Information-Guided Selective Modality-Interest Alignment for Multimodal Recommendation

Wenze Ma
mawenze991226@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Chenyu Sun
scy1520253537@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Yanmin Zhu*
yzhu@cs.sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Qiwen Gu
qwgu@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Xuhao Zhao
zhaoxuhao@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Abstract

Multimodal recommendation (MMRec) aims to enhance recommendation performance by leveraging rich item content from multiple modalities. However, directly incorporating all modality information does not necessarily lead to better preference modeling, since user interests are often more related to a subset of modality signals, while other signals may be weakly aligned with user preferences or even introduce noise. Although recent MMRec methods improve modality utilization through invariant learning, attention mechanisms, graph refinement, or contrastive learning, their alignment processes are often implicit or heuristic and lack a clear objective for selecting modality signals that better match user interests. In this paper, we propose AMUR, an information-guided selective modality-interest alignment framework for multimodal recommendation. Inspired by an information-theoretic view, AMUR aims to enhance modality information that is more related to user interests while reducing the influence of less aligned signals. Specifically, AMUR first refines modality graph structures towards user behavior, and then selectively aligns shared interest-related semantics across modalities. This enables AMUR to improve modality-interest alignment while preserving useful modality-specific complementary information. Extensive experiments on three real-world datasets demonstrate the effectiveness of AMUR over competitive baselines. *The codes are available here.*

CCS Concepts

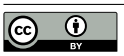
• **Information systems** → **Recommender systems**.

Keywords

multi-modal recommendation; modality-interest alignment; information-theoretic guidance

ACM Reference Format:
Wenze Ma, Chenyu Sun, Yanmin Zhu, Qiwen Gu, and Xuhao Zhao. 2026. Information-Guided Selective Modality-Interest Alignment for Multimodal Recommendation. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3799682.3840808

*Corresponding author.



CIKM '26, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840808

1 Introduction

Multimodal recommendation (MMRec) has attracted increasing attention in recent years, as items in modern recommender systems are often associated with rich auxiliary content, such as images, textual descriptions. Compared with ID-based collaborative signals, multimodal information provides fine-grained semantics about item attributes, styles, and functions, which can help alleviate data sparsity and improve user preference modeling. By incorporating such modality information, recommender systems are expected to better capture why users prefer certain items.

Existing studies have explored various ways to incorporate modality information into recommender systems, including extending matrix factorization with visual features [8], propagating modality information over graph structures [29, 39, 41, 45], and enhancing representation learning with self-supervised or contrastive objectives [18, 26, 31, 46]. These methods have demonstrated the usefulness of modality information in alleviating interaction sparsity and enriching item representations.

However, directly incorporating all modality information does not necessarily lead to better user preference modeling. In real-world recommendation scenarios, users often make decisions based on only part of the modality content. As illustrated in Figure 1, a user may be interested in Nike sneakers and new arrivals, while other modality details, such as color or promotional information, may be less aligned with the current preference. Similarly, an item image may contain background objects or decorative details that are not useful for recommendation. If such modality signals are indiscriminately propagated, they may introduce irrelevant information into user and item representations, leading to suboptimal recommendations.

This observation highlights the importance of aligning modality information with user interests. However, achieving such alignment is challenging, because the preference-aligned part of modality information is not directly observable. In real-world MMRec datasets, user interests are only reflected through coarse item-level interactions, such as clicks or purchases. These interactions indicate that a user likes an item, but do not reveal which visual regions, textual

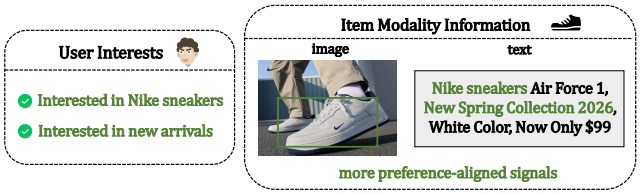


Figure 1: Illustration of modality-interest misalignment. The highlighted green regions are more aligned with the user's interests, whereas other regions are less relevant to this user and may introduce preference-weak signals.

phrases, or semantic attributes actually contribute to the decision. As a result, models have to infer preference-aligned modality information from implicit feedback, making selective modality-interest alignment difficult in practice.

Meanwhile, the same user interest may be expressed differently across modalities, which further complicates modality-interest alignment. For example, a preference for a specific brand may appear as a logo in images and as a brand-related phrase in text. These modality-specific expressions introduce natural discrepancies that make the two modalities of the same item diverge in the representation space. Without modeling such cross-modal correspondence, the resulting multi-modal representations can easily create semantic gaps between modalities and become inconsistent with user interests.

Several recent methods attempt to achieve modality–interest alignment through invariant representation learning [5], attention mechanisms [11], graph refinement [39, 41], or contrastive alignment [18, 31, 33]. However, their alignment processes are often implicit or heuristic. In particular, they usually lack a clear objective for explaining why certain modality structures should be emphasized or suppressed. As a result, modality signals that are weakly aligned with user preferences may still be propagated or fused into the final representations.

Moreover, many alignment strategies tend to encourage global consistency across modalities. However, different modalities naturally contain both shared semantics and modality-specific complementary details. Directly aligning the full modality representations may over-suppress such modality-specific information, which can also be useful for recommendation. Therefore, an effective MMRec model should not only refine modality information toward user interests, but also perform cross-modal alignment in a selective manner, enhancing shared-interest related semantics while preserving modality-specific complementary information.

To address these issues, we propose **AMUR**, an **information-guided selective modality-interest alignment** framework for MMRec. The core idea of AMUR is to use an information-theoretic view to guide the selective use of modality information, rather than treating all modality signals as equally useful. Specifically, we view modality information from two complementary aspects: the mutual information part [1] that is more aligned with user interests, and the residual part that is weakly related to user interests.

Based on this view, AMUR contains two key alignment stages. The first stage is *selective modality-interest refinement*. For each modality, AMUR first builds an item graph from raw modality features, which provides a candidate structure for modality propagation. Since visual or textual similarity does not always match user preference, we further use behavioral signals to calibrate this graph. Specifically, we use a KL regularizer to make the learned edge distribution closer to the item co-occurrence patterns observed from user interactions. In this way, edges supported by user behavior are more likely to be retained while others are weakened. After that, an interest-aware contrastive objective brings the refined modality representations closer to the users, encouraging the representations to capture semantics aligned with user interests.

The second stage is *selective cross-modal shared-interest alignment*. After each modality is refined toward user interests, AMUR further exploits the shared-interest related semantics across modalities. Instead of directly aligning the full visual and textual representations, AMUR learns a dimension-wise selection gate to obtain shared-interest subspace representations. Cross-modal consistency maximization and discrepancy regularization are then applied only to the selected subspace. In this way, AMUR enhances cross-modal shared semantics that are useful for preference modeling, while preserving modality-specific complementary information for final recommendation.

The main contributions of this work are summarized as follows:

- We identify the modality-interest misalignment problem in MMRec and propose **AMUR**, an information-guided selective modality-interest alignment framework. Instead of indiscriminately using all modality signals, AMUR aims to emphasize preference-aligned modality semantics while reducing the influence of less aligned signals.
- We design a selective modality-interest refinement module to calibrate modality information toward user interests. It refines modality-induced item relations with behavioral signals through a KL-divergence regularizer, and enhances user-aligned modality representations via interest-aware contrastive learning.
- We propose a selective cross-modal shared-interest alignment module to exploit shared semantics across modalities. It learns a shared-interest subspace and applies cross-modal alignment only within the selected subspace, thereby enhancing shared-interest related semantics while preserving modality-specific complementary information.
- We conduct extensive experiments on three real-world MMRec datasets. The results demonstrate the effectiveness of AMUR compared to competitive baselines.

