ID-Guided Multimodal Experts with Contrastive Diffusion for Sequential Recommendation

Anonymous

***Abstract*—Multimodal sequential recommendation has emerged as a promising direction to enrich user-item interaction representations by incorporating diverse modalities such as text and images. However, existing methods often overlook the inherent inconsistencies between different modalities and fail to effectively filter redundant noise within modality-specific features, leading to suboptimal recommendation performance. To address these issues, we propose a novel framework named ID-Guided Multimodal Experts with Contrastive Diffusion for Sequential Recommendation (IMECD). Specifically, IMECD introduces a novel ID-guided multimodal mixture of experts module, which uniquely leverages long-term user preferences encoded in ID embeddings to dynamically guide the extraction of text and image features. This module helps resolve cross-modal semantic inconsistency and suppresses irrelevant signals, thereby improving the quality of multimodal representations. To further mitigate noise in user interaction sequences, we introduce a modality-specific vector quantization module that denoises sequential features by independently quantizing each modality. Moreover, we propose a contrastive diffusion generation module, which conditions the diffusion process on sequence representations and employs a contrastive loss to alleviate generation bias. Extensive experiments on four benchmark datasets demonstrate that IMECD consistently outperforms state-of-the-art baselines. Our code is available at https://anonymous.4open.science/r/IMECD-LYH.**

***Index Terms*—mixture of experts, diffusion model, sequential recommendation, multimodal.**

I. Introduction

Nowadays, recommender system plays an important role of discovering preferred items from vast inventory and presenting them to potential users [1], [2]. Sequential recommendation (SR) aims to predict the next item a user will interact with based on their historical behavior sequence. Traditional sequential approaches [3]–[5] predominantly rely on ID to model interaction sequences. While ID-based methods benefit from simplicity and scalability, they often suffer from limitations in cold-start scenarios [6] and ignore the abundant multimodal semantic item descriptions [7].

To overcome the limitations of ID-only models, multimodal sequential recommendation has emerged as a promising direction. A number of studies [8]–[10] have integrated text and image modalities into sequence modeling to learn effective sequential representations. Recently, several studies [11], [12] have attempted to explore adaptive fusion mechanisms to achieve more flexible multimodal modeling. By leveraging multimodal semantic signals, existing methods have shown notable improvements in recommendation performance, enabling fine-grained item understanding and more accurate user interest modeling.

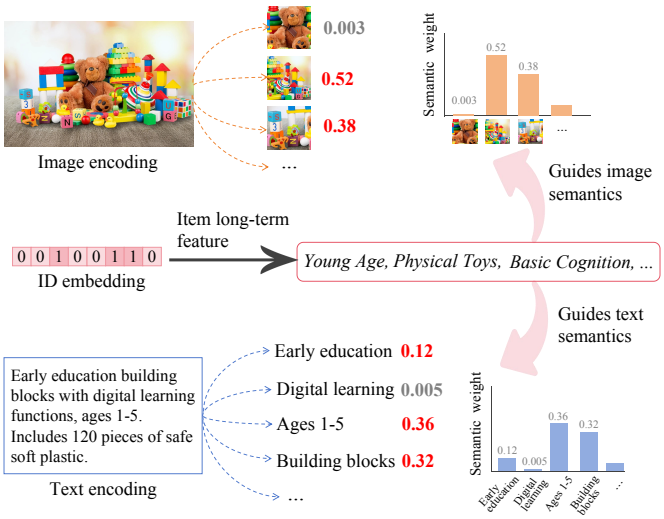

Fig. 1. Challenges of Multimodal Features in Sequential Recommendation.

Despite the progress in multimodal sequential recommendation, existing methods often assume that different modalities like product descriptions and images provide complementary and semantically aligned information suitable for direct fusion to enhance item representations [8], [10], [13]. However, this assumption frequently breaks down in practical e-commerce scenarios. On many commercial platforms, visual and textual content for the same product is often generated independently and optimized for distinct purposes. Product images are typically curated to emphasize aesthetic appeal or lifestyle context, while text descriptions focus on keyword optimization for search engines or highlight product functionalities. This discrepancy often leads to misaligned semantics across modalities. As illustrated in Figure 1, consider a children's building block toy, where the image modality depicts a teddy bear and colorful blocks, while the text modality mentions terms such as "Digital learning" and "Early education". This clearly introduces semantic inconsistency between the modalities. Additionally, noise may be present in each modality, such as irrelevant or misleading image features or imprecise textual descriptions. This disparity and noise in the information conveyed by the two modalities can lead to misalignment, which may confuse the recommendation model if these modalities are simply fused without considering their differences. This motivates the need for a more refined approach that does not treat all modalities equally, but instead dynamically assesses the alignment and relevance of each modality to the user's

evolving preferences.

