Empowering Large Language Model for Sequential Recommendation via Multimodal Embeddings and Semantic IDs

Yuhao Wang
City University of Hong Kong
Hong Kong, China
yhwang25-c@my.cityu.edu.hk

Junwei Pan
Tencent Inc.
Shenzhen, China
jonaspan@tencent.com

Xinhang Li
Tsinghua University
Beijing, China
xh-li20@mailstsinghua.edu.cn

Maolin Wang
City University of Hong Kong
Hong Kong, China
morin.wang@my.cityu.edu.hk

Yuan Wang
Tencent Inc.
Shenzhen, China
leoyuanwang@tencent.com

Yue Liu
Tencent Inc.
Shenzhen, China
herculesliu@tencent.com

Dapeng Liu
Tencent Inc.
Shenzhen, China
rocliu@tencent.com

Jie Jiang
Tencent Inc.
Shenzhen, China
zeus@tencent.com

Xiangyu Zhao ✉
City University of Hong Kong
Hong Kong, China
xianzhao@cityu.edu.hk

Abstract

Sequential recommendation (SR) aims to capture users' dynamic interests and sequential patterns based on their historical interactions. Recently, the powerful capabilities of large language models (LLMs) have driven their adoption in SR. However, we identify two critical challenges in existing LLM-based SR methods: 1) embedding collapse when incorporating pre-trained collaborative embeddings and 2) catastrophic forgetting of quantized embeddings when utilizing semantic IDs. These issues dampen the model scalability and lead to suboptimal recommendation performance. Therefore, based on LLMs like Llama3-8B-instruct, we introduce a novel SR framework named MME-SID, which integrates multimodal embeddings and quantized embeddings to mitigate embedding collapse. Additionally, we propose a Multimodal Residual Quantized Variational Autoencoder (MM-RQ-VAE) with maximum mean discrepancy as the reconstruction loss and contrastive learning for alignment, which effectively preserve intra-modal distance information and capture inter-modal correlations, respectively. To further alleviate catastrophic forgetting, we initialize the model with the trained multimodal code embeddings. Finally, we fine-tune the LLM efficiently using LoRA in a multimodal frequency-aware fusion manner. Extensive experiments on three public datasets validate the superior performance of MME-SID thanks to its capability to mitigate embedding collapse and catastrophic forgetting. The implementation code and datasets are publicly available for reproduction[1].

[1]https://github.com/Applied-Machine-Learning-Lab/MME-SID

✉Corresponding author.

CIKM '25, November 10–14, 2025, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761169

CCS Concepts

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

Keywords

Sequential Recommendation, Multimodal Recommendation, Recommender System, Large Language Model, Semantic IDs

ACM Reference Format:
Yuhao Wang, Junwei Pan, Xinhang Li, Maolin Wang, Yuan Wang, Yue Liu, Dapeng Liu, Jie Jiang, and Xiangyu Zhao ✉. 2025. Empowering Large Language Model for Sequential Recommendation via Multimodal Embeddings and Semantic IDs . In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761169

1 Introduction

In recent decades, the rapid development of web applications, such as short video platforms and e-commerce services, has significantly increased the importance of recommender systems in driving profits and enhancing user engagement [46]. In the recommendation community, sequential recommendation (SR) aims to model sequential patterns and capture users' dynamic interests by leveraging their historical interactions [5]. Traditional SR methods primarily rely on collaborative modality, *i.e.*, using only item IDs. However, these approaches are particularly vulnerable to the cold-start problem, which arises with new users, items, and business scenarios [47, 50].

Recently, large language models (LLMs) have demonstrated remarkable capabilities in comprehending semantic data in natural language format [1]. Consequently, an increasing number of studies have explored the use of LLMs for sequential recommendation (LLM4SR). For instance, some works like TALLRec [2] formulates SR as a text generation task and applies instruction tuning on LLMs. Meanwhile, to enable LLMs to perform generative recommendation or retrieval, a parallel line of research [35, 40, 66] introduces semantic IDs to represent items. Specifically, these methods learn to transform the item embedding into semantic IDs, which are then treated

as new generative LLM tokens. For example, as shown in Fig. 2, the textual embedding of item is encoded into a sequence of semantic IDs or codes as (1, 2). This scheme is referred to as quantization and Residual Quantized Variational Autoencoder (RQ-VAE) [7] is a representative quantization model which will be detailed in Sec. 2.2.

However, we identify two key challenges in existing LLM4SR [19, 20, 41] models which lead to suboptimal performance as follows:

- **Embedding Collapse.** Also known as dimensional collapse [4], this phenomenon indicates the embedding matrix is nearly low-rank with mostly significantly small singular values. In that circumstance, the embedding matrix only occupies a low-dimensional subspace, leading to inefficient use of model capacity and limited scalability [32]. Existing works [4, 32] found that in traditional recommender system this issue is caused by the interaction between low-dimensional embeddings and possibly high-dimensional embeddings of other feature fields. By contrast, we also observe embedding collapse in LLM4SR. Sec. 5.3 shows over **98%** dimensions of embedding matrix collapse in experiment. Sec. 3.1 shows the cause is simply mapping low-dimensional collaborative embeddings from pre-trained recommendation models into high-dimensional LLM representation space.
- **Catastrophic Forgetting.** It usually refers to the lost of previously learned knowledge when incorporating information relevant to the current task. Typically, existing studies [35, 40, 43, 49, 66] simply discard the learned code embeddings after training quantization model. They only maintain the assigned semantic IDs and train their embeddings from scratch on the downstream retrieval or recommendation task. Nonetheless, even if the hierarchical structure of semantic IDs is preserved, vast majority of information in the original code embeddings (*e.g.,* over **94%** of the partial order information of distance as shown in Sec. 3.2) is lost and cannot be retained, indicating catastrophic forgetting.

Moreover, we highlight a fundamental dilemma that simultaneously addressing embedding collapse and catastrophic forgetting poses a significant challenge in LLM4SR: **(i)** Relying solely on pre-trained low-dimensional collaborative embeddings inevitably leads to collapse. Though one can increase the embedding dimension of conventional SR model, blindly enlarging it can negatively impact model performance [4]. Meanwhile, the common solutions on tackling embedding collapse in traditional recommendation model like multi-embedding [15, 39] all fail in LLM-based recommendation framework, which will be shown in Sec. 5. **(ii)** Although randomly initialized embedding matrices are less prone to collapse [37, 38], training a new embedding table incurs high computational costs, particularly in industrial-scale SR systems that involve billions of users and items [3, 8, 14, 21, 23, 45, 60, 61, 64]. Moreover, these newly trained embeddings fail to retain previously acquired knowledge.

To address these challenges, we propose MME-SID, a novel framework that enhances large language models for sequential recommendation with multimodal embeddings and semantic IDs. Specifically, we introduce a Multimodal Residual Quantized Variational Autoencoder (MM-RQ-VAE) to generate multimodal semantic IDs. Notably, to better preserve distance information and alleviating forgetting, it incorporates a characteristic-kernel-based maximum mean discrepancy as the reconstruction loss. Besides, a contrastive learning objective is adopted to capture inter-modal correlations. On the one hand, to alleviate embedding collapse, we propose to simultaneously leverage the original embedding and the embedding of semantic IDs in collaborative, textual, and visual modalities to obtain an informative multimodal embedding for each item. On the other hand, to mitigate catastrophic forgetting, we initialize the embeddings of multimodal semantic IDs using the trained code embeddings from MM-RQ-VAE. Finally, we fine-tune the LLM in a multimodal frequency-aware and efficient manner using LoRA. We will comprehensively analyze the advantage of MME-SID in Sec. 4.4 to justify its advantage and profound impact in potential.

The key contributions of this paper are summarized as follows:

- To the best of our knowledge, it is the first work to identify and systematically address the embedding collapse and catastrophic forgetting issue in large language model for recommendation.
- We provide innovative perspectives on: 1) HOW multimodal information contributes to reducing collapse and improving recommendation performance, thus truly unleashing the potential of LLM for recommendation. 2) HOW to better preserve the distance information in quantized embeddings to mitigate forgetting. 3) WHAT is a better way to use semantic IDs.
- We conduct extensive experiments on three public datasets of Amazon, demonstrating the superior recommendation performance of MME-SID and providing in-depth analyses of its ability to address embedding collapse and catastrophic forgetting.

2 Background

In this section, we first demonstrate the problem formulation and introduce Residual Quantized Variational Autoencoder (RQ-VAE).