2 Methodology

2.1 Problem Formulation

Let $\mathcal{U}$ and $\mathcal{I}$ be the user set and the item set. The user-item interaction matrix is denoted as $\mathbf{R} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $\mathbf{R}_{ui} = 1$ indicates that user u has interacted with item i, and $\mathbf{R}_{ui} = 0$ otherwise. For each user u and item i, we have their id embeddings $e_u^{id}, e_i^{id} \in \mathbb{R}^d$ from embedding layers $\mathbf{E}_{\mathrm{U}}^{id}$ and $\mathbf{E}_{\mathrm{I}}^{id}$. Under MMRec settings, each item is associated with modality features $e_i^m \in \mathbb{R}^{d_m}$, where $m \in \{v, t\}$, including visual feature e_i^v and textual feature e_i^t.

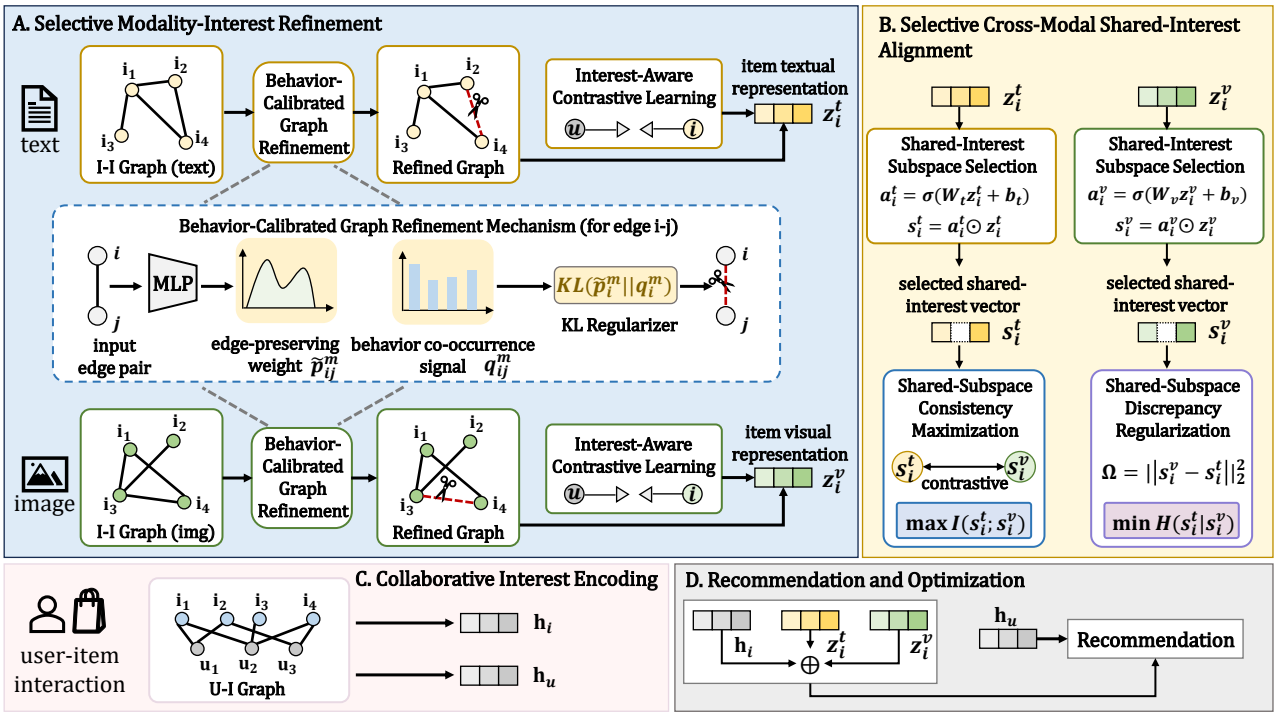


Figure 2: The overall framework of AMUR. AMUR first encodes collaborative interests from user–item interactions, then performs selective modality-interest refinement to obtain user-aligned modality representations. It further conducts selective cross-modal shared-interest alignment, and finally integrates collaborative and modality representations for recommendation.

The goal of MMRec is to predict the preference score $\hat{y}_{ui}$:

$$\hat{y}_{ui} = f\left(u, i \mid e_u^{id}, e_i^{id}, e_i^v, e_i^t, \mathbf{R}\right), \tag{1}$$

and recommend a top-k list of items that u might be interested in. In this work, rather than indiscriminately using all modality information, we focus on information-guided selective modality alignment, which aims to preserve preference-relevant modality semantics while reducing the influence of modality signals that are weakly related to user interests.

2.2 Overview

The overall framework of AMUR is illustrated in Figure 2. AMUR aims to selectively align item modality information with user interests under an information-guided view. Instead of directly using all modality signals, AMUR first extracts collaborative interests from user–item interactions, and then uses them to guide modality refinement.

Specifically, AMUR contains two key alignment stages. The first stage performs selective modality-interest refinement, where raw modality graphs are refined by behavior-calibrated signals and the learned modality representations are pulled closer to user interests. The second stage performs selective cross-modal shared-interest alignment, where only the selected shared-interest subspace is aligned across visual and textual modalities. This avoids forcing all modality-specific information to be identical while still enhancing cross-modal shared semantics.

Finally, AMUR integrates collaborative representations and refined modality representations for recommendation, and jointly optimizes all objectives.

2.3 Collaborative Interest Encoding

User-item interactions provide direct evidence of user preferences. Therefore, we first encode collaborative signals from the interaction graph and use them as preference anchors for subsequent modality refinement. Specifically, we construct a user-item graph based on the interaction matrix $\mathbf{R}$:

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{R} \\ \mathbf{R}^{\mathrm{T}} & 0 \end{bmatrix}. \tag{2}$$

Following LightGCN [9], we normalize $\mathbf{A}$ as $\hat{\mathbf{A}}$ and propagate ID embeddings $\mathbf{H}^0 = \begin{bmatrix} \mathbf{E}_{\mathrm{U}}^{id} \\ \mathbf{E}_{\mathrm{I}}^{id} \end{bmatrix}$ over the interaction graph:

$$\mathbf{H}^{l+1} = \hat{\mathbf{A}}\mathbf{H}^l, \quad l \in \{0, \cdots, L-1\}, \tag{3}$$

$$\mathbf{H} = \frac{1}{L+1} \sum_{l=0}^{L} \mathbf{H}^l. \tag{4}$$

After propagation, we obtain the user collaborative representation $\mathbf{h}_u$ and item collaborative representation $\mathbf{h}_i$. These representations summarize high-order behavioral patterns from observed interactions and serve as preference anchors for calibrating modality representations toward user interests.

2.4 Selective Modality-Interest Refinement

As discussed above, user interests often focus on only a subset of item modality contents. Therefore, directly propagating or aligning all modality information may inject preference-irrelevant signals into user preference modeling [41].

To alleviate this issue, existing MMRec methods often learn invariant representations across modalities [5], incorporate attention mechanisms [11] or refining modality representations and structures [18, 31, 41]. Although effective, these methods usually rely on implicit or heuristic strategies, making it difficult to interpret which part of modality information is encouraged to match user interests and which part should be suppressed.

