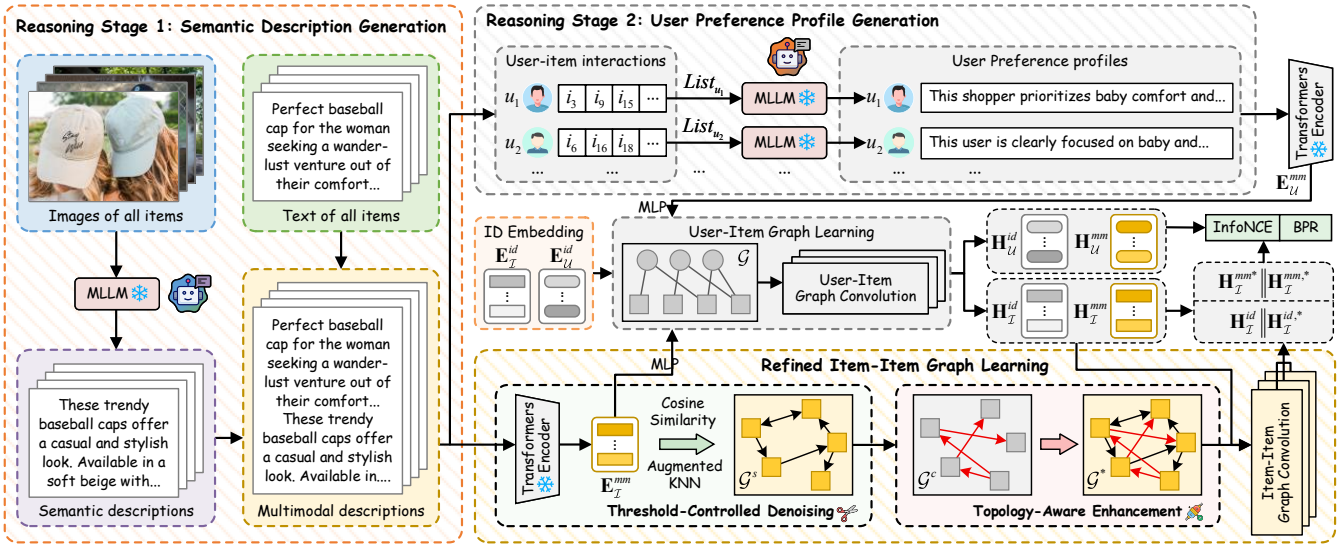


Figure 2: The overall framework of MLLMRec. First, the MLLM transforms the images into the semantic descriptions, which are fused with the text to yield the multimodal descriptions. Next, MLLMRec constructs the behavioral description lists for users, which are fed into the MLLM to reason about the user preference profiles. Then, the reasoned profiles and the multimodal descriptions, together with the ID embeddings, are propagated over the user-item interaction graph to absorb the collaborative signals. Meanwhile, MLLMRec refines the item-item graph through two designed strategies, which are applied to both the multimodal and ID item representations. Finally, the collaborative and multimodal representations are concatenated and aligned via the InfoNCE loss, and optimized by the BPR loss.

about the complex behavioral patterns from the user-item historical interactions, thereby distilling the purified user preference profiles that contain the latent interaction intents.

4.1.1 Semantic Description Generation. Leveraging the sophisticated cross-modal understanding of MLLMs, MLLMRec generates the high-quality semantic descriptions for item images, which is formulated as follows:

$$i^s = \text{MLLM}\left(\text{Prompt}_1, i^v\right), \tag{1}$$

where $i^s \in \mathcal{D}^s$ denotes the generated semantic description of image $i^v \in \mathcal{D}^v$, and the specific template of Prompt_1 is as follows:

Prompt_1: Semantic Description Generation

{*image*}: Please convert the given image into an accurate and concise textual description relevant to the {dataset name}, focusing on extracting key attributes that can influence the buying behavior of users, such as color, material, style, functionality, etc. To generate the textual description using a one-paragraph natural language overview in no more than 100 words.

Input: {*An Image*}
Output: {A paragraph of semantic descriptions for this image}

This stage is pivotal for two reasons. First, conventional visual encoders (e.g., VGG16 [24], ResNet [9]) often capture general features that may be tangential to the specific recommendation contexts. By contrast, MLLMRec designs a task-oriented prompt to guide the MLLM in selectively extracting the decision-relevant visual cues, such as aesthetic style. This yields the semantic descriptions that are aligned with the recommendation task and filter out the noise of visual modality. Second, this generative process projects the visual information into the linguistic space, effectively unifying the data formats of visual and textual modalities. Such homogenization directly addresses the modality imbalance problem, which is a common bottleneck in the multimodal learning where models disproportionately rely on textual data due to the high noise-to-signal ratio and the semantic deficiency of raw visual features [14, 20]. In summary, by transforming the images into the structured semantic description, we ensure that visual signals contribute equitably alongside textual signals to the preference learning.