2.1 Problem Formulation

In sequential recommendation, denote user set and item set as $\mathcal{U}$ and $\mathcal{I}$, we can obtain the behavioral item sequence $\{h_u\}$, target item x_u, and true label y_u of each user $u \in U$. A conventional sequential recommender system (SRS) f_θ usually takes $\{h_u\}$ as input and the prediction result $\hat{y}$ is obtained by multiplying its output and the target item embedding through dot product. Finally, the binary cross entropy (BCE) loss is usually optimized [11, 16, 27, 44, 51, 63, 65].

$$\min_\theta \mathcal{L} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \text{BCE}\left(f_\theta\left(\{h_u\}, x_u\right), y_u\right) \tag{1}$$

2.2 RQ-VAE

Residual Quantized Variational Autoencoder (RQ-VAE) [7] aims to tokenize and generate the semantic IDs of the original embedding in a hierarchical manner. Specifically, the original embedding $\boldsymbol{s}$ is encoded into an encoded into the latent semantic embedding $\boldsymbol{z}$, which is further quantized into the codes (or the so-called semantic IDs) through L-level codebooks. Specifically, for each code level $l = 1, \ldots, L$, there is a codebook $C_l = \left\{\boldsymbol{CE}_j\right\}_{j=1}^{S}$, where $\boldsymbol{CE}_j \in \mathbb{R}^d$ are learnable code embeddings and S denotes the codebook size. Furthermore, the residual quantization is formulated as

$$\begin{aligned} SID^l &= \arg\min_j \left\|\boldsymbol{r}_{l-1} - \boldsymbol{CE}_j\right\|^2 \\ \boldsymbol{r}_l &= \boldsymbol{r}_{l-1} - \boldsymbol{CE}_{SID^l} \end{aligned} \tag{2}$$

where SID^l is the assigned semantic ID at the l-th level codebook, $\boldsymbol{r}_{l-1}$ is the residual from the last level, $\boldsymbol{r}_0 = \boldsymbol{z}$, and $\|\cdot\|$ is $L2$

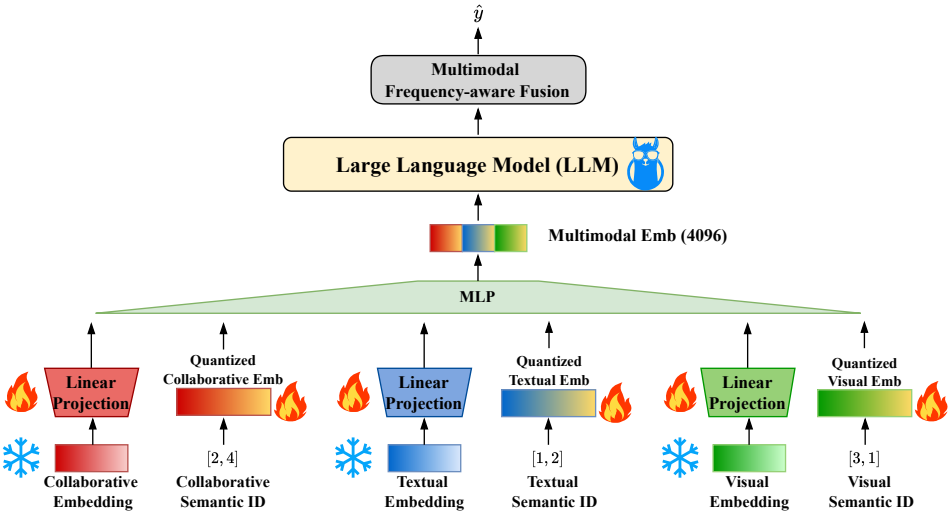

Figure 1: The overall framework of MME-SID.

norm. Finally, the semantic IDs are $\{SID^1, \ldots, SID^L\}$ and the quantized embedding $\hat{z} = \sum_{l=1}^{L} CE_{SID^l}$ is further decoded into $\hat{s}$ to reconstruct s. Denote SG as the stop gradient operation and α as a hyper-parameter, the overall loss function is

$$\mathcal{L} = \mathcal{L}_{\text{Recon}} + \mathcal{L}_{\text{RQ-VAE}} \tag{3}$$

$$\mathcal{L}_{\text{Recon}} = \|s - \hat{s}\|^2 \tag{4}$$

$$\mathcal{L}_{\text{RQ-VAE}} = \sum_{l=1}^{L} \left\|\text{SG}\left(\boldsymbol{r}_{l-1}\right) - \boldsymbol{CE}_{SID^l}\right\|^2 + \alpha \left\|\boldsymbol{r}_{l-1} - \text{SG}\left(\boldsymbol{CE}_{SID^l}\right)\right\|^2 \tag{5}$$

3 Preliminary Analysis

In this section, we investigate the embedding collapse and catastrophic forgetting phenomena theoretically and empirically in large language model for sequential recommendation (LLM4SR).

3.1 Embedding Collapse

Existing LLM4SR methods [10, 13, 59] usually extract collaborative information from the pre-trained collaborative embedding $\boldsymbol{E_c}$ by mapping it into LLM token space. Suppose there is a matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, then the following formula holds:

$$\text{rank}(\boldsymbol{A} \cdot \boldsymbol{B}) \leq \min\{\text{rank}(\boldsymbol{A}), \text{rank}(\boldsymbol{B})\} \tag{6}$$

$$\text{rank}(\boldsymbol{A} + \boldsymbol{B}) < \text{rank}(\boldsymbol{A}) + \text{rank}(\boldsymbol{B}) \tag{7}$$

Therefore, taking linear projection as a common example, we can derive that the rank of the projected embedding satisfies:

$$\begin{aligned} \text{rank}(\boldsymbol{W} \cdot \boldsymbol{E_c} + b) &< \text{rank}(\boldsymbol{W} \cdot \boldsymbol{E_c}) + \text{rank}(b) \\ &\leq \min\{\text{rank}(\boldsymbol{W}), \text{rank}(\boldsymbol{E_c})\} + 1 \\ &\leq \text{rank}(\boldsymbol{E_c}) + 1 \end{aligned} \tag{8}$$

where $\boldsymbol{E_c} \in \mathbb{R}^{M \times D}$ is the embedding table, $W \in \mathbb{R}^{D' \times D}$ and $b \in \mathbb{R}^{D' \times 1}$ denotes the weight and bias of the linear projection. D and D' denotes the dimension of the original and projected embedding. Consequently, since $\boldsymbol{E_c}$ is usually low-rank (*e.g.*, 64 or 128 in traditional SRS), we can find that after the transformation of linear projection, the pre-trained low-dimensional collaborative embedding is only mapped into a low-dimensional sub-space of the LLM token embedding space, leading to embedding collapse.

Besides, for nonlinear mappings it is difficult to draw a unified conclusion of their impact on matrix rank through theoretical analysis. Thus we empirically calculate the singular value of the embeddings in different methods and analyze the results in Sec. 5.3.

3.2 Catastrophic Forgetting

We adopt Kendall's tau [6] to measure forgetting, *i.e.*, how much distance information is lost or preserved. For example, consider a user's behavioral items are $\{i_1, i_2\}$ and the target item is i_3 in the historical interactions. A recommendation model f_θ first maps the items into embedding $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, and $\boldsymbol{e}_3$ and then the distance $\langle \cdot, \cdot \rangle$ between each pair of behavioral and target item embedding is computed. This results in the variable $\{\langle \boldsymbol{e}_1, \boldsymbol{e}_2 \rangle, \langle \boldsymbol{e}_1, \boldsymbol{e}_3 \rangle\}$. For another model $f_{\theta'}$ its distance variable is $\{\langle \boldsymbol{e}'_1, \boldsymbol{e}'_2 \rangle, \langle \boldsymbol{e}'_1, \boldsymbol{e}'_3 \rangle\}$. Subsequently, Kendall's tau can be utilized to assess the concordance between the distance variable of the two models, which is defined as

$$\tau = \frac{\#(\text{concordant pairs}) - \#(\text{disconcordant pairs})}{\#(\text{pairs})} \tag{9}$$

where # denotes the count. Specifically, a pair of samples is deemed concordant if the sorting order is consistent, *i.e.*, both $\langle \boldsymbol{e}_1, \boldsymbol{e}_2 \rangle < \langle \boldsymbol{e}_1, \boldsymbol{e}_3 \rangle$ and $\langle \boldsymbol{e}'_1, \boldsymbol{e}'_2 \rangle < \langle \boldsymbol{e}'_1, \boldsymbol{e}'_3 \rangle$ are true, or both $\langle \boldsymbol{e}_1, \boldsymbol{e}_2 \rangle > \langle \boldsymbol{e}_1, \boldsymbol{e}_3 \rangle$ and $\langle \boldsymbol{e}'_1, \boldsymbol{e}'_2 \rangle > \langle \boldsymbol{e}'_1, \boldsymbol{e}'_3 \rangle$ are true.