In this paper, we use an information-guided view to motivate a selective modality-interest refinement process.

2.4.1 Information-Guided Motivation. Ideally, a desirable modality representation should preserve information that is useful for explaining user preferences, while reducing the influence of modality details that are weakly related to user interests. We describe this goal from an information-theoretic perspective.

For each modality $m \in \{v, t\}$, let $\boldsymbol{z}_i^m \in \mathbb{R}^d$ denote the modality representation of item i encoded[1] from raw modality features. Let $\mathbf{h}_u$ denote the collaborative representation of user u learned from historical interactions. The entropy of $\boldsymbol{z}_i^m$ can be decomposed as:

$$H(\boldsymbol{z}_i^m) = \underbrace{I(\boldsymbol{z}_i^m; \mathbf{h}_u)}_{\text{preference-relevant part}} + \underbrace{H(\boldsymbol{z}_i^m | \mathbf{h}_u)}_{\text{unexplained residual part}}. \tag{5}$$

This decomposition provides an intuitive view of modality-interest refinement.

- The mutual information term $I(\boldsymbol{z}_i^m; \mathbf{h}_u)$ measures how much information in the modality representation is related to the user preference representation. A larger value means that the modality representation contains more preference-relevant semantics.
- The conditional entropy term $H(\boldsymbol{z}_i^m | \mathbf{h}_u)$ measures the remaining uncertainty in the modality representation after the user preference representation is known. In our scenario, this residual part may include modality details that are weakly related or irrelevant to the current user interest.

Based on this view, AMUR has two optimization directions: increasing the preference-relevant part by enhancing $I(\boldsymbol{z}_i^m; \mathbf{h}_u)$, and reducing the influence of the unexplained residual part associated with $H(\boldsymbol{z}_i^m | \mathbf{h}_u)$. However, directly estimating mutual information and conditional entropy is generally intractable in large-scale recommendation scenarios [1]. Therefore, we instantiate this information-guided objective with two tractable modules: a behavior-calibrated graph refinement module to suppress noisy modality structures, and an interest-aware contrastive learning module to enhance user-aligned modality semantics.

2.4.2 Modality-Initialized Graph Construction. Following [24, 31, 39], we first construct a modality-initialized item–item graph as the candidate structure for each modality. Specifically, for each modality $m \in \{v, t\}$, we compute the cosine similarity between item i and item j based on their raw modality features $\boldsymbol{e}_i^m$ and $\boldsymbol{e}_j^m$:

$$o_{i,j}^m = \frac{(\boldsymbol{e}_i^m)^\top \boldsymbol{e}_j^m}{\|\boldsymbol{e}_i^m\| \cdot \|\boldsymbol{e}_j^m\|}. \tag{6}$$

Then, each item is connected to its top-K most similar items under modality m:

$$s_{i,j}^m = \begin{cases} 1, & o_{i,j}^m \in \text{TopK}(\{o_{i,k}^m \mid k \in \mathcal{I}\}), \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $s_{i,j}^m$ denotes the entry at the i-th row and j-th column of the modality-initialized item–item graph $\boldsymbol{S}^m \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$. It provides a lightweight structural prior that organizes items into local neighborhoods based on their modality features, facilitating subsequent feature extraction. Such a graph learning approach is widely adopted by previous work, which has shown effectiveness in MMRec models.

2.4.3 Behavior-Calibrated Graph Refinement. The modality-initialized graph $\boldsymbol{S}^m$ captures item similarity in the modality feature space, but such similarity does not necessarily reflect user-interest relations. For example, two smartphones may appear visually similar, but appeal to different user groups. Therefore, directly propagating over the raw modality graph may introduce preference-inconsistent signals. To address this issue, we refine $\boldsymbol{S}^m$ by learning which modality edges should be preserved under the guidance of collaborative behavior.

For each candidate edge (i, j) in $\boldsymbol{S}^m$, we first estimate an edge-preservation probability:

$$p_{ij}^m = \sigma\left(\text{MLP}_\theta^m([e_i^{id}, e_j^{id}, o_{ij}^m])\right), \tag{8}$$

where $[e_i^{id}, e_j^{id}, o_{ij}^m]$ denotes the concatenation of the ID embeddings of items i and j, together with their raw modality similarity o_{ij}^m. The MLP maps the input to a scalar score, and the sigmoid function converts it into $p_{ij}^m \in (0, 1)$. A larger p_{ij}^m indicates that edge (i, j) is more likely to be useful for preference-aware modality propagation.

Based on p_{ij}^m, we sample a binary edge mask:

$$\beta_{ij}^m \sim \text{Bern}(p_{ij}^m), \tag{9}$$

where $\text{Bern}(p_{ij}^m)$ denotes a Bernoulli distribution with success probability p_{ij}^m. The sampled value β_{ij}^m acts as an edge switch: $\beta_{ij}^m = 1$ means that the edge is preserved, while $\beta_{ij}^m = 0$ means that the edge is filtered out. Here we adopt the Gumbel-softmax trick [13] to obtain differentiable binary masks. After sampling masks for all candidate edges, the refined modality graph is obtained by:

$$\hat{\boldsymbol{S}}^m = \boldsymbol{\beta}^m \odot \boldsymbol{S}^m, \tag{10}$$

where $\odot$ denotes the Hadamard product. Thus, $\hat{\boldsymbol{S}}^m$ is a preference-calibrated subgraph of $\boldsymbol{S}^m$, where more preference-consistent modality edges are more likely to be retained.

To further guide the refinement process, we introduce behavioral co-occurrence as a preference-aware reference. Specifically, we compute:

$$f_{ij} = [\boldsymbol{R}^\top \boldsymbol{R}]_{ij}, \tag{11}$$

where f_{ij} counts how many users have interacted with both items i and j. A larger f_{ij} indicates that the two items are more likely to be close from the perspective of collaborative preference.

[1] We will illustrate details of the encoder in Equation (15).

We then convert the behavior co-occurrence signal into a reference distribution over the candidate modality neighbors :

$$q_{ij}^m = \frac{s_{ij}^m f_{ij}}{\sum_k s_{ik}^m f_{ik} + \epsilon}, \tag{12}$$

where s_{ij}^m ensures that the reference distribution is defined only over modality-derived candidate edges, and ϵ is a small constant for numerical stability.

Similarly, we normalize the learned edge-preservation probabilities over the candidate neighbors:

$$\tilde{p}_{ij}^m = \frac{s_{ij}^m p_{ij}^m}{\sum_k s_{ik}^m p_{ik}^m + \epsilon}. \tag{13}$$

We then regularize the learned modality-neighbor distribution toward the behavior-informed reference distribution by minimizing their KL divergence:

$$\begin{aligned} \mathcal{L}_{br}^m &= \sum_i D_{\mathrm{KL}}\left(\tilde{\boldsymbol{p}}_i^m \| \boldsymbol{q}_i^m\right) \\ &= \sum_i \sum_j \tilde{p}_{ij}^m \log \frac{\tilde{p}_{ij}^m}{q_{ij}^m + \epsilon}. \end{aligned} \tag{14}$$

From the information-guided view, this surrogate discourages modality structures that are weakly supported by user preference signals, thereby reducing the influence of the unexplained residual part associated with $H(\boldsymbol{z}_i^m|\mathbf{h}_u)$.