Subsequently, we integrate i^s with its raw textual metadata i^t to construct a comprehensive multimodal description for each item. The fusion process is formulated as follows:

$$i^m = i^t \oplus i^s, \tag{2}$$

where $i^m \in \mathcal{D}^m$ denotes the multimodal description of item i, and $\oplus$ represents the concatenation operation. By integrating abstract visual descriptions with explicit textual metadata, facilitates the seamless fusion of multimodal information and effectively bridges the cross-modal semantic gap. Consequently, it establishes a purified and enriched multimodal knowledge base that serves as a robust foundation for the subsequent module.

4.1.2 User Preference Profile Generation. Distinguishing from conventional paradigms that rely on the behavior-agnostic random embeddings or the noise-contaminated feature aggregation, MLLMRec capitalizes on the advanced cognitive reasoning of MLLMs to transform the raw interaction data into the purified user preference

profiles. These profiles provide a high-fidelity semantic foundation for robust user representation learning. Specifically, we first synthesize a behavioral descriptions list for each user by chronologically aggregating the multimodal descriptions of their historically interacted items. This process is formulated as follows:

$$List_u = [i^m], \quad r_{u,i} = 1, \tag{3}$$

where $List_u$ denotes the behavioral description list of user u, and i^m is the multimodal description of item i that user u has interacted with. Subsequently, we embed $List_u$ into a meticulously designed prompt to guide the MLLM in reasoning about the user's profiles. The generative process is formulated as:

$$u^p = \text{MLLM}\left(\text{Prompt}_2, List_u\right), \tag{4}$$

where $u^p \in \mathcal{D}^p$ denotes the generated user preference profiles for user u, and the specific template of Prompt_2 is as follows:

Prompt_2: User Preference Profile Generation

Please reason about the user preferences based on the following list of item descriptions that he or she has interacted with. The list is: {*behavioral description*}. To generate the user preferences using a one-paragraph natural language in no more than 100 words.

Input: {*User's behavioral description list*}
Output: {User's preference profiles in natural language}

By utilizing the extensive world knowledge inherent in MLLMs, the resulting user preference profile set $\mathcal{D}^p$ offers a more real user interests compared to the conventional paradigms. It is noteworthy that since $List_u$ is derived from the interaction matrix $\mathbf{R}$, $\mathcal{D}^p$ already captures the latent motivations behind the user-item interactions. This inherent information richness provides a high-fidelity semantic foundation, which is further enhanced by the collaborative signals introduced in the subsequent learning stage.

4.2 Collaborative Representation Learning

To inject the collaborative signals into the preference reasoning paradigm, we design a dual-stream representation learning that jointly models the ID-based collaborative stream and the MLLM-driven multimodal stream, and then propagates both streams over the user-item interaction graph to obtain the high-order representations. Specifically, we initialize the trainable ID embeddings for users and items, denoted as $\mathbf{E}^{id}_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{U}|\times d}$ and $\mathbf{E}^{id}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|\times d}$, respectively, via the Xavier initialization, which serve as the initial representations of the collaborative stream:

$$\mathbf{H}^{id,(0)}_{\mathcal{U}} = \mathbf{E}^{id}_{\mathcal{U}}, \quad \mathbf{H}^{id,(0)}_{\mathcal{I}} = \mathbf{E}^{id}_{\mathcal{I}}. \tag{5}$$

Meanwhile, for the multimodal stream, we first leverage the pre-trained text encoder to project the user preference profiles $\mathcal{D}^p$ and the item multimodal descriptions $\mathcal{D}^m$ into the latent space, which is formulated as follows:

$$\mathbf{E}^{mm}_{\mathcal{U}} = \text{Encoder}\left(\mathcal{D}^p\right), \quad \mathbf{E}^{mm}_{\mathcal{I}} = \text{Encoder}\left(\mathcal{D}^m\right), \tag{6}$$

where $\mathbf{E}^{mm}_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{U}|\times d_t}$ and $\mathbf{E}^{mm}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|\times d_t}$ denote the extracted semantic features of users and items, respectively. To compress these high-dimensional features into a compact space, we further employ the multilayer perceptrons (MLPs) for the non-linear transformation, obtaining the initial representations of the multimodal stream. Taking the user side as an example:

$$\mathbf{H}^{mm,(0)}_{\mathcal{U}} = \sigma\left(\mathbf{E}^{mm}_{\mathcal{U}}\mathbf{W}_1 + \mathbf{b}_1\right)\mathbf{W}_2 + \mathbf{b}_2, \tag{7}$$

where $\mathbf{H}^{mm,(0)}_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{U}|\times d}$ denotes the initial user representations of the multimodal stream, $\sigma(\cdot)$ is the LeakyReLU activation function. The weight matrices $\mathbf{W}_1 \in \mathbb{R}^{d_t\times d_1}$, $\mathbf{W}_2 \in \mathbb{R}^{d_1\times d}$ and bias vectors $\mathbf{b}_1 \in \mathbb{R}^{d_1}$, $\mathbf{b}_2 \in \mathbb{R}^{d}$ are trainable parameters, where d_1 represents the dimension of the hidden layer, respectively. Similarly, the item representations $\mathbf{H}^{mm,(0)}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|\times d}$ are derived by applying an analogous projection to the item features $\mathbf{E}^{mm}_{\mathcal{I}}$.

