MoToRec: Sparse-Regularized Multimodal Tokenization for Cold-Start Recommendation

Jialin Liu[1], **Zhaorui Zhang**[2*], **Ray C.C. Cheung**[1]

[1] City University of Hong Kong
[2] The Hong Kong Polytechnic University
camilla.liu@my.cityu.edu.hk, zhaorui.zhang@polyu.edu.hk, r.cheung@cityu.edu.hk

Abstract

Graph neural networks (GNNs) have revolutionized recommender systems by effectively modeling complex user-item interactions, yet data sparsity and the item cold-start problem significantly impair performance, particularly for new items with limited or no interaction history. While multimodal content offers a promising solution, existing methods result in suboptimal representations for new items due to noise and entanglement in sparse data. To address this, we transform multimodal recommendation into discrete semantic tokenization. We present Sparse-Regularized **M**ultimodal **To**kenization for Cold-Start **Rec**ommendation (**MoToRec**), a framework centered on a sparsely-regularized Residual Quantized Variational Autoencoder (RQ-VAE) that generates a compositional semantic code of discrete, interpretable tokens, promoting disentangled representations. MoToRec's architecture is enhanced by three synergistic components: (1) a sparsely-regularized RQ-VAE that promotes disentangled representations, (2) a novel adaptive rarity amplification that promotes prioritized learning for cold-start items, and (3) a hierarchical multi-source graph encoder for robust signal fusion with collaborative signals. Extensive experiments on three large-scale datasets demonstrate MoToRec's superiority over state-of-the-art methods in both overall and cold-start scenarios. Our work validates that discrete tokenization provides an effective and scalable alternative for mitigating the long-standing cold-start challenge.

Code — https://github.com/Camilla-jl/MoToRec

Introduction

Graph Neural Networks (GNNs) have become the cornerstone of modern recommender systems, achieving state-of-the-art performance by modeling the rich connectivity of user-item interaction graphs (He et al. 2020; Wu et al. 2021; Mo et al. 2024). However, their success relies on dense historical data, exposing a critical vulnerability that manifests as a sharp performance decline in the face of data sparsity, particularly the persistent item cold-start problem (Schein et al. 2002; Li and Lu 2024). To mitigate this, multimodal information offers a promising solution (Lu and Yin 2025; Cui et al. 2025a,b). Early methods simply concatenated content features with ID embeddings (He and McAuley 2016). More recent GNN-based methods have grown in sophistication: for instance, MMGCN (Wei et al. 2019) builds modality-specific interaction graphs, LATTICE (Zhang et al. 2021) introduces item-item semantic graphs to capture latent correlations, and state-of-the-art approaches like FREEDOM (Zhou and Shen 2023) and BM3 (Zhou et al. 2023b) further employ self-supervised contrastive learning to bridge the modality gap.

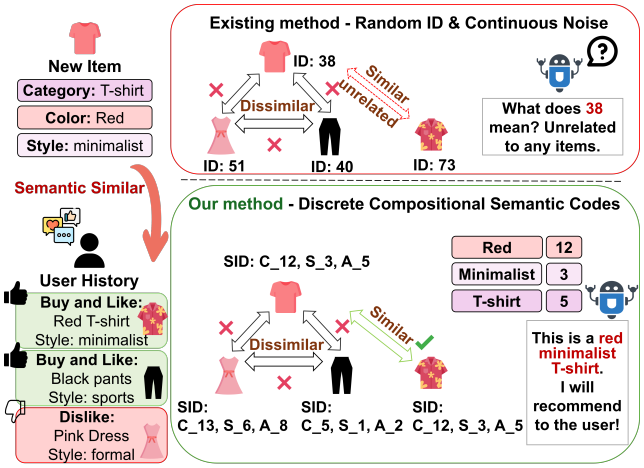

Figure 1: From continuous alignment to discrete compositional codes. Top: Existing methods struggle with noisy alignment and uninformative IDs. Bottom: MoToRec generates robust and interpretable codes for effective cold-start recommendation.

Despite their architectural diversity and progress, these approaches are all fundamentally hampered by a shared challenge: the inherent ambiguity and noise in high-dimensional vector alignment. We term this phenomenon the **"semantic fog"**. It makes mapping a concept like a "red T-shirt" from pixel-based and text-based vectors into a single, coherent point in a high-dimensional space a noise-sensitive and unreliable task. This core issue persists even in the most recent advances leveraging powerful Large Language Models (LLMs) as feature extractors (Geng et al. 2022; Bao et al. 2023; Lu, Fang, and Shi 2020; Ye et al. 2025; Liu et al.

*Corresponding author.

2025a). Aligning the noisy, continuous embeddings from these models often leads to suboptimal, out-of-distribution (OOD) representations, especially for cold-start items. As shown in Figure 1 above, this reliance on noisy alignment in a continuous space is a critical bottleneck limiting the effectiveness of existing methods.

To cut through this "semantic fog", we introduce a novel approach centered on discrete semantic tokenization. We present Sparse-Regularized **M**ultimodal **To**kenization for Cold-Start **Rec**ommendation (**MoToRec**), a framework that learns to convert raw multimodal features into a compositional semantic code. As shown in Figure 1 below, this code consists of a structured sequence of discrete tokens drawn from a learnable codebook, where each token represents a disentangled semantic concept (e.g., style: minimalist, color: red).