2.4.4 Interest-Aware Contrastive Learning. After obtaining the refined modality graph $\hat{S}^m$, we further extract modality representations and align them with user interests. The graph refinement module suppresses preference-inconsistent modality relations, while this module enhances the preference-relevant part by pulling modality representations closer to the collaborative interests of users who have interacted with them.

Specifically, we encode the modality representation of items by propagating modality features over the refined graph:

$$Z^m = \hat{S}^m \left(\mathbf{E}_{\mathrm{I}}^{id} \odot \mathrm{MLP}_\gamma^m(\mathbf{E}^m)\right), \tag{15}$$

where $\mathbf{E}^m$ denotes raw modality features under modality m, $\mathrm{MLP}_\gamma^m(\cdot)$ projects raw modality features into the latent space, and $\odot$ denotes element-wise multiplication. We use $\boldsymbol{z}_i^m$ to denote the i-th row of Z^m.

Then, for each observed user–item interaction (u, i), we regard $(\mathbf{h}_u, \boldsymbol{z}_i^m)$ as a positive pair, since item i reflects the preference of user u in historical behavior. Other items in the batch are treated as negative samples. The interest-aware contrastive loss is defined as:

$$\mathcal{L}_{cl}^m = -\frac{1}{|\mathcal{B}|} \sum_{(u,i)\in\mathcal{B}} \log \frac{\exp\left(s(\mathbf{h}_u, \boldsymbol{z}_i^m)/\tau\right)}{\sum_{j\in\mathcal{B}_I} \exp\left(s(\mathbf{h}_u, \boldsymbol{z}_j^m)/\tau\right)}, \tag{16}$$

where $\mathcal{B}$ denotes a mini-batch of observed interactions, $\mathcal{B}_I$ denotes the item set in the batch, $s(\cdot,\cdot)$ is the cosine similarity function, and τ is the temperature parameter.

This contrastive objective encourages the modality representation of the interacted item to be closer to the user collaborative representation than other items. From the information-guided view, it serves as a tractable surrogate for enhancing the preference-relevant term $I(\boldsymbol{z}_i^m; \mathbf{h}_u)$. Following the standard InfoNCE formulation [22], we have:

$$I(\boldsymbol{z}_i^m; \mathbf{h}_u) \geq \log|\mathcal{B}_I| - \mathcal{L}_{cl}^m. \tag{17}$$

Therefore, minimizing $\mathcal{L}_{cl}^m$ maximizes a lower-bound surrogate of the mutual information between modality representations and user interests. This helps the learned modality representations preserve user-aligned semantics after behavior-calibrated graph refinement.

Combining the behavior-calibrated graph refinement and interest-aware contrastive learning, the objective of modality-interest refinement is:

$$\mathcal{L}_{mir} = \sum_{m\in\{v,t\}} \left(\mathcal{L}_{br}^m + \mathcal{L}_{cl}^m\right). \tag{18}$$

2.5 Selective Cross-Modal Shared-Interest Alignment

After modality-interest refinement, each modality representation $\boldsymbol{z}_i^m$ has been calibrated toward user interests. However, different modalities may still express the same interest-related semantics in different forms. For example, a user's preference for a brand may appear as a logo in the visual modality and as a brand name in the textual modality. Such cross-modal shared semantics can further improve recommendation robustness.

A straightforward solution is to directly align the full visual and textual representations. However, this may over-suppress modality-specific complementary information, since visual and textual modalities naturally contain different but useful signals. Therefore, instead of enforcing full-space alignment, we propose a selective shared-interest alignment module. The key idea is to first select the dimensions that are more suitable for cross-modal shared semantic alignment, and then apply cross-modal objectives only within the selected shared-interest subspace.

2.5.1 Shared-Interest Subspace Selection. For each modality representation $\boldsymbol{z}_i^m$, we learn a dimension-wise soft selection gate:

$$\boldsymbol{a}_i^m = \sigma\left(\mathbf{W}_m \boldsymbol{z}_i^m + \mathbf{b}_m\right), \tag{19}$$

where $\boldsymbol{a}_i^m \in (0,1)^d$ denotes the selection weights for different dimensions, and $\sigma(\cdot)$ is the sigmoid function. Then, we obtain the selected shared space by:

$$\boldsymbol{s}_i^m = \boldsymbol{a}_i^m \odot \boldsymbol{z}_i^m, \tag{20}$$

where $\odot$ denotes element-wise multiplication.

The gate $\boldsymbol{a}_i^m$ softly reweights the modality representation and highlights dimensions that are more suitable for cross-modal shared semantic alignment. In this way, cross-modal objectives are not directly imposed on the entire modality representation $\boldsymbol{z}_i^m$, but only on the selected shared space $\boldsymbol{s}_i^m$. This design helps enhance shared semantics across modalities while reducing the risk of forcing modality-specific complementary information to be identical.

2.5.2 Information-Guided Shared-Subspace Alignment. From an information-guided view, the selected representation $\boldsymbol{s}_i^m$ should contain shared semantics that can be explained by another modality, while reducing unnecessary discrepancy in the selected subspace. For two different modalities $m, m' \in \{v, t\}$, the information

contained in s_i^m can be written as:

$$H(s_i^m) = \underbrace{I(s_i^m; s_i^{m'})}_{\text{shared semantic consistency}} + \underbrace{H(s_i^m | s_i^{m'})}_{\text{shared-subspace discrepancy}}. \quad (21)$$

This decomposition provides two optimization directions. First, the mutual information term $I(s_i^m; s_i^{m'})$ should be enhanced, so that the selected subspaces from different modalities preserve consistent item semantics. Second, the conditional entropy term $H(s_i^m | s_i^{m'})$ should be reduced, so that unnecessary discrepancy within the selected shared-interest subspace can be regularized.

2.5.3 Shared-Subspace Consistency Maximization. To enhance cross-modal shared semantics, we maximize the agreement between the selected representations of the same item. Specifically, for item i, we treat (s_i^v, s_i^t) as a positive pair, and (s_i^v, s_j^t) with $j \neq i$ as negative pairs. The shared-subspace contrastive loss is defined as:

$$\mathcal{L}_{sc} = -\frac{1}{|\mathcal{B}_I|} \sum_{i \in \mathcal{B}_I} \log \frac{\exp\left(s(s_i^v, s_i^t)/\tau\right)}{\sum_{j \in \mathcal{B}_I} \exp\left(s(s_i^v, s_j^t)/\tau\right)}, \quad (22)$$

where $\mathcal{B}_I$ denotes the item set in a mini-batch, $s(\cdot,\cdot)$ is the cosine similarity function, and τ is the temperature parameter.

Following the standard InfoNCE formulation, minimizing $\mathcal{L}_{sc}$ maximizes a lower-bound surrogate of $I(s_i^v; s_i^t)$. Therefore, this objective encourages the selected visual and textual subspaces to capture consistent and discriminative shared semantics.