To address this issue, we propose using item ID embeddings to extract features from both text and image representations that align with the user's preferences. Item ID embeddings inherently represent the long-term, stable characteristics of an item in the latent space, while text and image modalities capture more transient, short-term information. As shown in Figure 1, the item ID embedding can provide long-term preference features such as *young age, physical toys*, and *basic cognition*. Based on these long-term preference signals, we assign different importance to various features of the text and image modalities, guiding them to reflect the user's enduring interests. This approach effectively filters out noise and cross-modal semantic inconsistencies, ensuring that the generated multimodal representations are more aligned with the user's long-term preferences.

Based on this, we propose a novel framework named ID-Guided Multimodal Experts with Contrastive Diffusion for Sequential Recommendation (IMECD). Firstly, we propose the ID-Guided Multimodal Mixture of Experts (ID-MoE) module, which leverages ID embeddings as gating signals to selectively enhance the text and image representations. This design enables the model to extract item representations that align with long-term user preferences, filtering out noisy or inconsistent signals from the text and image modalities. Additionally, to further denoise the user's interaction sequences, we introduce the Modality-Specific Vector Quantization (MS-VQ) module. This module performs denoising on multimodal sequence representations by independently maintaining learnable codebooks for each modality. This approach effectively reduces noise in the sequences, ensuring cleaner representations that improve sequential predictions. Lastly, to model the temporal dynamics of user interests, we introduce a Contrastive Diffusion Generation (CDG) module. Inspired by recent advances in diffusion models (DMs), which have shown promising generative capabilities in sequential modeling tasks, DMs perform denoising over latent user trajectories to generate the next item representation. However, diffusion-based generation may suffer from mode collapse or bias due to data imbalance. To address this, we incorporate a contrastive learning objective that explicitly maximizes the representation distance between the next items from different sequences, thereby encouraging more diverse and accurate prediction. To the best of our knowledge, this work represents an early exploration of applying diffusion models to multimodal sequential recommendation. The contributions of this paper are concluded as follows:

1) We propose a novel ID-Guided Multimodal Mixture of Experts module, where ID embeddings are used as gating signals to selectively extract long-term user preferences information from other modal inputs. This approach alleviates inconsistencies between modalities and reduces noise, ensuring that the extracted features better reflect the user's long-term preferences.
2) We propose a method called ID-Guided Multimodal Experts with Contrastive Diffusion for Sequential Recommendation (IMECD). The ID-MoE module reduces inconsistencies and noise in item multimodal features, while the MS-VQ module addresses noise in the user's interaction sequences. Additionally, we use a contrastive loss to mitigate bias in the diffusion generation process.
3) We conduct comprehensive experiments on the Amazon review dataset to validate the effectiveness and adaptability of the IMECD model. The experimental results demonstrate that our model significantly outperforms existing methods.

The rest of our paper is organized as follows. We briefly review the related work in Section II. In Section III, the proposed IMECD is described in detail. The experimental results are reported in Section IV. At last, we draw the conclusion in Section V.

II. Related work

A. Traditional Sequential Recommendation

In the early stages of sequential recommendation, a number of methods rely on markov chains [14], [15], which assume that the next item depends only on the previous item in the sequence. With the advent of deep learning, Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and self-attention are widely adopted to model sequential behaviors, allowing for better representation of complex user interests over time. For instance, GRU4Rec [16], Caser [17], and SASRec [3] utilize RNN, CNN, and self-attention structures as their basic encoders, respectively. Besides, interest modeling [18]–[20] has been another popular methodology for SR, where user interests were usually implemented by attention or clustering. Despite their success, these traditional methods primarily focus on modeling interaction histories based on ID embeddings, often ignoring the rich multimodal content associated with items, which could provide valuable additional signals for recommendation.

B. Multimodal Sequential Recommendation

The fusion of multimodal information in sequential recommendation has gained increasing attention. Recent studies in multimodal sequential recommendation have focused on integrating various modalities. MV-RNN [9] employs several fusion strategies, including addition, concatenation, and reconstruction, to combine different modal data for SR. UniSRec [8] introduces a MoE framework that facilitates the transfer of semantic information from text representations into the ID embeddings. MMMLP [10] implements multimodal sequential recommendation via an MLP architecture, enabling feature fusion and linear-complexity prediction. Building on these ideas, many subsequent works have designed adaptive fusion modules to improve the effectiveness of multimodal integration. MISSRec [11] proposes a lightweight fusion mechanism that dynamically adjusts user attention across modalities. MMSR [12] incorporates heterogeneous GNNs for adaptive fusion, allowing the model to flexibly exploit the relationships between different modalities. HM4SR [7] develops a two-level hierarchical MoE to integrate explicit temporal information into multimodal learning. However, these approaches have