In particular, the MoToRec framework operationalizes this vision through several synergistic components. At its core, we leverage a Residual Quantized Variational Autoencoder (RQ-VAE) (Lee et al. 2022) to generate the token sequence. To ensure these tokens are semantically meaningful, we devise a novel sparsity-inducing regularization to promote disentangled representations. Furthermore, to address the data imbalance inherent in recommendation, we introduce an adaptive rarity amplification mechanism to prioritize learning on less frequent items. Finally, a hierarchical multi-source graph encoder robustly fuses these newly created semantic codes with pure collaborative signals, aligning content-based understanding with user interaction patterns.

In summary, our main contributions are summarized as follows:

- We propose a novel approach to multimodal recommendation, reframing it as a discrete semantic tokenization task to explicitly tackle the "semantic fog" and OOD issues prevalent in cold-start scenarios.
- We design MoToRec, an end-to-end architecture that synergistically integrates a sparsely-regularized RQ-VAE tokenizer, adaptive rarity amplification, and multi-source graph encoding for effective and robust signal fusion.
- We perform comprehensive experiments on three large-scale datasets to validate the effectiveness of our proposed approach, demonstrating significant improvements over state-of-the-art methods, particularly in cold-start scenarios.

Related Work

Graph-based Multimodal Recommendation

Graph Neural Networks (GNNs) (Kipf and Welling 2017) have advanced recommender systems. Early graph-based multimodal methods directly inherited the message propagation mechanism to incorporate side information. For example, VBPR (He and McAuley 2016) concatenated content features with ID embeddings, while MMGCN (Wei et al. 2019) built modality-specific user-item graphs to learn distinct representations. LATTICE (Zhang et al. 2021), for instance, constructs an item-item semantic graph to capture latent content correlations. More recently, contrastive learning has been widely adopted to enhance graph-based multimodal recommendations. State-of-the-art approaches like FREEDOM (Zhou and Shen 2023) and BM3 (Zhou et al. 2023b) design sophisticated cross-modal contrastive tasks to better align representations from different modalities. However, by operating in a high-dimensional continuous space, these methods are susceptible to alignment noise, an issue exacerbated in sparse, cold-start scenarios. This limitation motivates our exploration of discrete representations.

Vector Quantization in Recommendation

Vector Quantization (VQ), with its roots in generative modeling (Van Den Oord, Vinyals, and Kavukcuoglu 2017; Zeghidour et al. 2022), has been explored in recommender systems for its efficiency and noise-resilience. Early studies in this area primarily focused on embedding compression, where VQ is used to quantize large embedding tables to reduce memory footprint (Lian et al. 2020). Later, VQ was adopted for generative sequence modeling. VQ-Rec (Hou et al. 2023), for example, treats recommendation as a language modeling task over a discrete codebook of item prototypes, achieving strong performance in sequential recommendation. Unlike these works, which prioritize compression or sequence generation, our work is the first to leverage it to learn compositional, disentangled representations from multimodal content, specifically to address the item cold-start challenge.

Cold-Start Recommendation

The item cold-start problem, where new items have few or no interactions, is a long-standing challenge in recommender systems (Schein et al. 2002; Bobadilla et al. 2013). Meta-learning offers one solution. Inspired by MAML (Finn, Abbeel, and Levine 2017), methods like MeLU (Lee et al. 2019) learn model initializations for rapid adaptation to new items with only a few examples. This "few-shot" approach, however, is ill-suited for the "zero-shot" scenario where no interactions exist. Another powerful direction leverages Large Language Models (LLMs), which have been explored as zero-shot feature extractors or even direct recommenders (Geng et al. 2022; Bao et al. 2023; Lu, Fang, and Shi 2020; Ye et al. 2025). Despite their strong semantic capabilities, LLMs face two major hurdles in practice: (1) their continuous embeddings still suffer from the "semantic fog" alignment issue, and (2) their high computational cost challenges practical deployment. Our approach is motivated by this gap, aiming to distill knowledge into efficient, discrete codes for a scalable and robust zero-shot solution.

Methodology

As illustrated in Figure 2, MoToRec is a comprehensive framework designed to mitigate data sparsity and the item cold-start problem. It achieves this through three core components: an adaptive rarity amplification mechanism, a sparsely-regularized multimodal tokenizer, and a hierarchical multi-source graph encoder. We detail these components below.