2.5.4 Shared-Subspace Discrepancy Regularization. Besides maximizing shared semantic consistency, we further regularize the discrepancy between selected subspaces. Directly estimating $H(s_i^m | s_i^{m'})$ is intractable. Following a variational view [16], we introduce a conditional distribution $\Omega(s_i^m | s_i^{m'})$. Then we have:

$$H(s_i^m | s_i^{m'}) \leq -\mathbb{E}_{P(s_i^m, s_i^{m'})} \log \Omega(s_i^m | s_i^{m'}). \quad (23)$$

We instantiate $\Omega(s_i^m | s_i^{m'})$ as an isotropic Gaussian distribution:

$$\Omega(s_i^m | s_i^{m'}) = \mathcal{N}\left(s_i^m \mid s_i^{m'}, \gamma \mathbf{I}\right), \quad (24)$$

where $\gamma \mathbf{I}$ is the covariance matrix. Under this instantiation, minimizing the negative log-likelihood is equivalent to minimizing the squared distance between the selected subspaces. Therefore, we define the shared-subspace discrepancy regularization as:

$$\mathcal{L}_{sd} = \sum_{i \in \mathcal{B}_I} \left\| s_i^v - s_i^t \right\|_2^2. \quad (25)$$

This term reduces unnecessary discrepancy only in the selected shared-interest subspace, rather than forcing the full visual and textual representations to be identical.

The objective of selective cross-modal shared-interest alignment is:

$$\mathcal{L}_{sia} = \mathcal{L}_{sc} + \mathcal{L}_{sd}. \quad (26)$$

Although related, these two terms play *complementary* roles in aligning semantics.

- The contrastive term $\mathcal{L}_{sc}$ imposes *discriminative* alignment by separating mismatched modality pairs, preventing trivial solutions (e.g., all modalities collapsing to the same point) and preserving item-level distinctions.
- The discrepancy term $\mathcal{L}_{sd}$ enforces *local alignment* by penalizing systematic deviation between modalities of the same item, ensuring that different views collapse toward a common semantic anchor.

By applying cross-modal objectives on s_i^v and s_i^t, AMUR selectively enhances shared-interest related semantics across modalities while preserving modality-specific information in the original representations z_i^v and z_i^t for final recommendation.

Table 1: Statistics of three widely-used multi-modal recommendation datasets.

Dataset	Baby		Sports		Clothing	
Modality	V	T	V	T	V	T
Embed Dim	4096	384	4096	384	4096	384
User	19445		35598		39387	
Item	7050		18357		23033	
Interactions	160792		296337		278677	

2.6 Recommendation and Optimization

After obtaining the collaborative representations and modality representations, we integrate them for final recommendation. Specifically, for each item i, we combine its collaborative representation $\mathbf{h}_i$ and the refined modality representations z_i^v and z_i^t as:

$$\mathbf{h}_i^* = \mathbf{h}_i + z_i^v + z_i^t. \quad (27)$$

It is worth noting that the selected shared-interest representations s_i^v and s_i^t are only used for cross-modal shared-interest alignment. The final recommendation still uses the complete modality representations z_i^v and z_i^t, so that modality-specific complementary information can be preserved.

Then, the preference score between user u and item i is calculated by:

$$\hat{y}_{ui} = \mathbf{h}_u^{\mathrm{T}} \mathbf{h}_i^*. \quad (28)$$

We adopt the Bayesian Personalized Ranking (BPR) loss [23] for recommendation:

$$\mathcal{L}_{bpr} = -\sum_{(u,i,j) \in \mathcal{D}} \log \sigma\left(\hat{y}_{ui} - \hat{y}_{uj}\right), \quad (29)$$

where (u, i, j) denotes a training triplet, in which item i is an observed positive item for user u, and item j is a sampled negative item.

Finally, AMUR is optimized by jointly considering recommendation loss, modality-interest refinement loss, and selective shared-interest alignment loss:

$$\mathcal{L} = \mathcal{L}_{bpr} + \alpha \mathcal{L}_{mir} + \beta \mathcal{L}_{sia}, \quad (30)$$

where $\mathcal{L}_{mir}$ denotes the objective of selective modality-interest refinement, $\mathcal{L}_{sia}$ denotes the objective of selective cross-modal shared-interest alignment, α and β are trade-off hyperparameters.

3 Experiments

3.1 Experimental Setup

3.1.1 Datasets. Following prior multi-modal recommendation studies [7, 39, 44, 45], we evaluate our model on three widely-adopted

benchmark datasets: *Baby*, *Sports*, and *Clothing*. These datasets contain rich multi-modal information, including item images, textual descriptions, and user–item interaction records, and have been extensively used to assess multi-modal representation learning in recommendation systems. They span three distinct domains with heterogeneous visual and textual characteristics, providing diverse scenarios for evaluating multi-modal alignment.

To ensure fair comparison and reproducibility, we follow the standard preprocessing and training protocols used in previous work [43, 44]. Table 1 summarizes the dataset statistics.

3.1.2 Baseline Methods and Evaluation Metrics. We compare AMUR with representative MMRec baselines from different categories: **MF-based method**: VBPR [8]; **invariant and causal learning-based methods**: InvRL [5] and ISOLATOR [32]; **graph and topology-based methods**: MMGCN [29], LATTICE [41], FREEDOM [45], MGCN [39], LGMRec [7], DA-MRS [33], and TMLP [12]; **SSL and alignment-based methods**: SLMRec [26], BM3 [46], and MENTOR [31]; **diffusion-based methods**: DiffMM [15] and CCDRec [36]; and **Transformer/MLLM-based methods**: MIG-GT [11] and BeFA [6]. We adopt Recall@K and NDCG@K with $K = \{10, 20\}$ as evaluation metrics.

3.1.3 Parameter Settings. Following [44], we fix the batch size to 2048, the embedding dimension d to 64, and the learning rate to 0.001 for all methods. For each baseline, we tune the key hyperparameters according to the ranges suggested in the original papers or released codes. For AMUR, we search the coefficient α of selective modality-interest refinement and the coefficient β of selective cross-modal shared-interest alignment in $\{0, 0.001, 0.01, 0.1\}$. The influence of these two coefficients is further analyzed in the hyperparameter study.

3.2 Overall Performance

Table 2 reports the overall performance on three real-world multimodal recommendation datasets. AMUR achieves the best results across all datasets and evaluation metrics, showing the effectiveness of information-guided selective modality-interest alignment. We summarize the key observations as follows:

- **Overall effectiveness.** AMUR consistently outperforms representative baselines from different categories, including MF-based methods, invariant/causal learning methods, graph-based methods, SSL/alignment-based methods, diffusion-based methods, and Transformer/MLLM-based methods. Compared with the strongest baseline on each metric, AMUR improves Recall@20 by 1.56%, 3.96%, and 2.26% on Baby, Sports, and Clothing, respectively, and improves NDCG@20 by 2.24%, 7.78%, and 3.65%. These consistent improvements indicate that selectively aligning modality information with user interests is beneficial for multimodal recommendation.
- **Advantages over recent denoising and alignment baselines.** AMUR also outperforms strong recent baselines such as DA-MRS, MENTOR, DiffMM, CCDRec, TMLP, MIG-GT, and BeFA. This shows that existing denoising, alignment, diffusion-based, or Transformer-based designs may still suffer when modality information is not selectively aligned with user interests. By combining selective modality-interest refinement and selective cross-modal shared-interest alignment, AMUR better emphasizes preference-aligned modality semantics while preserving useful modality-specific information.

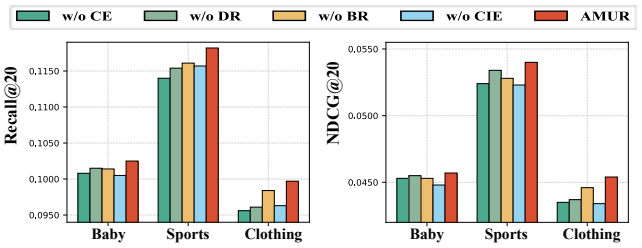


Figure 3: Ablation results of different components.

