Joint Multi-Modal Multi-Interest Profiling and Preference-Grounded Reasoning for Explainable Recommendation

Anqi Wang[1,2], **Jianye Xie**[1,2], **Weiming Liu**[3], **Rong Jiang**[4], **Lianyong Qi**[1,2*], **Haolong Xiang**[5], **Xiaolong Xu**[5], **Wenmin Lin**[6], **Yang Zhang**[7] and **Xiaokang Zhou**[8,9]
[1]College of Computer Science and Technology, China University of Petroleum (East China)
[2]Shandong Key Laboratory of Intelligent Oil and Gas Industrial Software
[3]College of Computer Science, Zhejiang University
[4]Yunnan Key Laboratory of Service Computing, Yunnan University of Finance and Economics
[5]School of Software, Nanjing University of Information Science and Technology
[6]Institute of VR and Intelligent Systems, Hangzhou Normal University
[7]Anuradha and Vikas Sinha Department of Data Science, University of North Texas
[8]Faculty of Business and Data Science, Kansai University
[9]RIKEN Center for Advanced Intelligence Project
wanganqi0618@gmail.com, jianyexie96@gmail.com, 21831010@zju.edu.cn, jiangrong@ynufe.edu.cn, lianyongqi@upc.edu.cn, hlxiang@nuist.edu.cn, xlxu@nuist.edu.cn, wenmin.lin@hznu.edu.cn, yang.zhang@unt.edu, zhou@kansai-u.ac.jp

Abstract

Explainable Recommendation (ER) aims to enhance recommendation transparency and prediction accuracy by providing faithful and persuasive explanations. However, Multi-Modal Multi-Interest Explainable Recommendation (MMER) is particularly challenging in two aspects: effectively utilizing diverse multi-modal information to construct reliable semantic evidence and precisely identifying the most relevant user interest to generate faithful explanations. Most previous methods fail to effectively exploit visual information to provide reliable evidence and primarily rely on a single unified representation of user preferences, making it difficult to distinguish diverse user interests and generate faithful explanations. To fill this gap, we propose Joint Multi-Modal Multi-Interest Profiling and Preference-Grounded Reasoning (**PRIME**) for solving the MMER problem. Specifically, **PRIME** first organizes items into interest-aware clusters to facilitate interest disentanglement. Based on these clusters, it constructs explicit multi-modal multi-interest user profiles and a comprehensive global item profile to capture fine-grained preferences and distinguish diverse user interests. Furthermore, **PRIME** performs preference-grounded reasoning by retrieving the most relevant user interest for each item, enabling faithful explanations and accurate predictions. Our experiments on three real-world datasets demonstrate that **PRIME** outperforms the state-of-the-art methods.

*Corresponding author.

1 Introduction

Recommender systems have become increasingly essential with the rapid growth of online platforms and data [Song *et al.*, 2025; Liu *et al.*, 2025c; Liu *et al.*, 2025d; Xie *et al.*, 2026b]. In addition to accurate recommendation, recent studies have increasingly emphasized explainability, aiming to provide faithful and persuasive explanations that enhance recommendation transparency and user trust. In practical scenarios, users often exhibit multiple interests, and items contain heterogeneous multi-modal information, such as textual and visual information. However, existing explainable recommendation methods primarily rely on a single unified representation of users' preferences, entangling multiple interests and hindering the identification of the specific interest behind each interaction. Thus, how to enhance model performance and explainability by effectively leveraging multi-modal information and precisely identifying the most relevant user interest still needs more investigation.

In this paper, we focus on the Multi-Modal Multi-Interest Explainable Recommendation (MMER) problem, as illustrated in Figure 1. That is, users often exhibit multi-faceted preferences within their interaction histories. Meanwhile, items are enriched with multi-modal textual and visual information. Item visual information is rather valuable as it provides additional signals beyond textual information and more directly reflects users' tastes. For instance, a user may exhibit multi-faceted interests in both casual and formal styles, with each interest derived from textual and visual information, as shown in Figure 1. There are two main challenges for solving MMER: **(CH1)**: How to construct reliable semantic evidence by effectively utilizing diverse multi-modal item information? **(CH2)**: How to generate faithful explanations by precisely identifying the most relevant user interest?

However, previous methods cannot solve the MMER prob-

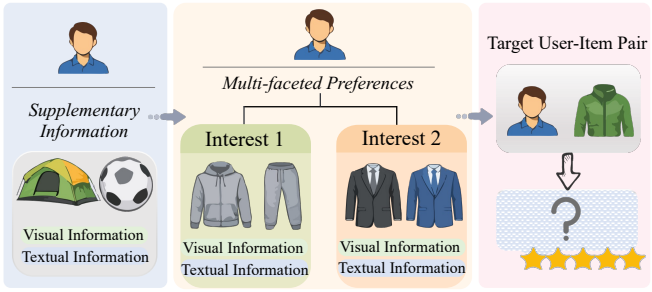


Figure 1: The problem definition of MMER.

lem well. On the one hand, previous explainable recommendation models [Li *et al.*, 2023; Takama *et al.*, 2024; Sabouri *et al.*, 2025] primarily rely on textual information to construct user representations, while neglecting the visual information, resulting in suboptimal performance. Moreover, conventional approaches [Sun *et al.*, 2025; Hao *et al.*, 2025] fail to consider the supplementary interaction signals from the user's broader interaction history, which are crucial for enriching reliable semantic evidence in explainability. Therefore, traditional explainable recommendation models cannot solve **CH1** well. On the other hand, recent explainable recommendation models [Ma *et al.*, 2024; Li *et al.*, 2025; Kim *et al.*, 2025; Fang *et al.*, 2025] leverage large language models (LLMs) to construct a single unified user profile, which provides a comprehensive representation of personalized tastes and preferences. However, these models fail to distinguish diverse user interests, leading to coarse user profiles and limiting the faithfulness of the generated explanations. Therefore, these LLM-based explainable recommendation models cannot solve **CH2** well. In conclusion, previous recommendation models cannot provide satisfactory results on tackling MMER problem.