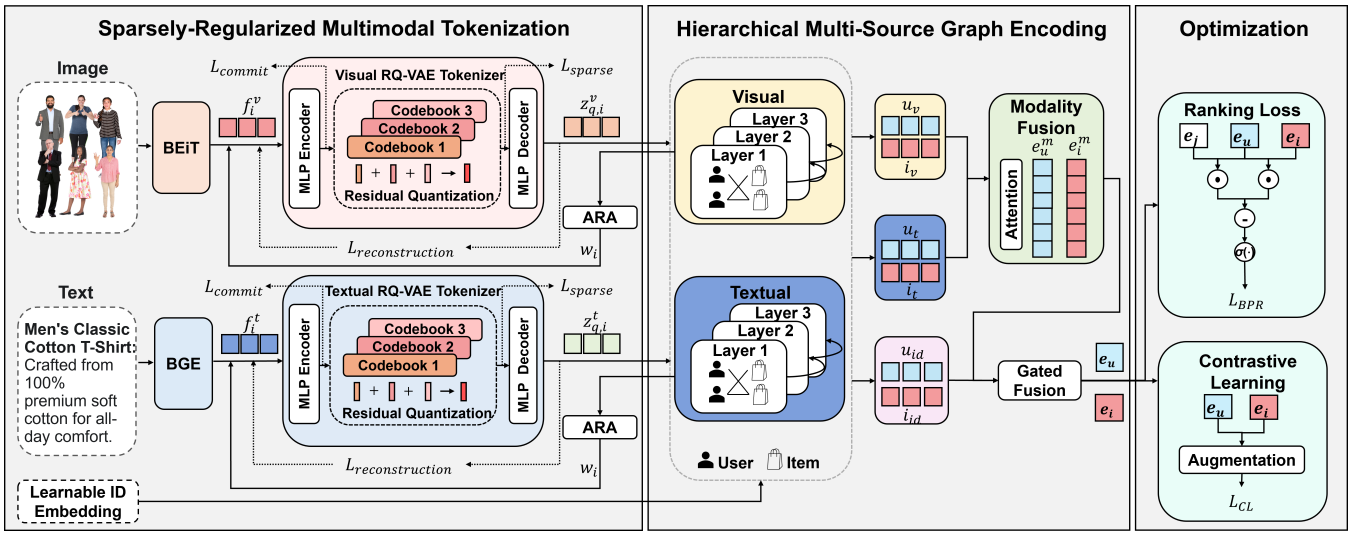

Figure 2: The overall architecture of MoToRec. It consists of three main stages: (1)a sparsely-regularized multimodal tokenization module that converts raw features into discrete codes using RQ-VAEs; (2) a hierarchical multi-source graph encoding module to learn and fuse preferences; and (3) an optimization module. The optimization is guided by both ranking and self-supervised contrastive losses, and is made rarity-aware through a dynamic weighting scheme.

Problem Formulation

Let $\mathcal{U}$ denote the set of users and $\mathcal{I}$ denote the set of items. The historical user-item interactions are represented by a sparse binary matrix $\mathbf{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $R_{ui} = 1$ signifies an implicit feedback (e.g., a click or purchase) from user u to item i. Each item $i \in \mathcal{I}$ is endowed with rich multimodal content features: a visual feature vector $\mathbf{f}_i^v \in \mathbb{R}^{d_v}$ extracted from a pre-trained vision transformer BEiT (Wang et al. 2023b), and a textual feature vector $\mathbf{f}_i^t \in \mathbb{R}^{d_t}$ derived from a state-of-the-art sentence embedding model BGE (Su et al. 2023).

The primary objective is to learn a mapping function that produces expressive, low-dimensional embeddings for users, $\mathbf{e}_u \in \mathbb{R}^d$, and items, $\mathbf{e}_i \in \mathbb{R}^d$. These embeddings are then used to compute a relevance score via dot product $\hat{y}_{ui} = \mathbf{e}_u^\top \mathbf{e}_i$, to rank items for each user. Our central challenge is to ensure the quality of $\mathbf{e}_i$ for all items, especially those in the cold-start regime where interaction data is minimal. This disconnection is particularly acute for cold-start users, where abstract preferences (e.g., style or genre) must be inferred from content alone.

Adaptive Rarity Amplification

To directly combat the popularity bias inherent in recommendation datasets, which causes models to neglect rare items and underperform in representing items transitioning from *cold* to *warm*, we devise a dynamic, degree-aware weighting scheme. The goal of this component is to amplify the learning signal for less frequent items, ensuring the model pays sufficient attention to the very items that define the cold-start challenge.

First, we stratify items into *cold* and *warm* sets. The interaction degree of each item i is computed as $d_i = \sum_{u\in\mathcal{U}} R_{ui}$. An item is designated as cold-start if its degree d_i falls below a domain-specific threshold τ:

$$c_i = \mathbb{I}(d_i < \tau), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Subsequently, we formulate an item-specific weight, w_i, designed to amplify the learning signal for items that are rare but not entirely devoid of interactions. The weight is defined by an inverse logarithmic relationship with the item's degree:

$$w_i = \begin{cases} (\log_2(d_i + 2))^{-1} & \text{if } c_i = 1 \text{ and } d_i > 0, \\ 1.0 & \text{otherwise.} \end{cases} \tag{2}$$

This inverse logarithmic weighting (Schnabel et al. 2016; Joachims, Swaminathan, and Schnabel 2017) compresses degree ranges and stabilizes small values with a $+2$ offset. The resulting item-specific weights w_i amplify the learning signal for less frequent items and are systematically integrated into the main learning objective.

Sparsely-Regularized Multimodal Tokenization

To overcome the "semantic fog" of continuous feature alignment and create robust semantic representations for cold-start items, we propose to transform raw multimodal features into a structured, discrete vocabulary of semantic tokens. This is implemented via a Residual Quantized Variational Autoencoder (RQ-VAE) (Lee et al. 2022), which bridges the modality gap between continuous features and discrete collaborative IDs.

Residual Quantization. For each modality $m \in \{v, t\}$, a modality-specific encoder E_m (an MLP) projects the raw feature $\mathbf{f}_i^m$ into a latent space: $\mathbf{z}_{e,i}^m = E_m(\mathbf{f}_i^m)$. This latent vector is then quantized by a cascade of N_q quantizers. At the first stage ($k = 1$), the vector quantizer finds the closest

prototype $\mathbf{q}_i^{(1)}$ from a codebook $\mathcal{C}_m^{(1)}$. The residual, $\mathbf{r}_i^{(1)} = \mathbf{z}_{e,i}^m - \mathbf{q}_i^{(1)}$, is then passed to the second stage. This process repeats iteratively:

$$q_i^{(k)} = \arg\min_{c \in C_m^{(k)}} \|r_i^{(k-1)} - c\|_2^2, \tag{3}$$

$$r_i^{(k)} = r_i^{(k-1)} - q_i^{(k)}, \tag{4}$$

The final quantized representation, $\mathbf{z}_{q,i}^m = \sum_{k=1}^{N_q} \mathbf{q}_i^{(k)}$, is a composite vector that represents the item's multimodal characteristics as a combination of learned semantic primitives.