3.3 Ablation Study

To evaluate the contribution of each component in AMUR, we construct four ablated variants:

- **w/o BR**. This variant removes the behavior-calibrated graph refinement module. The modality graph is directly used without being guided by behavioral co-occurrence signals.
- **w/o CIE**. This variant removes the interest-aware contrastive learning objective, so modality representations are not explicitly pulled toward the collaborative interests of interacted users.
- **w/o DR**. This variant removes the shared-subspace discrepancy regularization, weakening the constraint that reduces unnecessary discrepancy between selected visual and textual shared-interest representations.
- **w/o CE**. This variant removes the shared-subspace consistency maximization objective, thus discarding the cross-modal contrastive signal for enhancing shared interest-related semantics.

The results are shown in Figure 3. We have the following observations. **First**, removing any component leads to performance degradation across datasets, demonstrating that all modules contribute to the final recommendation performance. Specifically, BR helps refine modality structures with behavioral signals, CIE encourages modality representations to be aligned with user interests, DR reduces unnecessary discrepancy in the selected shared-interest subspace, and CE enhances discriminative shared semantics across modalities.

Second, the variants w/o CE and w/o DR generally suffer clear performance drops, indicating the importance of selective cross-modal shared-interest alignment. This shows that, after modality representations are refined toward user interests, exploiting shared interest-related semantics across modalities can further improve recommendation quality.

Third, removing BR or CIE also hurts performance, especially on datasets where modality signals are more diverse or noisy. This verifies the necessity of selective modality-interest refinement: behavioral graph refinement helps reduce the influence of less aligned modality relations, while interest-aware contrastive learning further guides modality representations toward user preferences.

Table 2: Performance comparisons of different approaches over three datasets. The best results of each metric are marked in bold, while the best results of baselines is underlined. * represents that the improvements of AMUR compared with best baselines are statistically significant for $p < 0.05$.

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
VBPR	0.0424	0.0664	0.0223	0.0285	0.0559	0.0858	0.0307	0.0384	0.0281	0.0410	0.0157	0.0190
InvRL	0.0320	0.0519	0.0171	0.0222	0.0324	0.0504	0.0176	0.0223	0.0227	0.0363	0.0122	0.0157
ISOLATOR	0.0628	0.0957	0.0346	0.0430	0.0730	0.1112	0.0395	0.0493	0.0659	0.0972	0.0358	0.0437
MMGCN	0.0397	0.0644	0.0209	0.0272	0.0381	0.0615	0.0200	0.0260	0.0221	0.0362	0.0113	0.0149
LATTICE	0.0536	0.0858	0.0287	0.0370	0.0618	0.0950	0.0337	0.0423	0.0459	0.0702	0.0253	0.0306
FREEDOM	0.0619	0.0987	0.0326	0.0420	0.0711	0.1090	0.0385	0.0482	0.0639	0.0940	0.0344	0.0421
MGCN	0.0614	0.0963	0.0327	0.0417	0.0732	0.1105	0.0397	0.0493	0.0659	0.0967	0.0360	0.0438
LGMRec	0.0650	0.0989	0.0349	0.0436	0.0723	0.1090	0.0392	0.0487	0.0552	0.0825	0.0301	0.0371
DA-MRS	0.0617	0.0960	0.0337	0.0425	0.0706	0.1073	0.0381	0.0476	0.0603	0.0902	0.0328	0.0404
TMLP	0.0658	0.1009	0.0356	0.0447	0.0751	0.1137	0.0402	0.0501	0.0654	0.0957	0.0360	0.0437
SLMRec	0.0540	0.0810	0.0285	0.0357	0.0676	0.1017	0.0374	0.0462	0.0452	0.0675	0.0247	0.0303
BM3	0.0547	0.0866	0.0290	0.0373	0.0646	0.0981	0.0353	0.0439	0.0429	0.0630	0.0232	0.0283
MENTOR	0.0616	0.0995	0.0336	0.0433	0.0718	0.1092	0.0389	0.0485	0.0606	0.0890	0.0325	0.0397
DiffMM	0.0591	0.0924	0.0320	0.0405	0.0690	0.1049	0.0371	0.0463	0.0569	0.0897	0.0313	0.0397
CCDRec	0.0661	0.1006	0.0352	0.0441	0.0730	0.1108	0.0392	0.0489	0.0661	0.0975	0.0358	0.0438
MIG-GT	0.0656	0.1009	0.0353	0.0442	0.0742	0.1085	0.0403	0.0491	0.0619	0.0904	0.0338	0.0410
BeFA	0.0555	0.0884	0.0299	0.0383	0.0649	0.0985	0.0346	0.0432	0.0568	0.0857	0.0307	0.0381
AMUR	**0.0692***	**0.1025***	**0.0371***	**0.0457***	**0.0791***	**0.1182***	**0.0439***	**0.0540***	**0.0686***	**0.0997***	**0.0375***	**0.0454***
%Improv.	4.69%	1.56%	4.21%	2.24%	5.33%	3.96%	8.93%	7.78%	3.78%	2.26%	4.17%	3.65%

Table 3: Effectiveness of behavior-guided graph refinement and shared-interest subspace selection.

Variants	Baby		Sports		Clothing	
	R@20	N@20	R@20	N@20	R@20	N@20
MM emb	0.0995	0.0443	0.1104	0.0490	0.0921	0.0409
Full space	0.0992	0.0441	0.1143	0.0520	0.0974	0.0437
AMUR	**0.1025**	**0.0457**	**0.1182**	**0.0540**	**0.0997**	**0.0454**

Table 3 further validates two key designs of AMUR. The **MM emb** variant replaces behavior-guided graph refinement with multimodal embedding-based edge estimation, while **Full space** removes shared-interest subspace selection and aligns the full modality representations. AMUR consistently outperforms both variants across all datasets. This demonstrates that behavioral signals are more reliable than raw multimodal similarity for preference-aware graph refinement, and that selective shared-subspace alignment is more effective than full-space cross-modal alignment.

3.4 In-depth Analysis

To further understand how AMUR improves modality-interest alignment, we conduct several in-depth analyses, including visualization, feature-importance study, information-theoretic trend analysis, robustness analysis, and hyperparameter study.

Visualization of Modality–Interest Alignment. We visualize the t-SNE projections of item modality representations and their corresponding user representations in Figure 4. Compared with DA-MRS, AMUR produces more compact clusters between users and their interacted items, indicating that the learned modality representations are more closely related to user interests.

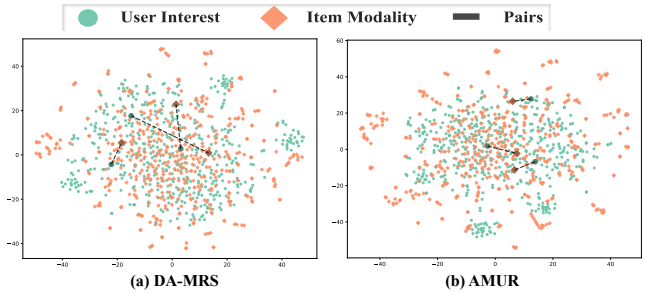


Figure 4: Visualization of modality-interest alignment.

Visualization of Aligned and Residual Representations. We further compare the aligned modality representations with the residual representations, where the residual is computed as

$$r^m = z^m_{\text{init}} - z^m_{\text{aligned}}.$$

As shown in Figure 5, the aligned representations form more compact clusters, while the residual representations are more dispersed. This suggests that AMUR concentrates user-aligned modality semantics into the aligned representations and reduces the influence of less structured residual signals.

