MusicRec: Multi-modal Semantic-Enhanced Identifier with Collaborative Signals for Generative Recommendation

Yuqiu Zhao[1], Lei Shi[1]*, Yan Zhong[2], Feifei Kou[3], Pengfei Zhang[4], Jiwei Zhang[3]
Mingying Xu[5], Yanchao Liu[1]

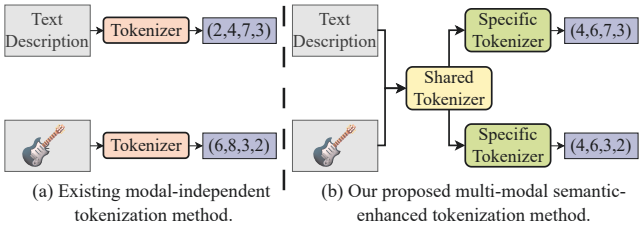
[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, China
[2]School of Mathematical Sciences, Peking University, China
[3]School of Computer Science (National Pilot School of Software Engineering), BUPT, China
[4] School of Computer Science and Engineering, Anhui University of Science and Technology, China
[5] School of Artificial Intelligence and Computer Science, North China University of Technology, China
yuqiuzhao@mails.cuc.edu.cn; leiky_shi@cuc.edu.cn; zhongyan@stu.pku.edu.cn; koufeifei000@bupt.edu.cn;
zpf.bupt@bupt.cn; jwzhang666@bupt.edu.cn; xumingying@ncut.edu.cn; yanchaoliu@cuc.edu.cn

Abstract

Generative recommendation as a new paradigm is influencing the current development of recommender systems. It aims to assign identifiers that capture richer semantic and collaborative information to items, and subsequently predict item identifiers via autoregressive generation using Large Language Models (LLMs). Existing approaches primarily tokenize item text into codebooks with preserved semantic IDs through RQ-VAE, or separately tokenize different modality features of items. However, existing tokenization methods face two major challenges: **(1)** Learning decoupled multi-modal features limits the quality of the semantic representation. **(2)** Ignoring collaborative signals from interaction history limits the comprehensiveness of identifiers. To address these limitations, we propose a **mu**lti-modal **s**emantic-enhanced **i**dentifier with **c**ollaborative signals for generative **rec**ommendation, named MusicRec. In MusicRec, we propose a tokenization approach based on shared-specific modal fusion, enabling the generated identifiers to preserve semantic information more comprehensively from all modalities. In addition, we incorporate collaborative signals from user interactions to guide identifier generation, preserving collaborative patterns in the semantic representation space. Extensive experiments on three public datasets demonstrate that MusicRec achieves state-of-the-art performance compared to existing baseline methods.

Introduction

Recommender systems play a crucial role in exploring personalized content across different scenarios, such as video platforms (Dong et al. 2024), products (Shi et al. 2024), and movies (Chee et al. 2024). Recently, generative recommendation has emerged as a promising paradigm in recommender systems, exceeding the representational limitations of traditional discriminative approaches by encoding items into token sequences and autoregressively generating the next interacted item (Rajput et al. 2023; Tan et al. 2024).



(a) Existing modal-independent tokenization method.

(b) Our proposed multi-modal semantic-enhanced tokenization method.

Figure 1: Existing multi-modal item tokenization methods vs. Ours.

The representational capacity of item identifiers directly determines the upper bound of the recommendation generation performance. Effective identifiers must retain sufficient semantic information while incorporating collaborative signals from user interactions to capture latent preferences accurately. Recent approaches (Rajput et al. 2023; Wang et al. 2024a) have employed cascading semantic codes from RQ-VAE (Zeghidour et al. 2021) for better compression and semantic extraction. However, these methods mainly rely on single-modal textual information, while emerging evidence demonstrates the superior effectiveness of multi-modal representations for generalizability (Zhao et al. 2025). To overcome this limitation, several studies have integrated multi-modal information during tokenization, generating identifiers through independent or joint processing of different modal features (Liu et al. 2024; Zhai et al. 2025).

Recent advances in generative recommendation have increasingly integrated semantic content with collaborative signals to enhance personalization (Zhang et al. 2025). Existing approaches can be categorized into several strategies: enriching semantic representations with auxiliary collaborative signals to bridge high-level semantics and behavioral patterns (Wang et al. 2024a,b; Xiao et al. 2025), developing separate models that combine semantic and collaborative signals as independent token sequences (Wang et al. 2024c; Hong et al. 2025), and unifying semantic search information with collaborative recommendation signals (Shi et al.

*Corresponding author

2025). Despite these efforts, significant challenges remain when applying multi-modal semantic and collaborative signals to generative recommendation.

Firstly, current semantic identifier generation relies on single-modal information and employs decoupled quantizers (Figure 1(a)), which inadequately capture the rich cross-modal semantic relationships essential for comprehensive item representation. The isolation of different modalities prevents the model from learning unified semantic representations that leverage complementary information. **Secondly**, the identifier exclusively focuses on semantic attributes while overlooking collaborative signals derived from user interaction history. This separation fails to preserve co-occurring patterns between semantic understanding and behaviors. Semantic preferences inherently drive user interactions, while collaborative signals can reciprocally enhance semantic comprehension, yet existing methods independently model these two crucial information sources.

To this end, we propose **MusicRec**, a novel approach that enables **Mu**lti-modal **s**emantic-enhanced **i**dentifier learning with **c**ollaborative signal supervision for generative **Rec**ommendation. To better integrate multi-modal item semantics, MusicRec first extracts the shared quantization code between different modalities, then learns modality-specific quantization code for each modality based on shared representations (Figure 1(b)), enabling the generated identifiers to capture richer cross-modal semantic representations. To effectively leverage collaborative information in identifier generation, we introduce a collaborative-semantic alignment mechanism during identifier learning. This approach ensures that semantic quantized representations maintain distributional alignment with collaborative embeddings, enabling more effective discovery of user interaction preferences while preserving proximity between items with similar interaction patterns in the quantized space.