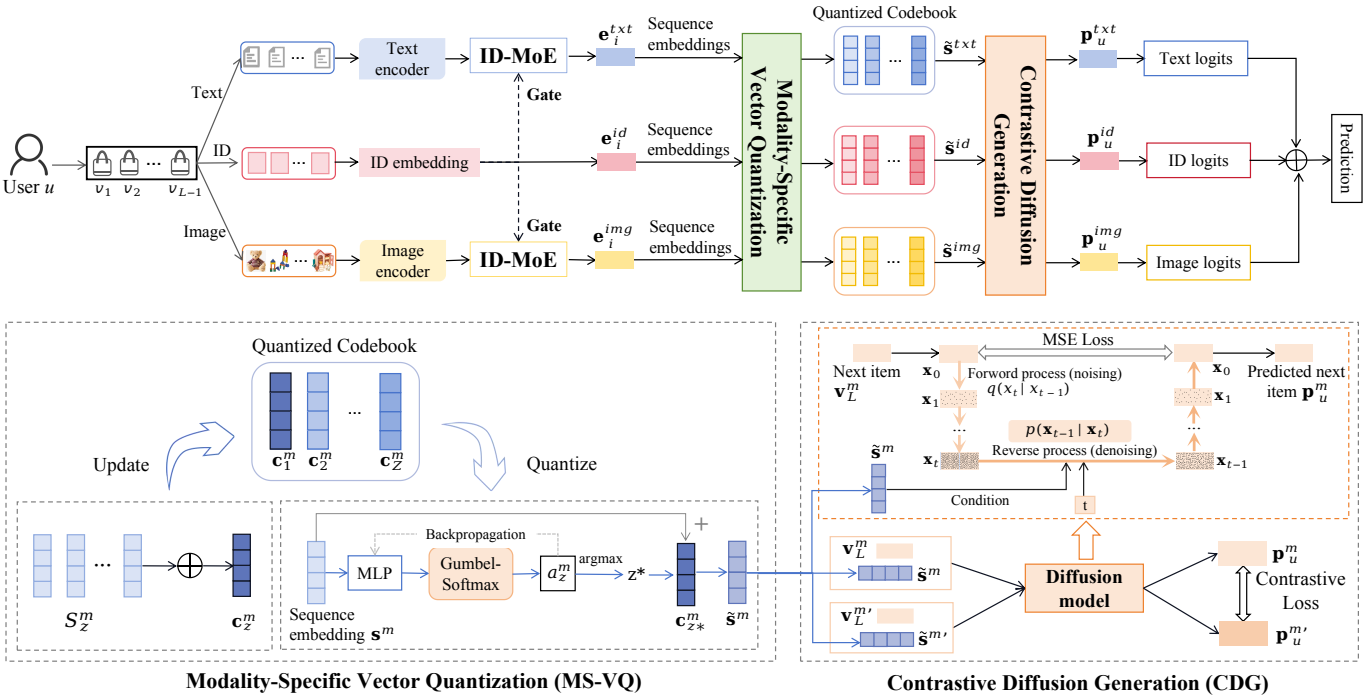

Fig. 2. The overall framework of the proposed IMECD.

limited capability in capturing items' latent aspects and users' diverse preferences [21]. Besides, they often overlook the challenges posed by semantic inconsistency between modalities.

C. Diffusion Models in Sequential Recommendation

Diffusion models have made remarkable success in computer vision, natural language processing, and many other fields [22]–[24]. Nowadays, DMs have recently been explored in sequential recommendation. By corrupting the next-item representation with Gaussian noise and gradually denoising it under the guidance of historical sequences, DMs effectively model complex user preferences and latent item semantics [25]. DiffuRec [21] and DreamRec [26] utilize DMs to model item distribution, generating the next item through a denoising process guided by interaction sequences. DiffRec [27] generates globally similar but personalized collaborative signals during the denoising process. Some other works use DMs to improve recommendation performance through data augmentation. DiffuASR [28] designs a novel data augmentation framework based on DM for sequential recommendation. CaDiRec [29] leverages DMs for context-consistent item replacement, enabling high-fidelity view augmentation. DiffKG [30] integrates DMs with a data augmentation paradigm, enabling robust knowledge graph representation learning. Due to the advancements of DMs in SR, we believe that integrating them into multimodal sequential recommendation is a promising direction.

III. Method

Figure 2 illustrates the proposed IMECD framework. The model consists of three key components: the ID-Guided Multimodal Mixture of Experts (ID-MoE) module, the Modality-Specific Vector Quantization (MS-VQ) module, and the Contrastive Diffusion Generation (CDG) module. The ID-MoE module utilizes the item ID embeddings as gating signals to selectively extract preference features from both the text and image modalities. And then, the MS-VQ module maintains a codebook for each modality, and performs sequence denoising through vector quantization. Finally, the CDG module performs denoising of the next item representation under the guidance of the sequence representation. To mitigate the biases introduced by the DM, we maximize the difference between the predicted next item representations from different sequences using a contrastive loss.

A. Problem Formulation

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively. $s_u = [v_1, v_2, \cdots, v_{L-1}]$ denote the interaction sequence for user $u \in \mathcal{U}$, v_L be the ground-truth next item that the user will interact with, where $v_k \in \mathcal{I}$ is the k-th interaction in the chronological sequence. Given historical interaction sequence s_u, sequential recommendation systems predict the next item v_L that optimally matches user preferences.