Thereafter, we propagate both streams over the user-item interaction graph following the method in [41] to capture the high-order collaborative signals. Formally, we construct the normalized adjacency matrix $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{A} \in \{0,1\}^{(|\mathcal{U}|+|\mathcal{I}|)\times(|\mathcal{U}|+|\mathcal{I}|)}$ denotes the bipartite user-item adjacency matrix and $\mathbf{D}$ is the corresponding degree matrix. The graph convolution at the l-th layer is formulated as follows:

$$\left[\mathbf{H}^{o,(l)}_{\mathcal{U}}; \mathbf{H}^{o,(l)}_{\mathcal{I}}\right] = \tilde{\mathbf{A}}\left[\mathbf{H}^{o,(l-1)}_{\mathcal{U}}; \mathbf{H}^{o,(l-1)}_{\mathcal{I}}\right], \tag{8}$$

where $o \in \{id, mm\}$ indexes the two streams and $[\cdot;\cdot]$ denotes the vertical stacking operation. After L_{ui} layers, we aggregate the multi-layer representations via summation as follows:

$$\mathbf{H}^{o}_{\mathcal{U}} = \sum_{l=0}^{L_{ui}} \mathbf{H}^{o,(l)}_{\mathcal{U}}, \quad \mathbf{H}^{o}_{\mathcal{I}} = \sum_{l=0}^{L_{ui}} \mathbf{H}^{o,(l)}_{\mathcal{I}}. \tag{9}$$

4.3 Refined Item-Item Graph Learning

To mitigate the semantic drift in the item representation learning caused by the suboptimal item-item graph topology, we design two graph refinement strategies comprising (1) threshold-controlled denoising and (2) topology-aware enhancement.

4.3.1 Threshold-Controlled Denoising. Based on the extracted item multimodal features $\mathbf{E}^{mm}_{\mathcal{I}}$ obtained from Eq. (6), following the established practice [41], we evaluate the semantic affinities between items by computing the cosine similarity of their multimodal features. This process is formulated as follows:

$$s_{a,b} = \frac{\left(\mathbf{e}^{mm}_{i_a}\right)^{\mathrm{T}}\mathbf{e}^{mm}_{i_b}}{\left\|\mathbf{e}^{mm}_{i_a}\right\| \cdot \left\|\mathbf{e}^{mm}_{i_b}\right\|}, \tag{10}$$

where $s_{a,b}$ denotes the similarity score between item a and item b, $\|\cdot\|$ represents the vector norm computation, and $(\cdot)^{\mathrm{T}}$ is the matrix transpose operation.

While the conventional KNN strategy [2] is widely adopted to sparsify the similarity matrix, it suffer from a rigid fixed-degree constraint. This constraint forces items lacking the genuine semantic neighbors to establish the spurious connections with the low-similarity nodes. As demonstrated by our empirical analysis in Figure 1(b), this phenomenon leads to an obvious presence of the false-positive edges (exceeding 10% in the Baby dataset). To rectify this, we introduce a threshold-controlled denoising strategy that combines the top-K selection with a similarity constraint α. This ensures that only highly similar neighbors are retained for each

item. This strategy is formulated as follows:

$$\tilde{s}_{a,b} = \begin{cases} 1, & s_{a,b} \in \text{top-}K_s(\mathbf{s}_{a,:}) \text{ and } s_{a,b} \geq \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

where vector $\mathbf{s}_{a,:} \in \mathbb{R}^{|\mathcal{I}|}$ represents the similarity scores between item i_a and all items, the hyper-parameters K_s and α control the top-K value and the pruning intensity, respectively. The resulting matrix $\tilde{\mathbf{S}} = [\tilde{s}_{a,b}] \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{I}|}$ serves as the adjacency matrix for the denoised item-item semantic affinity graph $\mathcal{G}^s = \{\mathcal{I}, \mathcal{E}^s\}$, where $\mathcal{I}$ is the set of item nodes and $\mathcal{E}^s = \{\langle i_a, i_b\rangle \,|\tilde{s}_{a,b} = 1\}$ denotes the edge set. This formulation effectively filters out structural noise while preserving high-confidence semantic correlations.

4.3.2 Topology-Aware Enhancement. We argue that the semantic graph $\mathcal{G}^s$ constructed exclusively from the feature similarities fails to capture the item-item co-occurrence correlations that are essential collaborative signals embedded within the user-item interactions [19]. To bridge this gap, we introduce a topology-aware enhancement strategy to recover these missing behavioral links. Specifically, we quantify the audience co-occurrence between item pairs by computing the Jaccard similarity coefficient over their respective interacted user sets. For any two items i_a and i_b, their co-occurrence score $c_{a,b}$ is formulated as follows:

$$c_{a,b} = J\left(\mathbf{r}_{:,a}, \mathbf{r}_{:,b}\right) = \frac{\left|\mathbf{r}_{:,a} \cap \mathbf{r}_{:,b}\right|}{\left|\mathbf{r}_{:,a} \cup \mathbf{r}_{:,b}\right|}, \tag{12}$$