To sum up, the contributions of this work are as follows:

- We propose a share-specific modal fusion tokenization approach to better capture cross-modal semantics, thus improving the quality of item semantic ID representation under multi-modal generative recommendation.
- We introduce a collaborative-semantic alignment mechanism to supervise the semantic identifier learning process and preserve collaborative relationships between items, enhancing the model's ability to capture user preferences and item associations.
- We conduct extensive experiments on three datasets to demonstrate the superior performance of MusicRec over existing traditional and generative recommendation methods, performing in-depth analyses under various settings to validate its effectiveness and robustness.

Related Works

- **Generative Recommendation.** As a new paradigm distinct from discriminative recommendation, generative recommendation typically involves two core stages: item tokenization and recommendation generation. In the first phase, items are converted into semantic IDs (SIDs) that serve as model-processable identifiers through various encoding strategies, such as using semantic content (Hua et al. 2023; Lin et al. 2024), product quantization (Petrov and Macdonald 2024), and residual quantization (Rajput et al. 2023; Wang et al. 2024a; Lin et al. 2025). For example, LETTER (Wang et al. 2024a) introduces a learnable tokenization that incorporates hierarchical semantics and collaborative information using RQ-VAE. For the second phase, several works utilize LLMs to produce recommendations by autoregressively predicting the next item token. For example, EAGER (Wang et al. 2024c) employs parallel processing of semantic and collaborative streams, while LC-Rec (Zheng et al. 2024) integrates LLMs for enhanced semantic understanding. However, existing works rely on limited semantic information and may not fully exploit collaborative signals during tokenization, leading to suboptimal performance.
- **Multi-modal Recommendation.** Traditional sequential recommendation models (Hidasi et al. 2015; Kang and McAuley 2018) only focus on item ID and category. Recent works attempt to incorporate multi-modal features by diffusion-based models (Ma et al. 2024; Jiang et al. 2024) and GNN-based models (Wang et al. 2022; Yi and Chen 2021; Wei et al. 2019). However, these methods often struggle with capturing fine-grained semantic relationships across modalities due to fixed feature representations. With the advancement of pre-trained models (Radford et al. 2021; Li et al. 2023), recent approaches (Geng et al. 2023; Liu et al. 2024; Zhai et al. 2025) have sought to integrate powerful representation learning to better capture multi-modal semantics in many scenarios. More recently, a few studies (Ye et al. 2025; Zhou et al. 2025; Zhang et al. 2024) have tried to use multi-modal large language models (MLLMs) to enhance the generalization of recommendation models, while high computational demands hinder widespread deployment.

Methodology

In this section, we present a comprehensive introduction to our proposed multi-model semantic-enhanced identifier with collaborative signals generative recommendation (MusicRec) approach.

Preliminaries and Overview

Problem Definition We formulate the sequential recommendation task as follows. Let $\mathcal{U} = u_1, u_2, ..., u_{|\mathcal{U}|}$ be the set of users and $\mathcal{I} = i_1, i_2, ..., i_{|\mathcal{I}|}$ be the set of items. For each user $u \in \mathcal{U}$, we have a chronologically ordered interaction sequence $S_u = i_1^u, i_2^u, ..., i_{|S_u|}^u$, where $i_t^u \in \mathcal{I}$ represents the item that user u interacted with at the time step t. Given the historical interaction sequence $S_u^{<t} = i_1^u, i_2^u, ..., i_{t-1}^u$, the goal is to predict the next item i_t^u that user u is likely to interact with, which is formulated as:

$$i_t^u = \arg\max_{i \in \mathcal{I}} P(i|S_u^{<t}). \quad (1)$$

Overview As shown in Figure 2, our proposed MusicRec approach consists of two key modules in the tokenization stage: the multi-modal semantic-enhanced tokenization, which incorporates different modality semantic information