Sparsity-Inducing Regularization. A key challenge with learned codebooks is their susceptibility to the "semantic fog", producing entangled representations. To address this, we introduce a novel sparsity constraint on the codebook usage. This encourages the model to represent each item using a small, specialized subset of codebook vectors, making the resulting codes more explainable. We achieve this by imposing a KL-divergence penalty that drives the aggregate posterior distribution of codebook usage towards a sparse prior, specifically a Bernoulli distribution with a small mean ρ. Let $\hat{\rho}_j$ be the average activation probability of the j-th codeword across a mini-batch. The sparsity loss is:

$$\mathcal{L}_{\text{sparse}} = \sum_{j=1}^{K} \text{KL}(\rho \| \hat{\rho}_j) = \sum_{j=1}^{K} \Big(\rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \Big). \tag{5}$$

This loss term encourages an efficient semantic code by penalizing high entropy in the average codebook usage. Theoretically, this KL penalty fosters disentangled representations by serving as a proxy for minimizing mutual information between codebook activations (Higgins et al. 2017), driving the codebook usage towards a sparse prior. This process is analogous to nonlinear Independent Component Analysis in the discrete latent space, yielding compositional representations crucial for robust generalization to cold-start items.

Tokenizer Training Objective. The RQ-VAE for each modality is trained with a composite objective that includes a reconstruction term, a commitment term to stabilize codebook learning, and our novel sparsity term:

$$\mathcal{L}_{\text{RQ-VAE}}^m = \underbrace{\|\mathbf{f}_i^m - D_m(\mathbf{z}_{q,i}^m)\|_2^2}_{\text{Reconstruction}} + \beta \underbrace{\|\mathbf{z}_{e,i}^m - \text{sg}(\mathbf{z}_{q,i}^m)\|_2^2}_{\text{Commitment}} + \gamma \underbrace{\mathcal{L}_{\text{sparse}}}_{\text{Sparsity}}, \tag{6}$$

where D_m is the decoder, $\text{sg}(\cdot)$ is the stop-gradient operator, and β, γ are hyperparameters. This entire loss is weighted by w_i during global optimization, ensuring high-fidelity tokenization for cold-start items.

Hierarchical Multi-Source Encoding and Fusion

Having generated high-fidelity semantic codes from content, the subsequent critical step is to align these codes with users' collaborative preferences, thereby mitigating the OOD representation problem for cold-start items. To achieve this, we design a hierarchical graph encoding architecture to synthesize these diverse signals, built upon the efficient LightGCN (He et al. 2020).

Intra-Modal Disentangled Propagation. Before fusing signals, it is critical to preserve the semantic purity of each information source. To this end, we maintain three parallel, disentangled propagation channels on the user-item graph $\mathcal{G}_{ui}$, allowing us to learn modality-specific collaborative patterns without premature interference. The visual channel is initialized with tokenized visual embeddings $\{\mathbf{z}_{q,i}^v\}_{i \in \mathcal{I}}$ to capture aesthetic preferences. The textual channel uses $\{\mathbf{z}_{q,i}^t\}_{i \in \mathcal{I}}$ to learn from item attributes. Crucially, a separate collaborative channel, initialized with standard learnable ID embeddings, exclusively models pure collaborative signals, untainted by content. Within each channel, we apply the LightGCN propagation rule to refine embeddings over L layers:

$$\mathbf{E}^{(l+1)} = (\mathbf{D}^{-1/2} \tilde{\mathbf{A}} \mathbf{D}^{-1/2}) \mathbf{E}^{(l)}, \tag{7}$$

where $\tilde{\mathbf{A}}$ is the adjacency matrix of $\mathcal{G}_{ui}$ with self-loops and $\mathbf{D}$ is the diagonal degree matrix. Aggregating layer embeddings yields three specialized representations: $(\mathbf{u}_v, \mathbf{i}_v)$, $(\mathbf{u}_t, \mathbf{i}_t)$, and $(\mathbf{u}_{id}, \mathbf{i}_{id})$.

Cross-Source Fusion and Enhancement. To form the final representations, we fuse the specialized representations learned in the previous stage, designed to integrate content-based features with pure collaborative signals. We employ a hybrid fusion strategy:

$$\mathbf{e}_i^m = \alpha \cdot \text{CONCAT}(\mathbf{i}_v, \mathbf{i}_t) + (1-\alpha) \cdot \text{Attention}(\mathbf{i}_v, \mathbf{i}_t), \tag{8}$$

where the hyperparameter α balances static feature preservation with dynamic, context-aware re-weighting. This unified multimodal content embedding $\mathbf{e}_i^m$ is then integrated with the collaborative embedding $\mathbf{i}_{id}$ using a gated residual connection. The final user embedding $\mathbf{e}_{u,\text{final}}$ is derived similarly from the user-side representations. This process yields the final predictive embeddings, $\mathbf{e}_{u,\text{final}}$ and $\mathbf{e}_{i,\text{final}}$.

Model Optimization

The entire MoToRec framework is trained end-to-end by minimizing a composite objective function, prioritizing representation quality and high-fidelity tokenization for cold-start items. The final prediction score for a user-item pair is the dot product of their final embeddings: $\hat{y}_{ui} = (\mathbf{e}_{u,\text{final}})^\top \mathbf{e}_{i,\text{final}}$.

