RESEARCH-ARTICLE

DeCoRec: Decoupled Collaborative Refinement for Multi-Modal Sequential Recommendations

ZHAOQI CHEN, Zhejiang University, Hangzhou, Zhejiang, China

WANNI XU, Zhejiang University, Hangzhou, Zhejiang, China

YUNFENG ZHANG

YAWEI HOU

ZHENYU WEN, Zhejiang University of Technology, Hangzhou, Zhejiang, China

CONG WANG, Zhejiang University, Hangzhou, Zhejiang, China

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

DeCoRec: Decoupled Collaborative Refinement for Multi-Modal Sequential Recommendations

Zhaoqi Chen
Zhejiang University
Hangzhou, China
22360332@zju.edu.cn

Wanni Xu
Zhejiang University
Hangzhou, China
22360280@zju.edu.cn

Yunfeng Zhang
Hello Inc.
Shanghai, China
zhangyunfeng594@hellobike.com

Yawei Hou
Hello Inc.
Shanghai, China
houyawei376@hellobike.com

Zhenyu Wen
Zhejiang University of Technology
Hangzhou, China
zhenyuwen@zjut.edu.cn

Cong Wang*
Zhejiang University
Hangzhou, China
cwang85@zju.edu.cn

Abstract

While multi-modal features offer rich semantic signals to enhance sequential recommendation systems, their integration with ID-based embeddings remains challenging. Conventional fusion strategies often degrade performance despite the semantic potential of multimodal data. Through empirical analysis, we identify asymmetric convergence dynamics between rapidly adapting ID embeddings and slowly evolving modality representations as the fundamental barrier. To address this, we propose *DeCoRec*, a novel framework to decouple ID and modality optimization trajectories to prevent gradient interference. To further reconcile ID and multi-modal data, we introduce modality-aware interest clustering and cross-modal contrastive learning to align semantic neighborhoods with behavioral patterns. Extensive experiments demonstrate 5-7% improvements in NDCG/HiT metrics against the existing schemes and particular robustness in cold-start scenarios. The code is available: https://github.com/KIKIENAO/decorec

CCS Concepts

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

Keywords

Sequential Recommendation, ID-Based Recommendation, Modality-Based Recommendation

ACM Reference Format:
Zhaoqi Chen, Wanni Xu, Yunfeng Zhang, Yawei Hou, Zhenyu Wen, and Cong Wang*. 2025. DeCoRec: Decoupled Collaborative Refinement for Multi-Modal Sequential Recommendations. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), Oct. 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755475

*Correspondence to Cong Wang (cwang85@zju.edu.cn).

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755475

1 Introduction

Sequential recommendation (SR) systems aim to predict a user's next interaction by modeling dynamic behavioral patterns from historical interaction sequences [12, 13, 23, 42]. Unlike collaborative filtering, which relies on aggregated user-item preferences [11], SR leverages temporal dependencies to capture evolving user interests [4, 19, 31]. This temporal awareness has made SR an indispensable part for real-world platforms from e-commerce (personalized product feeds such as Amazon and Ebay) to streaming services (next-video recommendations such as Youtube and TikTok).

Modern SR predominantly rely on ID-based recommendation [36], where unique user/item IDs are converted into learnable embeddings. Despite their long-standing dominance, ID-centric approaches suffer from a serial limitations of cold-start scenarios with sparse interaction [6, 35], popularity bias of unfairness [1, 40], cross-platform transferability due to non-shareable IDs [5, 32], and the isolation from the advance of foundational models such as Transformers [2]. On the other hand, the recent advance in foundational models has transformed modality representation with semantically rich features to surpass ID embeddings, which allows models to infer relationships beyond interaction patterns [14, 27, 33, 38, 39]. A plethora of recent research aims to replace IDs with modal features and improve their performance [33, 41], whereas in warm-start scenarios, ID-based approaches still outperform their modality-only counterparts since IDs provide an explicit guideance of collaborative relations [26].

A natural solution is to combine ID and modality features [7, 9, 10, 41]. Surprisingly, however, naive fusion strategies degrade performance compared to ID-only or modality-only baselines, which are attributed to the inconsistency between ID and modal embeddings [37], or their distinct rationales as symbolic identifier for ID and fine-grained preferences for modal embeddings [39]. Different from these views, our empirical study unveils a new insight: naive multi-modal fusion introduces interference that disrupts collaborative signals, which could reduce accuracy by 3-10%. These findings contradict the expectation that richer modality features should enhance recommendations, which raises a critical question: *If multi-modal features are semantically powerful, why do they fail to enhance SR when fused with IDs?*

Different from [37], we identify *asymmetric convergence* as the root cause. Intuitively, low-dimensional ID embeddings, distilled from dense interaction patterns with structural cues [26], converge

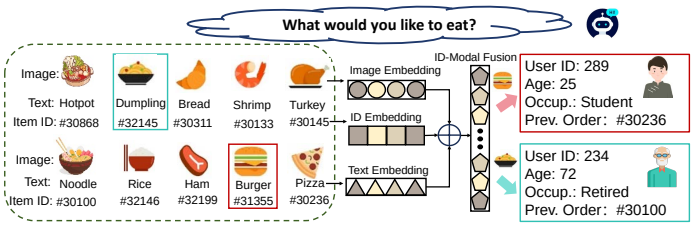

Figure 1: An example of fusing modality and ID embeddings.

rapidly to stable collaborative representations. In contrast, modality encoders – processing high-dimensional content like images or text – require prolonged training to stabilize semantic relationships. Joint optimization forces the model to reconcile conflicting objectives: preserving fast-converging collaborative signals while adapting to slow-evolving semantic patterns.

To address this, we propose Decoupled Collaborative Refinement for Multi-Modal Sequential Recommendation (DeCoRec), a framework designed to disentangle modality and ID features in SR. Departing from the prior attempts that fix pre-trained modality features [37], DeCoRec separates the optimization of modality and ID parameters into distinct phases. The first phase trains modality encoders exclusively on semantic sequences to build stable cross-item relationships; the second phase freezes modal parameters and refines ID embeddings with collaborative signals, thus preventing gradient interference across modalities. To further bridge semantic and structural gaps, we also introduce modality-aware interest clusters through k-means grouping of enriched embeddings, followed by cross-modal contrastive learning. This aligns semantic neighborhoods with behavioral patterns and effectively guides the model to prioritize recommendations within coherent interest domains and avoid semantically inconsistent suggestions. Our main contributions are summarized as follows:

- ✧ **Motivation.** We identify the gradient mismatch between ID and modality features as the primary barrier to effective fusion, which is overlooked in disentangling ID and modality [39]. To our best knowledge, this is one of the few efforts that aim to reconcile ID and multi-modality.
- ✧ **Methodology.** We propose DeCoRec, a decoupled training framework that isolates modality and ID-based semantics into a two-stage phased training. The framework supports synergistic integration of modality-interest clustering and cross-modal retrieval to align modal interaction-based interest clusters and leverage cross-modal information to enrich collaborative signals.
- ✧ **Evaluation.** Extensive experiments on diverse Amazon datasets demonstrate that DeCoRec improves the performance up to 5-7% in terms of NDCG/HiT metrics compared to the benchmarks, with particular robustness in cold-start scenarios compared to UnisRec [15] and MissRec [33]. We also present ablation studies for the modular design and visualization providing insights into the interaction between collaborative and semantic signals.