To address the aforementioned issues, in this paper, we propose Joint Multi-Modal Multi-Interest Profiling and Preference-Grounded Reasoning (**PRIME**), a multi-modal multi-interest explainable recommendation model for solving MMER problem. **PRIME** includes three main modules, i.e., *interest-aware item clustering module*, *multi-modal multi-interest profiling module* and *preference-grounded reasoning module*. In the interest-aware item clustering module, we aim to disentangle user interests and enrich semantic evidence by organizing items into interest-aware clusters and merging relevant supplementary items from the user's broader interaction history. The multi-modal multi-interest profiling module constructs multiple user profiles and a comprehensive global item profile by utilizing visual and textual information, capturing fine-grained preferences and distinguishing diverse user interests to address **CH1**. In the preference-grounded reasoning module, we propose a relevant interest retrieval strategy to explicitly ground the reasoning process in the user's most relevant interest for the target item, thereby addressing **CH2**. By integrating these three modules, **PRIME** effectively leverages diverse multi-modal item information and precisely identifies the most relevant user interest to enhance both recommendation explainability and accuracy.

We summarize the main contributions as follows: (1) We propose a novel framework, e.g., **PRIME**, for solving MMER problem, which contains interest-aware item clustering module, multi-modal multi-interest profiling module, and preference-grounded reasoning module. (2) We propose interest-aware item clustering and multi-modal multi-interest profiling to generate multiple user profiles and a global item profile by leveraging textual and visual information, distinguishing diverse user interests and constructing reliable semantic evidence. (3) We introduce a preference-grounded reasoning approach that retrieves the most relevant user interest for a target item to explicitly ground the reasoning process, thereby enhancing the explainability of recommendations. (4) Extensive experiments on three real-world datasets demonstrate that the proposed **PRIME** significantly outperforms existing recommendation models.

2 Related Work

Large Language Models. Large Language Models have demonstrated strong capabilities in natural language understanding and generation, leading to rapid progress across various AI tasks, including recommendation systems [Xie *et al.*, 2026a; Liu *et al.*, 2026; Xie *et al.*, 2026c]. Early Transformer-based models established large-scale pre-training on extensive text corpora, laying the foundation for improved reasoning and generalization [Hurst *et al.*, 2024]. Recent studies increasingly adopt open-source LLMs such as LLaMA [Touvron *et al.*, 2023] and Qwen [Bai *et al.*, 2023], together with parameter-efficient fine-tuning methods like QLoRA [Dettmers *et al.*, 2024], to enable domain adaptation. Moreover, multi-modal LLMs (e.g., GPT-4o [Hurst *et al.*, 2024]) further extend these capabilities by integrating visual information [Zhou *et al.*, 2025]. Despite these advances, many existing approaches still employ LLMs primarily as black-box semantic extractors or zero-shot predictors [Xie *et al.*, 2021], lacking explicit mechanisms to ground reasoning in structured, domain-specific evidence, which limits their effectiveness in complex recommendation scenarios.

Explainable Recommendation. Explainable Recommendation has been widely investigated to achieve accurate recommendations while providing faithful and persuasive explanations that enhance system transparency [Wardatzky *et al.*, 2025]. Early methods such as NRT [Li *et al.*, 2017] jointly modeled rating prediction and explanation generation, while later works leveraged Transformer-based architectures and personalized prompt learning to produce context-aware explanations [Li *et al.*, 2021; Li *et al.*, 2023]. Recent models, such as XRec [Ma *et al.*, 2024] and G-ReFer [Li *et al.*, 2025], leverage LLMs to generate explanations, incorporating graph-based structures and reasoning-enhanced mechanisms. Models like Reason4Rec [Fang *et al.*, 2025] and EXP3RT [Kim *et al.*, 2025] further refine this process by adopting deliberative preference reasoning and knowledge distillation techniques to improve the consistency between the generated explanations and predicted ratings. However, most existing methods mainly rely on textual information and model user preferences with a single unified profile, limiting fine-grained interest modeling and explanation grounding.

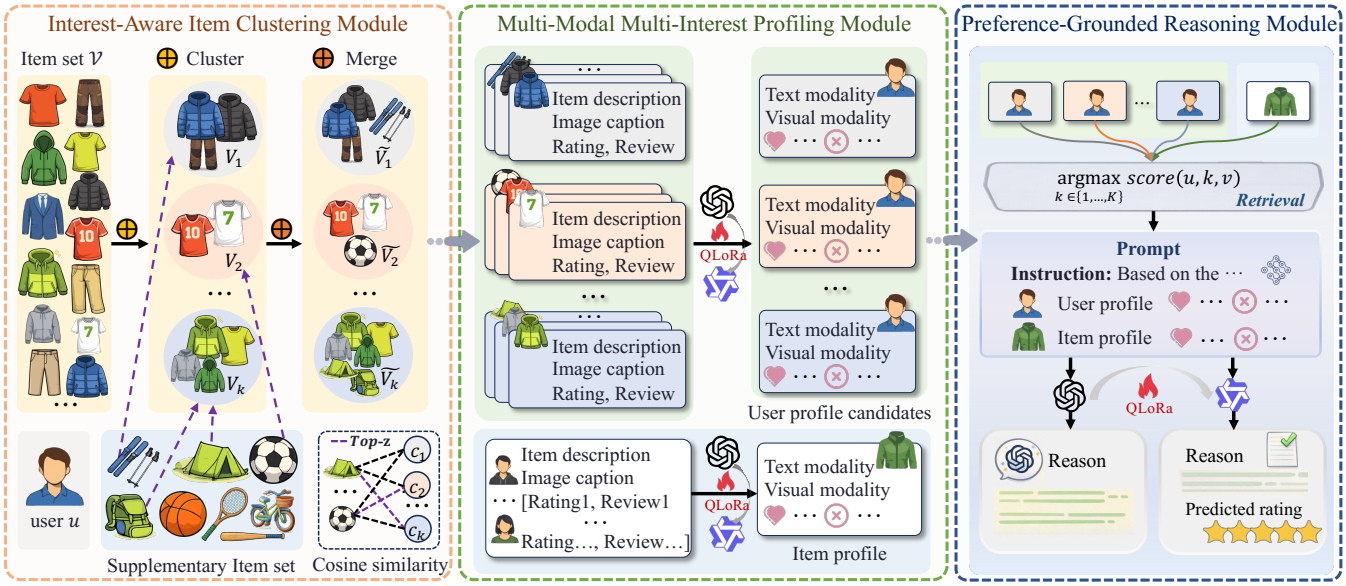


Figure 2: The model framework of proposed PRIME for solving MMER problem.