The primary objective is the Bayesian Personalized Ranking (BPR) loss, which optimizes for the relative ranking of items over observed interactions (Rendle et al. 2009):

$$\mathcal{L}_{\text{BPR}} = -\sum_{(u,i,j) \in \mathcal{D}} \ln \sigma(\hat{y}_{ui} - \hat{y}_{uj}), \tag{9}$$

where $\mathcal{D}$ is the set of training triplets where user u interacted with item i (positive) but not item j (negative).

To improve embedding quality, we incorporate the InfoNCE contrastive loss (Van Den Oord, Li, and Vinyals

Dataset	#Users	#Items	#Inters.	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	278,677	99.97%

Table 1: Statistics of the experimental datasets. The high sparsity highlights the challenge.

2018), which pulls together augmented positive views of the same node while pushing apart negative samples:

$$\mathcal{L}_{\text{CL}} = -\sum_{k \in \mathcal{B}} \log \frac{\exp(\text{sim}(\mathbf{e}_k^{(1)}, \mathbf{e}_k^{(2)})/\tau_{cl})}{\sum_j \exp(\text{sim}(\mathbf{e}_k^{(1)}, \mathbf{e}_j^{(2)})/\tau_{cl})}, \quad (10)$$

where $\mathcal{B}$ is the mini-batch, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

The final loss function integrates the ranking objective, contrastive loss, weighted multimodal tokenization loss, and a standard L2 regularization on all model parameters Θ:

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \lambda_{cl}\mathcal{L}_{\text{CL}} + \lambda_{rq} \sum_{m \in \{v,t\}} \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \cdot \mathcal{L}^m_{\text{RQ-VAE},i} + \lambda_{reg}\|\Theta\|_2^2, \quad (11)$$

where $\lambda_{cl}, \lambda_{rq}, \lambda_{reg}$ are hyperparameters that balance the different components. The rarity-amplification weight w_i ensures prioritized optimization for accurately tokenizing and representing cold-start items, which is central to our approach.

Experiments

Experimental Setup

Datasets. We evaluate our model on three public Amazon review datasets(McAuley et al. 2015): Baby, Sports, and Clothing. To ensure a fair comparison, we follow the same data processing and filtering settings as in prior works that use these common benchmark datasets (Zhou and Shen 2023; Zhou et al. 2023b). We utilize their rich user-item interaction, visual, and textual features. Detailed statistics are presented in Table 1.

Baselines. We compare MoToRec against a comprehensive suite of baseline models, which can be categorized into two groups:

- **Traditional**: MF-BPR (Rendle et al. 2009), LightGCN (He et al. 2020), SimGCL (You et al. 2020), LayerGCN (Zhou et al. 2023a).
- **Multimodal**: VBPR (He and McAuley 2016), MMGCN (Wei et al. 2019), DualGNN (Wang et al. 2023a), SLMRec (Tao et al. 2023), LATTICE (Zhang et al. 2021), FREEDOM (Zhou and Shen 2023), BM3 (Zhou et al. 2023b), LGMRec (Guo et al. 2024), LPIC (Liu et al. 2025b).

Evaluation Protocol. We use an 8:1:1 train/validation/test split for all interactions. The test set is then divided into overall and cold-start groups. Following common practice (Schein et al. 2002; Zhou and Shen 2023), the cold-start group contains test items with fewer than 10 interactions in the training set. For performance evaluation, we adopt two widely-used ranking metrics: Recall@N (R@N) and NDCG@N (N@N) (He et al. 2015). We report results for $N \in \{10, 20\}$, averaged over all test users.

Implementation Details. All models are implemented in the MMRec framework (Zhou 2023) with 64-dimensional embeddings and the Adam optimizer. For MoToRec, we conduct a grid search over key hyperparameters. For the sparse RQ-VAE module, we tune the number of quantizers $N_q \in \{4, 6, 8\}$ and codebook size $K \in \{256, 512, 1024\}$. The learning rate is searched in $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$. We optimize regularization weights: sparsity coefficient $\gamma \in \{0.01, 0.05, 0.1, 0.2\}$, RQ-VAE loss weight $\lambda_{rq} \in \{0.1, 0.5, 1.0, 2.0\}$, and contrastive loss weight $\lambda_{cl} \in \{0.01, 0.05, 0.1\}$. The number of GCN layers is fixed at $L = 2$, and the cold-start threshold is set to $\tau = 10$, a standard value for defining sparse items (Schein et al. 2002). Early stopping is applied with a patience of 20 epochs, monitoring R@20 on the validation set.

Performance Comparison

The main performance comparison is summarized in Table 2, with key findings as follows: (1) MoToRec's superiority over all baselines. MoToRec consistently outperforms all baselines, with gains up to 88% over ID-only models (MF-BPR, LightGCN) and 11.57% over state-of-the-art multimodal methods (LGMRec, LPIC). This highlights the value of multimodal content and, more critically, the superiority of our discrete representation. (2) Effectiveness in mitigating the cold-start problem. MoToRec's advantage is most pronounced in the critical item cold-start scenario, as shown in Figure 3. It achieves a remarkable uplift of up to 12.58% in N@20 on items with the fewest interactions. This provides strong evidence that our discrete tokenization enables superior generalization by representing novel items as a composition of known concepts. (3) Impact of sparse-regularized tokenization. The performance gap over strong baselines like LGMRec suggests that transforming features into discrete codes via our sparse-regularized RQ-VAE is key. This approach mitigates the "alignment haze" inherent in continuous fusion, creating more robust modality-aware representations. The importance of our content modeling is further corroborated by our ablation study in Table 3.