2 Related Works

2.1 Pure ID-Based SR

Sequential recommendation aims to predict a user's next interaction based on their historical behavior. Early approaches such as Markov chain-based models [10, 18] focus on capturing immediate item-to-item transitions, whereas these methods are restricted to low-order sequential patterns, and fall short to model complex, long-range dependencies in user sessions. Thus, the subsequent research shifts to RNN-based models [12, 13, 23] to encode sequential dynamics. While RNNs improve performance by modeling higher-order associations, their emphasis on strict temporal order makes them susceptible to overfitting noisy or inherent randomness in recommendation sequences.

Recently, transformer-based models such as Transformers4Rec, BERT4Rec, and SASRec [4, 19, 31] are proposed. By leveraging self-attention to capture long-range dependencies, these methods become the mainstream SR due to their capacity to model complex user behaviors effectively. Other approaches include graph [3, 25, 29] and diffusion-based models [24, 34]. However, purely ID-based methods heavily depend on interaction frequency, sequence patterns and are subject to problems of cold-start items or users, so their performance is often bottlenecked by their inability to leverage rich contextual signals.

2.2 Multi-Modal SR

The integration of multi-modal features (e.g., text, images, categories) has opened up new opportunities for enhancing SR. Modal information not only provides rich semantic signals that complement sparse interaction data, but also enables inferring item relationships beyond co-occurrence patterns. LATTICE utilizes the latent structural similarity across modalities to refine item embeddings and enhance item representations [38]. Multi-modal data enables finer-grained modeling of user preferences by introducing descriptive attributes [39], which helps mitigate ambiguity in interaction sequences. MISSRec leverages multi-modal features instead of relying solely on sparse, non-transferable ID features via an interest-aware encoder-decoder [33]. UniSRec utilizes text information to learn transferable representations across different recommendation scenarios [14]. These efforts bypass the reliance on explicit user IDs by representing items through their modal features and predicting the next item based on semantic patterns rather than item IDs alone.

Arguebly, modality-driven approaches still face limitations, since item IDs remain critical for capturing latent information derived from historical user-item interactions. The study in [37] demonstrates that under warm-start scenarios, ID-based models like SASRec [19], DSSM [16] outperform their modality-only counterparts, as IDs carry inherent collaborative relationships [26]. This highlights the indispensable role of ID embeddings.

2.3 Integrating ID and Modal Features for SR

Current research on integrating ID embeddings and multi-modal features remains limited, with most methods adopting either decoupled or loosely coupled strategies [7, 26, 27, 39]. DIMO disentangles ID and modal features through counterfactual inference and proxy learning, decomposing recommendations into separate collaborative (ID-driven) and semantic (modality-driven) pathways [39]. IDSF views ID embedding as subtle features to supplement multimodal recommendation and employs attention mechanisms to fuse structural (ID-based) and content-based (modal) signals [26].

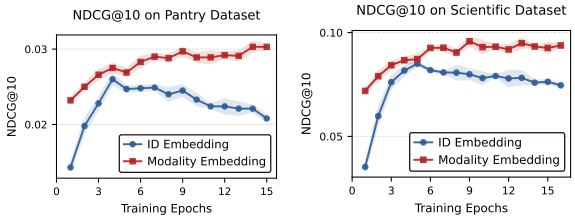

Figure 2: Divergent convergence dynamics between modal and ID embeddings on Pantry and Scientific datasets.

ALIGNRec aligns modal and ID representations via cross-modal projection to unify their feature spaces [27]. However, these methods overlook a critical factor of the disparity in convergence dynamics between ID and modal parameters. Different from these works that mostly focus on the feature level, we combine architectural and feature-level enhancement to synchronize the convergence trajectories of ID and modal parameters. The most relevant work to ours is LGMRec [7] whereas their primary solution is an architectural method. We posit asymmetric convergence as a more fundamental underlying reason for why joint optimization is challenging.

3 Divergence between Modal and ID

Before we describe the details of our design, we first present the key insights of the divergence between modal and ID information as the primary cause of performance degradation. As illustrated in Fig. 2, modality-only and ID-only models exhibit fundamentally different optimization trajectories: ID-based models converge rapidly within early iterations, reaching the peak performance around 5 iterations followed by overfitting, while modality-based counterparts demonstrate sustained performance gains throughout extended training. This asymmetry in learning paces creates divergent optimization objectives during the joint training process. To further validate this, we examine through the lens of gradient dynamics.

Gradient Dynamics. Fig. 3 presents the trace of cosine similarity in the process of training between the ID and modality embeddings. Fig. 3(a) looks into the first 50 training iterations of the gradient similarity in the training of model. In contrast to the global view of Fig. 3(b), we can see that ID and modal gradients turn from the opposite directions to partially correlated as the training progresses. Fig. 3(c) also validates the prior observation that ID embeddings tend to converge fast within a few epochs as the gradient magnitude approches zero within 50 iterations while modality embeddings converge much slower with the gradient magnitude hovering above 0.25.

Joint Modality-ID Training. As a result of gradient divergence, Fig. 4 shows that incorporating ID embeddings with modality features degrades performance compared to modality-only inputs, which contradicts conventional wisdom that ID features with structural cues should theoretically enhance multimodal representations [26]. We posit that the faster-converging ID component effectively "locks" the model into suboptimal regions of the parameter space before modality representations mature, thereby limiting the overall performance. In addition, Fig. 4 also indicates that simple linear projection layers are inadequate for modeling complex user behavioral patterns (Modal+Linear) and a more delicate design is needed to enhance expressiveness. In contrast to the previous

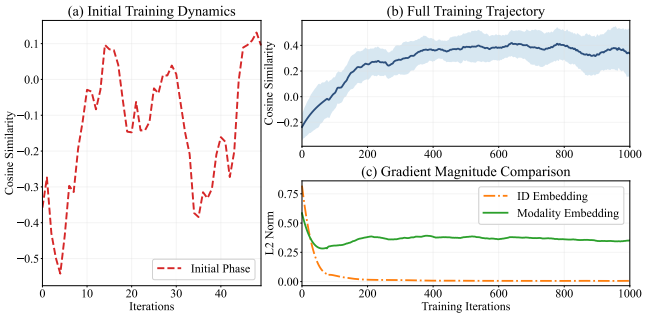

Figure 3: Gradient divergence between modal-only and ID-only models: a) examining the cosine similarity in the first 50 epochs; b) cosine similarity over the entire course of training; c) trace of gradient magnitudes.

studies, we attempt to leverage more complex adaptive structures to fully exploit the potential of modality features. Next, we propose a new framework to decouple the mutual impact of gradient interference between ID and modal embeddings.