B. MultiModal Feature Encoding

We obtain item initial representations from three modalities in real-world recommendation scenarios: ID, text, and image. For text and image, we employ pre-trained models to extract rich representations. Specifically, for item $i \in \mathcal{I}$, we utilize a pre-trained RoBERTa [31] model to obtain initial textual representations $\bar{\mathbf{e}}_i^{txt} \in \mathbb{R}^{d_{txt}}$ of the item's descriptive content, and a pre-trained Vision Transformer (ViT) [32] model to derive

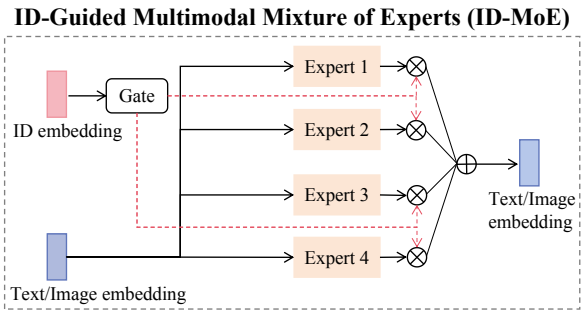

Fig. 3. The detailed implementation of ID-MoE module.

initial visual representations $\bar{\mathbf{e}}_i^{img} \in \mathbb{R}^{d_{img}}$ from the item's image. As for ID embedding, we initialize it as $\mathbf{e}_i^{id} \in \mathbb{R}^d$. Moreover, we utilize two linear projection layers to align the dimensionalities of both text and image embeddings with the ID embedding space:

$$\mathbf{e}_i^{txt} = \mathbf{W}_{txt}\bar{\mathbf{e}}_i^{txt} + \mathbf{b}_{txt}, \tag{1}$$

$$\mathbf{e}_i^{img} = \mathbf{W}_{img}\bar{\mathbf{e}}_i^{img} + \mathbf{b}_{img}, \tag{2}$$

where $\mathbf{W}_{txt} \in \mathbb{R}^{d_{txt}\times d}$, $\mathbf{W}_{img} \in \mathbb{R}^{d_{img}\times d}$, $\mathbf{b}_{txt} \in \mathbb{R}^d$ and $\mathbf{b}_{img} \in \mathbb{R}^d$ are trainable parameters. In this way, the representations of text and image modalities $\mathbf{e}_i^{txt} \in \mathbb{R}^d$ and $\mathbf{e}_i^{img} \in \mathbb{R}^d$ can be aligned to the ID embedding space.

C. ID-Guided Multimodal Mixture of Experts

Multimodal sequential recommendation methods often struggle with several inherent challenges, including semantic inconsistency across modalities, insufficient alignment with user preferences, and the introduction of noisy or irrelevant information during multimodal fusion. To address these limitations, we propose an ID-Guided Multimodal Mixture of Experts (ID-MoE) module, which extracts modality-specific item representations that are more closely aligned with user behavior patterns. As illustrated in Figure 3, the ID-MoE architecture consists of multiple expert networks for different modalities (e.g., text and image), each responsible for encoding modality-specific semantic information. A gating network is then used to adaptively weight and combine expert outputs. Unlike previous works [7], [8] where gating decisions are derived from the input modality itself, we introduce item ID embeddings as an external, stable signal to guide the gating process. ID embeddings capture long-term user-item interaction semantics and serve as preference indicators, enabling the gating network to suppress noisy features and select information that better aligns with user interests.

Specifically, given the text and image input representation $\mathbf{e}_i^{txt}$ and $\mathbf{e}_i^{img}$ of item i, we design an MoE to adaptively select expert outputs under the guidance of ID embeddings $\mathbf{e}_i^{id}$. We first compute the gating router vector $\mathbf{g}^{txt}$ and $\mathbf{g}^{img}$ over N experts using the ID embedding:

$$\mathbf{g}^{txt} = \text{Softmax}\left(\mathbf{W}_1\mathbf{e}_i^{id} + \mathbf{b}_1\right), \tag{3}$$

$$\mathbf{g}^{img} = \text{Softmax}\left(\mathbf{W}_2\mathbf{e}_i^{id} + \mathbf{b}_2\right), \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times N}, \mathbf{W}_2 \in \mathbb{R}^{d\times N}, \mathbf{b}_1 \in \mathbb{R}^N$ and $\mathbf{b}_2 \in \mathbb{R}^N$ are the learnable weight and bias of the gating routers.

Given the expert routing weights $\mathbf{g}^{txt}$ and $\mathbf{g}^{img}$, the process of constructing the MoE-enhanced adaptor for both text and image modalities is as follows:

$$\mathbf{e}_i^{txt} = \sum_{n=1}^{N} \mathbf{g}_n^{txt} \cdot \left(\mathbf{W}_n^{txt}\mathbf{e}_i^{txt} + \mathbf{b}_n^{txt}\right), \tag{5}$$

$$\mathbf{e}_i^{img} = \sum_{n=1}^{N} \mathbf{g}_n^{img} \cdot \left(\mathbf{W}_n^{img}\mathbf{e}_i^{img} + \mathbf{b}_n^{img}\right), \tag{6}$$

where $\mathbf{W}_n^{txt} \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_n^{txt} \in \mathbb{R}^d$ are the learnable weight and bias of the n-th expert for text modality. Similary, $\mathbf{W}_n^{img} \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_n^{img} \in \mathbb{R}^d$ represent the learnable weight and bias of the n-th expert for the image modality. $\mathbf{g}_n^{txt}$ and $\mathbf{g}_n^{img}$ represents the routing weight of the n-th expert for the text and image modality, respectively.