Ablation Study

To validate the contribution of each key component, we conduct a comprehensive ablation study with results detailed in Table 3. The results clearly show that the full MoToRec model substantially outperforms all variants, confirming our design choices contribute synergistically. The most severe performance degradation stems from removing the entire RQ-VAE (*w/o RQ-VAE*), providing strong evidence for our core thesis that discrete semantic tokenization is superior

Datasets	Baby				Sports				Clothing			
Model	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MF-BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0187	0.0279	0.0103	0.0126
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0340	0.0526	0.0188	0.0236
SimGCL	0.0513	0.0804	0.0273	0.0350	0.0601	0.0919	0.0327	0.0414	0.0356	0.0549	0.0195	0.0244
LayerGCN	0.0529	0.0820	0.0281	0.0355	0.0594	0.0916	0.0323	0.0406	0.0371	0.0566	0.0200	0.0247
Improv.	**34.03%**	**31.34%**	**33.81%**	**33.24%**	**30.45%**	**26.55%**	**32.42%**	**27.78%**	**85.44%**	**79.15%**	**88.00%**	**84.62%**
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
MMGCN	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0218	0.0345	0.0110	0.0142
DualGNN	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0454	0.0683	0.0241	0.0299
SLMRec	0.0529	0.0775	0.0290	0.0353	0.0663	0.0990	0.0365	0.0450	0.0452	0.0675	0.0247	0.0303
LATTICE	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
FREEDOM	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0628	0.0941	0.0341	0.0420
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
LGMRec	0.0639	0.0989	0.0337	0.0430	0.0719	0.1068	0.0387	0.0477	0.0555	0.0828	0.0302	0.0371
LPIC	0.0634	0.0977	0.0337	0.0422	0.0737	0.1113	0.0398	0.0485	0.0627	0.0928	0.0338	0.0405
MoToRec	**0.0709**	**0.1077**	**0.0376**	**0.0473**	**0.0784**	**0.1163**	**0.0433**	**0.0529**	**0.0688**	**0.1014**	**0.0376**	**0.0456**
Improv.	**10.95%**	**8.57%**	**11.57%**	**10.00%**	**6.38%**	**4.49%**	**8.79%**	**9.07%**	**9.55%**	**7.76%**	**10.26%**	**8.57%**

Table 2: Overall performance comparison on three datasets. Best and second-best (best baseline) results are in **bold** and underlined, respectively. The "Improv." row shows the improvement of MoToRec over the best-performing baseline in each group. The t-tests validate the significance of performance improvements with p-value ≤ 0.05.

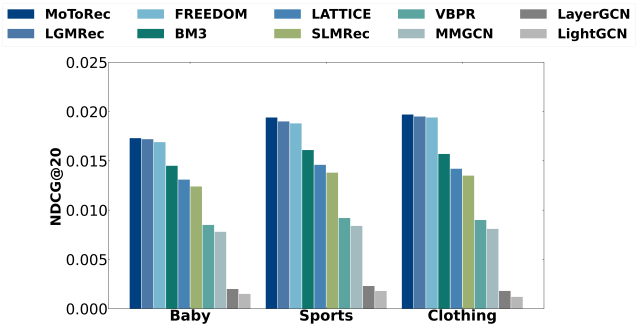

Figure 3: Performance comparison on the cold-start item set.

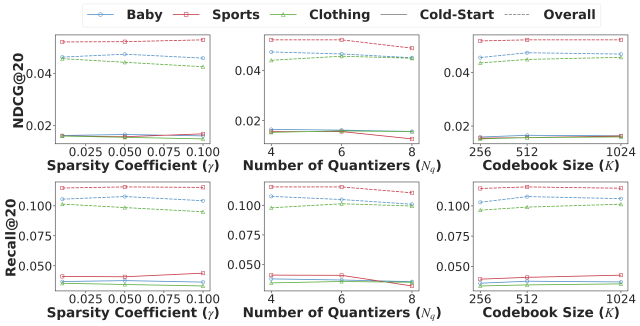

Figure 4: Individual hyperparameter sensitivity analysis for N@20 and R@20 across all datasets for key parameters.

to continuous feature mapping, especially in cold-start scenarios. The components designed to combat data sparsity, Adaptive Rarity Amplification (*w/o ARA*) and *Sparsity* regularization, are also proven critical, as their removal significantly harms cold-start performance by failing to amplify rare signals and learn disentangled tokens. Finally, the consistent decline after removing the Contrastive Loss (*w/o CL*), Hybrid Fusion (*w/o HF*), and Homogeneous Graph Enhancement (*w/o HGE*) verifies their respective importance in creating a robust embedding space, synthesizing multimodal signals, and capturing higher-order graph structures.

Hyperparameter Study

We analyze hyperparameter sensitivity in Figures 4 and 5. Results indicate that optimal configurations hinge on dataset characteristics: while sparse Baby dataset favors moderate sparsity ($\gamma = 0.05$) and compact codebooks ($K = 512$) for denoising, the visually rich Clothing dataset demands lower sparsity ($\gamma = 0.01$) and larger capacity ($K = 1024$) to preserve fine-grained semantics. Furthermore, pairwise interactions on Sports reveal that unlike the robust overall performance, cold-start results are highly sensitive to sparsity deviations, underscoring the necessity of precise tokenizer calibration for new items.

Qualitative Analysis

Embedding Space Visualization. To intuitively understand the quality of the learned representations, we visualize the item embedding space by sampling 500 items from the Sports dataset and projecting their embeddings into two dimensions using t-SNE (Maaten and Hinton 2008), as shown in Figure 6. It confirms that the full MoToRec model (c) learns a significantly more organized semantic manifold compared to variants w/o RQ-VAE (a) or sparsity (b). Crucially, cold-start items (red) are no longer isolated outliers but are seamlessly integrated within this structure, proving our model's ability to position novel items near their semantic neighbors.