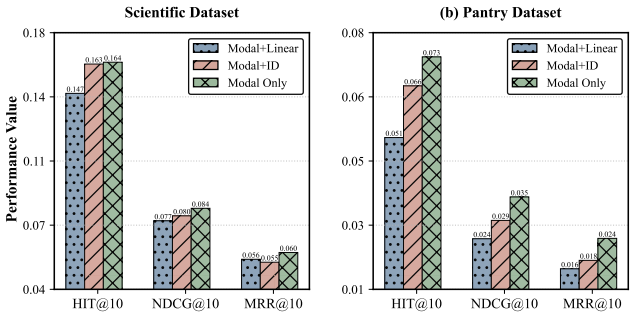

Figure 4: Naive integration of modal and ID embeddings undermines the overall performance across different metrics of HIT@10, NDCG@10 and MRR@10.

4 Designs of DeCoRec

In this section, we describe the designs of DeCoRec as formulated in the following. Define a user set $\mathcal{U}$ and an item set $\mathcal{X}$; and a modality set $\mathcal{M} = \{v, t\}$. Here, v and t represent different modal features, such as visual and textual features. The item ID embedding is represented as $\mathbf{E}^{id} = \{\mathbf{e}_1^{id}, \ldots, \mathbf{e}_i^{id}, \ldots, \mathbf{e}_{|\mathcal{X}|}^{id}\}$, with dimensions of $\mathbb{R}^{|\mathcal{X}| \times d}$, where d denotes the embedding dimension. The item modal characteristics are represented as $\mathbf{E}^{m} = \{\mathbf{e}_1^{m}, \ldots, \mathbf{e}_i^{m}, \ldots, \mathbf{e}_{|\mathcal{X}|}^{m}\}$, with dimensions of $\mathbb{R}^{|\mathcal{X}| \times d_m}$, where $m \in \mathcal{M}$ represents different modal characteristics and d_m is the dimension of the modal characteristic.

4.1 Design Overview

Fig. 5 illustrates the main components of the proposed framework: ❶ *Co-occurrence Representation*, which injects collaborative information into different modal data (including image/text/IDs); ❷ *Collaborative Interest Clustering*, which refines modal-collaborative information interest points through k-means clustering; ❸ *Phased Training*, that decouples gradient interference among different modalities; ❹ *Multi-modal Fusion* to form the final sequence representation for retrieving the next item of most interest to the user.

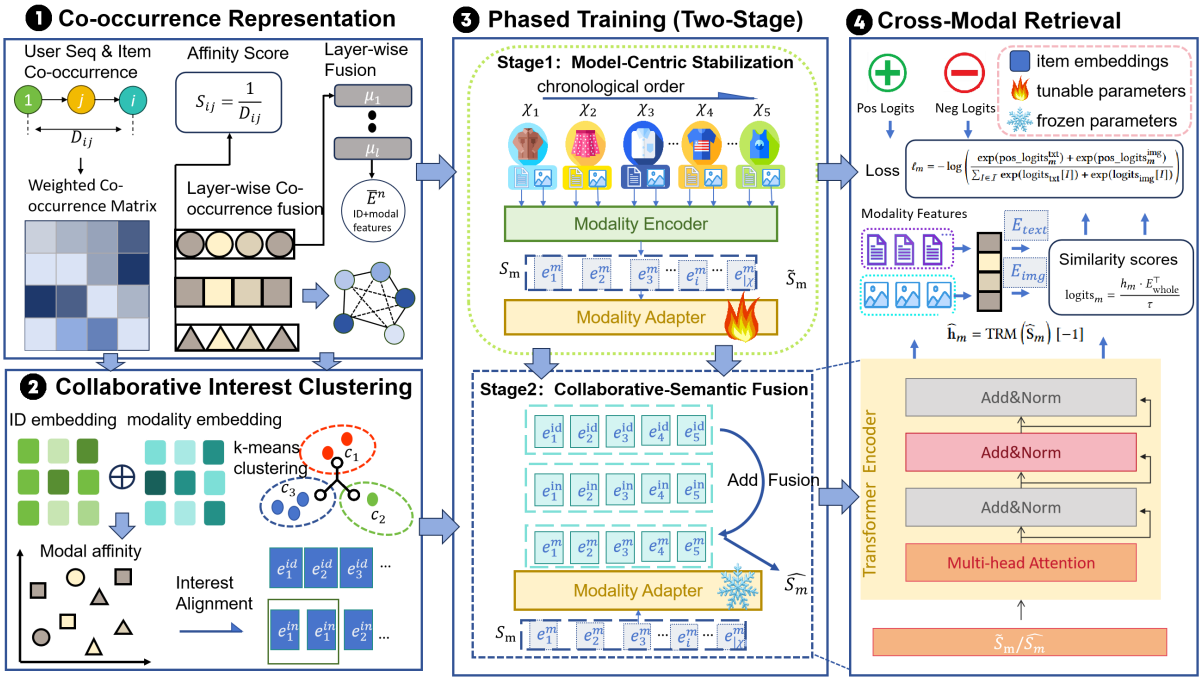

Figure 5: The main procedures of DeCoRec: ❶ Co-occurrence Representation injects collaborative information into different modal data (including image/text/IDs); ❷ Collaborative Interest Clustering refines modal-collaborative information interest points through k-means clustering; ❸ (Two-Stage) Phased training decouples gradient interference among different modalities that in the first stage, the ID embeddings are frozen and the modality embeddings are trained, and vice versa for the second stage; ❹ Cross-Modal Retrieval forms the final sequence representation for retrieving the next item of most interest to the user.