where $\mathbf{r}_{:,a} \in \mathbb{R}^{|\mathcal{U}|}$ and $\mathbf{r}_{:,b} \in \mathbb{R}^{|\mathcal{U}|}$ are column vectors of the interaction matrix $\mathbf{R}$, denoting the sets of users who have interacted with item i_a and i_b, respectively. The resulting matrix $\mathbf{C} = [c_{a,b}] \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ represents the audience co-occurrence matrix. Given the sparsity of interaction data, most item pairs exhibit limited co-occurrence. Consequently, to ensure structural robustness, we utilize a weighted KNN strategy to retain only the top-K neighbors with the highest co-occurrence scores for each item, which is formulated as follows:

$$\tilde{c}_{a,b} = \begin{cases} c_{a,b}, & c_{a,b} \in \text{top-}K_c(\mathbf{c}_{a,:}), \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where vector $\mathbf{c}_{a,:} \in \mathbb{R}^{|\mathcal{I}|}$ represents the audience co-occurrence scores between item i_a and all items, and hyper-parameter K_c controls the density of co-occurrence connections. This resulting weighted matrix $\tilde{\mathbf{C}} = [\tilde{c}_{a,b}] \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ serves as the adjacency matrix for the item-item audience co-occurrence graph $\mathcal{G}^c = \{\mathcal{I}, \mathcal{E}^c\}$, where $\mathcal{I}$ is the set of item nodes and $\mathcal{E}^c = \{\langle i_a, i_b\rangle \,|\, \tilde{c}_{a,b} > 0\}$ denotes the edge set with non-zero co-occurrence scores.

Finally, to establish a holistic representation of item-side relationships, we perform a graph fusion by integrating the semantic affinity graph with the audience co-occurrence graph, which is formulated as follows:

$$\mathbf{S}^* = \tilde{\mathbf{S}} + \tilde{\mathbf{C}}, \tag{14}$$

where $\mathbf{S}^* = [s^*_{a,b}] \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ serves as the adjacency matrix of the topology-enhanced item-item graph $\mathcal{G}^* = \{\mathcal{I}, \mathcal{E}^*\}$. The fused edge set $\mathcal{E}^* = \mathcal{E}^s \cup \mathcal{E}^c$ effectively integrates the multimodal semantics with the collaborative signals.

4.3.3 Item-Item Graph Convolution. Building upon the refined graph $\mathcal{G}^*$, we employ LightGCN [11], a simplified and effective variant of GCNs, to perform the neighborhood aggregation on the item representations of both streams obtained from the user-item graph convolution.

This process enables MLLMRec to capture the high-order correlations among items. Specifically, the graph convolution operation at the l-th layer is formulated as follows:

$$\mathbf{H}_{\mathcal{I}}^{o,(l)} = \left((\mathbf{N})^{-\frac{1}{2}}\mathbf{S}^*(\mathbf{N})^{-\frac{1}{2}}\right)\mathbf{H}_{\mathcal{I}}^{o,(l-1)}, \tag{15}$$

where $o \in \{id, mm\}$ indexes the two streams, $\mathbf{H}_{\mathcal{I}}^{o,(l)} \in \mathbb{R}^{|\mathcal{I}|\times d}$ represents the item representations of stream o at the l-th layer, $\mathbf{H}_{\mathcal{I}}^{o,(0)} = \mathbf{H}_{\mathcal{I}}^{o}$ denotes the item representations output by the user-item graph convolution, $\mathbf{N}$ denotes the degree matrix of graph $\mathcal{G}^*$ and $\mathbf{N}_{aa} = \sum_b s^*_{a,b}$. After that, we aggregate the representations across all hidden layers through a summation function, which is formulated as follows:

$$\mathbf{H}_{\mathcal{I}}^{o,*} = \sum_{l=0}^{L_{ii}} \mathbf{H}_{\mathcal{I}}^{o,(l)}, \tag{16}$$

where $\mathbf{H}_{\mathcal{I}}^{o,*} \in \mathbb{R}^{|\mathcal{I}|\times d}$ denotes the refined high-order item representations of stream o, and hyper-parameter L_{ii} controls the depth of the graph convolution.

4.4 Fusion and Optimization

To combine the complementary signals from the two streams, we concatenate the collaborative and multimodal features to form the final user and item embeddings, which is formulated as follows:

$$\mathbf{H}_{\mathcal{U}} = \left[\mathbf{H}_{\mathcal{U}}^{id} \,\|\, \beta \cdot \mathbf{H}_{\mathcal{U}}^{mm}\right], \quad \mathbf{H}_{\mathcal{I}} = \left[\mathbf{H}_{\mathcal{I}}^{id,*} \,\|\, \beta \cdot \mathbf{H}_{\mathcal{I}}^{mm,*}\right], \tag{17}$$

where $\mathbf{H}_{\mathcal{U}}, \mathbf{H}_{\mathcal{I}} \in \mathbb{R}^{2d}$ denote the final representations of users and items, respectively, $\|$ denotes the concatenation operation, and β is a hyper-parameter used to control the weight of the multimodal stream. To further align the two views, we introduce an InfoNCE contrastive loss [8] between the collaborative and multimodal representations. Taking the user side as an example, the contrastive objective is formulated as follows:

$$\mathcal{L}_{cl}^{u} = -\frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \log \frac{\exp\left(\mathbf{H}_{\mathcal{U}}^{id}[b] \cdot \mathbf{H}_{\mathcal{U}}^{mm}[b]/\tau\right)}{\sum_{b'\in\mathcal{B}} \exp\left(\mathbf{H}_{\mathcal{U}}^{id}[b] \cdot \mathbf{H}_{\mathcal{U}}^{mm}[b']/\tau\right)}, \tag{18}$$

where $\mathcal{B}$ denotes the training batch, τ is the temperature hyper-parameter, and $\mathbf{H}_{\mathcal{U}}^{id}[b]$ and $\mathbf{H}_{\mathcal{U}}^{mm}[b]$ are the collaborative and multimodal representations of the b-th user, respectively. The item-side contrastive loss $\mathcal{L}_{cl}^{i}$ is defined analogously, and the overall contrastive loss is $\mathcal{L}_{cl} = \mathcal{L}_{cl}^{u} + \mathcal{L}_{cl}^{i}$. Finally, we optimize the model parameters via the Bayesian Personalized Ranking (BPR) loss [23], which encourages the model to assign higher preference scores to the observed interactions than to unobserved ones:

$$\mathcal{L}_{BPR} = -\sum_{(u,i_a,i_b)\in\mathcal{T}} \log\left(\sigma\left(\hat{y}_{ui_a} - \hat{y}_{ui_b}\right)\right), \tag{19}$$

where $\hat{y}_{ui} = \mathbf{h}_u^{\mathrm{T}} \cdot \mathbf{h}_i$ represents the predicted preference score for user u on item i, and $\mathcal{T} = \{(u, i_a, i_b) \mid r_{u,i_a} = 1, r_{u,i_b} = 0\}$ denotes the set of training triplets, where i_a is an observed item and i_b is

Table 1: Statistics of the experimental datasets.

Datasets	#Users	#Items	#Interactions	Density
Baby	19,445	7,050	160,792	0.117%
Sports	35,598	18,357	296,337	0.045%
Clothing	39,387	23,033	278,677	0.031%

a negative sample randomly drawn from the unobserved set. The overall training loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda_{cl}\mathcal{L}_{cl}, \tag{20}$$

where λ_{cl} is the hyper-parameter balancing the contrastive loss. Ultimately, items are ranked in descending order of $\hat{y}_{u,i}$ to generate the top-K recommendation list for each user, where K denotes the length of the recommendation list.

5 EXPERIMENTS

5.1 Experiment Setup

5.1.1 Datasets. Following [34, 41], we conduct experiments on three widely used datasets from the Amazon product repository[1]: (i) Baby, (ii) Sports and Outdoors, and (iii) Clothing, Shoes, and Jewelry (referred to as Baby, Sports, and Clothing, respectively). These datasets cover diverse product categories, interaction sparsities, and scales of user and items, alongside rich multimodal metadata (i.e., item images and textual descriptions). To ensure the reliability of collaborative signals, we apply the 5-core filtering strategy to filter out the users and items with fewer than five interactions. The statistics of the datasets after filtering are summarized in Table 1. For model training, each dataset is randomly partitioned into training, validation and testing sets with a ratio of 8:1:1.

5.1.2 Evaluation Metrics. We adopt two widely recognized top-K recommendation metrics to quantify model performance: Recall (R@K) and Normalized Discounted Cumulative Gain (NDCG, N@K). Specifically, Recall measures the fraction of the testing items successfully recommended, while NDCG accounts for ranking quality by assigning higher weights to relevant items positioned earlier in the list. Consistent with prior works, we report the average scores across all users in the testing set under $K \in \{10, 20\}$.

5.1.3 Baselines. To demonstrate the superiority of MLLMRec, we compare it against the following three categories of representative baselines: CF-based recommenders, general multimodal recommenders, and MLLMs-based multimodal recommenders. A brief description of each baseline is provided below:

i) CF-based recommenders:

- **BPR** [23] establishes a foundational Bayesian ranking framework that utilizes the pairwise comparison of items to capture the personalized user preferences from the implicit feedback.
- **LightGCN** [11] simplifies the vanilla GCN by omitting the redundant non-linear activations and feature transformations.

ii) general multimodal recommenders:

- **VBPR** [10] integrates the visual features from item images into the BPR model, enabling the capture of visual preferences.
- **LATTICE** [40] constructs modality-aware graphs to mine the latent item-item semantic from multimodal features.
- **SLMRec** [26] employs multimodal data augmentation and contrastive learning to capture the cross-modal patterns.
- **FREEDOM** [41] freezes the item-item graph to improve efficiency and incorporates a degree-sensitive pruning strategy to denoise the user-item interaction graph.
- **LGMRec** [7] combines local topological learning with a global hyper-graph module to capture the high-order dependencies.
- **GUME** [16] utilizes mutual information maximization to align explicit interaction features with extended interest profiles.
- **ModalSync** [18] synchronizes multimodal feature extraction with user behaviors using a staged co-training strategy.
- **COHESION** [34] integrates modality fusion and representation learning via a dual-stage fusion strategy and a composite GCN.

iii) MLLMs-based multimodal recommenders:

- **LLM-CF** [25] introduce LLMs into collaborative filtering through a novel instruction-tuning method and in-context learning.
- **DOGE** [20] leverages MLLMs to generate cross-modal representations by aligning the image and text information.
- **NEGGEN** [14] uses MLLMs to construct robust negative samples via tailored prompting, providing better contrastive signals.