3 Methodology

In this section, we provide a detailed introduction to our proposed **PRIME** for solving MMER problem. Let $\mathcal{U} = \{u_1, \dots, u_{N_U}\}$ and $\mathcal{V} = \{v_1, \dots, v_{N_V}\}$ denote the sets of users and items, respectively. The historical user-item interactions are represented by the matrix $\boldsymbol{I} \in \mathbb{R}^{N_U \times N_V}$, where each entry $i_{u,v} \in \{0, 1\}$ indicates whether user u has interacted with item v. Meanwhile, let $r_{u,v}$ denote the observed rating score. We adopt the pre-trained CLIP model [Radford *et al.*, 2021] to extract $\boldsymbol{x}^{\text{visual}}$ and $\boldsymbol{x}^{\text{text}}$ from raw images $\mathcal{G}$ and textual descriptions $\mathcal{D}$, respectively. For a target user-item pair (u, v), we leverage user reviews $\mathcal{W}$, image captions $\mathcal{N}$, and textual descriptions $\mathcal{D}$ to construct the multi-interest user profile set $\mathcal{P}_u = \{p_{u,1}, \dots, p_{u,K}\}$ and the item profile p_v. Our goal is to leverage a fine-tuned large language model to generate a reason $\hat{E}_{u,v}$ and predict the rating $\hat{r}_{u,v}$ for a given target user-item pair (u, v).

The framework of **PRIME** is shown in Figure 2. **PRIME** has three modules: *interest-aware item clustering module*, *multi-modal multi-interest profiling module*, and *preference-grounded reasoning module*. Interest-aware item clustering module organizes items into interest-aware semantic clusters. Multi-modal multi-interest profiling module is designed to construct multi-interest user profiles and a comprehensive item profile based on heterogeneous multi-modal information. The preference-grounded reasoning module retrieves the most relevant user profile and performs grounded reasoning with an LLM to generate explanations and predict ratings.

3.1 Interest-Aware Item Clustering Module

To start with, we introduce the interest-aware item clustering module in **PRIME**. Interest-aware item clustering module includes two main steps. Multi-modal item clustering aims to partition items into distinct item clusters by leveraging visual and textual information. Global item retrieval aims to incorporate supplementary items from the user's broader interaction history to enrich user interest profiles.

Multi-Modal Item Clustering. Following previous multi-modal works [Liu *et al.*, 2025a; Hua *et al.*, 2025], we first discretize high-dimensional CLIP features $\boldsymbol{x}^m$ into compact embeddings $\boldsymbol{e}_v^m$ via residual quantization. To handle heterogeneous modalities, we compute modality-specific attention weights via a learnable multi-layer perceptron $\eta(\cdot)$:

$$\alpha_v^m = \frac{\exp(\eta(\boldsymbol{e}_v^m))}{\sum_{m' \in \{\text{visual,text}\}} \exp(\eta(\boldsymbol{e}_v^{m'}))}, \tag{1}$$

where $m \in \{\text{visual}, \text{text}\}$ denotes the modality. We aggregate the weighted features to obtain the unified multi-modal representations $\boldsymbol{h}_v = \sum_m \alpha_v^m \boldsymbol{e}_v^m$. To disentangle user multi-faceted interests, we introduce learnable prototypes $\mathcal{C} = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_K\}$ as cluster centroids. The relevance between item v and prototype $\boldsymbol{c}_k$ is measured by $s_{v,k} = \text{sim}(\boldsymbol{h}_v, \boldsymbol{c}_k)$, where $\text{sim}(\cdot)$ denotes the cosine similarity. To address the non-differentiability of hard assignment, we employ the straight-through Gumbel-Softmax estimator [Jang *et al.*, 2016] to compute the soft assignment probability:

$$\pi_{v,k} = \frac{\exp((s_{v,k} + g_k)/\tau)}{\sum_{j=1}^{K} \exp((s_{v,j} + g_j)/\tau)}, \tag{2}$$

where $g_k \sim \text{Gumbel}(0, 1)$ and τ is the temperature. During the forward pass, we obtain the cluster index $k_v^* = \arg\max_{k \in \{1,\dots,K\}} \pi_{v,k}$. Consequently, we obtain a set of item clusters $\{\mathcal{V}_1, \dots, \mathcal{V}_K\}$, where $\mathcal{V}_k \subseteq \mathcal{V}$. Therefore, this step organizes items into interest-aware clusters that are consistent with users' different preferences.

Global Item Retrieval. Although the initial clustering effectively partitions items into diverse interest-aware clusters, relying solely on a limited view of a user's local interactions

may not fully capture the full scope of a particular interest. Therefore, we treat the learned prototypes $\mathcal{C}$ as global semantic reference points to query the supplementary item pool. For the k-th prototype $\boldsymbol{c}_k$ and each candidate item v from the user's broader history, we measure their similarity as $w_{k,v} = \mathrm{sim}(\boldsymbol{c}_k, \boldsymbol{h}_v)$. We select the top-$z$ most relevant items, thereby filtering out irrelevant items, to form a reliable supplementary item set $\mathcal{V}_k^{\text{sup}}$ for the k-th prototype. We construct an enriched interest-aware item set $\tilde{\mathcal{V}}_k$ by merging the primary interest item set $\mathcal{V}_k$ and the supplementary item set:

$$\tilde{\mathcal{V}}_k = \mathcal{V}_k \cup \mathcal{V}_k^{\text{sup}}, \tag{3}$$

where $\cup$ denotes the set union operation. Consequently, the enriched item sets $\{\tilde{\mathcal{V}}_1, \ldots, \tilde{\mathcal{V}}_K\}$ integrate the user's primary interactions with highly relevant supplementary evidence, enabling more robust user preference modeling.

Optimization. We optimize the model using a variational objective inspired by the evidence lower bound [Hoffman and Johnson, 2016]. Specifically, the reconstruction loss $\mathcal{L}_{rec}$ for implicit feedback is defined as:

$$\mathcal{L}_{rec} = -\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} \Big(i_{u,v} \log \hat{i}_{u,v} + (1 - i_{u,v}) \log(1 - \hat{i}_{u,v}) \Big), \tag{4}$$

where $\hat{i}_{u,v} = \sigma\left(\frac{\boldsymbol{z}_{u,k_v^*}^\top \boldsymbol{h}_v}{\|\boldsymbol{z}_{u,k_v^*}\| \cdot \|\boldsymbol{h}_v\|}\right)$ denotes the predicted probability and $\sigma(\cdot)$ denotes the sigmoid function. Here, $\boldsymbol{z}_{u,k_v^*}$ is the user's specific latent interest inferred from the subset of their history corresponding to cluster k_v^*. The term $\mathcal{L}_{KL}$ regularizes the latent space [Liu *et al.*, 2025b]:

$$\mathcal{L}_{KL} = \sum_{u\in\mathcal{U}}\sum_{k=1}^{K} D_{KL}\left(q_\phi(\boldsymbol{z}_{u,k} \mid \mathcal{V}_{u,k}) \parallel \mathcal{N}(\mathbf{0}, \mathbb{I})\right), \tag{5}$$

where q_ϕ denotes the variational distribution and $\mathcal{V}_{u,k}$ is the subset of user interactions associated with cluster k. The overall training objective is then formulated as:

$$\min_{\Theta} \mathcal{L} = \mathcal{L}_{rec} + \beta \mathcal{L}_{KL} + \lambda \|\Theta\|_2^2, \tag{6}$$

where β balances the reconstruction and regularization terms, and λ controls the L_2 regularization strength. Minimizing $\mathcal{L}$ encourages the learned prototypes and user representations to capture robust and disentangled interest structures.