4.2 Co-occurrence Representation

Sequential co-occurrence relationships refer to the collaborative information related to user behavior. Since raw ID embeddings lack explicit behavioral context and modal features ignore collaborative patterns, we follow [21, 39] to construct a co-occurance scheme that encodes how frequently items appear together in sequences, modulated by their positional proximity. This is done via injecting behavioral signals into the ID embeddings of items. For two items i, j, the behavioral affinity score S_{ij}^u between them for user u decays with their positional distance D_{ij},

$$S_{ij}^u = \begin{cases} \frac{1}{D_{ij}} & \text{if } x_i, x_j \in \mathcal{S}^u, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

We sum up the behavioral affinity scores S_{ij}^u between items i and j across all user sequences and item pairs, $\mathrm{S}_{ij} = \sum_{u \in \mathcal{U}} S_{ij}^u$. For layers $l \in \mathcal{L}$, the injection strength μ_l^n is adaptively tuned to balance behavioral and semantic signals,

$$\mu_l^n = \sigma(\mathbf{W}_\mu \cdot \texttt{concat}(\mathbf{E}_l^n, \mathrm{S}_l)), n \in \{\text{id}\} \cup \mathcal{M} \tag{2}$$

in which $\sigma(\cdot)$ is the sigmoid activation and $\mathbf{W}_\mu$ are the model parameters to learn the influence of co-occurence patterns $\mathcal{S}_l$ for layer l. Unlike static fusion [33, 41], layer-wise gating prevents overly noisy co-occurrence signals such that the higher layers (capturing abstract patterns) receive smaller μ_l^n and vice versa. Then, the enriched embedding $\overline{\mathbf{E}}^n$ is computed as,

$$\overline{\mathbf{E}}^n = \mathbf{E}^n + \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mu_l^n \mathrm{S}_l \mathbf{E}_l^n, \quad n \in \{\text{id}\} \cup \mathcal{M}. \tag{3}$$

μ_l^n is an adaptive gating paramter controlling cross-layer signal fusion. $S_l \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ is the normalized co-occurance matrix for layer l. $\mathbf{E}_l^n \in \mathbb{R}^{|\mathcal{X}| \times d}$ is the input feature with co-occurence cues. The modal features $\overline{\mathbf{E}}^m (m \in \mathcal{M})$ construct item-item graphs where edges encode both semantic similarity and behavioral co-occurrence and the ID embedding $\overline{\mathbf{E}}^{\text{id}}$ integrate collaborative signals, that enables robust next-item prediction even for cold-start items via behavioral propagation [21].

4.3 Collaborative Interest Clustering

The enriched modal representations $\overline{\mathbf{E}}_m \in \mathbb{R}^{|\mathcal{X}| \times d}$ exhibit item similarities driven by two complementary factors: 1) *Modal Affinity.* Items share intrinsic semantic attributes (e.g., visual appearance, textual descriptions); 2) *Collaborative Correlation.* Items co-occur frequently in user interaction sequences.

Cluster Centroid Computation. To exploit this duality, we cluster $\overline{\mathbf{E}}_m$ to identify prototypical user interests that unify modal and behavioral similarity. The integration of clustering aligns with the core recommendation paradigm of suggesting items that are either semantically analogous to interacted content or behaviorally affiliated through historical co-occurrence [14]. The clusters act as pseudo-labels that guide the model to focus on intra-interest item

relationships during training,

$$\mathbf{C} = \texttt{K-Means}\left(\overline{\mathbf{E}}^{m}, k\right), \quad \mathbf{C} \in \mathbb{R}^{k \times d}. \tag{4}$$

We apply k-means clustering [8] to $\overline{\mathbf{E}}_m$ and obtain k interest prototypes. Each centroid $\mathbf{C}_j$ represents a latent interest combination of modal and collaborative similarity.

Interest Alignment. For item x_i, we assign it to the nearest cluster centroid via,

$$\mathbf{C}_{\mathbf{i}}^{\text{interest}} = \arg \min_{j \in \{1,2,\ldots,k\}} \left\|\mathbf{E}_i - \mathbf{C}_j\right\|_2. \tag{5}$$

Each item's cluster assignment is mapped to a dense interest embedding,

$$\mathbf{e}_i^{\text{interest}} = \texttt{MLP}(\mathbf{C}_i^{\text{interest}}), \tag{6}$$

where $\texttt{MLP}(\cdot)$ is the multi-layer perceptron that nonlinearly projects cluster indices to the recommendation space. For cold-start items with sparse interactions, interest assignments inferred from modal features could also stabilize collaborative signal propagation.

Visualization. Fig. 6 visualizes the effectiveness of clustering as the introduction of clusters facilitates the discrimination of embeddings in representation space. Without clustering, those predictions represented as the triangles become more difficult to distinguish than the clustered ones. More evaluations are available in Sec. 5.3.

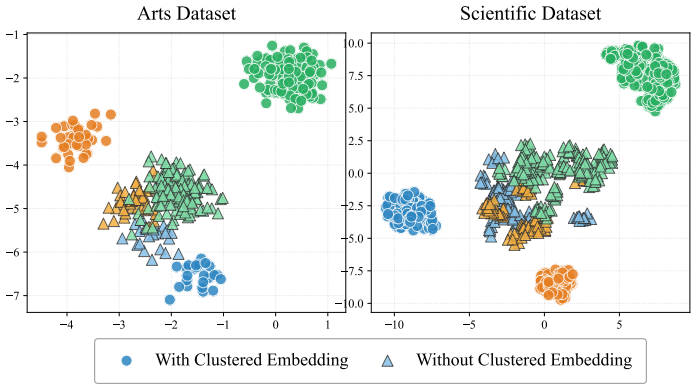

Figure 6: Visualization of the effectiveness of interest clustering with T-SNE. In contrast to the "with clustering" approach, the triangles represent the approach "without clustering", which becomes more difficult to distinguish.

4.4 Phased Training for Convergence Alignment

A key novelty of our framework is the two-stage training scheme that resolves the convergence speed mismatch between modality encoders (slow, high-dimensional parameters) and ID embeddings (fast, low-dimensional). Unlike joint training in prior work [37, 41], which might risk modal gradient dominance destabilizing ID representations, our phased approach explicitly separates modal stabilization from collaborative fusion and prevents gradient conflict.

Stage 1: Model-Centric Stabilization. In the first stage, we only use the modal data to anchor features for user behavior dynamics in order to avoid interference from the volatile ID signals.

$$\mathbf{S}_m = \left[\mathbf{e}_1^m, \mathbf{e}_2^m, \ldots, \mathbf{e}_n^m\right], \quad \mathbf{e}_i^m \in \mathbb{R}^d, \tag{7}$$

where $\mathbf{e}_i^m$ is the modal feature of item x_i. To ensure the transition of features from the semantic space to the user behavior space, we employ a modal adapter to project raw features into a latent space and align with behavioral sequences via a `Linear-ReLU-Linear` structure,

$$\widetilde{\mathbf{S}}_m = \mathbf{W}_2 \cdot \text{ReLU}\left(\mathbf{W}_1 \mathbf{S}_m + \mathbf{b}_1\right) + \mathbf{b}_2, \tag{8}$$

in which $\mathbf{W}_1 \in \mathbb{R}^{d \times d_h}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_h \times d}$ $(d_h < d)$ are weight matrices to enforce dimensionality reduction and capture discriminative behavioral patterns. $\mathbf{b}_1$ and $\mathbf{b}_2$ are bias terms. After adaptation, we use a transformer encoder $\texttt{TRM}(\cdot)$ to process the adapted sequence,

$$\overline{\mathbf{S}}_m = \texttt{TRM}\left(\widetilde{\mathbf{S}}_m\right), \quad \mathbf{h}_m = \overline{\mathbf{S}}_m[-1] \in \mathbb{R}^d, \tag{9}$$

$\mathbf{h}_m$ is the hidden state capturing sequence-level intent.