Based on Llama3-8B-instruct model, we conduct a preliminary experiment on Amazon Beauty dataset using semantic IDs. Specifically, an RQ-VAE model is adopted on the collaborative embedding $\boldsymbol{E_c}$ obtained from a pre-trained SASRec model. It generates the semantic IDs of $\boldsymbol{E_c}$ and its quantized embedding is $\hat{\boldsymbol{z}}$. Moreover, for both $\boldsymbol{E_c}$ and $\hat{\boldsymbol{z}}$ we calculate the distance between each behavioral and target item embedding adopting Euclidean distance and it achieves $\tau = 0.3714$. This indicates that the quantized embedding trained preserves 37.14% of the original information (*i.e.*, partial order of distance) in $\boldsymbol{E_c}$. By contrast, we also randomly initialize the code embeddings and fine-tune the LLM to conduct sequential recommendation on the same training data as SASRec. However,

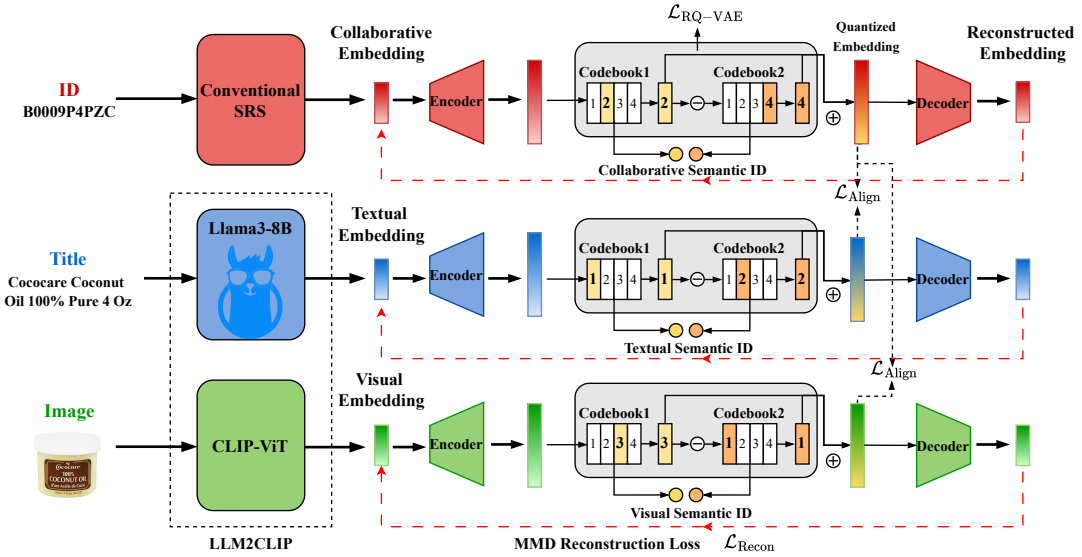

Figure 2: The model architecture of MM-RQ-VAE, which consists of an RQ-VAE for each modality. Specifically, the black solid arrow denotes data flow, the red dashed arrow denotes alignment, and the red dashed arrow denotes reconstruction.

the fine-tuned quantized embedding only achieves $\tau = 0.0550$, indicating that *94.50% of the previously learned information is forgotten.* Therefore, this result shows the catastrophic forgetting when randomly initializing the code embeddings on the downstream tasks, which provides preliminary validation for our conjecture.

4 Method

We first provide an overview of the proposed framework, then detail it into two stages. Finally the advantage and potential significance of MME-SID is discussed.

4.1 Overview

To alleviate embedding collapse and catastrophic forgetting phenomena, we propose to leverage both multimodal embeddings and semantic IDs with the trained code embeddings. The overview of MME-SID is depicted in Fig. 1, which consists of two stages: encoding and fine-tuning. Specifically, the encoding stage aims to obtain the multimodal embeddings and their semantic IDs. Next, the fine-tuning stage aims to efficiently tune the LLM to conduct SR task in a multimodal frequency-aware manner. Additionally, the pseudo-code is provided in Appendix. A.

4.2 Encoding Stage

In the encoding stage, first the item embeddings in the collaborative, textual, and visual modality are obtained, which are further quantized and transformed into multimodal semantic IDs by our proposed multimodal RQ-VAE model. In the following parts, the collaborative, textual, and visual embedding of item are denoted as $\boldsymbol{E}_c$, $\boldsymbol{E}_t$, and $\boldsymbol{E}_v$, respectively.

4.2.1 ***Multimodal Embedding Encoding.*** Existing works on multimodal recommendation [12, 17, 24] either leverage a combination of individual vision encoder and text encoder or adopt a multimodal encoder like BEiT3 [42] to transform the original multimodal data into multimodal embeddings. However, there are two limitations of these methods. First, embeddings from the individual vision encoder and text encoder are not in the same representational space, which requires additional cost of alignment and even leads to semantic loss [48]. Second, most existing multimodal encoders like CLIP [34] have limited capability of processing long and complex texts, which can not meet the demand of handling the textual information of items like title, descriptions, and review.

Therefore, we adopt LLM2CLIP [52] as the multimodal encoder which enhances the original CLIP model by replacing the text encoder with a more powerful LLM like Llama3-8B. Specifically, LLM2CLIP takes the multimodal attribute of items as input and outputs the textual and visual embedding $\boldsymbol{E}_t \in \mathbb{R}^{D_t \times |\mathcal{I}|}$ and $\boldsymbol{E}_v \in \mathbb{R}^{D_v \times |\mathcal{I}|}$, where D_t and D_v denotes the embedding size of textual and visual embedding, respectively. Meanwhile, a traditional SRS like SASRec [5] is trained on collaborative data (*i.e.*, item ID only) and its embedding table $\boldsymbol{E}_c \in \mathbb{R}^{D_c \times |\mathcal{I}|}$ is extracted where D_c is the collaborative embedding size.

4.2.2 ***Multimodal Embedding Quantization.*** Existing works on using semantic IDs for recommendation suffer from two drawbacks. First, as shown in Eq. 4, they simply adopt mean squared error (MSE) as the reconstruction loss, which does not explicitly preserve the information of distance distribution. This is because minimizing the MSE between the decoded quantized embedding and the original embedding is equivalent to minimizing their distance in Euclidean space. Second, existing methods [35, 66] usually use the semantic IDs of only textual embedding to represent each item and fine-tune the corresponding embeddings on the downstream task, which can not capture the distinction across modalities.

To address them, we propose a multimodal Residual Quantized Variational Autoencoder named MM-RQ-VAE and its model architecture is shown in Fig. 2. Specifically, in each modality $j \in \{c, t, v\}$, the original embedding $\boldsymbol{s}_j \in \boldsymbol{E}_j$ is encoded into semantic embedding

$\boldsymbol{z}_j$, then the semantic IDs $\{SID_j^1, \ldots, SID_j^L\}$, the quantized embedding $\hat{\boldsymbol{z}}_j$, and the decoded quantized embedding $\hat{\boldsymbol{s}}_j$ are generated through L-level codebook.

First and foremost, to explicitly improve the ability of the quantized embedding $\hat{\boldsymbol{z}}_j$ to preserve the information in the original embedding $\boldsymbol{s}_j$, we propose to minimize the maximum mean discrepancy (MMD) between $\hat{\boldsymbol{s}}_j$ and $\boldsymbol{s}_j$ as the reconstruction loss. Specifically, MMD [28, 36] measures the distance between any probability distribution P and Q which is defined as

$$\mathrm{MMD}_K(P, Q) \triangleq \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|_{\mathcal{H}_K} \tag{10}$$

where $k(\cdot, \cdot)$ is a symmetric positive-definite kernel with its unique reproducing kernel Hilbert space $\mathcal{H}_K$, $\boldsymbol{\mu}$ represents the mean embedding of distribution, and $\|\cdot\|_{\mathcal{H}_K}$ is the norm of $\mathcal{H}_K$. Notably, the kernel mean of characteristic kernel can preserve all statistics of the distribution [36]. We will validate the advantage of MMD over MSE in the subsequent analytical experiments in Sec. 5.4.

Second, to grasp inter-modality connection, we propose to align the quantized collaborative embedding $\hat{\boldsymbol{z}}_c$ with quantized textual and visual embedding $\hat{\boldsymbol{z}}_t$ and $\hat{\boldsymbol{z}}_v$ via contrastive objective like Info_NCE loss. Notably, MM-RQ-VAE does not need to align the visual and textual modality because LLMCLIP has already mapped the visual and textual information into the same embedding space. Overall, the trade-off between the reconstruction and alignment loss enables the quantized embedding to simultaneously learn the intra-modal and inter-modal correlations of multimodal embeddings.