Finally, we extract the text representation $\mathbf{e}_i^{txt}$ and image representation $\mathbf{e}_i^{img}$ that are most relevant to user behavior preferences for item i. By aligning text and image representations with the user's long-term preferences, we believe that the inconsistency issues within the item modalities can be effectively addressed, thereby improving recommendation performance.

D. Modality-Specific Vector Quantization

In recommendation systems, user-item interactions often contain noise that can degrade the quality of learned representations, particularly when multimodal features are involved. Inspired by [25], we introduce the Modality-Specific Vector Quantization (MS-VQ) module, which aims to denoise multimodal features by quantizing them into discrete codebook vectors.

To begin with, given a interaction sequence $s_u = [v_1, v_2, \ldots, v_{L-1}]$ of user u, we can represent its multimodal sequence representation as $\mathbf{s}^m = [\mathbf{e}_{v_1}^m, \mathbf{e}_{v_2}^m, \ldots, \mathbf{e}_{v_{L-1}}^m] \in \mathbb{R}^{(L-1)\times d}$, where $m \in \{\text{id, txt, img}\}$ denotes the modality. Each item's multimodal representation $\mathbf{e}_{v_k}^m$ in the sequence is obtained through the multimodal feature encoding and ID-MoE module presented above. And the next item is represented as $\mathbf{e}_{v_L}^m$. We define the semantic codebook for each modality m as $\mathbf{C}^m = \{\mathbf{c}_z^m\}_{z=1}^Z$, where each code vector $\mathbf{c}_z^m \in \mathbb{R}^{(L-1)\times d}$ matches the size of the sequence representation, and Z is the number of discrete code vectors in the codebook.

To quantize the sequence representation, we implement a code selection model with a Multi-Layer Perceptron (MLP) to compute the Z-dimensional logits for each sequence representation $\mathbf{s}^m$ for modality m. Formally, we have:

$$\mathbf{o}^m = \text{MLP}\left(\mathbf{s}^m\right), \tag{7}$$

where $\mathbf{o}^m \in \mathbb{R}^Z$ represents the output logits generated by the MLP, which correspond to the importance scores for each code vector in the codebook.

Rather than performing a deterministic selection of the nearest code vector, which would introduce non-differentiability, we adopt a Gumbel-Softmax [33]–[35] approach for differentiable sampling. This enables the selection of the code vector in a stochastic manner, facilitating end-to-end training of the quantization process. The Gumbel-Softmax formula is as follows:

$$a_z^m = \frac{\exp((o_z^m + n_z)/\tau)}{\sum_{z'=1}^{Z} \exp((o_{z'}^m + n_{z'})/\tau)}, \tag{8}$$

where o_z^m represents the logit for the z-th code vector for modality m, n_z is the Gumbel noise, and τ is the temperature parameter that controls the smoothness of the sampling. The output $a_z^m \in [0, 1]$ corresponds to the probability of selecting the z-th code vector for quantizing the input sequence.

In the forward propagation of training, we adopt $z^* = \text{argmax}_z a_z^m$ to select the z^*-th code vector $\mathbf{c}_{z^*}^m$ for quantizing the sequence s^m. During training, we utilize the gradient from the Gumbel-Softmax to further backpropagate towards the code selection MLP.

After obtaining the quantized representation $\mathbf{c}_{z^*}^m$ for the sequence s^m, we integrate it with the original sequence representation to form the final sequence representation $\tilde{\mathbf{s}}^m$:

$$\tilde{\mathbf{s}}^m = \lambda_q \mathbf{c}_{z^*}^m + \mathbf{s}^m, \tag{9}$$

where λ_q controls the injection strength of the quantized representation $\mathbf{c}_{z^*}^m$.

To further improve the quantization process, the codebook is updated using an expectation-maximization procedure, commonly utilized in clustering algorithms. Specifically, for each code vector, the sequences selecting that particular vector are aggregated, and the corresponding code vector is updated as the mean of these sequences:

$$\mathbf{c}_z^m = \frac{1}{|S_z^m|} \sum_{\mathbf{s}^m \in S_z^m} \mathbf{s}^m, \tag{10}$$

where S_z^m denotes the set of sequences in the batch that select the z-th code vector. This procedure allows the codebook to evolve iteratively, ensuring that the code vectors are updated to better represent the underlying structure of the data and contribute to improved feature denoising.