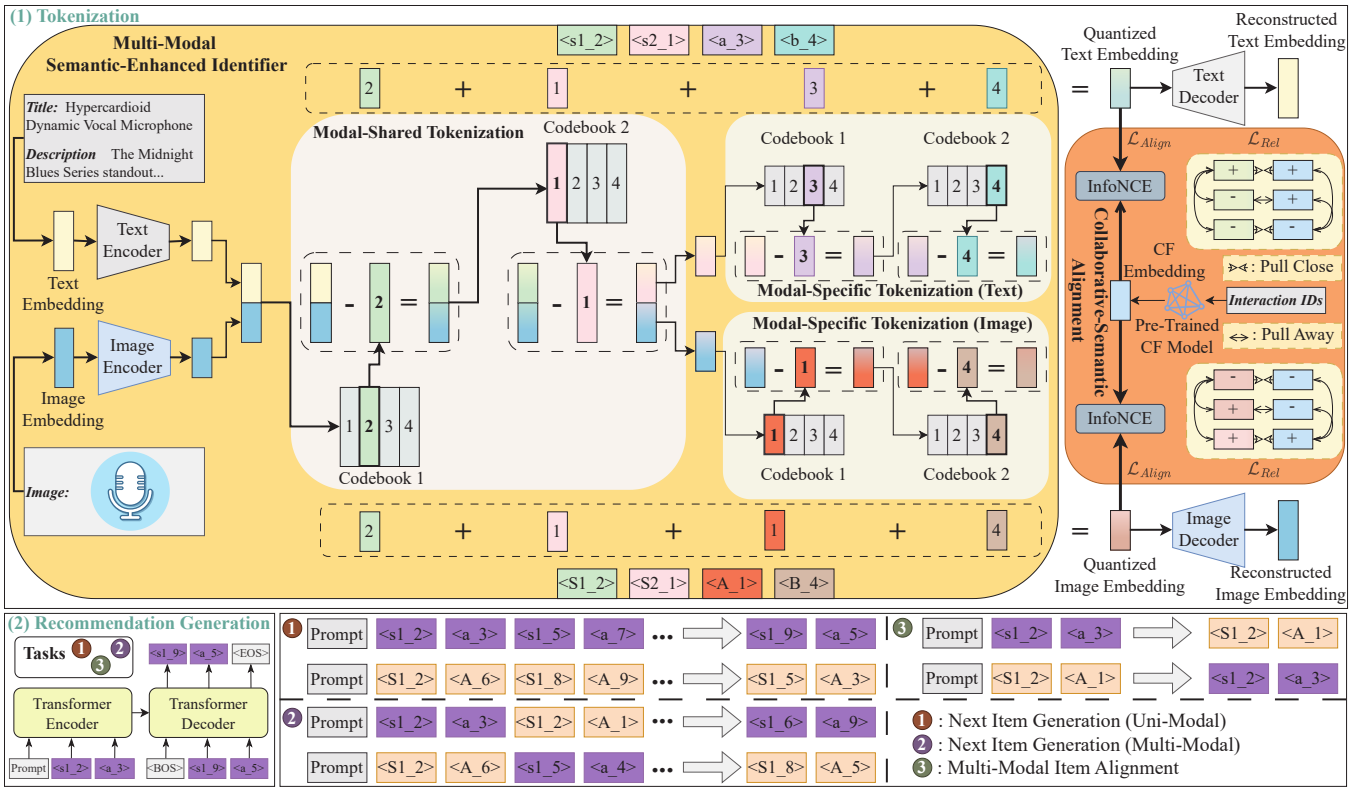


Figure 2: The overview of our proposed MusicRec. MusicRec consists of two stages: (1) Tokenization, which incorporates a multi-modal semantic-enhanced identifier component to integrate multi-modal item semantics and employs a collaborative-semantic alignment mechanism to preserve collaborative relationships; and (2) Recommendation Generation, which introduces three specialized tasks to effectively leverage the generated identifiers during recommendation training.

into the generation of identifiers using RQ-VAE, and the collaborative-semantic alignment module, which supervises semantic identifiers learning with collaborative signals. During the recommendation generation stage, there are three core training tasks to enhance the performance using our proposed identifiers, including next item generation with different modalities and multi-modal item alignment. The following sections provide a detailed explanation of each component. Details of each part are introduced below.

Multi-Modal Semantic-Enhanced Identifier

To fully utilize rich multi-modal semantic information, we introduce a shared-specific semantic tokenizer to obtain item identifiers. It quantizes the semantic information from text and image modalities into a structured set of identifiers that capture common and unique semantic features.

Semantic Embedding Extraction Given an item with textual information (e.g., titles and descriptions) and visual information (e.g., images), we first extract textual semantic embeddings $\boldsymbol{x}_t \in \mathbb{R}^{d_t}$ and visual semantic embeddings $\boldsymbol{x}_i \in \mathbb{R}^{d_i}$ by leveraging pre-trained models (e.g., LLaMA and ViT). These embeddings are then compressed into the same-dimensional latent semantic embedding $\boldsymbol{z}_t \in \mathbb{R}^d$ and $\boldsymbol{z}_i \in \mathbb{R}^d$ through Encoder_t and Encoder_i:

$$\boldsymbol{z}_t = \text{Encoder}_t(\boldsymbol{x}_t), \boldsymbol{z}_i = \text{Encoder}_i(\boldsymbol{x}_i), \tag{2}$$

where d denotes the dimension of the latent embeddings, $\text{Encoder}_t(\cdot)$ and $\text{Encoder}_i(\cdot)$ are implemented as Multi-Layer Perceptrons (MLPs).

To effectively capture shared and modality-specific representations for text and image data, we propose a hierarchical quantization approach built upon the foundation of RQ-VAE with two stages: Modal-Shared Tokenization and Modal-Specific Tokenization.

Modal-Shared Tokenization To combine two modalities, the hidden representations $\boldsymbol{z}_t$ and $\boldsymbol{z}_i$ are concatenated and passed through a linear projection layer to create a joint latent representation $\boldsymbol{z} = [\boldsymbol{z}_t; \boldsymbol{z}_i]$, where $\boldsymbol{z} \in \mathbb{R}^{2d}$. The shared representation z is then quantized into the code sequence through L_e-level codebooks, with L_e corresponding to the length of the shared identifier. For each level $l \in \{1, ...L_e\}$, we have a codebook $C_l^e = \{\boldsymbol{e}_i\}_{i=1}^N$, where $\boldsymbol{e}_i \in \mathbb{R}^{2d}$ is a learnable embedding representing a codeword, and N denotes the codebook size. The residual quantization process can be formulated as:

$$\begin{cases} c_l^e = \arg\min_i ||\boldsymbol{r}_{l-1}^e - \boldsymbol{e}_i||_2^2, \quad \boldsymbol{e}_i \in C_l^e, \\ \boldsymbol{r}_l^e = \boldsymbol{r}_{l-1}^e - \boldsymbol{e}_{c_l^e}, \end{cases} \tag{3}$$

where c_l^e denotes the codeword selected at the l-th level from the shared codebook, $\boldsymbol{r}_{l-1}^e$ is the shared semantic residual of the l-1-th level, and we set $\boldsymbol{r}_0^e = \boldsymbol{z}$.

Modal-Specific Tokenization After shared tokenization, we extract modal-specific semantic information using separate residual quantizers. The residual embeddings from the last shared level $\boldsymbol{r}^e_{L_e} \in \mathbb{R}^{2d}$ serve as input to modality-specific quantizers, which is linearly split into two equal-dimensional parts: $\boldsymbol{r}^T_0 \in \mathbb{R}^d$ for text semantics and $\boldsymbol{r}^I_0 \in \mathbb{R}^d$ for image semantics, where d represents the embedding dimension for each modality. The specific embeddings for text and image are then quantized into the code sequence through L_c-level codebooks, and L_c is the length of the specific identifier. For each level $l \in \{1, ...L_c\}$, we have text and image codebooks C^T_l and C^I_l. The text-specific residual quantization process can be formulated as:

$$\begin{cases} c^T_l = \arg\min_i ||\boldsymbol{r}^T_{l-1} - \boldsymbol{e}_i||^2_2, \quad \boldsymbol{e}_i \in C^T_l, \\ \boldsymbol{r}^T_l = \boldsymbol{r}^T_{l-1} - \boldsymbol{e}_{c^T_l}. \end{cases} \tag{4}$$

Similarly, the image-specific residual quantization process can be formulated as:

$$\begin{cases} c^I_l = \arg\min_i ||\boldsymbol{r}^I_{l-1} - \boldsymbol{e}_i||^2_2, \quad \boldsymbol{e}_i \in C^I_l, \\ \boldsymbol{r}^I_l = \boldsymbol{r}^I_{l-1} - \boldsymbol{e}_{c^I_l}, \end{cases} \tag{5}$$

where c^T_l and c^I_l denote the codeword selected at the l-th level from the specific codebooks for text and image modalities, respectively.