Stage 2: Collaborative-Semantic Fusion. At this stage, we integrate ID-driven collaborative signals without disrupting stabilized modal representations: freeze the modal adapter and unfreeze the ID embeddings. This combines stabilized modal features with ID/interest signals, while ensuring behavioral semantics from Stage 1 remain intact, whereas end-to-end finetuning may cause inevitable feature drift of learned modal embeddings [17, 27].

$$\widehat{\mathbf{S}_{\mathbf{m}}} = \underbrace{\widetilde{\mathbf{S}}_{\mathbf{m}}}_{\text{Stabilized Modality}} + \underbrace{\mathbf{S}_{\mathbf{id}}}_{\text{Collaborative}} + \underbrace{\mathbf{S}_{\mathbf{interest}}}_{\text{Hybrid Intent}} \tag{10}$$

where $\mathbf{S}_{\text{id}} = [\mathbf{e}_1^{\text{id}}, \cdots, \mathbf{e}_n^{\text{id}}]$ and $\mathbf{S}_{\text{interest}}$ are from Section 4.3. After processing through the transformer, we obtain the final representation of a single modality:

$$\widehat{\mathbf{h}}_m = \text{TRM}\left(\widehat{\mathbf{S}}_m\right)[-1], \tag{11}$$

with the modal adapter frozen to preserve Stage 1 semantics.

4.5 Cross-Modal Information Retrieval

Existing fusion strategies focus on intra-item alignment [14], but neglect intra-item collaborative patterns across modalities, which limits their ability to generalize to sparse or cold-start scenarios. To address this, we propose a cross-modal co-occurrence retrieval. Unlike the traditional methods [10, 13, 18], our approach explicitly models how modalities interact at the system level to enrich collaborative signals: 1) *Cross-Modal Co-Occurrence*. By retrieving complementary information across modalities, we capture latent user interest patterns that transcend single-modality representations; 2) *Sparsity Mitigation*. Cross-modal retrieval compensates for missing or sparse data in one modality (e.g., few item images) by leveraging richer signals from another (e.g., detailed textual descriptions), effectively densifying the interaction graph. Next, we present the loss formulation for cross-modal retrieval.

Loss Formulation. We propose a new contrastive-alike loss function to reward the model for retrieving the correct item across modalities while penalizing irrelevant items. For a user's interaction sequence $\mathbf{S}_m$ (modality $m \in \{\text{text}, \text{img}\}$), we compute similarity scores between $\mathcal{S}_m$ and all candidate item's text/image embeddings E_{whole} represent $[E_{\text{text}}, E_{\text{img}}]$ in the Model-Centric Stabilization stage or $[\hat{E}_{\text{text}}, \hat{E}_{\text{img}}]$ in Collaborative-Semantic Fusion stage,

$$\text{logits}_m = \frac{h_m \cdot E_{\text{whole}}^{\top}}{\tau}, \tag{12}$$

where τ (temperature) controls modality alignment strictness. Lower τ enforces hard alignment between exact matches, while higher ones permit softer matches across semantically related items. Define the target item I^* (the next interacted item) by the user as the positive samples. $E_{\text{txt}}(I^*)$ and $E_{\text{img}}(I^*)$ provide embeddings from both modalities of text and image representations and all other items' embeddings are negative samples in E_{whole},

$$\text{pos\_logits}_m^{\text{txt|img}} = \text{logits}_m[E_{\text{txt|img}}(I^*)]. \tag{13}$$

To prioritize alignment within the current modality while still encouraging cross-modal co-occurrence, we downweight non-current modality positives by a factor α,

$$\text{pos\_logits}_m^{\neg m} = \alpha \cdot \text{pos\_logits}_m^{\neg m}, \quad \alpha \in [0, 1]. \tag{14}$$

If $m = \text{txt}$, the image-positive score $\text{pos\_logits}_m^{\neg m}$ is scaled by α. The loss for modality m combines weighted positive scores and normalizes over all items $\mathcal{I}$:

$$\ell_m = -\log\left(\frac{\exp(\text{pos\_logits}_m^{\text{txt}}) + \exp(\text{pos\_logits}_m^{\text{img}})}{\sum_{I\in\mathcal{I}} \exp(\text{logits}_{\text{txt}}[I]) + \exp(\text{logits}_{\text{img}}[I])}\right) \tag{15}$$

and the total loss aggregates contributions from both modalities, $\ell = \sum_{I\in\mathcal{I}} \ell_m$. The effectiveness of cross-modal retrieval is evaluated in Section 5.4.

Inference Stage. During the inference stage, we obtain the final prediction score for an item by fusing inference results from different modalities,

$$\text{score}_m = \text{softmax}(h_m, E_{whole}, \tau) \tag{16}$$

$$\text{score}_m(I) = \text{logits}_m[I_m] + \alpha \cdot \text{logits}_m[I_{\neg m}], \tag{17}$$

$$\text{score}(I) = \text{score}_{\text{text}}(I) + \text{score}_{\text{img}}(I). \tag{18}$$

5 Experiments

The experiment section aims to answer the following question:

- ✧ **RQ1**: How does DeCoRec compared to the existing techniques?
- ✧ **RQ2 & RQ3**: Are the interest clustering (RQ2) and cross-modal retrieval (RQ3) effective?
- ✧ **RQ4**: Does our modular design consistently enhance the model performance?

5.1 Experimental Setup

Our model is evaluated on five datasets, including "Industrial and Scientific" (Scientific), "Prime Pantry" (Pantry), "Crafts and Sewing" (Arts), "Musical Instruments" (Instruments), and "Office Products" (Office), obtained from the Amazon review dataset [28], which is a large-scale corpora of user-generated product reviews on Amazon and widely utilized for recommendation system research. For multi-modal inputs, we follow [33] to crawl relevant item images and extract features as the input for the image modality. To avoid interference, we perform filtering on the original dataset and exclud entries with missing modalities. The relevant statistics of the dataset is shown in Table 1. Similar to [15, 33, 42], we adopt two standard metrics of Recall (R@k) and NDCG (N@k) to evaluate the retrieval performance and set k to 10, 50.