E. Contrastive Diffusion Generation

1) Training Phase: Similar to the classical DMs, the diffusion process in our framework comprises both the forward perturbation and the reverse denoising. In SR, the forward process gradually adds Gaussian noise to perturb the ground-truth next item representation. Conversely, the reverse process restores the perturbed representations from a disordered state back to the representation space. We adopt a conditional Denoising Diffusion Probabilistic Model (DDPM) [36] to train the denoising model.

Specifically, we randomly add t steps of Gaussian noise to the ground-truth next item representation $\mathbf{e}_{v_L}^m$ of modality m, where $m \in \{\text{id, txt, img}\}$. We initialize the diffusion process with $\mathbf{x}_0 = \mathbf{e}_{v_L}^m$. The forward process can be formalized as follows:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \tag{11}$$

where $t \in \{1, 2, \ldots, T\}$ represents the diffusion timestep, T is the upper limit of the diffusion step, $\beta_t \in (0, 1)$ denotes the added Gaussian noise scale at step t, $\mathbf{I}$ is an identity matrix, and $\mathcal{N}$ denotes the Gaussian distribution.

By applying the reparameterization trick and leveraging the additive property of independent Gaussian noise, $\mathbf{x}_t$ can be derived directly from $\mathbf{x}_0$, as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{12}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{t'=1}^{t} \alpha_{t'}$. This formulation allows direct sampling of $\mathbf{x}_t$ at any timestep t from the clean multimodal representation $\mathbf{x}_0$, without the need to iterate over all intermediate steps.

Next, we choose Transformer as the backbone of the denoising model to reconstruct the clean next item representation $\hat{\mathbf{x}}_0$ under the guidance of sequence representation $\tilde{\mathbf{s}}^m$, which can be formulated as:

$$\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, t, \tilde{\mathbf{s}}^m). \tag{13}$$

Then, the model is optimized to minimize the Mean-Squared Error (MSE) loss, which simplifies to:

$$\mathcal{L}_r = \mathbb{E}_{t, \mathbf{x}_0, \tilde{\mathbf{s}}^m, \epsilon} \left[||\mathbf{x}_0 - f_\theta(\mathbf{x}_t, t, \tilde{\mathbf{s}}^m)||^2 \right], \tag{14}$$

where $\mathbf{x}_0 = \mathbf{e}_{v_L}^m$ is the ground-truth next item representation of modality m. $\mathcal{L}_r$ denotes the reconstruction loss.

Despite the strong representational capacity of DMs, they can be susceptible to prediction biases caused by imbalances or inconsistencies in the training data, leading to similar next item predictions even for different input sequences. To alleviate this issue and encourage diversity-aware generation, we incorporate a contrastive learning objective to explicitly enlarge the representational gap between denoised outputs from distinct user sequences.

Let $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{x}}_0'$ denote the denoised next item representation generated from two different user sequences within the same mini-batch $\mathcal{B}$. To enhance the distinctiveness of denoised representations and mitigate mode collapse, we define the contrastive loss $\mathcal{L}_c$ as:

$$\mathcal{L}_c = \mathbb{E}_{\hat{\mathbf{x}}_0} \left[\log \sum_{\hat{\mathbf{x}}_0' \in \mathcal{B}} \left[\exp(\text{sim}(\hat{\mathbf{x}}_0, \hat{\mathbf{x}}_0')) \right] \right], \tag{15}$$

where $\text{sim}(\cdot)$ denotes the cosine similarity function. By minimizing $\mathcal{L}_c$, the model is encouraged to reduce the similarity between denoised outputs of different sequences, thereby enhancing the ability of the diffusion model to reflect personalized interest patterns across users.

2) Inference Phase: During the inference phase, we gradually denoise from the standard Gaussian representation $\mathbf{x}_T$ and use the sequence representation $\tilde{\mathbf{s}}^m$ of modality m as guidance to iteratively perform reverse denoising through the denoiser. Specifically, we approximate the real representation $\mathbf{x}_T \rightarrow \mathbf{x}_{T-1} \rightarrow \ldots \rightarrow \mathbf{x}_0$ step-by-step.

Specifically, at each reverse step t, the reverse process can be defined as:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{16}$$

where μ_θ and Σ_θ are parameterized functions that predict the mean and variance of the denoised representation, respectively. θ represents the learnable parameters of the model. Based on the Gaussian distribution parameterization, we can set $\Sigma_\theta = \sigma_t^2 \mathbf{I}$ as constants, where $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. The mean μ_θ is parameterized as:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t, \tag{17}$$

where $\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, t, \tilde{\mathbf{s}}^m)$ is implemented by a Transformer model that has been trained during the training phase. The corresponding stepwise output is computed as:

$$\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} f_\theta(\mathbf{x}_t, t, \tilde{\mathbf{s}}^m) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \sigma_t \epsilon, \tag{18}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. After T reverse steps, we take $\mathbf{x}_0$ as the predicted next item representation of user u, denoted as $\mathbf{p}_u^m$, where $m \in \{\text{id, txt, img}\}$.