Multi-Modal Identifier Construction After the residual quantization process, we use the selected codewords to reconstruct the quantized embeddings for each modality. Specifically, we create the complete identifiers for each item modality by concatenating the shared and specific indices. The quantized embedding from the shared stage $\hat{z}_e = \sum_{l=1}^{L_e} e_{c^e_l}$ is split into modality-specific components $\hat{z}_e = [\hat{z}_{e,t}; \hat{z}_{e,i}]$. The final quantized text and image embeddings $\hat{\boldsymbol{z}}_t$ and $\hat{\boldsymbol{z}}_i$ are constructed as the sum of their corresponding shared and specific quantized components:

$$\hat{\boldsymbol{z}}_t = \hat{z}_{e,t} + \sum_{l=1}^{L_c} e_{c^T_l}, \hat{\boldsymbol{z}}_i = \hat{z}_{e,i} + \sum_{l=1}^{L_c} e_{c^I_l}, \tag{6}$$

where $e_{c^T_l}$ and $e_{c^I_l}$ denote specific code embeddings for text and image modalities, respectively. Finally, we utilize two separate MLP decoders to reconstruct the original input features from $\hat{\boldsymbol{z}}_t$ and $\hat{\boldsymbol{z}}_i$:

$$\hat{\boldsymbol{x}}_t = \text{Decoder}_t(\hat{\boldsymbol{z}}_t), \hat{\boldsymbol{x}}_i = \text{Decoder}_i(\hat{\boldsymbol{z}}_i). \tag{7}$$

The loss for residual quantization is defined as:

$$\begin{aligned} \mathcal{L}^e_{\text{RQ}} &= \sum_{i=1}^{Le} ||\text{sg}[\mathbf{r}^e_{l-1}] - \mathbf{e}_{c^e_l}||^2_2 + \mu||\mathbf{r}^e_{l-1} - \text{sg}[\mathbf{e}_{c^e_l}]||^2_2, \\ \mathcal{L}^T_{\text{RQ}} &= \sum_{i=1}^{Lc} ||\text{sg}[\mathbf{r}^T_{l-1}] - \mathbf{e}_{c^T_l}||^2_2 + \mu||\mathbf{r}^T_{l-1} - \text{sg}[\mathbf{e}_{c^T_l}]||^2_2, \\ \mathcal{L}^I_{\text{RQ}} &= \sum_{i=1}^{Lc} ||\text{sg}[\mathbf{r}^I_{l-1}] - \mathbf{e}_{c^I_l}||^2_2 + \mu||\mathbf{r}^I_{l-1} - \text{sg}[\mathbf{e}_{c^I_l}]||^2_2, \end{aligned} \tag{8}$$

where $\text{sg}[\cdot]$ is the stop-gradient operation (van den Oord, Vinyals, and Kavukcuoglu 2017), μ is the coefficient to balance the strength between the optimization of code embeddings and the encoder, and the total loss function of the semantic identifier $\mathcal{L}_{\text{Sem}}$ training can be formalized as:

$$\begin{cases} \mathcal{L}_{\text{Recon}} = ||\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t||^2_2 + ||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i||^2_2, \\ \mathcal{L}_{\text{RQ}} = \mathcal{L}^e_{\text{RQ}} + \mathcal{L}^T_{\text{RQ}} + \mathcal{L}^I_{\text{RQ}}, \\ \mathcal{L}_{\text{Sem}} = \mathcal{L}_{\text{Recon}} + \mathcal{L}_{\text{RQ}}, \end{cases} \tag{9}$$

where $\mathcal{L}_{\text{Recon}}$ is the reconstruction loss, and $\mathcal{L}_{\text{RQ}}$ is used to minimize the discrepancy between the residual vectors and their corresponding codebook embeddings.

To construct a structured semantic identifier for each item, we introduce a token prefix to distinguish items based on trained codebooks. For shared codewords, we use the prefix "s" with numeric suffixes indicating the level. For specific codewords, we use distinct alphabetic prefixes for modalities. Text-specific levels are denoted by {a, b, ...} and image-specific levels by {A, B, ...}. The full semantic-enhanced identifier is formed by concatenating shared and specific codewords as new tokens. For example, a text identifier may be <s1_1><s2_2><a_3><b_4>, while the corresponding image identifier is <S1_1><S2_2><A_5><B_6>.

Collaborative-Semantic Alignment

To bridge the gap between semantic features and collaborative signals, we utilize a pre-trained sequential recommender model (e.g., SASRec (Kang and McAuley 2018)) to obtain item Collaborative Filtering (CF) embedding $z_{cf} \in \mathbb{R}^{d_{cf}}$, where d_{cf} is the hidden dimension of the CF embeddings.

Collaborative-aware Alignment Loss We use InfoNCE (van den Oord, Li, and Vinyals 2018) loss to pull the quantized representation $\hat{\boldsymbol{z}}$ of an item closer to $\boldsymbol{z}_{cf}$, while pushing it away from the embeddings of other irrelevant items in the batch, and both representations are L2-normalized. The collaborative alignment loss is defined as follows:

$$\mathcal{L}_{align} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\text{sim}(\hat{z}_i, z^i_{cf})/\tau)}{\sum_{j \in \mathcal{B}} \exp(\text{sim}(\hat{z}_i, z^j_{cf})/\tau)}, \tag{10}$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity function, τ is a temperature hyperparameter, and $\mathcal{B}$ is the batch size.

Collaborative Relational Preservation Loss To maintain the collaborative relational structure encoded in the CF space, we introduce a collaborative relational preservation loss that enforces structural consistency between the quantized and collaborative embedding spaces. Specifically, given a batch $\mathcal{B} = \{i_1, i_2, ...i_{|\mathcal{B}|}\}$ of items, we compute the pairwise similarity matrices for both spaces:

$$\mathcal{L}_{relation} = \frac{1}{|\mathcal{B}|^2} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \left(\text{Sim}^{quant}_{ij} - \text{Sim}^{cf}_{ij} \right)^2, \tag{11}$$

where Sim^{quant} and Sim^{cf} denote the cosine similarities between item quantized representations and collaborative embeddings, respectively.