Table 2: The optimal hyper-parameter settings of MLLMRec.

Dataset	α	K_c	β	λ_{cl}
Baby	0.7	15	0.0001	0.01
Sports	0.7	10	0.01	0.001
Clothing	0.7	15	0.01	0.001

5.1.4 Implementation Details. In the reasoning phase, we employ Qwen2.5vl-7b as our MLLM, accessed via the Ollama framework. For fair comparisons with the existing multimodal baselines, we utilize Sentence-Transformer [22] as the text encoder. Specifically, the training batch size is 2048 and the learning rate is 0.0001. All trainable parameters are initialized via the Xavier method and optimized by the Adam optimizer. We set the output dimension d to 512, the hidden layer dimension d_1 to 2,048, the semantic neighbor count K_s to 10, and the convolution depth L_{ui} and L_{ii} to 2 and 1, respectively. For the similarity threshold α, we perform a grid search in $\{0.5, 0.6, 0.7, 0.8\}$, and the co-occurrence neighbor count K_c is searched over $\{5, 10, 15, 20\}$. The multimodal stream weight β and the contrastive loss weight λ_{cl} are tuned in $\{0.0001, 0.001, 0.01, 0.1\}$, and the temperature τ is set to 0.2. In addition, to prevent overfitting, the training process is terminated if R@20 on the validation set does not improve for 20 consecutive epochs, with a cap training limit of 1,000 epochs. All experiments are implemented using PyTorch and evaluated on a 24GB NVIDIA RTX4090 GPU.

The optimal values of the dataset-specific hyper-parameters (i.e., α, K_c, β, and λ_{cl}) are summarized in Table 2.

5.2 Overall Performance

Table 3 presents the experimental results of the baselines and MLLMRec on three datasets, with the following key findings:

(1) Multimodal signals enhance the preference modeling. Integrating the modal representations substantially improves the

[1] http://jmcauley.ucsd.edu/data/amazon

Table 3: Overall performance of the baselines and MLLMRec on three datasets, with the best and second-best results highlighted by bold and underline, respectively.

Models	Baby				Sports				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 (UAI'09)	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0187	0.0279	0.0103	0.0126
LightGCN (SIGIR'20)	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0340	0.0526	0.0188	0.0236
VBPR (AAAI'16)	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 (MM'21)	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 (TMM'23)	0.0540	0.0810	0.0285	0.0357	0.0676	0.1017	0.0374	0.0462	0.0452	0.0675	0.0247	0.0303
FREEDOM (MM'23)	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0629	0.0941	0.0341	0.0420
LGMRec (AAAI'24)	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0555	0.0828	0.0302	0.0371
GUME (CIKM'24)	0.0673	0.1042	0.0365	0.0460	0.0778	0.1165	0.0427	0.0527	0.0703	0.1024	0.0384	0.0466
ModalSync (WWW'25)	0.0693	0.1045	0.0373	0.0462	<u>0.0821</u>	<u>0.1189</u>	**0.0457**	<u>0.0551</u>	<u>0.0738</u>	<u>0.1058</u>	<u>0.0406</u>	<u>0.0488</u>
COHESION (SIGIR'25)	0.0680	0.1052	0.0354	0.0454	0.0752	0.1137	0.0409	0.0503	0.0665	0.0983	0.0358	0.0438
LLM-CF (CIKM'24)	0.0593	0.0874	0.0312	0.0409	0.0698	0.1057	0.0376	0.0465	0.0596	0.0886	0.0329	0.0402
DOGE (AAAI'25)	<u>0.0718</u>	<u>0.1096</u>	<u>0.0391</u>	<u>0.0482</u>	0.0792	0.1185	0.0435	0.0536	0.0712	0.1036	0.0389	0.0468
NEGGEN (MM'25)	0.0701	0.1065	0.0342	0.0438	0.0763	0.1114	0.0411	0.0506	0.0654	0.0961	0.0350	0.0441
MLLMRec (Ours)	**0.0722**	**0.1105**	**0.0396**	**0.0491**	**0.0823**	**0.1223**	<u>0.0454</u>	**0.0575**	**0.0742**	**0.1075**	**0.0410**	**0.0493**

performance of CF-based recommenders. A direct comparison between BPR and its modality extension VBPR reveals that the latter achieves superior results across all datasets. This confirms that relying solely on the collaborative signals is insufficient to capture the fine-grained user preferences due to the interaction sparsity. Multimodal information, such as visual cues, provides a critical semantic supplement, effectively enriching the representation space.