The overall loss function of MM-RQ-VAE is

$$\mathcal{L}_{\text{MM-RQ-VAE}} = \mathcal{L}_{\text{Recon}} + \beta \cdot \mathcal{L}_{\text{Align}} + \gamma \cdot \sum_{j \in \{c,t,v\}} \mathcal{L}_{\text{RQ-VAE}} \tag{11}$$

$$\mathcal{L}_{\text{Recon}} = \sum_{b \subset \mathcal{I}} \sum_{j \in \{c,t,v\}} \mathrm{MMD}_K^2 \left(\mathtt{SG}(\boldsymbol{s}_j^b), \hat{\boldsymbol{s}}_j^b \right) \tag{12}$$

$$\mathcal{L}_{\text{Align}} = \mathcal{L}_{c-t} + \mathcal{L}_{c-v} \tag{13}$$

$$\mathcal{L}_{c-t} = -\frac{1}{|\mathcal{I}|} \sum_{i=1}^{|\mathcal{I}|} \log \frac{\exp\left(\langle \hat{\boldsymbol{z}}_c^i, \hat{\boldsymbol{z}}_t^i \rangle / \epsilon\right)}{\exp\left(\langle \hat{\boldsymbol{z}}_c^i, \hat{\boldsymbol{z}}_t^i \rangle / \epsilon\right) + \sum_{i' \neq i} \exp\left(\langle \hat{\boldsymbol{z}}_c^i, \hat{\boldsymbol{z}}_t^{i'} \rangle / \epsilon\right)} \tag{14}$$

$$\mathcal{L}_{c-v} = -\frac{1}{|\mathcal{I}|} \sum_{i=1}^{|\mathcal{I}|} \log \frac{\exp\left(\langle \hat{\boldsymbol{z}}_c^i, \hat{\boldsymbol{z}}_v^i \rangle / \epsilon\right)}{\exp\left(\langle \hat{\boldsymbol{z}}_c^i, \hat{\boldsymbol{z}}_v^i \rangle / \epsilon\right) + \sum_{i' \neq i} \exp\left(\langle \hat{\boldsymbol{z}}_c^i, \hat{\boldsymbol{z}}_v^{i'} \rangle / \epsilon\right)} \tag{15}$$

where $\mathcal{L}_{\text{RQ-VAE}}$ is equivalent to Eq. 5, $\langle \cdot, \cdot \rangle$ denotes the similarity metric like cosine, b denotes a batch of samples, SG denotes stop gradient operation, β and γ are hyper-parameters, and ϵ is the temperature coefficient [18, 21, 22, 25, 26, 54, 55, 57, 62].

4.3 Fine-tuning Stage

As mentioned in Sec. 3.2, existing methods tend to only leverage the semantic IDs and discard the trained code embeddings, thus neglecting the impact of catastrophic forgetting. To address this issue, we propose to initialize the embeddings of semantic ID $\boldsymbol{E}_{SID_j^l}$ with code embeddings $\boldsymbol{CE}_{SID_j^l}$ from the trained MM-RQ-VAE, which preserve abundant intra-modal information (*i.e.,* distance between behavioral and target item embedding).

The prompt template is provided in the code. Specifically, the LLM input consists of '{Instruction}' and '{Behavioral Item Sequence}' in which '{Instruction}' denotes the instruction to conduct SR task while '{Behavioral Item Sequence}' is formulated as

$$f_{\text{MLP}}\left(\left[\boldsymbol{W}_j \cdot (\boldsymbol{X} \cdot \mathtt{SG}(\boldsymbol{E}_j)) + b_j, \sum_{l=1}^{L} \boldsymbol{E}_{SID_j^l} \right] \right), j \in \{c, t, v\} \tag{16}$$

where $\boldsymbol{X}$ denotes the one-hot vector of the behavioral item sequence. $\boldsymbol{W}_j$ and b_j denotes the weight and bias of linear projection of each modality. The square bracket denotes the concatenation operation and SG denotes the stop gradient operation. Generally, the linear projection of the original embeddings and sum of embeddings of semantic ID (*i.e.,* quantized embeddings) of different modalities are concatenated. Then it is fed into an MLP to convert the dimension into D_{LLM}, which denotes the dimension of token embeddings of LLM. Afterward, the subsequent LLM acts as the SR model. Notably, this input format of MME-SID is distinctive from that of all existing methods as illustrated in Tab. 2. It has the advantage of simultaneously preserving distance information of the original embedding and hierarchical structure of semantic IDs.

Meanwhile, existing SR models often ignore the fact that the importance of different modalities varies for cold or warm items, leading to suboptimal recommendation result. Therefore, we propose a multimodal frequency-aware fusion module to adaptively fuse the score between LLM output and item embeddings in different modalities. Specifically, the last hidden state of the LLM output $\boldsymbol{o}_{\text{LLM}}$ is leveraged. Besides, the frequency of each item i occurred in the training set is recorded as q_i. Given the observation that the user-item interaction data usually follows a highly-skewed long-tail distribution [33], the frequency q_i is first transformed into the feature q_i' and then normalized as q_i'':

$$\begin{aligned} q_i' &= \log(q_i + 1) \\ q_i'' &= \frac{q_i' - \min(q_i')}{\max(q_i') - \min(q_i')} \end{aligned} \tag{17}$$

Next, an MLP g takes q_i'' as the input feature and output the weight of fusion $\{w_x, w_c, w_t, w_v\}$ for each target item. Finally, the prediction score $\hat{y}$ for each target item is

$$w_x \odot (\boldsymbol{o}_{\text{LLM}} \cdot \boldsymbol{E}_x^\top) + \sum_j w_j \odot (\boldsymbol{o}_{\text{LLM}} \cdot (\boldsymbol{W}_j \cdot (\boldsymbol{X} \cdot \mathtt{SG}(\boldsymbol{E}_j)) + b_j)^\top) \tag{18}$$

where $j \in \{c, t, v\}$. $\odot$ and $\cdot$ denotes hadamard product and dot product. $\boldsymbol{E}_x \in \mathbb{R}^{D_{\text{LLM}} \times |\mathcal{I}|}$ denotes a new embedding table for target item aiming at relieving the potential collapse issue in the target item embedding. We will justify its necessity in Sec. 5.3. Finally, the BCE loss is calculated to update the LLM using $\hat{y}$ and y. Notably, only a small proportion (*e.g.,* only about **0.19%** in our experiments) of all parameters are updated efficiently using LoRA.

4.4 Discussions

Our primary motivation is to address the embedding collapse and catastrophic forgetting issues in LLM4SR and further enhance the performance of LLM on the SR task. Most significantly, the proposed solution MME-SID has the potential to subvert the common while

suboptimal practice on using semantic IDs in generative retrieval or generative recommendation. It has the following advantages:

- MME-SID is able to generate a ranking list on the whole item set and output the most relevant top-k items flexibly. By contrast, existing methods like TIGER [35] can only retrieve the most relevant item in an autoregressive manner (*i.e.,* code by code).
- MME-SID does not need to tackle collision [35], an issue that multiple items are mapped into the same sequence of semantic IDs. This is because multimodal data can naturally discriminate between different items. By contrast, existing methods like TIGER require extra cost of computation and storage to ensure that each item is mapped into a unique sequence of semantic IDs.
- MME-SID achieves higher inference efficiency than existing methods, *e.g.,* TIGER. Suppose the token embedding dimension of LLM is D_{LLM} and there are N behavioral items interacted by a user. Each item is further encoded into a sequence of L semantic IDs. Therefore, TIGER needs to take a $D_{\text{LLM}} \times N \times L$-dimensional vector of {Behavioral Item Sequence} as input. By contrast, MME-SID only takes a $D_{\text{LLM}} \times N$-dimensional vector as input because each item is efficiently represented as a less collapsed, less forgetting, and more informative multimodal embedding, thus improving inference efficiency.

5 Experiments

We conduct extensive experiments on three public datasets and answer the following research questions:

- **RQ1:** What is the performance of the proposed MME-SID compared with baseline methods?
- **RQ2:** Do multimodal embeddings and semantic IDs contribute to alleviating embedding collapse?
- **RQ3:** What is the effect of MMD-based reconstruction loss?
- **RQ4:** Does using trained code embeddings for initialization mitigate catastrophic forgetting?

5.1 Experimental Settings

5.1.1 ***Datasets.*** We experiment on three categories of Amazon[2] 5-core dataset [31] including Beauty, Toys & Games, and Sports & Outdoors, in which each user and item has at least 5 interactions. Specifically, this dataset is crawled from Amazon, an e-commerce platform. The task is to predict whether a user will give a rating (ranging from 1 to 5) higher than 3 to the target item. The dataset statistics are summarized in Tab. 1, in which the sparsity metric denotes the proportion of negative samples with label $y = 0$. Meanwhile, denote N as the length of historical interactions of a user, the $(N-1)$-th and N-th item are treated as the target item in the training and test set, respectively.

5.1.2 ***Evaluation Metrics.*** To conduct evaluation, the top-k Hit Ratio (HR@k) and top-k normalized Discounted Cumulative Gain (nDCG@k) are adopted with k = 5, 10, and 20.