Overall Tokenizer Optimization The semantic-enhanced tokenizer is optimized by jointly considering the semantic loss $\mathcal{L}_{\text{Sem}}$, collaborative alignment loss $\mathcal{L}_{align}$, and collaborative relational preservation loss $\mathcal{L}_{relation}$. The overall training loss can be denoted as follows:

$$\mathcal{L}_{\text{Tokenizer}} = \mathcal{L}_{Sem} + \alpha\mathcal{L}_{align} + \beta\mathcal{L}_{relation}, \quad (12)$$

where α and β are hyperparameters to control the strength of $\mathcal{L}_{align}$ and $\mathcal{L}_{relation}$, respectively.

Recommendation Generation Tasks

In the recommendation generation stage, we adopt a series of recommendation tasks to better leverage the advantages of multi-modal semantic information in identifiers, as shown at the bottom of Figure 2.

Next Item Generation (1) Uni-Modal Generation: In this setting, we predict the next item within a single modality. The model takes a sequence of historical item tokens from one modality as input, generates the token sequence for the next item of the same modality, and performs the analogous task for the image modality. This task allows the model to capture modality-specific patterns in user preferences.
(2) Multi-Modal Generation: To capture the interplay between modalities in users' behaviors, we use a unified input sequence containing interleaved text and image tokens from the user's interaction history. From this combined historical context, the model jointly predicts the next text item tokens and next image item tokens.

Multi-Modal Item Alignment While the generation tasks facilitate an implicit alignment, we introduce an explicit cross-modal alignment objective to enforce representational consistency. This objective comprises two symmetric subtasks: (1) Text-to-Image Alignment and (2) Image-to-Text Alignment. These tasks compel the model to learn an item's representation from one modality to its corresponding representation in the other.

Instantiation

We formulate the sequential recommendation task as a sequence generation problem using a T5 encoder-decoder architecture, with autoregressive generation and beam search during inference. As shown in Table 1, we briefly compare MusicRec with several representative traditional and generative recommendation baseline methods, demonstrating the differences and advantages of MusicRec.

Experiments

In this section, we perform extensive experiments on three datasets to answer the following research questions:

- **RQ1**: How does MusicRec perform compared to existing traditional and generative recommendation methods?
- **RQ2**: How do the components of MusicRec (e.g., multi-modal identifier, collaborative-semantic alignment, and multi-modal recommendation generation tasks) affect the performance?
- **RQ3**: How does MusicRec perform under diverse evaluation settings?

Methods	IMM	Fusion Type	Collaborative	GR
SASRec	✗	-	✔	✗
MISSRec	✔	discrete	✗	✗
TIGER	✗	-	✗	✔
LETTER	✗	-	✔	✔
MQL4GRec	✔	discrete	✗	✔
MusicRec	✔	combined	✔	✔

Table 1: Comparison of MusicRec with several related methods. "IMM" means the identifier with multi-modal semantics. "GR" means generative recommendation.

Experimental Settings

Datasets. We evaluate MusicRec on three public real-world datasets derived from the Amazon Product Reviews (Ni, Li, and McAuley 2019), which contains a large-scale collection of user reviews and ratings spanning multiple domains, offering rich behavioral data for recommender system evaluation. Specifically, we select three categories, including "**Instruments**", "**Arts**", and "**Games**" for sequential recommendation tasks.

Evaluation Metrics. To ensure a fair evaluation of sequential recommendation, we evaluate the recommendation performance using two widely adopted metrics: the top-K Hit Ratio (HR@K) and the top-K Normalized Discounted Cumulative Gain (NDCG@K), with K set to 5 and 10. To avoid evaluation bias, we conduct a full ranking over the entire item set rather than utilizing a sampling-based evaluation. For generation-based approaches employing beam search, we set the beam size to 20 during inference.

Baselines. We compare the performance of MusicRec against (1) traditional sequential recommendation methods, including GRU4Rec (Hidasi et al. 2015), BERT4Rec (Sun et al. 2019), SASRec (Kang and McAuley 2018), FDSA (Zhang et al. 2019), S^3-Rec (Zhou et al. 2020), VQ-Rec (Hou et al. 2023), and MISSRec (Wang et al. 2023), and (2) generative recommendation methods, including P5-CID (Hua et al. 2023), TIGER (Rajput et al. 2023), LETTER (Wang et al. 2024a), VIP5 (Geng et al. 2023), and MQL4GRec (Zhai et al. 2025).

Implementation Details. For item tokenization, we employ LLaMA to encode item titles and descriptions into textual embeddings and use CLIP (Radford et al. 2021) with ViT-L/14 as the backbone to obtain visual embeddings. Both the encoder and decoder of the RQ-VAE are implemented using MLPs equipped with ReLU activation functions. The model employs a hierarchical codebook structure consisting of 4 levels, where each level contains 256 embedding vectors of dimension 32. We obtain 32-dimensional item collaborative embeddings from a pre-trained SASRec (Kang and McAuley 2018).

During training, we instantiate MusicRec on the T5-small (Raffel et al. 2020) architecture. We follow the previous work (Rajput et al. 2023) to set μ as 0.25, and search α in