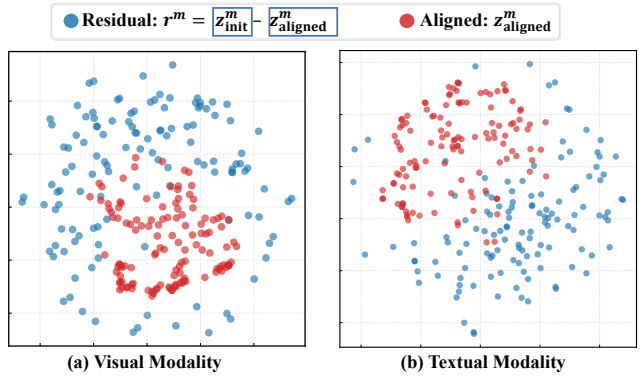


Figure 5: Visualization of aligned and residual modality representations.

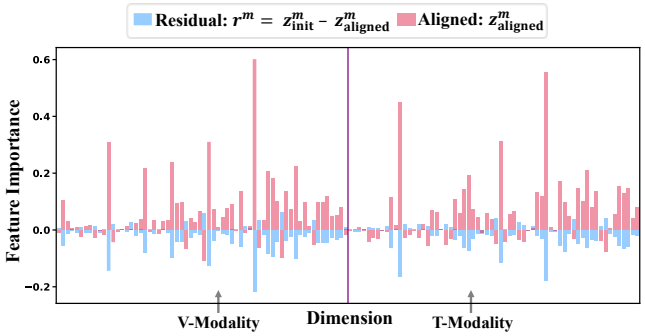


Figure 6: Feature-importance analysis of aligned and residual modality representations.

Feature-Importance Analysis. For each dimension k, we mask it as zero and measure the resulting decrease in the user–item score:

$$\Delta_k = \text{Score}_{\text{full}} - \text{Score}_{\text{without dim } k}.$$

As shown in Figure 6, the aligned representations generally show higher importance values than the residual representations. This indicates that AMUR preserves more prediction-useful information in the aligned representations, while the residual part contains more weakly useful or potentially distracting signals.

Information-Theoretic Trend Analysis. We estimate mutual information with MINE [2] and conditional entropy with a neural variational estimator. As shown in Figure 7, $I(\boldsymbol{z}_i^v; \mathbf{h}_u)$ and $I(\boldsymbol{z}_i^t; \mathbf{h}_u)$ gradually increase, while $H(\boldsymbol{z}_i^v|\mathbf{h}_u)$ and $H(\boldsymbol{z}_i^t|\mathbf{h}_u)$ decrease. For cross-modal shared-interest alignment, $I(\boldsymbol{s}_i^v; \boldsymbol{s}_i^t)$ increases and $H(\boldsymbol{s}_i^v|\boldsymbol{s}_i^t)$ decreases. These trends are consistent with the information-guided motivation of AMUR.

Robustness to Corrupted Modality Content. We randomly corrupt a portion of item modality content by replacing it with the modality content of another item. In Figure 8, AMUR consistently outperforms the compared baselines under different corruption ratios, demonstrating its robustness to noisy modality signals.

Hyperparameter Analysis. We study the influence of α and β, which control selective modality-interest refinement and selective cross-modal shared-interest alignment, respectively. As shown in Figure 9, performance generally improves as each coefficient increases from 0 and reaches the best result around 0.01. Further increasing the coefficient may degrade performance, indicating the need to balance recommendation supervision and auxiliary alignment objectives.

4 Related Work

4.1 Multimodal Recommendation

Multimodal recommendation aims to improve user preference modeling by leveraging rich modality content such as images and text. Early studies extend matrix factorization with modality features, such as VBPR [8] and CKE [40]. With the development of graph neural networks, graph-based methods have been widely explored for modeling collaborative and multimodal signals [11, 12, 14, 24, 27, 29, 30, 45]. Representative methods include MMGCN [29], which propagates modality-specific representations over user–item graphs, and LGMRec [7], which models local and global user interests with multimodal graph learning.

Self-supervised methods further improve multimodal representation learning through contrastive or reconstruction objectives [17, 26, 34, 37, 46], such as MGCN [39] and BM3 [46]. Recently, diffusion-based methods have also been introduced into multimodal recommendation [4, 10, 15, 20], such as DiffMM [15], MCDRec [19], and LD4MRec [38].

Although these methods have achieved promising performance, they mainly focus on how to incorporate modality information into recommendation models, while paying less attention to whether the introduced modality signals are well aligned with user interests. In contrast, our work focuses on selective modality-interest alignment, aiming to emphasize modality semantics that better match user preferences while reducing the influence of less aligned signals.

4.2 Denoising and Aligning Modality Information for Recommendation

Noisy and misaligned modality information is an important challenge in multimodal recommendation. Existing methods address this issue from different perspectives, including modality denoising [21, 35, 39], graph refinement [3, 25, 28, 41, 42], diffusion-based representation refinement [19, 38], invariant representation learning [5], and contrastive or self-supervised alignment [18, 31, 33].

However, existing denoising and alignment strategies are often implicit or heuristic. They usually lack a clear objective for selectively aligning modality information with user interests, making it difficult to explain why certain modality signals should be emphasized or weakened. Moreover, many cross-modal alignment strategies encourage consistency across full modality representations, which may overlook useful modality-specific complementary information. Different from these methods, AMUR uses an information-guided view to refine modality information toward user interests and align only the selected shared-interest subspace across modalities.

4.3 Runtime Analysis

We further evaluate the training efficiency of AMUR on the Sports dataset using an NVIDIA GeForce RTX 3060. Table 4 reports the

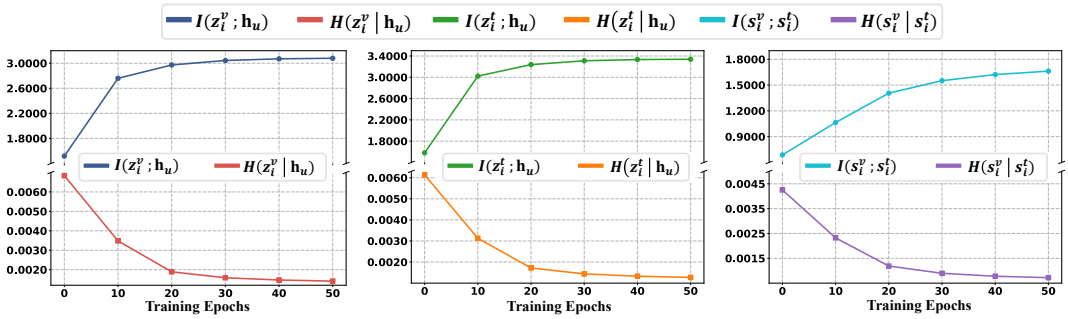


Figure 7: Estimated mutual information and conditional entropy during training.

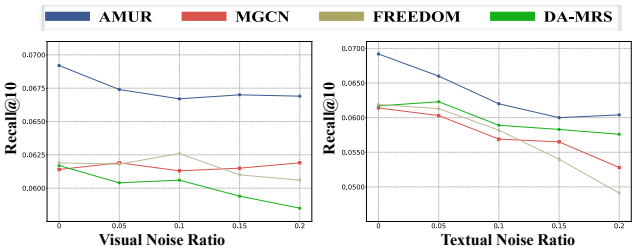


Figure 8: Performance under corrupted modality content.