3.2 Multi-Modal Multi-Interest Profiling Module

After we obtain the enriched interest item clusters $\{\tilde{\mathcal{V}}_1, \ldots, \tilde{\mathcal{V}}_K\}$, we leverage these structured item collections to build comprehensive multi-interest user profiles. Therefore, we propose the multi-modal multi-interest profiling module to model user preferences and global item attributes.

Multi-modal Profile Generator. To bridge the semantic gap between visual and textual modalities, we adopt the strategy of leveraging Multi-modal Large Language Models (MLLMs) as semantic enhancers to transform visual information into textual form [Zhou *et al.*, 2025]. Specifically, the MLLM generates descriptive image captions that encapsulate the visual content of items, thereby enriching the textual metadata with visual knowledge. Following this paradigm, for each item $v \in \mathcal{V}$, we employ an MLLM to convert its raw image $g_v \in \mathcal{G}$ into a descriptive image caption $n_v \in \mathcal{N}$:

$$n_v = \text{MLLM}(g_v), \tag{7}$$

where n_v translates latent visual characteristics into a structured textual form, enabling the LLM to interpret visual cues together with textual metadata. We then leverage an LLM to construct structured profiles for both users and items. By incorporating carefully designed system prompt $S_{u/v}$ with desired output formats (i.e., organizing *textual* and *visual* modalities under `[Like]` and `[Dislike]` tags), the profiles are generated as follows:

$$p_{u,k} = \text{LLM}(S_u, T_{u,k}), \quad p_v = \text{LLM}(S_v, T_v), \tag{8}$$

where $T_{u,k}$ and T_v represent the input prompts for users and items, respectively. The generated profiles $p_{u,k}$ and p_v explicitly incorporate fine-grained information extracted from the visual modality, providing essential evidence for subsequent preference-grounded reasoning. To implement the profile generator, we fine-tune a student LLM with a QLoRA adapter [Dettmers *et al.*, 2024], using the generated profiles by a teacher LLM as supervision for supervised fine-tuning.

Multi-interest User Profile Prompt Construction. To capture users' fine-grained preferences, we construct K distinct user profiles $\mathcal{P}_u = \{p_{u,1}, \ldots, p_{u,K}\}$ for each user u, representing diverse user interests. Specifically, we obtain the item subset $\tilde{\mathcal{V}}_{u,k}$ by selecting user-interacted items from the enriched interest cluster $\tilde{\mathcal{V}}_k$ defined in Equation (3). For each item $v \in \tilde{\mathcal{V}}_{u,k}$, we concatenate its textual description d_v, the image caption n_v as defined in Equation (7), and the user's specific feedback (e.g., rating $r_{u,v}$ and review $w_{u,v}$) as $\mathbf{s}_{u,v} = [d_v, n_v, (r_{u,v}, w_{u,v})]$. The input prompt $T_{u,k}$ for generating the k-th user profile is then formulated as:

$$T_{u,k} = f_u(\{\mathbf{s}_{u,v} \mid v \in \tilde{\mathcal{V}}_{u,k}\}), \tag{9}$$

where $f_u(\cdot)$ is a function specific to each user, which combines the heterogeneous multi-modal metadata into a single structured and coherent instruction string.

Item Profile Prompt Construction. In contrast to user profiling, the item profile p_v aims to encapsulate the global item attributes. To achieve this, we aggregate the item's intrinsic multi-modal metadata with collective feedback from its observed interaction history. Specifically, for each item $v \in \mathcal{V}$, we concatenate its original textual description d_v, the image caption n_v, and the set of historical feedback pairs associated with the item, $\mathcal{F}_v = \{(r_{1,v}, w_{1,v}), (r_{2,v}, w_{2,v}), \ldots\}$. The input prompt T_v, designed for item profile generation, is formulated as follows:

$$T_v = f_v([d_v, n_v, \mathcal{F}_v]), \tag{10}$$

where $f_v(\cdot)$ serves a similar purpose to $f_u(\cdot)$ by organizing multi-modal attributes into a coherent string.

3.3 Preference-Grounded Reasoning Module

Although the interest-aware item clustering module and the multi-modal multi-interest profiling module construct explicit

multi-interest user profiles and item profiles, they do not precisely identify which user interest is relevant to a given item. Therefore, we propose the preference-grounded reasoning module to perform interest selection and preference-grounded reasoning for explanation generation. Specifically, we first propose a *relevant interest retrieval* strategy to identify the most relevant user interest for the target item, which is essential for reliable preference-grounded reasoning. Each user profile $p_{u,k}$ is decomposed into positive and negative components, $p_{u,k} = (p^+_{u,k}, p^-_{u,k})$, corresponding to the `[Like]` and `[Dislike]` sections, respectively. During retrieval, the item profile p_v is treated as a unified description of global item attributes. We encode the user preference components and the item profile using a frozen CLIP text encoder: $\boldsymbol{q}^{\pm}_{u,k} = \mathrm{Enc}(p^{\pm}_{u,k})$ and $\boldsymbol{t}_v = \mathrm{Enc}(p_v)$. Based on these representations, the preference-grounded matching scores s_{pos} and s_{neg} are defined as:

$$s_{\text{pos}} = \mathrm{sim}(\boldsymbol{q}^+_{u,k}, \boldsymbol{t}_v), \quad s_{\text{neg}} = \mathrm{sim}(\boldsymbol{q}^-_{u,k}, \boldsymbol{t}_v), \tag{11}$$

which are then combined to obtain the final matching score:

$$\mathrm{score}(u, k, v) = s_{\text{pos}} - \mu \cdot s_{\text{neg}}, \tag{12}$$

where $\mu \geq 0$ controls the penalty strength. The optimal user interest profile is selected by:

$$k' = \underset{k \in \{1,\ldots,K\}}{\mathrm{argmax}} \ \mathrm{score}(u, k, v). \tag{13}$$

The retrieved user profile $p_{u,k'}$ ensures that the reasoning process is grounded in the most relevant user interest for recommendation. To train the reasoning model, we adopt a teacher-student learning paradigm. Specifically, a teacher LLM is first prompted with the adaptive input $T_{\text{match}} = f(p_{u,k'}, p_v)$ to generate high-quality reasons $E_{u,v}$ as supervision for explanation generation. A system prompt S_{match} is employed to guide the teacher to reason about the matching between the item profile and user interest profiles. We then fine-tune a pre-trained student LLM using QLoRA [Dettmers *et al.*, 2024] to efficiently learn preference-grounded reasoning and rating prediction. During training, the student model is prompted with augmented inputs $T_{train} = f(p_{u,k'}, p_v, \bar{r}_u, \bar{r}_v)$ and jointly optimized to generate both reasons and predict ratings, where the historical average ratings $\bar{r}_u$ and $\bar{r}_v$ serve as statistical priors to calibrate subjective scoring.