Next, to estimate the interaction probability between user u and the candidate item i, we compute the relevance score within each modality and then sum up these scores to obtain the final prediction $\hat{y}_{ui}$, which can be formulated as below:

$$\hat{y}_{ui} = \mathbf{p}_u^{id} \cdot \mathbf{e}_i^{id} + \mathbf{p}_u^{txt} \cdot \mathbf{e}_i^{txt} + \mathbf{p}_u^{img} \cdot \mathbf{e}_i^{img}. \tag{19}$$

F. Model Optimization

Based on the aforementioned network architecture, we obtain the predicted interaction score $\hat{y}_{ui}^m$, which represents the probability that item i will be selected as the next interacted item by the user u under modality m. To compute this probability, we first calculate a matching score between the predicted next item representation $\mathbf{p}_u^m$ and the embedding of each candidate item $\mathbf{e}_i^m$ via an inner product. These scores are then normalized across all candidates using the softmax function:

$$\hat{y}_{ui}^m = \frac{\exp(\mathbf{p}_u^m \cdot \mathbf{e}_i^m)}{\sum_{j \in \mathcal{I}} \exp(\mathbf{p}_u^m \cdot \mathbf{e}_j^m)}, \tag{20}$$

where j iterates over all candidate items in the set $\mathcal{I}$.

To optimize the model, we adopt the cross-entropy loss as the primary training objective, which quantifies the discrepancy between the predicted score $\hat{y}_{ui}^m$ and the ground-truth label y_{ui} for modality m.

$$\mathcal{L}_{ce} = -\sum_{i \in \mathcal{I}} y_{ui} \log(\hat{y}_{ui}^m). \tag{21}$$

To further enhance the quality of next item prediction and alleviate potential biases during generation, we incorporate two additional objectives derived from the contrastive diffusion module. Specifically, the denoising reconstruction loss $\mathcal{L}_r$ defined in Equation 14 is employed to guide the model in recovering the original item embedding from the noisy input through conditional generation, while the contrastive loss $\mathcal{L}_c$ defined in Equation 15 encourages distinguishable generation by maximizing the representational divergence among different sequences.

Finally, we integrate all objectives into a unified training framework. For each modality (i.e., ID, text, and image), we compute the cross-entropy loss $\mathcal{L}_{ce}^m$, the reconstruction loss $\mathcal{L}_r^m$, and the contrastive loss $\mathcal{L}_c^m$, and aggregate them to form the total optimization objective as follows:

$$\mathcal{L}_{total} = \sum_{m \in \{\text{id,txt,img}\}} \left(\mathcal{L}_{ce}^m + \mathcal{L}_r^m + \lambda_c \mathcal{L}_c^m\right), \tag{22}$$

where λ_c are hyperparameters that control the contributions of the contrastive loss. This joint optimization encourages the model to not only fit the observed interactions but also to generate robust and personalized representations for future item prediction.

TABLE I
STATISTICS OF THE FOUR DATASETS. "AVG.LENGTH" DENOTES THE AVERAGE LENGTH OF INTERACTION SEQUENCES

Dataset	#Users	#Items	#Interactions	#Avg.length
Toys	147356	49637	1258412	8.54
Games	39828	11364	343349	8.62
Instruments	18870	6619	152271	8.07
Food	107203	32896	946309	8.83

IV. EXPERIMENT

Extensive experiments are conducted to evaluate the effectiveness of the proposed model. The experiments aim to answer the following research questions:

RQ1: Does the proposed IMECD outperform the state-of-the-art SR methods?

RQ2: How do different components contribute to IMECD?

RQ3: How do different choices of hyper-parameters affect the performance of IMECD?

A. Experimental Setting

1) Datasets: Following existing methods [7], [11], we evaluate IMECD on a real-world recommendation dataset, namely the Amazon review dataset [1] [37]. Specifically, we select four types of datasets: Toys and Games (Toys), Video Games (Games), Musical Instruments (Instruments), and Grocery and Gourmet Food (Food). We use the 5-core subsets, in which all users and items have at least 5 reviews.

To support multimodal inputs, we extract item textual information from the metadata, including the item's title, brand, and category. Additionally, we obtain item images by parsing the metadata's URL to retrieve the corresponding product images. Some items in the dataset have missing text or image data,

[1] https://cseweb.ucsd.edu/ jmcauley/datasets/amazon_v2/

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT BASELINE METHODS.THE BEST AND THE SECOND-BEST PERFORMANCE IS BOLD AND UNDERLINED RESPECTIVELY.