(2) MLLMs promote the multimodal feature learning. Compared to the general multimodal recommenders, MLLMs-based methods exhibit competitive performance. For instance, DOGE achieves the best baseline results on the Baby dataset. This competitiveness is attributed to the exceptional cognitive reasoning and cross-modal alignment abilities of MLLMs, which facilitate a paradigm shift from the shallow feature extraction to the deep semantic understanding. By transforming the raw data into high-level descriptors, these models better capture the core features, leading to more precise matching in complex recommendation tasks.

(3) MLLMRec achieves the state-of-the-art performance. Specifically, MLLMRec achieves the best performance on 11 of the 12 evaluation metrics across the three datasets, and the second-best result on the remaining one. These results demonstrate that jointly modeling the MLLM-reasoned multimodal semantics and the collaborative signals, together with the item-item graph refinement, effectively enhances the recommendation performance.

(4) The collaborative and multimodal signals are complementary. Compared with the CF-based baselines (e.g., BPR and LightGCN) that rely solely on the collaborative signals, and the MLLM-based baselines (e.g., DOGE and NEGGEN) that exploit the multimodal semantics, MLLMRec integrates both streams through the user-item graph convolution and aligns them via the InfoNCE contrastive loss. The consistent gains over both categories of baselines demonstrate that the collaborative and multimodal signals are complementary, and their synergistic integration is beneficial for multimodal recommendation.

Table 4: Performance of MLLMRec and its variants in terms of R@20 and N@20 on three datasets.

Model Variant	Baby		Sports		Clothing	
	R@20	N@20	R@20	N@20	R@20	N@20
MLLMRec	**0.1105**	**0.0491**	**0.1223**	**0.0575**	**0.1075**	**0.0493**
w/o RS	0.0990	0.0438	0.1156	0.0524	0.1012	0.0459
w/o GD	0.1094	0.0487	0.1214	0.0550	0.1068	0.0486
w/o TE	0.1087	0.0480	0.1209	0.0546	0.1061	0.0477
w/o GCN_{II}	0.1037	0.0465	0.1030	0.0464	0.0957	0.0432
w/o CL	0.0998	0.0442	0.1174	0.0531	0.1023	0.0464

5.3 Ablation Study

To evaluate the performance contributions of each core component, we design the following five model variants: (1) w/o RS, which removes the reasoning strategy (i.e., the multimodal stream) and retains only the collaborative stream, (2) w/o GD, which excludes the threshold-controlled denoising strategy, (3) w/o TE, which abandons the topology-aware enhancement strategy, (4) w/o GCN_{II}, which removes the item-item graph convolution, and (5) w/o CL, which removes the InfoNCE contrastive loss. Table 4 lists the experimental results, and it reveals the following three observations:

(1) The absence of any core component leads to performance degradation. This confirms the necessity of the reasoning strategy, the two graph-refinement strategies, the item-item graph convolution, and the contrastive learning objective for the optimal performance of MLLMRec.

(2) The reasoning strategy and the item-item graph convolution are the two most critical components. Specifically, in terms of R@20, w/o RS suffers the largest degradation on Baby, whereas w/o GCN_{II} incurs the largest drop on Sports and Clothing. This validates the two pillars of MLLMRec: the MLLM-reasoned preference profiles provide high-quality user representations, and

the refined item-item graph captures the high-order relationships among items, both of which are indispensable for accurate preference modeling.

(3) The contrastive loss brings consistent gains, while the graph-refinement strategies play a supplementary role. w/o CL degrades the performance on all three datasets, demonstrating that aligning the collaborative and multimodal signals via the InfoNCE objective is beneficial. In contrast, w/o GD and w/o TE incur relatively smaller but steady drops, indicating that the threshold-controlled denoising and topology-aware enhancement strategies serve as effective complements that refine the item-item graph and further stabilize the representation learning.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose MLLMRec, a novel framework that redefines multimodal recommendation by synergizing the cognitive reasoning of MLLMs, the collaborative signals, and graph structural refinement. Specifically, we introduce a pioneering paradigm that leverages MLLMs to reason about high-fidelity user preference profiles, effectively capturing the latent interaction intents behind the user behaviors. These reasoned profiles, together with the multimodal descriptions of items and the ID embeddings, are propagated over the user-item interaction graph to absorb the high-order collaborative signals, and the two views are further aligned via an InfoNCE contrastive loss. Furthermore, to address the structural noise in the item-item graph, we design the threshold-controlled denoising and topology-aware enhancement strategies to refine the suboptimal item-item graph, which are applied to both the multimodal and ID item representations. Extensive experiments across three public datasets demonstrate that MLLMRec achieves the state-of-the-art performance on most evaluation metrics.

In the future, we aim to investigate temporally aware reasoning strategies to capture the evolving dynamics of user interests, while exploring the use of implicit reasoning tokens to bypass explicit textual steps, thereby improving the computational efficiency.

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces.

REFERENCES

[1] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 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*. 1007–1014.

[2] Jie Chen, Haw-ren Fang, and Yousef Saad. 2009. Fast Approximate *k*NN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *Journal of Machine Learning Research* 10 (2009), 1989–2012.