5.1.3 ***Baselines.*** We compare the proposed MME-SID with the following representative baseline methods and their inputs are formulated in Tab. 2. Notably, the first three methods only leverages item ID, *i.e.*, collaborative modality data while the last five baselines take multimodal data as input. For a fair comparison, Llama3-8B-instruct is adopted for all LLM-based methods and RQ-VAE is used to generate semantic IDs.

- **SASRec** [5] represents the original SASRec model using self-attention to model sequential pattern.
- **E4SRec** [10] adopts a linear projection of the pre-trained ID embeddings to tackle the out-of-range geneartion problem.
- **Multi Embedding (ME)** is a baseline we propose, which takes both the linear projection of pre-trained ID embedding $\boldsymbol{E}_c$ and a new set of randomly initialized ID embedding $\boldsymbol{E}'_c$ as input.
- **Concat** simply leverages a linear layer or MLP to map the pre-trained collaborative embedding to LLM token embedding space, then it is directly concatenated with token embedding. It is adopted in existing works like CoLLM [59] and LLaRA [13].
- **Concat&MLP** is a typical method of multimodal fusion [12, 53]. Specifically, the concatenation of collaborative, textual, and visual embedding of items is first fed into an MLP, whose output is then passed into LLM.
- **CTRL-MM** is adapted from CTRL [9]. It has the same input as Concat&MLP, while it explicitly aligns the collaborative embedding with textual and visual embedding using InfoNCE as the contrastive learning loss.
- **TIGER-MM** is a multimodal variant adapted from TIGER [35]. It only utilizes the semantic IDs of collaborative, textual, and visual embeddings to conduct generative retrieval. Specifically, it trains an RQ-VAE to generate semantic IDs for the embeddings in each modality separately.
- **MOTOR** [56] replaces the collaborative embedding with token embeddings of vision and text features, then adopts token cross

Table 1: The statistics of three categories of Amazon dataset: Beauty, Toys & Games, and Sports & Outdoors.

Category	Users	Items	Interactions	Sparsity
Beauty	22,332	12,086	198,215	99.93%
Toys & Games	19,121	11,757	165,221	99.93%
Sports & Outdoors	35,092	18,090	292,007	99.95%

Table 2: The formulation of {Behavioral Item Sequence} of baseline methods. X denotes one-hot vector of the historical interaction. E denotes embedding matrix. W and b denote the weight and bias of linear projection. SID^l denotes the semantic ID at the l-th codebook where $l = 1, \ldots, L$. The subscript c, t, and v denote collaborative, textual, and visual modality. The square bracket denotes the concatenation operation. SG denotes stop gradient operation.

Method	Input
SASRec	$\boldsymbol{X} \cdot \boldsymbol{E}_c$
E4SRec	$\boldsymbol{W}_c \cdot (\boldsymbol{X} \cdot \texttt{SG}(\boldsymbol{E}_c)) + b_c$
ME	$f_{\text{MLP}}(\left[\boldsymbol{W}_c \cdot (\boldsymbol{X} \cdot \texttt{SG}(\boldsymbol{E}_c)) + b_c, \boldsymbol{X} \cdot \boldsymbol{E}'_c\right])$
Concat	$\left[\boldsymbol{W}_j \cdot (\boldsymbol{X} \cdot \texttt{SG}(\boldsymbol{E}_j)) + b_j\right], j \in \{c, t, v\}$
Concat&MLP	$f_{\text{MLP}}(\left[\boldsymbol{W}_j \cdot (\boldsymbol{X} \cdot \texttt{SG}(\boldsymbol{E}_j)) + b_j\right]), j \in \{c, t, v\}$
CTRL-MM	$f_{\text{MLP}}(\left[\boldsymbol{W}_j \cdot (\boldsymbol{X} \cdot \texttt{SG}(\boldsymbol{E}_j)) + b_j\right]), j \in \{c, t, v\}$
TIGER-MM	$\left[SID^1_j, \ldots, SID^L_j\right], j \in \{c, t, v\}$
MOTOR	$\left[SID^1_j, \ldots, SID^L_j\right], j \in \{t, v\}$
LETTER	$\left[SID^1_j, \ldots, SID^L_j\right], j \in \{t\}$

[2]https://jmcauley.ucsd.edu/data/amazon/index_2014.html

Table 3: Overall performance comparison on Beauty, Toys & Games, and Sports & Outdoors dataset. Boldface denotes the highest value while underline indicates the second best result. 'Impr.' indicates our improvement against the second best baseline. ⋆ represents statistical significance with p-value < 0.05 in t-test compared with the best baseline.

Datasets	Metric	SASRec	E4SRec	ME	Concat	Concat&MLP	CTRL-MM	TIGER-MM	MOTOR	LETTER	Ours-full	Impr.
Beauty	HR@5	0.0368	0.0545	0.0567	0.0523	0.0581	0.0614	0.0471	0.0226	0.0415	**0.0675**⋆	9.93%
	HR@10	0.0578	0.0757	0.0787	0.0757	0.0830	0.0875	0.0668	0.0380	0.0654	**0.0955**⋆	9.14%
	HR@20	0.0903	0.1040	0.1046	0.1070	0.1177	0.1224	0.0945	0.0635	0.0833	**0.1342**⋆	9.64%
	nDCG@5	0.0243	0.0388	0.0402	0.0365	0.0404	0.0430	0.0329	0.0140	0.0262	**0.0475**⋆	10.47%
	nDCG@10	0.0310	0.0456	0.0473	0.0440	0.0484	0.0515	0.0393	0.0189	0.0351	**0.0566**⋆	9.90%
	nDCG@20	0.0392	0.0527	0.0538	0.0519	0.0571	0.0602	0.0463	0.0253	0.0408	**0.0663**⋆	10.13%
Toys & Games	HR@5	0.0508	0.0593	0.0598	0.0620	0.0623	0.0618	0.0486	0.0168	0.0471	**0.0653**⋆	4.82%
	HR@10	0.0713	0.0802	0.0827	0.0846	0.0871	0.0850	0.0667	0.0310	0.0650	**0.0909**⋆	4.36%
	HR@20	0.1022	0.1064	0.1120	0.1114	0.1184	0.1179	0.0889	0.0528	0.0852	**0.1223**⋆	3.29%
	nDCG@5	0.0357	0.0433	0.0435	0.0452	0.0444	0.0429	0.0354	0.0104	0.0343	**0.0472**⋆	4.42%
	nDCG@10	0.0422	0.0501	0.0509	0.0525	0.0524	0.0503	0.0412	0.0150	0.0399	**0.0555**⋆	5.71%
	nDCG@20	0.0500	0.0566	0.0582	0.0592	0.0602	0.0586	0.0468	0.0204	0.0449	**0.0634**⋆	5.32%
Sports & Outdoors	HR@5	0.0204	0.0316	0.0339	0.0287	0.0292	0.0270	0.0251	0.0154	0.0224	**0.0371**⋆	9.44%
	HR@10	0.0327	0.0456	0.0494	0.0431	0.0445	0.0424	0.0376	0.0253	0.0334	**0.0541**⋆	9.51%
	HR@20	0.0522	0.0650	0.0718	0.0658	0.0667	0.0652	0.0551	0.0426	0.0503	**0.0778**⋆	8.36%
	nDCG@5	0.0132	0.0218	0.0234	0.0191	0.0194	0.0181	0.0167	0.0100	0.0149	**0.0253**⋆	8.12%
	nDCG@10	0.0171	0.0263	0.0285	0.0237	0.0243	0.0230	0.0207	0.0131	0.0186	**0.0308**⋆	8.07%
	nDCG@20	0.0220	0.0312	0.0341	0.0294	0.0299	0.0287	0.0251	0.0174	0.0226	**0.0367**⋆	7.62%

network for interaction. Besides, we obtain the semantic IDs of visual and textual embeddings and adopt SASRec as the traditional downstream multimodal recommendation model.

- **LETTER** [43] adopts various regularization methods like diversity to achieve better item tokenization. We implement LETTER on TIGER as the backbone model of generative recommendation.

5.1.4 ***Implementation Details.*** For multimodal encoding, the product title and image are leveraged and the dimension of embeddings are $D_c = 64$ and $D_t = D_v = 1280$. Meanwhile, we adopt Gaussian kernel as $k(\boldsymbol{e}, \boldsymbol{e'}) = \exp(-\frac{\|\boldsymbol{e}-\boldsymbol{e'}\|^2}{2\sigma^2})$ which is characteristic. Besides, we adopt Llama3-8B-instruct ($D_{\text{LLM}} = 4096$) as recommender for better capability of following instructions compared with the original Llama3-8B. Besides, all experiments are conducted on A100 GPUs and the results shown are averaged over 3 runs. Detailed experimental settings are provided in Appendix. B.

5.2 Overall Performance (RQ1)

To answer **RQ1**, we compare the performance of MME-SID with different baseline methods in Sec. 5.1.3 and the overall performance is shown in Tab. 3. Specifically, we have the following observations.