Optimization. We provide the optimization details below. The generation loss $\mathcal{L}_{gen}$ is defined as the negative log-likelihood of the ground-truth explanation $E_{u,v}$. Let X denotes the input instruction and y denotes the target token sequence of length L. The loss is computed as:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{L} \log P(y_t \mid X, y_{<t}; \Theta), \tag{14}$$

where $P(\cdot)$ denotes the probability of token y_t conditioned on X and the preceding tokens $y_{<t}$, and Θ represents the trainable QLoRA parameters. The regression loss $\mathcal{L}_{reg}$ measures the error between the predicted rating $\hat{r}_{u,v}$ and the observed rating $r_{u,v}$ using mean squared error:

$$\mathcal{L}_{reg} = \frac{1}{|B|} \sum_{i=1}^{|B|} \left(\hat{r}^{(i)}_{u,v} - r^{(i)}_{u,v} \right)^2, \tag{15}$$

Dataset	# Users	# Items	# Train	# Valid	# Test
Sports	18,838	10,320	121,292	11,036	9,142
Clothing	19,807	12,133	108,553	10,413	9,234
Toys	8,400	5,282	52,748	4,054	3,092

Table 1: The statistics of datasets.

where $|B|$ denotes the batch size. The final training objective for model optimization is defined as:

$$\mathcal{L}_{match} = \mathcal{L}_{gen} + \gamma \mathcal{L}_{reg}, \tag{16}$$

where γ denotes the balancing hyper-parameter. By minimizing $\mathcal{L}_{match}$, the model jointly learns preference-grounded reasoning and accurate rating prediction.

3.4 Inference

For inference on an unobserved user-item pair (u, v), **PRIME** aims to generate a faithful reason along with an accurate rating prediction. We first apply Equation (13) to retrieve the most relevant user interest and construct the inference prompt T_{inf}. The fine-tuned LLM then produces:

$$(\hat{E}_{u,v}, \hat{r}_{u,v}) = \mathrm{LLM}_{Student}(S_{inf}, T_{inf}), \tag{17}$$

where $\hat{E}_{u,v}$ and $\hat{r}_{u,v}$ denote the generated reason and predicted rating, respectively. Here, S_{inf} denotes a multi-step system prompt. By combining relevant interest retrieval with LLM-based reasoning, **PRIME** enhances the explanation faithfulness and rating prediction accuracy.

4 Experiments

In this section, we conduct experiments on three widely used datasets to answer the following questions: (1) **RQ1:** How does **PRIME** perform compared with the state-of-the-art recommendation methods? (2) **RQ2:** How do different components of **PRIME** contribute to performance improvement? (3) **RQ3:** How does the performance of **PRIME** vary with different values of the hyper-parameters?

4.1 Experimental Settings

Datasets. We conduct extensive experiments on three widely used subsets of the Amazon dataset [Geng *et al.*, 2022; Tran and Lauw, 2025]: (a) Sports and Outdoors, (b) Clothing, Shoes and Jewelry, and (c) Toys and Games, which are abbreviated as *Sports*, *Clothing*, and *Toys*, respectively. We sample interactions from the most recent two years. The data are split into training, validation, and test sets with a ratio of 8:1:1 according to interaction timestamps, ensuring that all test interactions occur strictly after the training and validation data to avoid information leakage [Ji *et al.*, 2023; Zhang *et al.*, 2025]. Following prior work [Fang *et al.*, 2025], we apply a 5-core filtering strategy to remove users and items with fewer than five interactions, and exclude cold-start users and items that do not appear in the training set from the validation and test sets. The statistics of the processed datasets are reported in Table 1. We use users' cross-category interaction histories to build enriched interest-aware item clusters.

Model	Sports				Clothing				Toys			
	MAE ↓	RMSE ↓	GPTS ↑	BLEURT ↑	MAE ↓	RMSE ↓	GPTS ↑	BLEURT ↑	MAE ↓	RMSE ↓	GPTS ↑	BLEURT ↑
MF	0.6585	0.9482	-	-	0.8388	1.1143	-	-	0.6324	0.9031	-	-
NRT	0.7454	0.9679	75.01	0.4215	0.9113	1.3224	73.25	0.4018	0.7218	0.9297	74.88	0.4156
CoNet	0.6606	0.9471	-	-	0.8336	1.1158	-	-	0.6325	0.9035	-	-
PETER	0.6892	0.9501	76.55	0.4421	0.8521	1.1563	75.14	0.4234	0.6715	0.9150	76.02	0.4355
LATTICE	0.6459	0.9323	-	-	0.8126	1.0953	-	-	0.6258	0.8983	-	-
EMER	0.6723	0.9454	77.17	0.4516	0.8415	1.1420	76.59	0.4351	0.6654	0.9121	76.83	0.4412
PEPLER	0.6354	0.9213	78.29	0.4651	0.7956	1.0827	77.85	0.4525	0.6153	0.8952	78.05	0.4586
GPT-4o	0.7165	0.9611	78.50	0.4610	0.8677	1.2447	78.18	0.4555	0.7060	0.9153	78.25	0.4590
XRec	-	-	78.72	0.4681	-	-	78.94	0.4653	-	-	78.81	0.4667
EXP3RT	0.5924	0.8851	78.85	0.4783	0.7659	1.0628	79.56	0.4717	0.5854	0.8659	79.26	0.4750
Reason4Rec	0.5840	0.8723	78.91	0.4831	0.7587	1.0578	80.11	0.4819	0.5750	0.8573	79.88	0.4824
PRIME-E	0.5680	0.8553	81.14	0.5082	0.7322	1.0127	81.89	0.5021	0.5718	0.8486	80.25	0.4927
PRIME-V	0.5652	0.8483	81.51	0.5128	0.7524	1.0485	80.51	0.4880	0.5526	0.8283	81.52	0.5054
PRIME-S	0.5721	0.8613	80.51	0.5016	0.7354	1.0154	81.50	0.4983	0.5655	0.8425	80.87	0.4984
PRIME-M	0.5752	0.8651	-	-	0.7480	1.0259	-	-	0.5688	0.8453	-	-
PRIME	**0.5541**	**0.8353**	**82.45**	**0.5243**	**0.7125**	**0.9956**	**83.17**	**0.5311**	**0.5384**	**0.8123**	**82.22**	**0.5180**

Table 2: Experimental results on three datasets. Bold indicates the best results. "-" denotes not applicable. GPTS denotes GPTScore.