[3] Yuzhuo Dang, Zhiqiang Pan, Xin Zhang, Wanyu Chen, Fei Cai, and Honghui Chen. 2025. Discrepancy Learning Guided Hierarchical Fusion Network for Multi-modal Recommendation. *Knowledge-Based Systems* 317 (2025), 113496.

[4] Yuxiao Duan, Xiang Zhao, and Hao Guo. 2025. Sentiment-aware cross-modal semantic interaction model for harmful meme detection. *Decision Support Systems* 196 (2025), 114509.

[5] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. *arXiv preprint arXiv:2303.14524* (2023).

[6] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[7] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedingts of the 38th AAAI Conference on Artificial Intelligence*. 8454–8462.

[8] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th Artificial Intelligence and Statistics (AISTATS'10) (JMLR Proceedings, Vol. 9)*. JMLR.org, 297–304.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 29th IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. 144–150.

[11] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 639–648.

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the 10th International Conference on Learning Representations*.

[13] Zheng Hu, Zhe Li, Ziyun Jiao, Satoshi Nakagawa, Jiawen Deng, Shimin Cai, Tao Zhou, and Fuji Ren. 2025. Bridging the User-side Knowledge Gap in Knowledge-aware Recommendations with Large Language Models. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*. 11799–11807.

[14] Yanbiao Ji, Dan Luo, Chang Liu, Shaokai Wu, Jing Tong, Qichen He, Deyi Ji, Hongtao Lu, and Yue Ding. 2025. Generating Negative Samples for Multi-Modal Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6007–6016.

[15] Zhenyang Li, Fan Liu, Yinwei Wei, Zhiyong Cheng, Liqiang Nie, and Mohan S. Kankanhalli. 2024. Attribute-driven Disentangled Representation Learning for Multimodal Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 9660–9669.

[16] Guojiao Lin, Zhen Meng, Dongjie Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. 2024. GUME: Graphs and User Modalities Enhancement for Long-Tail Multimodal Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1400–1409.

[17] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2025. How Can Recommender Systems Benefit from Large Language Models: A Survey. *ACM Transactions on Information Systems* 43, 2 (2025), 28:1–28:47.

[18] Shiqin Liu, Chaozhuo Li, Minjun Zhao, Litian Zhang, and Jiajun Bu. 2025. ModalSync: Synchronizing User Behavior with Multimodal Features for Multimodal Pre-training Recommendation. In *Proceedings of the ACM Web Conference 2025*. 2278–2287.

[19] Yaokun Liu, Xiaowang Zhang, Minghui Zou, and Zhiyong Feng. 2023. Co-occurrence embedding enhancement for long-tail problem in multi-interest recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 820–825.

[20] Fanshen Meng, Zhenhua Meng, Ru Jin, Rongheng Lin, and Budan Wu. 2025. DOGE: LLMs-Enhanced Hyper-Knowledge Graph Recommender for Multimodal Recommendation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*. 12399–12407.

[21] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning Hybrid Behavior Patterns for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 376–384.

[22] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 3980–3990.

[23] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. 452–461.

[24] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[25] Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, and Jun Xu. 2024. Large language models enhanced collaborative filtering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2178–2188.

[26] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2023. Self-Supervised Learning for Multimedia Recommendation. *IEEE Transactions on Multimedia* 25 (2023), 5107–5116.

[27] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2023. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Transactions on Multimedia* 25 (2023), 1074–1084.

[28] Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Zhang, Qing Cui, Longfei Li, Jun Zhou, and Sheng Li. 2024. LLMRG: Improving Recommendations through Large Language Model Reasoning Graphs. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. 19189–19196.

[29] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024. RecMind: Large Language Model Powered Agent For Recommendation. In *Findings of the Association for Computational Linguistics*. 4351–4364.

[30] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.

[31] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. LLMRec: Large Language Models with Graph Augmentation for Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 806–815.

[32] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3541–3549.

[33] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.

[34] Jinfeng Xu, Zheyu Chen, Wei Wang, Xiping Hu, Sang-Wook Kim, and Edith CH Ngai. 2025. COHESION: Composite graph convolutional network with dual-stage fusion for multimodal recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1830–1839.

[35] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C. H. Ngai. 2025. MENTOR: Multi-level Self-supervised Learning for Multimodal Recommendation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*. 12908–12917.

[36] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig MacDonald. 2022. Multi-modal Graph Contrastive Learning for Micro-video Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1811.

[37] Jun Yin, Zhengxin Zeng, Mingzheng Li, Hao Yan, Chaozhuo Li, Weihao Han, Jianjin Zhang, Ruochen Liu, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, Shirui Pan, and Senzhang Wang. 2025. Unleash LLMs Potential for Sequential Recommendation by Coordinating Dual Dynamic Index Mechanism. In *Proceedings of the ACM on Web Conference 2025*. 216–227.

[38] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-View Graph Convolutional Network for Multimedia Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6576–6585.

[39] Zhenrui Yue, Sara Rabhi, Gabriel de Souza Pereira Moreira, Dong Wang, and Even Oldridge. 2023. LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking. *arXiv preprint arXiv:2311.02089* (2023).

[40] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[41] Xin Zhou and Zhiqi Shen. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 935–943.