Case Study. We verify that our model's discrete codes learn human-interpretable concepts. On the Clothing dataset, individual codes consistently capture disentangled

Datasets	Baby				Sports				Clothing			
Scenario	Overall		Cold-start		Overall		Cold-start		Overall		Cold-start	
Metrics	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20
MoToRec	**0.0473**	**0.1077**	**0.0147**	**0.0347**	**0.0529**	**0.1163**	**0.0183**	**0.0452**	**0.0456**	**0.1014**	**0.0170**	**0.0420**
w/o RQ-VAE	0.0398	0.0915	0.0092	0.0229	0.0422	0.0930	0.0115	0.0304	0.0362	0.0816	0.0097	0.0281
w/o ARA	0.0437	0.0977	0.0111	0.0281	0.0466	0.1027	0.0139	0.0367	0.0397	0.0894	0.0118	0.0342
w/o Sparsity	0.0430	0.0972	0.0109	0.0277	0.0455	0.1003	0.0137	0.0362	0.0389	0.0876	0.0116	0.0336
w/o CL	0.0455	0.1026	0.0118	0.0281	0.0515	0.1146	0.0153	0.0391	0.0438	0.0972	0.0129	0.0374
w/o HF	0.0449	0.1028	0.0120	0.0303	0.0468	0.1042	0.0150	0.0396	0.0401	0.0899	0.0127	0.0368
w/o HGE	0.0438	0.1002	0.0117	0.0295	0.0489	0.1106	0.0146	0.0384	0.0407	0.0917	0.0124	0.0357

Table 3: Ablation study of MoToRec on both overall and cold-start performance.

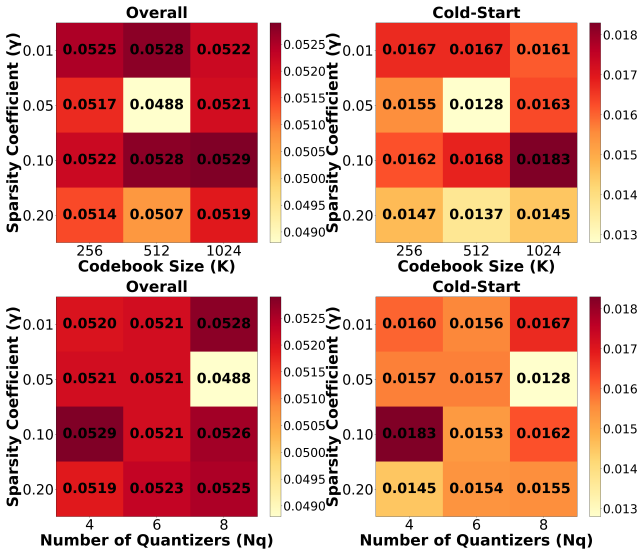

Figure 5: Pairwise hyperparameter study on the Sports dataset (N@20).

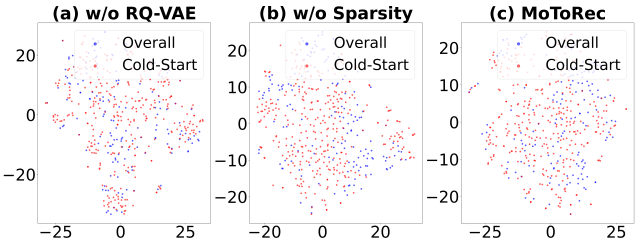

Figure 6: t-SNE visualization of item embeddings.

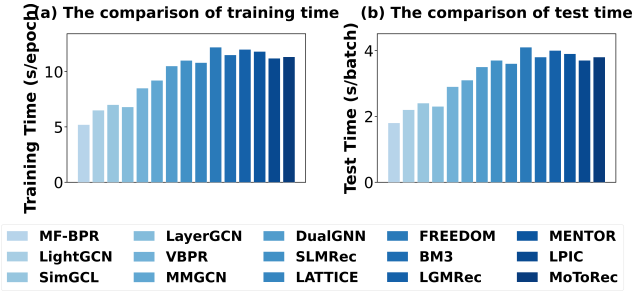

Figure 7: Runtime comparison between MoToRec and baselines. (a) Training time per epoch. (b) Test time per batch.

attributes; for example, code <c_121> reliably activates for the color 'red', while <a_34> corresponds to the category 'T-shirt'. This demonstrates our tokenizer learns a true semantic vocabulary. Critically, these codes are compositional: a cold-start item such as a 'red minimalist T-shirt' activates a logical combination of these codes: [Color: Red] (<c_121>), [Style: Minimalist] (<s_5>), and [Category: T-shirt] (<a_34>). This case study validates our model's ability to transform a single vector into a clear, compositional, and interpretable representation.

Efficiency Study

To assess the practical deployability of MoToRec, we evaluate its computational efficiency. Figure 7 illustrates a detailed comparison of (a) training time per epoch and (b) test time per batch against a wide range of baseline models on the Sports dataset. To validate MoToRec's practical viability, we benchmark its computational efficiency, with results presented in Figure 7. MoToRec demonstrates a highly competitive training time of 11.33s per epoch, outperforming strong multimodal baselines such as FREEDOM at 12.2s and LGMRec at 12.0s. While this represents a modest $\sim$74% overhead compared to the simpler ID-only LightGCN at 6.5s, this cost is solely attributed to our expressive tokenizer module; the graph propagation stage preserves the original efficiency of the LightGCN architecture. Critically, for inference, MoToRec remains highly practical with a test time of 3.8s per batch, on par with other high-performance models. This analysis confirms that MoToRec's architectural advancements do not impose a prohibitive computational cost, representing a favorable trade-off for its substantial accuracy gains, particularly on the item cold-start problem.