Method	Instruments				Arts				Games			
	HR@5	**HR@10**	**N@5**	**N@10**	**HR@5**	**HR@10**	**N@5**	**N@10**	**HR@5**	**HR@10**	**N@5**	**N@10**
GRU4Rec	0.0975	0.1207	0.0783	0.0857	0.0817	0.1088	0.0602	0.0690	0.0544	0.0895	0.0341	0.0453
BERT4Rec	0.0856	0.1081	0.0667	0.0739	0.0697	0.0922	0.0502	0.0575	0.0426	0.0725	0.0270	0.0366
SASRec	0.0946	0.1233	0.0654	0.0746	0.0951	0.1250	0.0610	0.0706	0.0587	0.0985	0.0333	0.0461
FDSA	0.0987	0.1249	0.0775	0.0859	0.0832	0.1190	0.0583	0.0695	0.0614	0.0988	0.0389	0.0509
S^3-Rec	0.0937	0.1123	0.0693	0.0743	0.0739	0.1030	0.0511	0.0630	0.0527	0.0903	0.0351	0.0468
VQ-Rec	0.1062	0.1357	0.0796	0.0891	0.1038	0.1386	0.0732	0.0844	0.0408	0.0679	0.0242	0.0329
MISSRec	0.1089	0.1361	0.0797	0.0880	0.1021	0.1321	0.0699	0.0815	0.0674	0.1048	0.0385	0.0499
P5-CID	0.0839	0.1119	0.0678	0.0704	0.0713	0.0994	0.0607	0.0662	0.0532	0.0824	0.0331	0.0454
TIGER	0.1007	0.1221	0.0882	0.0950	0.0894	0.1167	0.0718	0.0806	0.0523	0.0857	0.0345	0.0453
LETTER	0.1045	0.1271	0.0913	0.0986	0.0939	0.1202	0.0767	0.0852	0.0512	0.0855	0.0331	0.0441
VIP5	0.0892	0.1071	0.0815	0.0872	0.0704	0.0859	0.0586	0.0635	0.0480	0.0758	0.0328	0.0418
MQL4GRec	0.1071	0.1287	0.0927	0.0995	0.0928	0.1176	0.0748	0.0828	0.0576	0.0932	0.0377	0.0492
Ours*	**0.1257**	**0.1472**	**0.0993**	**0.1071**	**0.1121**	**0.1451**	**0.0895**	**0.1049**	**0.0749**	**0.1353**	**0.0488**	**0.0681**
Imp. (%)	+15.42	+8.15	+7.11	+7.63	+7.99	+4.68	+16.68	+23.12	+11.12	+29.10	+25.44	+33.79

Table 2: Overall performance comparison between baselines and MusicRec on three datasets. The best and second-best results are highlighted in bold and underlined, respectively. "Imp." denotes the relative improvement of MusicRec compared to the SOTA baseline. "*" indicates that the improvements are statistically significant based on a paired t-test (p-value < 0.05).

Variants	Instruments				Arts				Games			
	HR@5	**HR@10**	**N@5**	**N@10**	**HR@5**	**HR@10**	**N@5**	**N@10**	**HR@5**	**HR@10**	**N@5**	**N@10**
(0) MusicRec	**0.1257**	**0.1472**	**0.0993**	**0.1071**	**0.1121**	**0.1451**	**0.0895**	**0.1049**	**0.0749**	**0.1353**	**0.0488**	**0.0681**
(1) *w/o* I-Sem	0.0984	0.1172	0.0881	0.0941	0.0839	0.1038	0.0713	0.0776	0.0558	0.0908	0.0443	0.0548
(2) *w/o* T-Sem	0.0921	0.1081	0.0833	0.0884	0.0795	0.0964	0.0681	0.0735	0.0521	0.0847	0.0363	0.0476
(3) *w/o* $\mathcal{L}_{align}$	0.1173	0.1216	0.0914	0.1045	0.1093	0.1387	0.0835	0.0963	0.0731	0.1311	0.0395	0.0624
(4) *w/o* $\mathcal{L}_{rel}$	0.1099	0.1386	0.0957	0.1049	0.1109	0.1426	0.0853	0.1033	0.0678	0.1288	0.0339	0.0595
(5) *w/o* MMG	0.1216	0.1396	0.0951	0.1012	0.1029	0.1435	0.0801	0.0982	0.0728	0.1325	0.0421	0.0644
(6) *w/o* MMA	0.1108	0.1423	0.0956	0.1063	0.1079	0.1441	0.0869	0.1013	0.0741	0.1349	0.0471	0.0673

Table 3: Ablation study of MusicRec. The best results are highlighted in bold.

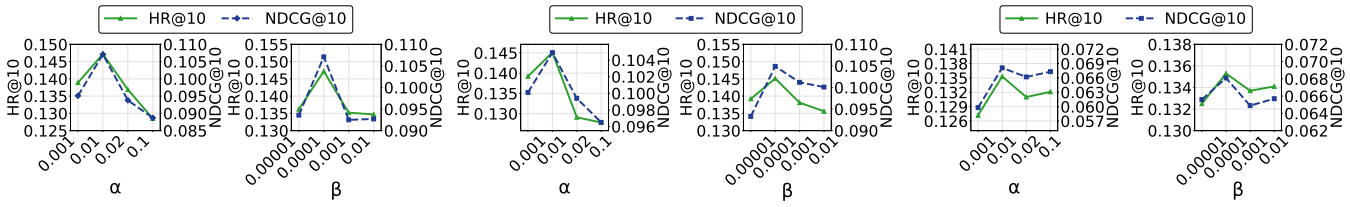
a range of {1e-1, 1e-2, 2e-2, 1e-3}, β in a range of {1e-2, 1e-3, 1e-4, 1e-5}. To ensure fair comparisons, we implement LETTER on TIGER following the original setup in the paper. For MQL4GRec, we directly fine-tune it without performing pre-training on source domain datasets. We conduct all experiments on 2 Tesla V100 GPUs.

Overall Performance (RQ1)

The performance comparison between traditional and generative recommendation baselines and MusicRec is shown in Table 2, from which we have the following observations. Among traditional sequential recommendation methods, MISSRec achieves the best performance on the Instruments and Games datasets of HR. This highlights the effectiveness of incorporating multi-modal item information. VQ-Rec performs best on the Arts dataset, except for NDCG@5. This suggests that vector quantization can lead to more efficient and compact representations. In addition, FDSA achieves the highest NDCG on the Games dataset, benefiting from its effective integration of additional textual feature embeddings. For generative recommendation methods, LETTER consistently outperforms TIGER and P5-CID across most metrics. CID underperforms due to its independent item indexing, which lacks semantic information. TIGER utilizes RQ-VAE to capture multi-granularity item semantics, but it lacks additional modality and collaborative signals. LETTER further enhances TIGER by incorporating collaborative information. MQL4GRec achieves the best overall performance on the Instruments and Games datasets by integrating multi-modal item information. However, the design of the unentangled modality identifier constrains its representation capacity. Compared with baseline approaches, MusicRec achieves the top performance on all datasets, which validates its overall effectiveness. This improvement can be attributed to integrating multi-modal semantic information during tokenization and using collaborative-semantic alignment in generating identifiers.