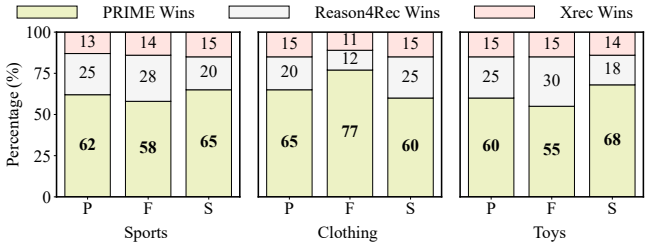


Figure 3: Human evaluation results.

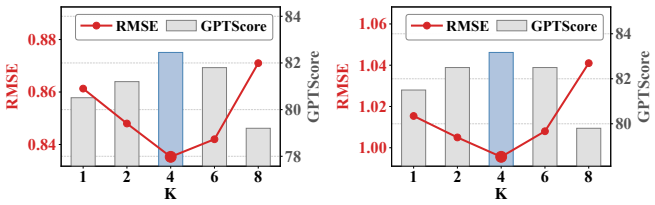
Evaluation Metrics. We adopt different evaluation metrics for each target task. For rating prediction, we use MAE and RMSE [Hurst *et al.*, 2024; Kim *et al.*, 2025]. For reason generation, we adopt semantic-aware metrics, including GPTScore [Li *et al.*, 2017] and BLEURT [Sellam *et al.*, 2020], to measure the semantic similarity between the generated reason and the corresponding item reviews. Beyond quantitative metrics, we further conduct a human evaluation on Persuasiveness (P), Faithfulness (F), and Scrutability (S). Specifically, we randomly sample 100 examples from each dataset and invite three human judges to select the best reason among those generated by XRec, Reason4Rec, and **PRIME** based on these criteria [Kim *et al.*, 2025].

Baselines. To verify the effectiveness of **PRIME**, we compare it with representative models: **MF** [Koren *et al.*, 2009], **NRT** [Li *et al.*, 2017], **CoNet** [Hu *et al.*, 2018], **PETER** [Li *et al.*, 2021], **LATTICE** [Zhang *et al.*, 2021], **EMER** [Zhu *et al.*, 2022], **PEPLER** [Li *et al.*, 2023], **GPT-4o** [Hurst *et al.*, 2024], **XRec** [Ma *et al.*, 2024], **EXP3RT** [Kim *et al.*, 2025], **Reason4Rec** [Fang *et al.*, 2025].

Implementation Details. We employ GPT-4o as both the teacher LLM and the MLLM, and use Qwen-3-8B as the student model. The number of interest clusters is set to $K = 4$, and the top-z retrieved items for interest-aware item clustering are fixed to $z = 5$. For the optimization of multi-modal item clustering, the temperature parameter is set to $\tau = 0.1$, while the KL divergence weight and the regularization coefficient are set to $\beta = 0.01$ and $\lambda = 1\text{e-}4$, respectively. In the preference-grounded reasoning module, we set $\mu = 1.0$ and $\gamma = 1.0$. To accelerate training and reduce memory consumption, we adopt the Unsloth framework. The student LLM is fine-tuned using QLoRA with rank $r_{lora} = 64$, scaling factor $\alpha_{lora} = 16$, and a dropout rate of 0.1. We optimize the model using AdamW [Loshchilov and Hutter, 2017] with a learning rate of 2e-4 and a batch size of $|B| = 16$ for 5 epochs. The best model checkpoint is selected based on validation performance. During inference, we set the temperature to 0.0 and the maximum number of generated tokens to 1024. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Each experiment is repeated five times with different random seeds, and the average results are reported. For reason evaluation, GPT-4o is used to compute GPTScore.

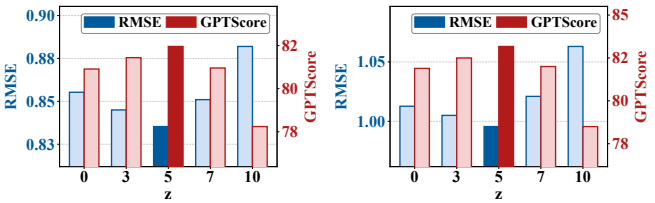
4.2 Recommendation Performance (RQ1)

The comparison results on different datasets are shown in Table 2 and further supported by Figure 3, from which we make the following observations: (1) **PRIME** outperforms the strongest baseline with relative improvements of 8.5%, 10.2%, and 7.4% in BLEURT on the Sports, Clothing, and Toys datasets, respectively. This demonstrates the effectiveness of our framework in jointly modeling multi-modal multi-interest profiles and preference-grounded reasoning. (2) **PRIME** significantly surpasses conventional baselines such as CoNet and LATTICE. This advantage is attributed to the interest-aware item clustering module, which effectively organizes different items into interest-aware item clusters and retrieves relevant items from the user's broader interaction history through these clusters, thereby reducing irrelevant items and improving user interest disentanglement. (3) **PRIME** also outperforms LLM-based approaches, including EXP3RT and Reason4Rec. Notably, although GPT-4o (zero-shot) produces high-quality explanations, it performs poorly on rating prediction due to the lack of collaborative signals. In contrast, **PRIME** leverages the preference-grounded reasoning module to fine-tune the student LLM. This enables the model to learn preference-grounded reasoning based on the retrieved user interest profile, achieving a balance be-



(a) Sports (b) Clothing

Figure 4: Effect of number of interest clusters K.