Generally, the performance of LLM-based methods are superior to the methods adopting traditional SRS including SASRec and MOTOR, indicating the potential of LLM4SR. On the one hand, for the single-modal methods, E4SRec consistently surpasses SASRec and multi-embedding (ME) paradigm brings improvement on E4SRec by maintaining a new collaborative embedding table of item. However, the enhancement on Beauty and Toys & Games dataset is not significant and we speculate this is because ME only leverages data in collaborative modality, which can not bring much additional information gain. More analysis are conducted in Sec. 5.3.

On the other hand, for multimodal methods, we surprisingly find that even if the multimodal data is introduced, the widely adopted Concat, Concat&MLP, and CTRL-MM achieve worse performance than E4SRec, meaning that these methods utilize multimodal data in a suboptimal manner. Meanwhile, TIGER-MM, MOTOR, and LETTER achieve the worst accuracy among the multimodal methods comparably, which challenges the common approach that only utilizes semantic IDs to conduct generative retrieval [35, 40].

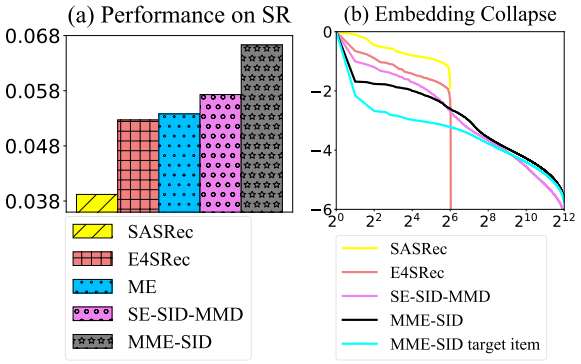

Figure 3: (a) Sequential recommendation performance where the y-axis is nDCG@20. (b) Embedding collapse Measurement. The x-axis is dimension index and y-axis is the logarithm of singular value (normalized by the maximum value) of embedding. They are both conducted on Beauty dataset.

By contrast, our proposed MME-SID achieves significant improvement on all three datasets and consistently surpasses all baseline methods, validating its efficacy. It beats the best performing baseline by 10.47%, 4.42%, and 8.12% on nDCG@5 on the three datasets. We will further investigate the reason for its superiority by analyzing its ability to tackle embedding collapse and catastrophic forgetting in the following sections.

5.3 Alleviating Embedding Collapse (RQ2)

To answer **RQ2**, we compare the following five methods: SASRec, E4SRec, ME, SE-SID-MMD, and MME-SID. Specifically, SE-SID-MMD takes $f_{\text{MLP}}(\left[\boldsymbol{W}_c \cdot (\boldsymbol{X} \cdot \texttt{SG}(\boldsymbol{E}_c)) + b_c, \sum_{l=1}^{L} \boldsymbol{E}_{SID_c^l}\right])$ as input, *i.e.,* only the collaborative modal. Notably, the SR performance is

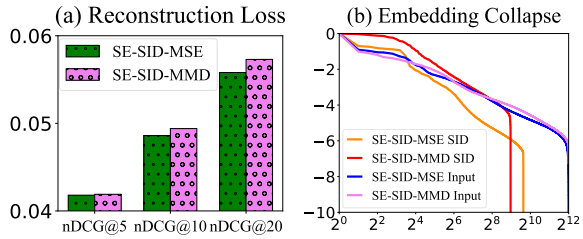

Figure 4: Comparison of MMD and MSE as the reconstruction loss on (a) sequential recommendation performance and (b) embedding collapse on Beauty dataset.

evaluated by nDCG@k. Embedding collapse is measured by the singular value of embedding table [4] in which a higher value indicates a lower degree of collapse. The results on Beauty dataset are shown in Fig. 3(a) and (b), in which 'MME-SID' and 'MME-SID target item' denote the input behavioral item embedding defined in Eq. 16 and target item embedding $\boldsymbol{E}_x$, respectively. It is clearly seen that, first, SASRec and E4SRec perform the worst and their 4096-dimensional embedding matrices drastically collapse after the 64−th dimension since $D_c = 64$. Second, our MME-SID obtains the best SR performance and it has the lowest degree of collapse from the 65-th to the last 4096-th dimension accounting for *over 98% of the dimensions of embedding matrix*. It indicates that introducing multimodal embeddings and semantic IDs effectively expanding the valid embedding space, thus enhancing model capacity. Third, the result on the target item of MME-SID shows that a new target item embedding table is necessary in alleviating embedding collapse.

To empirically analyze the effect of nonlinear mappings on singular value of embedding matrix, we take the most common activation function ReLU as an example and Llama3-8B-instruct as the LLM backbone. Specifically, compared with the model variant without ReLU, we found that 1) Embedding matrix rank is not significantly improved. 2) Recommendation accuracy degrades. 3) Catastrophic forgetting is still observed probably because the nonlinearity disrupts distance information in the original embedding.

> *Result 1. Solely relying on the pre-trained low-dimensional collaborative embeddings in LLM4SR leads to embedding collapse. By contrast, our proposed MME-SID alleviates this phenomenon and achieves better performance by adopting multimodal embeddings and semantic IDs.*

5.4 MMD-based Reconstruction Loss (RQ3)

To answer **RQ3**, we compare two model variants named 'SE-SID-MMD' and 'SE-SID-MSE'. Specifically, SE-SID-MMD trains an RQ-VAE with MMD as the reconstruction loss while SE-SID-MSE adopts an RQ-VAE with MSE reconstruction loss. Their performance of SR on Beauty dataset is shown in Fig. 4(a), suggesting that SE-SID-MMD performs better. Besides, similar to Sec. 3.2, to measure the forgetting in the input embedding of SE-SID-MMD, we first calculate the variable of Euclidean distance between each pair of behavioral-target item collaborative embedding. Next its τ between the distance variable from the pre-trained collaborative embedding $\boldsymbol{E}_c$ is calculated with a value of 0.4436. It is larger than $\tau = 0.3714$ of SE-SID-MSE, indicating less forgetting. Meanwhile, referring to the blue and violet line in Fig. 4(b), the input embedding of SE-SID-MSE

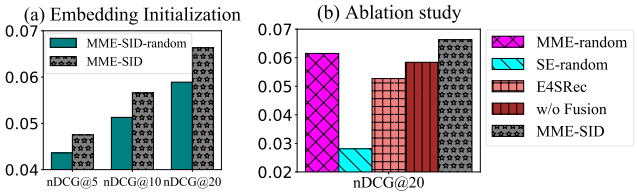

Figure 5: (a) Comparison of code embedding initialization where the y-axis denotes nDCG@k. (b) Ablation study on Beauty dataset where the y-axis denotes nDCG@20.

and SE-SID-MMD have comparable degree of collapse. Even if the semantic ID embeddings of SE-SID-MMD has a lower degree of collapse than that of SE-SID-MSE, the f_{MLP} only leverages information beneficial to SR. Therefore, it is more appropriate to ascribe the superiority of SE-SID-MMD to the mitigation of forgetting.

> *Result 2. Compared with Mean Squared Error as the reconstruction loss, the Maximum Mean Discrepancy reconstruction loss enables the quantized embedding to better preserve the information (*i.e.*, the partial order of behavioral-target item embedding distance), thus achieving better recommendation performance.*

5.5 Embedding Initialization (RQ4)

To answer **RQ4**, we compare MME-SID with 'MME-SID-random' which randomly initializes the embeddings of semantic IDs. Fig. 5(a) shows that MME-SID-random performs worse. Besides, we calculate the Euclidean distance of each behavioral-target item collaborative embedding after fine-tuning in MME-SID-random. Its τ between the distance variable of $\boldsymbol{E}_c$ is 0.0508, indicating catastrophic forgetting. By contrast, our MME-SID achieves $\tau = 0.2727$ after fine-tuning, which demonstrates a significant relief on forgetting.

> *Result 3. Simply discarding the pre-trained code embeddings and randomly initialize them on the downstream tasks would lead to catastrophic forgetting. By contrast, our proposed MME-SID mitigates this phenomenon by initializing with the trained code embeddings, thus preserving the distance information.*

5.6 Ablation Study

We conduct ablation study on Beauty dataset and the results are depicted in Fig. 5(b). Specifically, to demonstrate that the improvement in the performance of MME-SID does not simply stem from an increase in the number of input parameters, we experiment on the modal variant 'MME-random', which has the same number of parameters in input as our MME-SID. Specifically, it replaces the quantized embedding with a new embedding table with randomly initialization of each modality while it is inferior to MME-SID because it can not leverage the intra- and inter-modal correlation from our proposed MM-RQ-VAE. Besides, we also experiment on the single-modal model variant 'SE-random' whose input is only the randomly initialized item ID embeddings. It has the same number of input parameters as E4SRec but achieves a pretty worse performance due to forgetting. Finally, we experiment on the model variant 'w/o Fusion' which removes the multimodal frequency-aware fusion module. The performance decrease indicates the significance of the multimodal frequency-aware fusion module.