Dataset	# Items	# Interactions	Sparsity (%)
Arts	9,438	154,642	99.927
Instruments	6,290	121,163	99.889
Office	16,629	400,208	99.960
Pantry	4,588	109,517	99.812
Scientific	1,586	16,784	99.639

Table 1: Statistics of pre-processed Datasets.

5.1.1 Experimental Details. We use CLIP-B/32 to extract image and text features [30], and the corresponding `[cls]` token as the modality representation of the content. We divide the dataset using leave-one-out cross-validation into training, validation, and test sets. During the experiment, we set the maximum length of the user's historical sequence to 50. For the training set, we apply data augmentation using the standard sliding window technique and Adam as the optimizer.

For the phased training: in the first stage, we train the modality data until convergence, using the validation set to evaluate the model's performance. We employ early stopping when the loss on the validation set converges on the last 10 trailing epochs and the checkpoint is stored as the final model. In the second stage, we introduce ID embeddings and keep the parameters of the modality adapter frozen, which has been empirically validated to improve the overall performance.

5.1.2 Baselines. We compare DeCoRec with the following baselines, which include various advanced ID-based and modality-based SR models.

- ✧ **SASRec** [20]: Unidirectional self-attention architecture focusing on short-term patterns through causal attention masks.
- ✧ **BERT4Rec** [31]: Employs bidirectional Transformer layers for masked item prediction, capturing contextual dependencies in the user sequences.
- ✧ **UniSRec** [15]: Universal sequence encoder using item text descriptions with parameter whitening and Mixture-of-Expert adapters for cross-domain transfer learning.
- ✧ **MISSRec** [33]: Multi-modal pre-training framework integrating visual-textual features via interest-aware discovery modules and cross-platform contrastive alignment.
- ✧ **MMMLP** [41]: Multi-modal MLP framework combining ID-based co-occurrence patterns and modality-aware features.
- ✧ **DIMO** [22]: Session-based model disentangling ID effects (collaborative signals) and modality effects (content features) through dual-channel contrastive learning.

5.2 Main Results (RQ1)

Table 2 demonstrates DeCoRec's superiority over six state-of-the-art baselines across five Amazon datasets. Our model achieves the best performance in 16/20 comparisons and the second in another 4 comparisons (a total of 20/20). First, we observe that modality-enhanced models (the proposed DeCoRec, MISSRec [33], UnisRec [15]) consistently outperform ID-only baselines (SASRec [20], BERT4Rec [31]), which validates the necessity of modality features. The 22% average NDCG@10 gap between DeCoRec and SASRec underscores the important of semantic signals in mitigating interaction sparsity. While MISSRec incorporates both text and images, the

Dataset	Metric	SASRec [20]	BERT4Rec [31]	UnisRec [15]	MissRec [33]	MMMLP [41]	DIMO [22]	**DeCoRec***	Improv.
Scientific	Hit@10	0.1523	0.0751	0.1615	0.1660	0.1226	0.1506	0.1779	+7.16%
	Hit@50	0.2923	0.1933	0.3286	0.3200	0.2220	0.3214	0.3559	+8.31%
	NDCG@10	0.0724	0.0354	0.0809	0.0858	0.0786	0.0787	0.0897	+4.34%
	NDCG@50	0.1023	0.0603	0.1166	0.1189	0.0998	0.1153	0.1292	+8.66%
Arts	Hit@10	0.1217	0.0865	0.1617	0.1613	0.1366	0.1197	0.1626	+0.56%
	Hit@50	0.2163	0.1642	0.2839	0.2808	0.2291	0.1948	0.2872	+1.16%
	NDCG@10	0.0758	0.0550	0.0891	0.0980	0.0955	0.0745	0.0962	-1.92%
	NDCG@50	0.0960	0.0714	0.1153	0.1237	0.1152	0.0906	0.1193	-3.55%
Instruments	Hit@10	0.1269	0.1062	0.1543	0.1580	0.1405	0.1558	0.1585	+0.32%
	Hit@50	0.2158	0.1883	0.2658	0.2704	0.2335	0.2706	0.2737	+1.1%
	NDCG@10	0.0910	0.0796	0.0998	0.1112	0.1049	0.0933	0.1110	-0.1%
	NDCG@50	0.1100	0.0970	0.1237	0.1349	0.1246	0.1185	0.1336	-0.7%
Office	Hit@10	0.1290	0.0948	0.1394	0.1411	0.1335	0.1378	0.1415	+0.28%
	Hit@50	0.1958	0.1455	0.2161	0.2221	0.1991	0.2221	0.2252	+1.39%
	NDCG@10	0.0885	0.0673	0.0934	0.0911	0.0900	0.0936	0.0938	+0.21%
	NDCG@50	0.1028	0.0781	0.1098	0.1084	0.1000	0.1105	0.1118	+1.18%
Pantry	Hit@10	0.0488	0.0305	0.0755	0.0757	0.0467	0.0729	0.0797	+5.28%
	Hit@50	0.1366	0.1060	0.1872	0.1849	0.1293	0.1799	0.1920	+2.56%
	NDCG@10	0.0209	0.0145	0.0340	0.0357	0.0243	0.0345	0.0380	+6.44%
	NDCG@50	0.0395	0.0303	0.0577	0.0587	0.0417	0.0571	0.0618	+5.28%

Table 2: Performance comparison on different datasets. The Best and Second Best values are marked in Red and Blue. Improvements indicate the gap between the Best and the second Best schemes.

marginal gain over text-only UniSRec on the "Office" datasets suggests textual descriptions with rich brand/specification details provide stronger intent signals than visual features for tangible goods. This aligns with e-commerce user behavior where textual search predominates visual browsing. DIMO's inconsistent performance reveals the pitfalls of radical ID-modality separation. Our phased integration proves to be more effective as DeCoRec maintains 89% of MISSRec's visual advantages while adding 11% collaborative signal gains through controlled ID injection.

5.3 Collaborative Interest Clustering (RQ2)

We validate the design of modality-interest clustering in Section 4.3 by designing a recall model that uses sequence information to recall the next item a user may be interested in. We recall items using different perspectives of the model: 1) *Interest Perspective*. For the items in the interaction sequence, we recall the top-k items with the highest co-occurrence scores for each item; 2) *Modality Similarity Perspective*. For the items in the interaction sequence, we attempt to recall the top-k items that have the highest modality similarity to each item; 3) *Modality-Interest Perspective*. For the items in the interaction sequence, we recall items that satisfy both the interest perspective and modality similarity perspective simultaneously. By adjusting the top-k parameter to maintain a consistent number of predictions across experiments, we enable a direct comparison of prediction rates between different model variants under equivalent prediction volumes. Table 3 shows that the precision rate of the modality-interest perspective is 1.3-1.7× higher than the single interest and modality similarity perspective. This is because pure modality similarity suffers from semantic ambiguity that visually similar pantry items may serve divergent culinary purposes and it may also miss functionally equivalent substitutes. The intersection of modality-interest perspective could filter a large portion of false positives by enforcing both semantic relevance and co-occurrence likelihood. This explains the importance of combining both types of information for recommendation.