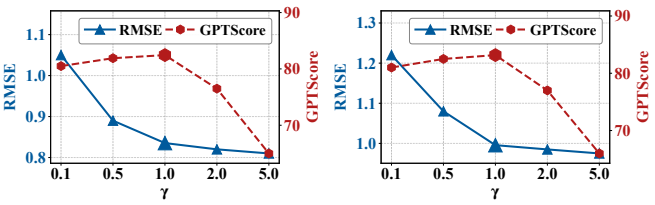
(a) Sports (b) Clothing

Figure 5: Effect of number of retrieved items z.

tween rating accuracy and explanation quality. (4) **PRIME** consistently achieves the highest human preference rate, significantly outperforming both XRec and Reason4Rec across all dimensions. Notably, **PRIME** shows remarkable performance in Faithfulness on the Clothing dataset, verifying that our multi-modal profiling effectively utilizes visual information (e.g., color). Furthermore, the substantial lead in Scrutability demonstrates that disentangling user interests leads to more faithful and interpretable reasons.

4.3 Ablation Study (RQ2)

To investigate how each component of **PRIME** contributes to the final performance, we compare **PRIME** with several of its variants, including **PRIME**-E, **PRIME**-V, **PRIME**-S, and **PRIME**-M. **PRIME**-E removes global item retrieval, relying solely on local interactions. **PRIME**-V excludes all visual information from interest-aware item clustering and profile generation. **PRIME**-S does not involve multi-modal item clustering, which means it generates a single global user profile. **PRIME**-M replaces the multi-step reason-based prediction with direct rating prediction. The comparison results are shown in Table 2. From it, we can observe that: (1) **PRIME** outperforms both **PRIME**-S and **PRIME**-E, indicating that our proposed interest-aware item clustering module effectively captures users' multi-faceted preferences and enriches the semantic evidence for better profiling. (2) By comparing the performance of **PRIME** and **PRIME**-E, it shows that **PRIME**-E performs worse as it fails to provide the supplementary items retrieved from the user's broader interaction history. This effect is particularly evident on the Toys dataset, indicating that merely utilizing local interactions cannot effectively support reliable semantic evidence. (3) **PRIME** achieves much better performance than **PRIME**-V, highlighting the importance of visual signals in modeling user preferences and global item attributes. Notably, the performance drop is largest on the Clothing dataset, demonstrating that the multi-modal profiling effectively utilizes diverse multi-modal information for tackling **CH1**. (4) The performance of **PRIME**-S is inferior to that of **PRIME**, showing that a unified user representation fails to disentangle users' multi-faceted interests. (5) **PRIME** outperforms **PRIME**-M, indicating that preference-grounded reasoning effectively guide the final recommendation and is critical for solving **CH2**.



(a) Sports (b) Clothing

Figure 6: Effect of the loss balance coefficient γ.

4.4 Hyperparameter Study (RQ3).

We further explore how some key hyper-parameters affect the performance of **PRIME** on the Sports and Clothing datasets. Specifically, we vary the number of interest clusters K in the range of $\{1, 2, 4, 6, 8\}$ and report the results of RMSE and GPTScore in Figure 4. We can observe that when K is smaller (e.g., $K = 1$), the single user profile is less effective for disentangling multi-faceted preferences. Meanwhile, when K is larger (e.g., $K = 8$), the over-segmentation leads to sparse clusters, resulting in deteriorated results. Therefore we set $K = 4$ empirically. Moreover, we conduct experiments on varying the number of retrieved supplementary items z in the range of $\{0, 3, 5, 7, 10\}$ and report the results in Figure 5. We can conclude that the model obtains better performance when z is moderate. When z is smaller (e.g., $z = 0$), the lack of supplementary information results in insufficient semantic evidence. However, when z increases (e.g., $z = 10$), the excessive low-affinity items introduce noise, compromising the overall accuracy. Therefore we set $z = 5$ empirically. As illustrated in Figure 6, we vary the loss balance coefficient γ in the range of $\{0.1, 0.5, 1.0, 2.0, 5.0\}$. We observe that when γ is smaller, the weak regression constraint leads to inaccurate rating predictions. Conversely, when γ is larger (e.g., $\gamma = 5.0$), the excessive focus on MSE negatively impacts the LLM's ability to generate faithful reasons. Therefore we set $\gamma = 1.0$ empirically for **PRIME**.

5 Conclusion

In this paper, we propose Joint Multi-Modal Multi-Interest Profiling and Preference-Grounded Reasoning (**PRIME**) to address the Multi-Modal Multi-Interest Explainable Recommendation (MMER) problem. Specifically, we organize items into interest-aware clusters, incorporating broader interaction history, and construct multi-modal multi-interest user and item profiles by utilizing textual and visual information. Then, we introduce a preference-grounded reasoning module that grounds the prediction on the most relevant user interest to generate both explanations and rating predictions. Finally, extensive experiments on three datasets demonstrate the superior performance of our proposed model.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No.62572486), Natural Science Foundation of Shandong Province (No.ZR2023MF007).

Contribution Statement

Anqi Wang and Jianye Xie are co-first authors.

References

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Dettmers *et al.*, 2024] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2024.

[Fang *et al.*, 2025] Yi Fang, Wenjie Wang, Yang Zhang, Fengbin Zhu, Qifan Wang, Fuli Feng, and Xiangnan He. Reason4rec: Large language models for recommendation with deliberative user preference alignment. *arXiv preprint arXiv:2502.02061*, 2025.

[Geng *et al.*, 2022] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM conference on recommender systems*, pages 299–315, 2022.

[Hao *et al.*, 2025] Qingbo Hao, Chundong Wang, Yingyuan Xiao, and Wenguang Zheng. Iregnn: Implicit review-enhanced graph neural network for explainable recommendation. *Knowledge-Based Systems*, 311:113113, 2025.

[Hoffman and Johnson, 2016] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

[Hu *et al.*, 2018] Guangneng Hu, Yu Zhang, and Qiang Yang. Conet: Collaborative cross networks for cross-domain recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 667–676, 2018.

[Hua *et al.*, 2025] Yu Hua, Weiming Liu, Gui Xu, Yaqing Hou, Yew-Soon Ong, and Qiang Zhang. Deterministic-to-stochastic diverse latent feature mapping for human motion synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22724–22734, 2025.

[Hurst *et al.*, 2024] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[Jang *et al.*, 2016] Eric Jang, Shixiang Shane Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ArXiv*, abs/1611.01144, 2016.

[Ji *et al.*, 2023] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems*, 41(3):1–27, 2023.

[Kim *et al.*, 2025] Jieyong Kim, Hyunseo Kim, Hyunjin Cho, SeongKu Kang, Buru Chang, Jinyoung Yeo, and Dongha Lee. Review-driven personalized preference reasoning with large language models for recommendation. corr, abs/2408.06276, 2024. doi: 10.48550. *arXiv preprint ARXIV.2408.06276*, 2025.