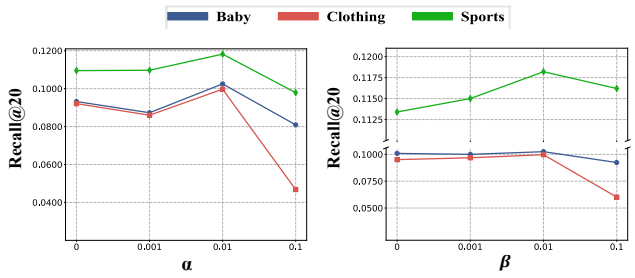


Figure 9: Hyperparameter study on coefficients α and β.

number of epochs, total training cost, and Recall@20 of different models. As shown in Table 4, AMUR achieves the best performance with a moderate training cost. Although MGCN is slightly faster, its performance is clearly lower than AMUR. Compared with strong baselines such as DA-MRS, MIG-GT, and TMLP, AMUR achieves higher Recall@20 while requiring less training time. These results show that AMUR provides a good trade-off between effectiveness and efficiency, indicating that the proposed selective alignment objectives do not introduce excessive training overhead.

5 Conclusions and Future Work

In this work, we proposed AMUR, an information-guided selective modality-interest alignment framework for multimodal recommendation. AMUR refines modality information toward user interests and further aligns shared interest-related semantics across modalities, while preserving modality-specific complementary information. Experiments on three real-world datasets demonstrate the effectiveness of AMUR, and further analyses verify the contribution of its key components.

In the future, we will explore more fine-grained preference-aligned modality modeling and extend the proposed selective alignment idea to broader recommendation scenarios.

Table 4: Runtime analysis on the Sports dataset.

Model	Epochs	Training Cost	R@20
BM3	246	28.4 min	0.0981
MGCN	76	**11.4 min**	0.1105
LGMRec	60	15.7 min	0.1090
DA-MRS	63	43.8 min	0.1073
MIG-GT	300	18.5 min	0.1085
TMLP	153	36.3 min	0.1137
AMUR	67	12.9 min	**0.1187**

6 GenAI Usage Disclosure

Generative AI tools were used solely for language polishing and improving the clarity of the manuscript. All research ideas, methodological designs, experiments, analyses, and conclusions were developed and verified by the authors.

Acknowledgments

This research is supported in part by National Science Foundation of China (No. 62472277, No. 62072304), Shanghai Municipal Science and Technology Commission (No. 21511104700), the Shanghai East Talents Program (2023-177).

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* (2016).
[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*. PMLR, 531–540.
[3] Yu Chen, Lingfei Wu, and Mohammed Zaki. 2020. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. *Advances in neural information processing systems* 33 (2020), 19314–19326.

[4] Xiaoxi Cui, Weihai Lu, Yu Tong, Yiheng Li, and Zhejun Zhao. 2025. Multi-Modal Multi-Behavior Sequential Recommendation with Conditional Diffusion-Based Feature Denoising. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1593–1602.

[5] Xiaoyu Du, Zike Wu, Fuli Feng, Xiangnan He, and Jinhui Tang. 2022. Invariant representation learning for multimedia recommendation. In *Proceedings of the 30th ACM international conference on multimedia*. 619–628.

[6] Qile Fan, Penghang Yu, Zhiyi Tan, Bing-Kun Bao, and Guanming Lu. 2025. BeFA: A General Behavior-driven Feature Adapter for Multimedia Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11634–11644.

[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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8454–8462.

[8] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[9] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong 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.

[10] Yue He, Jingxi Xie, Fengling Li, Lei Zhu, and Jingjing Li. 2025. Flip is Better than Noise: Unbiased Interest Generation for Multimedia Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6298–6306.

[11] Jun Hu, Bryan Hooi, Bingsheng He, and Yinwei Wei. 2025. Modality-Independent Graph Neural Networks with Global Transformers for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11790–11798.

[12] Junjie Huang, Jiarui Qin, Yong Yu, and Weinan Zhang. 2025. Beyond graph convolution: Multimodal recommendation with topology-aware mlps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11808–11816.

[13] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[14] Dongho Jeong, Taeri Kim, Donghyeon Cho, and Sang-Wook Kim. 2025. MELON: Learning Multi-Aspect Modality Preferences for Accurate Multimedia Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1850–1859.

[15] Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. 2024. Diffmm: Multi-modal diffusion model for recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7591–7599.

[16] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[17] Yi Li, Qingmeng Zhu, Changwen Zheng, and Jiangmeng Li. 2024. MSI: Multimodal Recommendation via Superfluous Semantics Discarding and Interaction Preserving. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*. 814–823.

[18] Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, and Weinan Zhang. 2024. An Aligning and Training Framework for Multimodal Recommendations. *arXiv preprint arXiv:2403.12384* (2024).

[19] Haokai Ma, Yimeng Yang, Lei Meng, Ruobing Xie, and Xiangxu Meng. 2024. Multimodal Conditioned Diffusion Model for Recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*. 1733–1740.

[20] Wenze Ma, Chenyu Sun, Yanmin Zhu, Zhaobo Wang, Xuhao Zhao, Mengyuan Jing, Jiadi Yu, and Feilong Tang. 2025. Generating Difficulty-aware Negative Samples via Conditional Diffusion for Multi-modal Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 979–988.

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

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[23] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[24] Hongzu Su, Jingjing Li, Fengling Li, Ke Lu, and Lei Zhu. 2024. SOIL: Contrastive Second-Order Interest Learning for Multimodal Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5838–5846.

[25] Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and S Yu Philip. 2022. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4165–4174.

[26] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* 25 (2022), 5107–5116.

[27] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* 25 (2021), 1074–1084.

[28] 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.

[29] 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.

[30] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, and Edith CH Ngai. 2025. The best is yet to come: Graph convolution in the testing phase for multimodal recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6325–6334.

[31] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith CH Ngai. 2025. Mentor: multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12908–12917.

[32] Guipeng Xv, Xinyu Li, Yi Liu, Chen Lin, and Xiaoli Wang. 2025. Unveiling the Impact of Multi-modal Content in Multi-modal Recommender Systems. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6093–6102.

[33] Guipeng Xv, Xinyu Li, Ruobing Xie, Chen Lin, Chong Liu, Feng Xia, Zhanhui Kang, and Leyu Lin. 2024. Improving Multi-modal Recommender Systems by Denoising and Aligning Multi-modal Content and User Feedback. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3645–3656.

[34] Wei Yang and Qingchen Yang. 2024. Multimodal-aware Multi-intention Learning for Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5663–5672.

[35] Wei Yang, Rui Zhong, Yiqun Chen, Shixuan Li, Heng Ping, Chi Lu, and Peng Jiang. 2025. FITMM: Adaptive Frequency-Aware Multimodal Recommendation via Information-Theoretic Representation Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6193–6202.

[36] Yimeng Yang, Haokai Ma, Lei Meng, Shuo Xu, Ruobing Xie, and Xiangxu Meng. 2025. Curriculum conditioned diffusion for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 13035–13043.

[37] 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.

[38] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. LD4MRec: Simplifying and Powering Diffusion Model for Multimedia Recommendation. *arXiv preprint arXiv:2309.15363* (2023).

[39] 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.

[40] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 353–362.

[41] 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.

[42] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9154–9167.

[43] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473* (2023).

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

[45] 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.

[46] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.