6 Related Work

We summarize the related works on semantic IDs, multimodal recommendation, and large language model for recommendation.

6.1 Semantic IDs for Recommendation

Semantic IDs denotes a sequence of tokens to represent users or items in recommendation. Existing works can be categorized into two groups. First, generative models like LLMs use semantic IDs to conduct recommendation or retrieval task in a generative manner [35, 40, 43, 66]. For example, TIGER [35] proposes to use the content information of item to generate sequence of semantic tokens, which are further adopted to train the transformer model on the SR task. Nonetheless, they mainly discard the trained code embeddings which are randomly initialize on the downstream tasks, leading to catastrophic forgetting.

Besides, some works treat semantic ID as auxiliary information to enhance the performance of traditional RS [30, 56, 58]. QARM [30] adopts both vector quantization and residual quantization to generate quantitative codes as new features of downstream recommendation model. However, their improvement achieved is usually limited due to the constraints imposed by the traditional model structure. By contrast, our proposed MME-SID unleashes the power of LLMs by adopting multimodal embeddings and the trained quantized embeddings, thus achieving significant improvement on SR.

6.2 LLM4Rec & Multimodal Recommendation

The early works on large language model for recommendation (LLM4Rec) like TALLRec [2] merely formulate the recommendation task in the natural language format and tune the LLM. Afterward, some works focus on leveraging additional data in different modalities. For example, to integrate textual and collaborative semantic, LC-Rec [66] first conduct item indexing and proposes different semantic alignment tasks. Nevertheless, they suffer from high inference latency of auto-regressive generation. By contrast, our MME-SID only leverages multimodal embeddings of items and directly calculate the score between output embedding and target item embeddings, which leads to high efficiency.

7 Conclusion

In this paper, we first identify the embedding collapse and catastrophic forgetting issues in the existing works on large language model for sequential recommendation. To tackle them, we propose a novel MME-SID framework by leveraging both multimodal embeddings and semantic IDs whose embeddings are initialized with the trained code embeddings. To better preserve distance information and learn inter-modal connections, we propose a multimodal Residual Quantized Variational Autoencoder (MM-RQ-VAE) using maximum mean discrepancy as the reconstruction loss and a contrastive learning objective. Extensive experiments on three public datasets of Amazon validate the efficacy of the proposed method.

A Pseudo-code

The procedure of MME-SID is shown in Alg. 1, which consists of two stages: (1) Encoding stage including Multimodal Embedding Encoding step (from Line 1 to 2) and Multimodal Embedding Quantization step (from Line 3 to 10); (2) Fine-tuning Stage.

Algorithm 1: Procedure of MME-SID

Input: User set $\mathcal{U}$; item set $\mathcal{I}$; historical interaction sequence $\{h_u\}$, target item x_u, and true label y_u;
Output: A trained LLM as sequential recommender system.
Stage 1: Encoding
1 Obtain the collaborative embedding from a pre-trained conventional sequential recommender system;
2 Obtain the textual and visual embedding using LLM2CLIP;
3 **while** *not converge* **do**
4 Sample a mini-batch data from $\mathcal{I}$;
5 Calculate the reconstruction loss $\mathcal{L}_{\text{Recon}}$;
6 Calculate the alignment loss $\mathcal{L}_{\text{Align}}$;
7 Calculate the RQ-VAE loss $\mathcal{L}_{\text{RQ-VAE}}$;
8 Take the gradient and update MM-RQ-VAE;
9 **end**
10 Obtain multimodal semantic IDs and code embeddings;
Stage 2: Fine-tuning
11 **while** *not converge* **do**
12 Sample a mini-batch data from $\mathcal{U}$;
13 Retrieve the multimodal embeddings and semantic IDs;
14 Obtain the last hidden state of LLM output;
15 Calculate the fusion weight of target item;
16 Calculate the prediction score of multimodal fusion;
17 Calculate the BCE loss;
18 Take the gradient and update LLM using LoRA;
19 **end**

Table 4: The hyper-parameter settings of experiments.

Dataset	Beauty	Toys & Games	Sports & Outdoors
Training epochs	3	3	2
Learning rate	3e-4	2e-4	2e-4
Batch size	16	16	16
LoRA rank	8	8	8
LoRA alpha	16	16	16
LoRA dropout	0.05	0.05	0.05
Warm-up steps	100	100	200
Number of codes	256	256	300
Level of codebooks	4	4	4

B Experimental Settings

For the data processing, we remove the items lacking title or image in the original dataset. For implementation, AdamW [29] optimizer is adopted and the hyper-parameters are shown in Tab. 4. we set $\alpha = 1$, $\beta =$ 1e-3, and $\gamma = 1$. Besides, the target modules of LoRA are [gate_proj, down_proj, up_proj]. Finally, only about 0.19% of all parameters are updated in our experiments.

ACKNOWLEDGEMENT

This research was partially supported by Hong Kong Research Grants Council's Research Impact Fund (No.R1015-23), Collaborative Research Fund (No.C1043-24GF), General Research Fund (No.11218325), Institute of Digital Medicine of City University of Hong Kong (No.9229503), Tencent (CCF-Tencent Open Fund, Tencent Rhino-Bird Focused Research Program), and National Natural Science Foundation of China (No.62502404).

GenAI Usage Disclosure

No GenAI tools were used in any stage of the research and writing.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 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*. 1007–1014.

[3] Jingtong Gao, Xiangyu Zhao, Muyang Li, Minghao Zhao, Runze Wu, Ruocheng Guo, Yiding Liu, and Dawei Yin. 2024. Smlp4rec: An efficient all-mlp architecture for sequential recommendations. *ACM Transactions on Information Systems* 42, 3 (2024), 1–23.

[4] Xingzhuo Guo, Junwei Pan, Ximei Wang, Baixu Chen, Jie Jiang, and Mingsheng Long. 2024. On the Embedding Collapse when Scaling up Recommendation Models. In *Proceedings of the 41st International Conference on Machine Learning*.

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

[6] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1-2 (1938), 81–93.

[7] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11523–11532.

[8] Chengxi Li, Yejing Wang, Qidong Liu, Xiangyu Zhao, Wanyu Wang, Yiqi Wang, Lixin Zou, Wenqi Fan, and Qing Li. 2023. STRec: Sparse transformer for sequential recommendations. In *Proceedings of the 17th ACM conference on recommender systems*. 101–111.

[9] Xiangyang Li, Bo Chen, Lu Hou, and Ruiming Tang. 2023. CTRL: Connect Collaborative and Language Model for CTR Prediction. *ACM Transactions on Recommender Systems* (2023).

[10] Xinhang Li, Chong Chen, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. 2023. E4srec: An elegant effective efficient extensible solution of large language models for sequential recommendation. *arXiv preprint arXiv:2312.02443* (2023).

[11] Xiaopeng Li, Fan Yan, Xiangyu Zhao, Yichao Wang, Bo Chen, Huifeng Guo, and Ruiming Tang. 2023. Hamur: Hyper adapter for multi-domain recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1268–1277.

[12] Youhua Li, Hanwen Du, Yongxin Ni, Yuanqi He, Junchen Fu, Xiangyan Liu, and Qi Guo. 2024. An Empirical Study of Training ID-Agnostic Multi-modal Sequential Recommenders. *arXiv preprint arXiv:2403.17372* (2024).

[13] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1785–1795.

[14] Weilin Lin, Xiangyu Zhao, Yejing Wang, Yuanshao Zhu, and Wanyu Wang. 2023. Autodenoise: Automatic data instance denoising for recommendations. In *Proceedings of the ACM Web Conference 2023*. 1003–1011.

[15] Zhutian Lin, Junwei Pan, Haibin Yu, Xi Xiao, Ximei Wang, Zhixiang Feng, Shifeng Wen, Shudong Huang, Lei Xiao, and Jie Jiang. 2024. Disentangled Representation with Cross Experts Covariance Loss for Multi-Domain Recommendation. *arXiv preprint arXiv:2405.12706* (2024).

[16] Langming Liu, Liu Cai, Chi Zhang, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Yifu Lv, Wenqi Fan, Yiqi Wang, Ming He, et al. 2023. Linrec: Linear attention mechanism for long-term sequential recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 289–299.

[17] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal recommender systems: A survey. *Comput. Surveys* 57, 2 (2024), 1–17.

[18] Qidong Liu, Xian Wu, Wanyu Wang, Yejing Wang, Yuanshao Zhu, Xiangyu Zhao, Feng Tian, and Yefeng Zheng. 2025. Llmemb: Large language model can be a good embedding generator for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12183–12191.