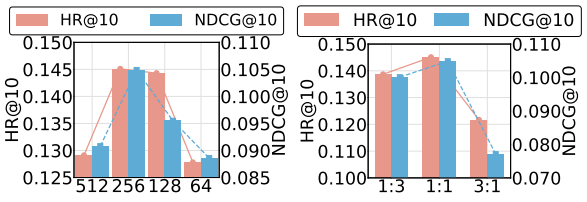
Ablation Study (RQ2).

We conduct a series of ablation studies to verify the design rationality of MusicRec following the same experimental setups, and the performance of the variants is depicted in Table 3. Removing image identifiers (*w/o* I-Sem) degrades performance across all datasets, confirming the importance



(a) Instruments (b) Arts (c) Games

Figure 3: Performance of MusicRec over different hyperparameters.



(a) The size of codebook (b) The contribution ratio of each codebook

Figure 4: Impact of different parameter settings in codebook and identifier on Arts.

of visual semantics in identifiers. The *w/o* T-Sem variant, which eliminates text identifiers, shows an even greater performance decline, indicating that the textual modality provides more indispensable semantic information than visual features. Removing the collaborative-aware alignment module (*w/o* $\mathcal{L}_{align}$) results in significant performance drops, demonstrating the critical role of integrating collaborative signals during identifier generation. The *w/o* $\mathcal{L}_{rel}$ variant, which eliminates collaborative relational preservation loss, shows a moderate decline, highlighting the importance of maintaining collaborative signal similarity. The *w/o* MMG variant, which eliminates the multi-modal next item generation task, shows a performance decline, demonstrating the importance of modal fusion in generating recommendation sequences. Finally, removing multi-modal fusion alignment tasks (*w/o* MMA) during sequence generation confirms that, beyond multi-modal identifier incorporation, multi-modal alignment in the generation stage is equally essential for optimal performance. In conclusion, each component of MusicRec is necessary to improve the performance.

Hyper-Parameter Analysis (RQ3).

(1) Strength of α and β. For collaborative-semantic alignment, we investigate it by varying the coefficient α in {1e-1, 1e-2, 2e-2, 1e-3}. As Figure 3 illustrates, increasing α beyond this optimal range could interfere with model learning and adversely affect performance. Our approach achieves the optimal results with α = 1e-2 for all datasets. To explore the influence of collaborative relational preservation, we tune β within the range in {1e-2, 1e-3, 1e-4, 1e-5} and observe similar trends to those seen with α, MisucRec performs best on all three datasets when β = 1e-4.

(2) Codebook size. We evaluate MusicRec with different codebook sizes in [64, 128, 256, 512], results in Figure 4(a) show that the codebook size substantially impacts model performance. Performance consistently improves as codebook size increases from 64 to 256, with the optimal results achieved at 256. Beyond this point, further expansion leads to performance degradation. This phenomenon can be attributed to two factors. Smaller codebooks suffer from insufficient representational capacity, limiting the model's ability to capture diverse item characteristics and resulting in poor discrimination. In contrast, excessively large codebooks introduce redundancy and make the model vulnerable to noisy features, causing it to learn spurious patterns that harm generalization. The 256-sized codebook strikes an optimal balance between expressiveness and robustness.

(3) Codebook contribution ratio. To evaluate the contribution of shared and specific codebooks in tokenization, we set the total number of codebooks to 4, then change the distribution ratio between them in [1:3, 1:1, 3:1]. Figure 4(b) demonstrates that the 1:1 ratio achieves the best performance as it balances modality-shared semantic and modality-specific semantic optimally. The 1:3 ratio shows slightly worse results since limited shared codebooks cannot capture sufficient cross-modal features. The 3:1 ratio exhibits a significant performance drop because excessive shared capacity lacks adequate specific semantics to preserve unique modality features, leading to information loss.

Conclusion

In this work, we proposed MusicRec, a novel multi-modal semantic-enhanced identifier approach with collaborative signals for generative recommendation. Unlike previous methods that relied solely on a single modality or two decoupled modalities for identifier generation, MusicRec effectively integrated semantic information across different modalities with a shared-specific item tokenization. To enhance the modeling of interactions among different modalities, we designed a novel alignment mechanism with collaborative signals in the identifier learning process. To enhance recommendation generation, we further proposed a series of multi-modal fusion and alignment tasks. Extensive experiments and comparative studies with advanced methods highlight the strong performance and robustness of MusicRec. In future work, we will explore combining more types of modality and transferring to generative recommendation models with other structures.

Acknowledgments

This work was supported by National Key Research and Development Program of China (No.2022YFC3302103), the Joint Fund Key Program of the National Natural Science Foundation of China(No.U23B2029), the Fundamental Research Funds for the Central Universities (No. CUC25SG013), the National Science Foundation of China (No.62002027) and the Youth Research Special Project of NCUT (No.2025NCUTYRSP012).

References

Chee, J.; Kalyanaraman, S.; Ernala, S. K.; Weinsberg, U.; Dean, S.; and Ioannidis, S. 2024. Harm mitigation in recommender systems under user preference dynamics. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 255–265.

Dong, Z.; Liu, X.; Chen, B.; Polak, P.; and Zhang, P. 2024. Musechat: A conversational music recommendation system for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12775–12785.

Geng, S.; Tan, J.; Liu, S.; Fu, Z.; and Zhang, Y. 2023. VIP5: Towards Multimodal Foundation Models for Recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9606–9620.

Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Hong, M.; Xia, Y.; Wang, Z.; Zhu, J.; Wang, Y.; Cai, S.; Yang, X.; Dai, Q.; Dong, Z.; Zhang, Z.; et al. 2025. EAGER-LLM: Enhancing Large Language Models as Recommenders through Exogenous Behavior-Semantic Integration. In *Proceedings of the ACM on Web Conference 2025*, 2754–2762.

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.