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

[Li *et al.*, 2017] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 345–354, 2017.

[Li *et al.*, 2021] Lei Li, Yongfeng Zhang, and Li Chen. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601*, 2021.

[Li *et al.*, 2023] Lei Li, Yongfeng Zhang, and Li Chen. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4):1–26, 2023.

[Li *et al.*, 2025] Yuhan Li, Xinni Zhang, Linhao Luo, Heng Chang, Yuxiang Ren, Irwin King, and Jia Li. G-refer: Graph retrieval-augmented large language model for explainable recommendation. In *Proceedings of the ACM on Web Conference 2025*, pages 240–251, 2025.

[Liu *et al.*, 2025a] Weiming Liu, Jun Dan, Fan Wang, Xinting Liao, Junhao Dong, Hua Yu, Shunjie Dong, and Lianyong Qi. Distinguish then exploit: Source-free open set domain adaptation via weight barcode estimation and sparse label assignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4927–4938, 2025.

[Liu *et al.*, 2025b] Weiming Liu, Xinting Liao, Jun Dan, Fan Wang, Hua Yu, Junhao Dong, Shunjie Dong, Lianyong Qi, and Yew-Soon Ong. Solving discrete (semi) unbalanced optimal transport with equivalent transformation mechanism and kkt-multiplier regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Liu *et al.*, 2025c] Yuwen Liu, Lianyong Qi, Xingyuan Mao, Weiming Liu, Shichao Pei, Fan Wang, Xuyun Zhang, Amin Beheshti, and Xiaokang Zhou. Variational graph auto-encoder driven graph enhancement for sequential recommendation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 3144–3152, 2025.

[Liu *et al.*, 2025d] Yuwen Liu, Lianyong Qi, Xingyuan Mao, Weiming Liu, Fan Wang, Xiaolong Xu, Xuyun Zhang, Wanchun Dou, Xiaokang Zhou, and Amin Beheshti. Hyperbolic variational graph auto-encoder for next poi recommendation. In *Proceedings of the ACM on Web Conference 2025*, pages 3267–3275, 2025.

[Liu *et al.*, 2026] Yuwen Liu, Lianyong Qi, Xingyuan Mao, Weiming Liu, Xuhui Fan, Qiang Ni, Xuyun Zhang, Yang Zhang, Yuan Tian, and Amin Beheshti. Hyperbolic-enhanced mixture-of-experts mamba for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 15403–15411, 2026.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Ma *et al.*, 2024] Qiyao Ma, Xubin Ren, and Chao Huang. Xrec: Large language models for explainable recommendation. *arXiv preprint arXiv:2406.02377*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Sabouri *et al.*, 2025] Milad Sabouri, Masoud Mansoury, Kun Lin, and Bamshad Mobasher. Towards explainable temporal user profiling with llms. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 219–227, 2025.

[Sellam *et al.*, 2020] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892, 2020.

[Song *et al.*, 2025] Te Song, Lianyong Qi, Weiming Liu, Fan Wang, Xiaolong Xu, Hongsheng Hu, Yang Cao, Xuyun Zhang, and Amin Beheshti. Boosting guided diffusion with large language models for multimodal sequential recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6203–6212, 2025.

[Sun *et al.*, 2025] Ying Sun, Yang Ji, Hengshu Zhu, Fuzhen Zhuang, Qing He, and Hui Xiong. Market-aware long-term job skill recommendation with explainable deep reinforcement learning. *ACM Transactions on Information Systems*, 43(2):1–35, 2025.

[Takama *et al.*, 2024] Yasufumi Takama, Makito Inada, and Hiroki Shibata. Proposal on virtual user profile generation for explainable recommendation. In *International Symposium on Computational Intelligence and Industrial Applications*, pages 200–213. Springer, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Tran and Lauw, 2025] Nhu-Thuat Tran and Hady W Lauw. Parameter-efficient variational autoencoder for multimodal multi-interest recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6412–6420, 2025.

[Wardatzky *et al.*, 2025] Kathrin Wardatzky, Oana Inel, Luca Rossetto, and Abraham Bernstein. Whom do explanations serve? a systematic literature survey of user characteristics in explainable recommender systems evaluation. *ACM Transactions on Recommender Systems*, 3(4):1–35, 2025.

[Xie *et al.*, 2021] Jianye Xie, Haotong Sun, Junsheng Zhou, Weiguang Qu, and Xinyu Dai. Event detection as graph parsing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1630–1640, 2021.

[Xie *et al.*, 2026a] Jianye Xie, Chunhua Hu, Lianyong Qi, Fan Wang, Xiaolong Xu, Haolong Xiang, Xuyun Zhang, Shichao Pei, Amin Beheshti, Wanchun Dou, et al. Plikd: Prompt learning with instance-aware knowledge distillation for web-scale semantic image classification. In *Proceedings of the ACM Web Conference 2026*, pages 3788–3799, 2026.

[Xie *et al.*, 2026b] Jianye Xie, Lianyong Qi, Weiming Liu, Anqi Wang, Xiaolong Xu, Haolong Xiang, Xuyun Zhang, Wenwen Gong, Yang Zhang, Amin Beheshti, et al. Joint similar user exploration and informative behavior guidance for multi-modal new item recommendation. In *Proceedings of the ACM Web Conference 2026*, pages 6218–6229, 2026.

[Xie *et al.*, 2026c] Jianye Xie, Lianyong Qi, Fan Wang, Anqi Wang, Wenjuan Gong, Danxin Wang, Wanchun Dou, Yang Cao, Shichao Pei, and Xiaokang Zhou. Retrieval-driven reasoning for deliberative visual classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11060–11068, 2026.

[Zhang *et al.*, 2021] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3872–3880, 2021.

[Zhang *et al.*, 2025] Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. Collm: Integrating collaborative embeddings into large language models for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Zhou *et al.*, 2025] Peilin Zhou, Chao Liu, Jing Ren, Xinfeng Zhou, Yueqi Xie, Meng Cao, Zhongtao Rao, You-Liang Huang, Dading Chong, Junling Liu, et al. When large vision language models meet multimodal sequential recommendation: An empirical study. In *Proceedings of the ACM on Web Conference 2025*, pages 275–292, 2025.

[Zhu *et al.*, 2022] Jihua Zhu, Yujiao He, Guoshuai Zhao, Xuxiao Bu, and Xueming Qian. Joint reason generation and rating prediction for explainable recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4940–4953, 2022.