Dataset	View	Recall Count	Predicted Count	Precision Rate
Scientific	Modal Similarity	255	33,149	0.00769
	Interest	320	29,898	0.01070
	Modality-Interest	354	26,854	**0.01318**
Pantry	Modal Similarity	594	147,981	0.00401
	Interest	631	141,906	0.00446
	Modality-Interest	845	143,559	**0.00589**

Table 3: Recall results for different views in the "Scientific" and "Pantry" datasets

5.4 Cross-Modal Information Retrieval (RQ3)

To validate the effectiveness of cross-modal retrieval in Section 4.5, we conduct additional experiments comparing 4 different variants: (1) image (visual embeddings only), (2) text (textual embeddings only), (3) intra (single-modal retrieval without cross-modal interaction), and (4) cross (multimodal retrieval-enhanced training). Fig.7 reveals that visual embeddings generally underperform compared to textual counterparts across most scenarios. Furthermore, naive summation of prediction logits from different modalities occasionally fails to deliver expected performance. These findings emphasize the critical importance of modality-aware information integration and empirically confirm the superiority of the proposed cross-modal retrieval mechanisms over the simplistic fusion approaches.

Uni-Modality Enhancement. We also find that the power of the proposed cross-modal retrieval could even enhance *single-modal inference* (text or image only). As shown in Fig. 8, both visual and textual modalities exhibit consistent improvements across all the datasets, with relative gains reaching up to 12.4% on the Instruments dataset. This enhancement arises from that fact that different modalities share latent representations of user preferences through common abstraction layers in our model. User interactions inherently reflect interest alignment with specific product characteristics, whether conveyed through uni-modal visual appearance or textual

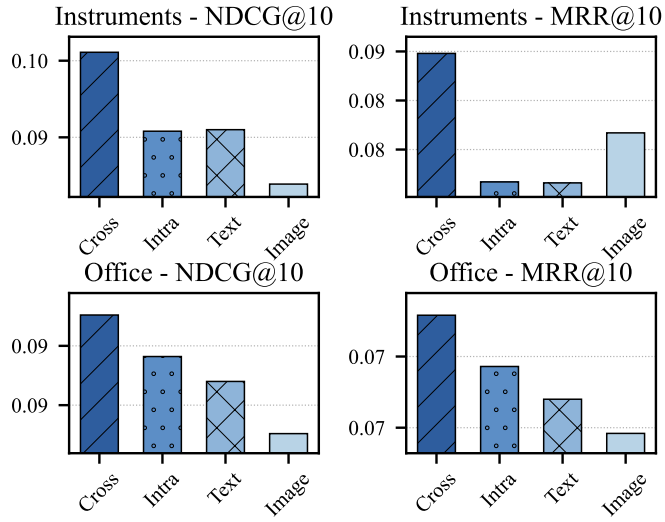

Figure 7: Comarison of cross-modal information retrieval against different variants and cross-modal retrieval.

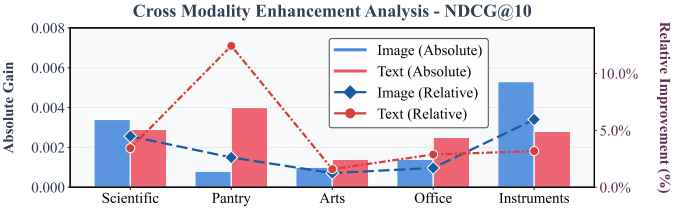

Figure 8: Enhancement on uni-modal retrieval.

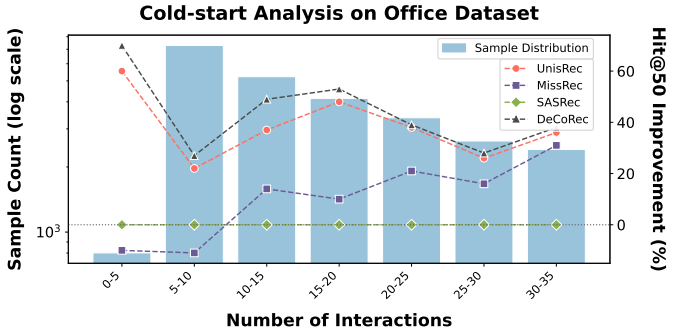

Figure 9: Cold-start analysis on the Office dataset.

descriptions. Our analysis reveals that strong negative perceptions towards either modality (e.g., unfavorable product images or descriptions) typically prevent interaction occurrences.

Cold-start Scenarios. To evaluate the multimodal utilization capabilities in cold-start scenarios, we conduct experiments comparing the performance improvement of different multimodal models using SASRec [20] as the baseline in Fig. 9. Cold users are defined as the bottom 40% quantile by interaction count. We did not artificially corrupt modal features or test varying modality missing rates, focusing on performance with naturally sparse user interaction data. The results demonstrate that DeCoRec exhibits observable advantages under cold-start conditions compared to UnisRec [15] and MissRec [33]. This is because: 1) our framework captures user interests from different perspectives using cross-modal retrieval, which boosts the migration of user interest signals across different modalities, thereby obtaining more representative sequential embeddings; 2) the architectural design with complete modality decoupling, which injects collaborative signals through contrastive learning, which further improves the cold-start performance.

5.5 Ablation Studies (RQ4)

Method	Scientific		Arts	
	Hit@10	NDCG@10	Hit@10	NDCG@10
Linear Adapter + w/o CI	0.1612	0.0817	0.1440	0.0831
Linear Adapter	0.1629	0.0816	0.1540	0.0858
w/o Collaborative Information	0.1680	0.0859	0.1544	0.0887
w/o Phase	0.1650	0.0805	0.1424	0.0778
w/o Frozen	0.1687	0.0867	0.1547	0.0865
w/o cluster embedding	0.1762	0.0878	0.1590	0.0925
DeCoRec*	0.1779	0.0897	0.1626	0.0962

Table 4: Ablation studies on Scientific and Arts datasets.