[19] Qidong Liu, Xian Wu, Yejing Wang, Zijian Zhang, Feng Tian, Yefeng Zheng, and Xiangyu Zhao. 2024. Llm-esr: Large language models enhancement for long-tailed sequential recommendation. *Advances in Neural Information Processing Systems* 37 (2024), 26701–26727.

[20] Qidong Liu, Xian Wu, Xiangyu Zhao, Yejing Wang, Zijian Zhang, Feng Tian, and Yefeng Zheng. 2024. Large language models enhanced sequential recommendation for long-tail user and item. *arXiv e-prints* (2024), arXiv–2405.

[21] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Zijian Zhang, Feng Tian, and Yefeng Zheng. 2024. Large language model distilling medication recommendation model. *arXiv preprint arXiv:2402.02803* (2024).

[22] Qidong Liu, Xiangyu Zhao, Yejing Wang, Zijian Zhang, Howard Zhong, Chong Chen, Xiang Li, Wei Huang, and Feng Tian. 2025. Bridge the Domains: Large Language Models Enhanced Cross-domain Sequential Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1582–1592.

[23] Shuchang Liu, Qingpeng Cai, Bowen Sun, Yuhao Wang, Ji Jiang, Dong Zheng, Peng Jiang, Kun Gai, Xiangyu Zhao, and Yongfeng Zhang. 2023. Exploration and regularization of the latent action space in recommendation. In *Proceedings of the ACM Web Conference 2023*. 833–844.

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

[25] Ziwei Liu, Qidong Liu, Yejing Wang, Wanyu Wang, Pengyue Jia, Maolin Wang, Zitao Liu, Yi Chang, and Xiangyu Zhao. 2024. Bidirectional gated mamba for sequential recommendation. *arXiv e-prints* (2024), arXiv–2408.

[26] Ziwei Liu, Qidong Liu, Yejing Wang, Wanyu Wang, Pengyue Jia, Maolin Wang, Zitao Liu, Yi Chang, and Xiangyu Zhao. 2025. SIGMA: Selective Gated Mamba for Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12264–12272.

[27] Ziru Liu, Shuchang Liu, Zijian Zhang, Qingpeng Cai, Xiangyu Zhao, Kesen Zhao, Lantao Hu, Peng Jiang, and Kun Gai. 2024. Sequential recommendation for optimizing both immediate feedback and long-term retention. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1872–1882.

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.

[29] Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101* (2017).

[30] Xinchen Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, et al. 2024. QARM: Quantitative Alignment Multi-Modal Recommendation at Kuaishou. *arXiv preprint arXiv:2411.11739* (2024).

[31] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[32] Junwei Pan, Wei Xue, Ximei Wang, Haibin Yu, Xun Liu, Shijie Quan, Xueming Qiu, Dapeng Liu, Lei Xiao, and Jie Jiang. 2024. Ads recommendation in a collapsed and entangled world. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5566–5577.

[33] Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*. 11–18.

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

[35] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 10299–10315.

[36] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The annals of statistics* (2013), 2263–2291.

[37] Anirvan M Sengupta and Partha P Mitra. 1999. Distributions of singular values for some random matrices. *Physical Review E* 60, 3 (1999), 3389.

[38] Jianhong Shen. 2001. On the singular values of Gaussian random matrices. *Linear Algebra Appl.* 326, 1-3 (2001), 1–14.

[39] Liangcai Su, Junwei Pan, Ximei Wang, Xi Xiao, Shijie Quan, Xihua Chen, and Jie Jiang. 2024. STEM: Unleashing the Power of Embeddings for Multi-task Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 9002–9010.

[40] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2024. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems* 36 (2024).

[41] Hanbing Wang, Xiaorui Liu, Wenqi Fan, Xiangyu Zhao, Venkataramana Kini, Devendra Yadav, Fei Wang, Zhen Wen, Jiliang Tang, and Hui Liu. 2024. Rethinking large language model architectures for sequential recommendations. *arXiv preprint arXiv:2402.09543* (2024).

[42] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442* (2022).

[43] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2400–2409.

[44] Yuhao Wang. 2024. Multi-Granularity Modeling in Recommendation: from the Multi-Scenario Perspective. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5491–5494.

[45] Yejing Wang, Zhaocheng Du, Xiangyu Zhao, Bo Chen, Huifeng Guo, Ruiming Tang, and Zhenhua Dong. 2023. Single-shot feature selection for multi-task recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 341–351.

[46] Yuhao Wang, Ha Tsz Lam, Yi Wong, Ziru Liu, Xiangyu Zhao, Yichao Wang, Bo Chen, Huifeng Guo, and Ruiming Tang. 2023. Multi-task deep recommender systems: A survey. *arXiv preprint arXiv:2302.03525* (2023).

[47] Yuhao Wang, Ziru Liu, Yichao Wang, Xiangyu Zhao, Bo Chen, Huifeng Guo, and Ruiming Tang. 2024. Diff-MSR: A diffusion model enhanced paradigm for cold-start multi-scenario recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 779–787.

[48] Yuhao Wang, Junwei Pan, Pengyue Jia, Wanyu Wang, Maolin Wang, Zhixiang Feng, Xiaotian Li, Jie Jiang, and Xiangyu Zhao. 2025. Pre-train, Align, and Disentangle: Empowering Sequential Recommendation with Large Language Models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1455–1465.

[49] Yidan Wang, Zhaochun Ren, Weiwei Sun, Jiyuan Yang, Zhixiang Liang, Xin Chen, Ruobing Xie, Su Yan, Xu Zhang, Pengjie Ren, et al. 2024. Content-Based Collaborative Generation for Recommender Systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2420–2430.

[50] Yuhao Wang, Yichao Wang, Zichuan Fu, Xiangyang Li, Wanyu Wang, Yuyang Ye, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024. LLM4MSR: An LLM-Enhanced Paradigm for Multi-Scenario Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2472–2481.

[51] Yuhao Wang, Xiangyu Zhao, Bo Chen, Qidong Liu, Huifeng Guo, Huanshuo Liu, Yichao Wang, Rui Zhang, and Ruiming Tang. 2023. PLATE: A prompt-enhanced paradigm for multi-scenario recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1498–1507.

[52] Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Chunyu Wang, Liang Hu, Xiyang Dai, Dongdong Chen, Chong Luo, Lili Qiu, et al. 2024. LLM2CLIP: Powerful Language Model Unlock Richer Visual Representation. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*.

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

[54] Chi Zhang, Yantong Du, Xiangyu Zhao, Qilong Han, Rui Chen, and Li Li. 2022. Hierarchical item inconsistency signal learning for sequence denoising in sequential recommendation. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 2508–2518.

[55] Chi Zhang, Qilong Han, Rui Chen, Xiangyu Zhao, Peng Tang, and Hongtao Song. 2024. Ssdrec: Self-augmented sequence denoising for sequential recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 803–815.

[56] Kangning Zhang, Jiarui Jin, Yingjie Qin, Ruilong Su, Jianghao Lin, Yong Yu, and Weinan Zhang. 2024. Learning ID-free Item Representation with Token Crossing for Multimodal Recommendation. *arXiv preprint arXiv:2410.19276* (2024).

[57] Sheng Zhang, Maolin Wang, Wanyu Wang, Jingtong Gao, Xiangyu Zhao, Yu Yang, Xuetao Wei, Zitao Liu, and Tong Xu. 2025. Glint-ru: Gated lightweight intelligent recurrent units for sequential recommender systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 1948–1959.

[58] Taolin Zhang, Junwei Pan, Jinpeng Wang, Yaohua Zha, Tao Dai, Bin Chen, Ruisheng Luo, Xiaoxiang Deng, Yuan Wang, Ming Yue, et al. 2024. Towards Scalable Semantic Representation for Recommendation. *arXiv preprint arXiv:2410.09560* (2024).

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

[60] Zijian Zhang, Shuchang Liu, Jiaao Yu, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Ziru Liu, Qidong Liu, Hongwei Zhao, Lantao Hu, et al. 2024. M3oe: Multi-domain multi-task mixture-of experts recommendation framework. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 893–902.

[61] Kesen Zhao, Shuchang Liu, Qingpeng Cai, Xiangyu Zhao, Ziru Liu, Dong Zheng, Peng Jiang, and Kun Gai. 2023. KuaiSim: A comprehensive simulator for recommender systems. *Advances in Neural Information Processing Systems* 36 (2023), 44880–44897.

[62] Kesen Zhao, Xiangyu Zhao, Zijian Zhang, and Muyang Li. 2022. Mae4rec: Storage-saving transformer for sequential recommendations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2681–2690.

[63] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. 2018. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 95–103.

[64] Xiangyu Zhao, Long Xia, Lixin Zou, Hui Liu, Dawei Yin, and Jiliang Tang. 2020. Whole-chain recommendations. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1883–1891.

[65] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. 2018. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1040–1048.

[66] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 1435–1448.