Hua, W.; Xu, S.; Ge, Y.; and Zhang, Y. 2023. How to index item ids for recommendation foundation models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 195–204.

Jiang, Y.; Xia, L.; Wei, W.; Luo, D.; Lin, K.; and Huang, C. 2024. Diffmm: Multi-modal diffusion model for recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7591–7599.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining*, 197–206. IEEE.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 19730–19742.

Lin, X.; Shi, H.; Wang, W.; Feng, F.; Wang, Q.; Ng, S.-K.; and Chua, T.-S. 2025. Order-agnostic Identifier for Large Language Model-based Generative Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1923–1933.

Lin, X.; Wang, W.; Li, Y.; Feng, F.; Ng, S.-K.; and Chua, T.-S. 2024. Bridging items and language: A transition paradigm for large language model-based recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1816–1826.

Liu, H.; Wei, Y.; Song, X.; Guan, W.; Li, Y.-F.; and Nie, L. 2024. Mmgrec: Multimodal generative recommendation with transformer model. *arXiv preprint arXiv:2404.16555*.

Ma, H.; Yang, Y.; Meng, L.; Xie, R.; and Meng, X. 2024. Multimodal conditioned diffusion model for recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, 1733–1740.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural L Processing and the 9th International Joint Conference on Natural Language Processing*, 188–197.

Petrov, A. V.; and Macdonald, C. 2024. RecJPQ: training large-catalogue sequential recommenders. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 538–547.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rajput, S.; Mehta, N.; Singh, A.; Keshavan, R.; Vu, T.; Heidt, L.; Hong, L.; Tay, Y.; Tran, V. Q.; Samost, J.; et al. 2023. Recommender systems with generative retrieval. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 10299–10315.

Shi, L.; Yang, J.; Lv, P.; Yuan, L.; Kou, F.; Luo, J.; and Xu, M. 2024. Self-derived knowledge graph contrastive learning for recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7571–7580.

Shi, T.; Xu, J.; Zhang, X.; Zang, X.; Zheng, K.; Song, Y.; and Yu, E. 2025. GenSAR: Unifying Balanced Search and Recommendation with Generative Retrieval. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 124–134.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 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.

Tan, J.; Xu, S.; Hua, W.; Ge, Y.; Li, Z.; and Zhang, Y. 2024. Idgenrec: Llm-recsys alignment with textual id learning. In *Proceedings of the 47th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, 355–364.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6309–6318.

Wang, J.; Zeng, Z.; Wang, Y.; Wang, Y.; Lu, X.; Li, T.; Yuan, J.; Zhang, R.; Zheng, H.-T.; and Xia, S.-T. 2023. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6548–6557.

Wang, W.; Bao, H.; Lin, X.; Zhang, J.; Li, Y.; Feng, F.; Ng, S.-K.; and Chua, T.-S. 2024a. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2400–2409.

Wang, Y.; Ren, Z.; Sun, W.; Yang, J.; Liang, Z.; Chen, X.; Xie, R.; Yan, S.; Zhang, X.; Ren, P.; et al. 2024b. Content-Based Collaborative Generation for Recommender Systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2420–2430.

Wang, Y.; Xun, J.; Hong, M.; Zhu, J.; Jin, T.; Lin, W.; Li, H.; Li, L.; Xia, Y.; Zhao, Z.; et al. 2024c. Eager: Two-stream generative recommender with behavior-semantic collaboration. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3245–3254.

Wang, Z.; Ye, W.; Chen, X.; Zhang, W.; Wang, Z.; Zou, L.; and Liu, W. 2022. Generative session-based recommendation. In *Proceedings of the ACM Web Conference 2022*, 2227–2235.

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-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1437–1445.

Xiao, L.; Wang, H.; Wang, C.; Ji, L.; Wang, Y.; Zhu, J.; Dong, Z.; Zhang, R.; and Li, R. 2025. Progressive Collaborative and Semantic Knowledge Fusion for Generative Recommendation. *arXiv preprint arXiv:2502.06269*.

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 39th AAAI Conference on Artificial Intelligence*, volume 39, 13069–13077.

Yi, J.; and Chen, Z. 2021. Multi-modal variational graph auto-encoder for recommendation systems. *IEEE Transactions on Multimedia*, 24: 1067–1079.

Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507.

Zhai, J.; Mai, Z.-F.; Wang, C.-D.; Yang, F.; Zheng, X.; Li, H.; and Tian, Y. 2025. Multimodal Quantitative Language for Generative Recommendation. In *Proceedings of the 13th International Conference on Learning Representations*, 88215–88232.

Zhang, C.; Wu, S.; Zhang, H.; Xu, T.; Gao, Y.; Hu, Y.; and Chen, E. 2024. Notellm: A retrievable large language model for note recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, 170–179.

Zhang, T.; Zhao, P.; Liu, Y.; Sheng, V. S.; Xu, J.; Wang, D.; Liu, G.; and Zhou, X. 2019. Feature-level deeper self-attention network for sequential recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 4320–4326.

Zhang, Y.; Feng, F.; Zhang, J.; Bao, K.; Wang, Q.; and He, X. 2025. CoLLM: Integrating Collaborative Embeddings Into Large Language Models for Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, (01): 1–12.

Zhao, Y.; Tan, C.; Shi, L.; Zhong, Y.; Kou, F.; Zhang, P.; Chen, W.; and Ma, C. 2025. Generative Recommender Systems: A Comprehensive Survey on Model, Framework, and Application. *Information Fusion*, 103919.

Zheng, B.; Hou, Y.; Lu, H.; Chen, Y.; Zhao, W. X.; Chen, M.; and Wen, J.-R. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering*, 1435–1448. IEEE.

Zhou, K.; Wang, H.; Zhao, W. X.; Zhu, Y.; Wang, S.; Zhang, F.; Wang, Z.; and Wen, J.-R. 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.

Zhou, P.; Liu, C.; Ren, J.; Zhou, X.; Xie, Y.; Cao, M.; Rao, Z.; Huang, Y.-L.; Chong, D.; Liu, J.; et al. 2025. When Large Vision Language Models Meet Multimodal Sequential Recommendation: An Empirical Study. In *Proceedings of the ACM on Web Conference 2025*, 275–292.