We evaluate the following architectural variants,

- `linear adapter+w/o collaborative information`: train the model using conventional linear modality adapters without collaborative signals.
- `linear adapter`: train the model with conventional linear modality adapters incorporating collaborative information.
- `w/o collaborative information`: train exclusively on modality data without collaborative signals.
- `w/o cluster embedding`: remove cluster features during second-stage training.
- `w/o frozen`: without freezing the parameters in modal adapter during the second stage.
- `w/o phase`: Baseline model trained without the phased two-stage training framework.

As shown in Table 4, compared with conventional linear adapters, more sophisticated adaptation structures enable superior transfer of modality information from semantic to user behavior space. Even when incorporating ID features – which substantially enhance model performance – variants employing linear adapters remain outperformed by counterparts utilizing more complex adaptation architectures. This empirical evidence substantiates the necessity of adopting the non-linear adaptation structures (i.e. Linear-ReLU-Linear in Eq. (8)).

Furthermore, the performance degradation observed in `w/o frozen` variants validates the criticality of our phased training paradigm. The inherent heterogeneity between modality and ID embeddings persists, even when employing phased integration of ID features. The experimental outcomes from `w/o cluster embedding` highlight the efficacy of modality-interest cluster features, while simultaneously reinforcing the importance of modeling both collaborative and modality signals jointly.

6 Conclusion

This paper proposes DeCoRec, a novel SR framework that effectively integrates ID-based signals and modality features by addressing their asymmetric convergence dynamics. Through the pipelines of co-occurrence-enhanced representation learning, two-phase decoupled training, and cross-modal interest alignment, DeCoRec achieves substantial improvement in recommendation accuracy over the benchmarks while significantly enhancing cold-start performance and training stability.

7 Acknowledgement

This work is supported in part by NSF of Zhejiang Province LZ25F020007, NSFC 62394341, the Fundamental Research Funds for the Central Universities 226202400182, the ZJUCSE-Enflame Cloud and Edge Intelligence Joint Laboratory and the Key Research and Development Program of Zhejiang Province (Grant No:2025C01064).

References

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM conference on recommender systems*. 726–731.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[3] Xingming Chen and Qing Li. 2024. Causality-driven user modeling for sequential recommendations over time. In *Companion Proceedings of the ACM Web Conference 2024*. 1400–1406.

[4] Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation. In *Proceedings of the 15th ACM conference on recommender systems*. 143–153.

[5] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-shot recommender systems. *arXiv preprint arXiv:2105.08318* (2021).

[6] Philip J Feng, Pingjun Pan, Tingting Zhou, Hongxiang Chen, and Chuanjiang Luo. 2021. Zero shot on the cold-start problem: Model-agnostic interest learning for recommender systems. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 474–483.

[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] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[9] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[10] 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.

[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[12] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 843–852.

[13] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *ICLR* (2015).

[14] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 585–593.

[15] Yupeng Hou, Wayne Zhang, Zihan Chen, Chen Wang, Ruobing Xie, and Dawei Yin. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 1–10.

[16] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2333–2338.

[17] Wei Ji, Xiangyan Liu, An Zhang, Yinwei Wei, Yongxin Ni, and Xiang Wang. 2023. Online distillation-enhanced multi-modal transformer for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 955–965.

[18] Santosh Kabbur, Xia Ning, and George Karypis. 2013. Fism: factored item similarity models for top-n recommender systems. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 659–667.

[19] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[20] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *IEEE International Conference on Data Mining (ICDM)*. 197–206.

[21] Changhong Li and Zhiqiang Guo. 2024. Multimodal Difference Learning for Sequential Recommendation. *arXiv preprint arXiv:2412.08103* (2024).

[22] Xiangyu Li, Qiang Zhang, Yang Liu, and Ming Zhou. 2024. Disentangling ID and Modality Effects for Session-Based Recommendations. In *Proceedings of the ACM Web Conference (WWW)*. 1–12.

[23] Xiang Lin, Shuzi Niu, Yiqiao Wang, and Yucheng Li. 2018. K-plet recurrent neural networks for sequential recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1057–1060.

[24] Qidong Liu, Fan Yan, Xiangyu Zhao, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Feng Tian. 2023. Diffusion augmentation for sequential recommendation. In *Proceedings of the 32nd ACM International conference on information and knowledge management*. 1576–1586.

[25] Sijia Liu, Jiahao Liu, Hansu Gu, Dongsheng Li, Tun Lu, Peng Zhang, and Ning Gu. 2023. Autoseqrec: Autoencoder for efficient sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1493–1502.

[26] Yuting Liu, Enneng Yang, Yizhou Dang, Guibing Guo, Qiang Liu, Yuliang Liang, Linying Jiang, and Xingwei Wang. 2023. ID Embedding as Subtle Features of Content and Structure for Multimodal Recommendation. *arXiv preprint arXiv:2311.05956* (2023).

[27] Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. 2024. AlignRec: Aligning and Training in Multimodal Recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1503–1512.

[28] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.

[29] Bo Peng, Ziqi Chen, Srinivasan Parthasarathy, and Xia Ning. 2024. Modeling sequences as star graphs to address over-smoothing in self-attentive sequential recommendation. *ACM Transactions on Knowledge Discovery from Data* 18, 8 (2024), 1–24.

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

[31] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[32] Junting Wang, Adit Krishnan, Hari Sundaram, and Yunzhe Li. 2023. Pre-trained neural recommenders: A transferable zero-shot framework for recommendation systems. *arXiv preprint arXiv:2309.01188* (2023).

[33] Zhenlei Wang, Wayne Zhang, Yupeng Hou, Zihan Chen, Dawei Yin, and Xiang Ren. 2023. MISSRec: Pre-training and Transferring Multi-Modal Interest-Aware Sequence Representation for Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*. 1–15.

[34] Zihao Wu, Xin Wang, Hong Chen, Kaidong Li, Yi Han, Lifeng Sun, and Wenwu Zhu. 2023. Diff4rec: Sequential recommendation with curriculum-scheduled diffusion augmentation. In *Proceedings of the 31st ACM international conference on multimedia*. 9329–9335.

[35] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. 2020. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1469–1478.

[36] Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, et al. 2022. Tenrec: A large-scale multipurpose benchmark dataset for recommender systems. *Advances in Neural Information Processing Systems* 35 (2022), 11480–11493.

[37] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? id- vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2639–2649.

[38] 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.

[39] Xiaokun Zhang, Bo Xu, Zhaochun Ren, Xiaochen Wang, Hongfei Lin, and Fenglong Ma. 2024. Disentangling id and modality effects for session-based recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*. 1883–1892.

[40] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th international ACM SIGIR*

conference on research and development in information retrieval. 11–20.

[41] Yiming Zhang, Xun Wang, Haokai Chen, and Wei Li. 2023. MMMLP: Multi-Modal Multilayer Perceptron for Sequential Recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.

[42] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1893–1902.