Conclusion

We propose MoToRec, which reframes recommendation as discrete semantic tokenization to learn robust multimodal representations from noisy raw data. Extensive experiments on three datasets demonstrate that MoToRec achieves state-of-the-art performance and substantial gains in cold-start scenarios. This validates discrete tokenization as a pivotal direction for future multimodal recommendations.

References

Bao, K.; Zhang, J.; Zhang, Y.; Wang, W.; Feng, F.; and He, X. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 1040–1045.

Bobadilla, J.; Ortega, F.; Hernando, A.; and Gutiérrez, A. 2013. Recommender systems survey. *Knowledge-Based Systems*, 109–132.

Cui, X.; Lu, W.; Tong, Y.; Li, Y.; and Zhao, Z. 2025a. Diffusion-based multi-modal synergy interest network for click-through rate prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 581–591.

Cui, X.; Lu, W.; Tong, Y.; Li, Y.; and Zhao, Z. 2025b. 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.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135.

Geng, S.; Liu, S.; Fu, Z.; Ge, Y.; and Zhang, Y. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt Predict Paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, 299–315.

Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; and Ruan, B. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of The Web Conference 2024*, 1709–1719.

He, R.; and McAuley, J. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 144–150.

He, X.; Chen, T.; Kan, M.-Y.; and Chen, X. 2015. TriRank: Review-aware Explainable Recommendation by Modeling Aspects. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM '15)*, 1661–1670.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 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.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*.

Hou, Y.; He, Z.; McAuley, J.; and Zhao, W. X. 2023. Learning Vector-Quantized Item Representation for Transferable Sequential Recommenders. In *Proceedings of the ACM Web Conference 2023*, 1162–1171.

Joachims, T.; Swaminathan, A.; and Schnabel, T. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, 781–789.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive Image Generation using Residual Quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11513–11522.

Lee, H.; Im, J.; Jang, S.-W.; Cho, H.; and Chung, S. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD '19)*, 1073–1082.

Li, Y.; and Lu, W. 2024. MHHCR: Multi-behavior Heterogeneous Hypergraph Contrastive Recommendation. In *International Conference on Web Information Systems Engineering*, 91–102.

Lian, D.; Wang, H.; Liu, Z.; Lian, J.; Chen, E.; and Xie, X. 2020. LightRec: A Memory and Search-Efficient Recommender System. In *Proceedings of The Web Conference 2020 (WWW '20)*, 695–705.

Liu, S.; Ding, R.; Lu, W.; Wang, J.; Yu, M.; Shi, X.; and Zhang, W. 2025a. Coherency Improved Explainable Recommendation via Large Language Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12201–12209.

Liu, X.; Song, Q.; Xiao, L.; Wang, C.; and Gao, X. 2025b. LPIC: Learnable Prompts and ID-guided Contrastive Learning for Multimodal Recommendation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(9): 1–16.

Lu, W.; and Yin, L. 2025. DMMD4SR: Diffusion Model-based Multi-level Multimodal Denoising for Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6363–6372.

Lu, Y.; Fang, Y.; and Shi, C. 2020. Meta-learning on Heterogeneous Information Networks for Cold-start Recommendation. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)*, 1563–1573.

Maaten, L. V. D.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

McAuley, J.; Targett, C.; Shi, Q.; and van den Hengel, A. 2015. Image-based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*, 43–52.

Mo, M.; Lu, W.; Xie, Q.; Lv, X.; Xiao, Z.; Yang, H.; and Zhang, Y. 2024. MIN: Multi-stage Interactive Network for Multimodal Recommendation. In *International Conference on Web Information Systems Engineering*, 191–205.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 452–461.

Schein, A. I.; Popescul, A.; Ungar, L. H.; and Pennock, D. M. 2002. Methods and Metrics for Cold-Start Recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 253–260.

Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; and Joachims, T. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML '16)*, volume 48, 1670–1679.

Su, H.; Shi, W.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.-t.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2023. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1102–1121.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T.-S. 2023. *IEEE Transactions on Multimedia*, 25: 5107–5116.

Van Den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*.

Van Den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30, 6306–6315.

Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2023a. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Transactions on Multimedia*, 25: 1074–1084.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; and Wei, F. 2023b. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19175–19186.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1437–1445.

Wu, J.; Wang, X.; Feng, F.; He, X.; Chen, L.; Lian, J.; and Xie, X. 2021. Self-supervised Graph Learning for Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 726–735.

Ye, Y.; Zheng, Z.; Shen, Y.; Wang, T.; Zhang, H.; Zhu, P.; Yu, R.; Zhang, K.; and Xiong, H. 2025. Harnessing Multimodal Large Language Models for Multimodal Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13069–13077.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph Contrastive Learning with Augmentations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 5812–5823.

Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2022. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining Latent Structures for Multimedia Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3872–3880.

Zhou, X. 2023. MMRec: Simplifying Multimodal Recommendation. In *Proceedings of the ACM Multimedia Asia Workshop*, 1–2.

Zhou, X.; Lin, D.; Liu, Y.; and Miao, C. 2023a. Layer-refined Graph Convolutional Networks for Recommendation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 1247–1259.

Zhou, X.; and Shen, Z. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, 935–943.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023b. Bootstrap Latent Representations for Multi-modal Recommendation. In *Proceedings of the ACM Web Conference 2023*, 845–854.