Dataset	Metric	GRU4Rec	SASRec	BERT4Rec	Caser	DifffuRec	DiQDiff	UniSRec	MISSRec	MMMLP	HM4SR	IMECD	Improve
Toys	H@5	0.0633	0.0732	0.0398	0.0155	0.0776	0.0779	0.0486	0.0791	0.0792	0.0794	**0.0825**	3.90%
	H@10	0.0768	0.0924	0.0515	0.0220	0.0961	0.0938	0.0638	0.0961	0.0935	0.0955	**0.0977**	1.66%
	N@5	0.0518	0.0547	0.0307	0.0109	0.0614	0.0632	0.0377	0.0586	0.0643	0.0642	**0.0679**	5.60%
	N@10	0.0561	0.0609	0.0345	0.0130	0.0673	0.0683	0.0426	0.0638	0.0696	0.0694	**0.0728**	4.60%
Games	H@5	0.0986	0.0948	0.0531	0.0326	0.1033	0.1032	0.0870	0.1004	0.1084	0.1075	**0.1133**	4.52%
	H@10	0.1361	0.1319	0.0813	0.0548	0.1404	0.1386	0.1240	0.1343	0.1434	0.1466	**0.1485**	1.30%
	N@5	0.0752	0.0701	0.0360	0.0200	0.0793	0.0813	0.0621	0.0715	0.0835	0.0830	**0.0883**	5.75%
	N@10	0.0872	0.0820	0.0451	0.0271	0.0912	0.0927	0.0740	0.0819	0.0955	0.0956	**0.0997**	4.29%
Instruments	H@5	0.1041	0.1061	0.0380	0.0603	0.1027	0.1041	0.0933	0.1076	0.1084	0.1094	**0.1120**	2.38%
	H@10	0.1238	0.1247	0.0632	0.0736	0.1190	0.1182	0.1167	0.1220	0.1280	0.1288	**0.1305**	1.32%
	N@5	0.0900	0.0894	0.0232	0.0519	0.0907	0.0920	0.0841	0.0916	0.0937	0.0955	**0.0975**	2.09%
	N@10	0.0964	0.0954	0.0314	0.0562	0.0959	0.0966	0.0916	0.0942	0.1000	0.1020	**0.1034**	1.37%
Food	H@5	0.1329	0.1324	0.0917	0.0655	0.1373	0.1364	0.1021	0.1324	0.1374	0.1356	**0.1401**	1.96%
	H@10	0.1488	0.1496	0.1060	0.0769	0.1525	0.1530	0.1218	0.1456	0.1534	0.1520	**0.1556**	1.43%
	N@5	0.1164	0.1126	0.0768	0.0505	0.1185	0.1202	0.0797	0.1152	0.1164	0.1162	**0.1216**	1.16%
	N@10	0.1215	0.1182	0.0815	0.0542	0.1230	0.1232	0.0860	0.1192	0.1222	0.1215	**0.1266**	2.76%

so we remove those with incomplete multimodal information for fairness. Finally, we retain items and users with more than 5 interactions, ensuring sufficient data for training and evaluation. Statistics of these datasets are listed in Table I.

2) Evaluation Metrics: Following previous works [13], [25], [38], we adopt two standard metrics, i.e., Hit Rate (H@K) and Normalized Discounted Cumulative Gain (N@K), to evaluate the recommendation performance. We set K to 5 and 10 for showcases.

3) Baselines: To verify the effectiveness of our method, we select the following representative and competitive baselines for sequential recommendation from three categories.

1) **Traditional Methods.** GRU4Rec [16] employs gated recurrent units to enhance standard RNN capabilities, partially addressing the vanishing gradient issue. SASRec [3] adopts self-attention mechanisms y to learn user sequential interest. BERT4Rec [39] is a bidirectional transformer model for SR. Caser [17] is a CNN-based model that captures both local and global sequential patterns for SR.
2) **Diffusion-based Methods.** DiffuRec [21] introduces DMs to sequential recommendation, generating adaptive item distributions to better capture user preferences. DiQDiff [25] enhances DMs by quantizing user sequences and contrastively diversifying item generation to overcome noise and popularity bias.
3) **MultiModal Methods.** UniSRec [8] uses MoE adapters to encode item texts for SR. MISSRec [11] designs an interest discovery module to grasp deep relations among items, modalities, and preferences. MMMLP [10] is a pure MLP-based SR model that efficiently integrates multimodal data through specialized feature mixer and fusion mixer layers. HM4SR [7] uses hierarchical MoEs to purify multimodal signals and model temporal dynamics for SR.

4) Implementation and Hyperparameter Setting: We implement our framework using PyTorch. The maximum length of each behavior sequence is limited to 50. The textual and visual features are extracted using pre-trained models, namely roberta-based[2] and vit-base-patch16-224[3], both obtained from the Hugging Face model repository. We employ the Adam optimizer, where the initial learning rate is 0.001. We set the training batch size as 1280, and the hidden size of all methods is 128. The dropout rates for both the denoising model and item embeddings are set to 0.1. The temperature τ of Gumbel-Softmax is set to 0.1. For the denoiser of the ID modality, the number of attention heads and self-attention layers in the Transformer are set to 4 and 4, respectively. For the denoisers of the text and image modalities, the number of attention heads and self-attention layers in the Transformer are set to 4 and 1, respectively. To balance efficiency and quality, we set $T = 32$ for DDPM's diffusion timesteps and adopt a truncated linear noise schedule. The expert number N for ID-MoE is selected from $\{2, 4, 6, 8, 10\}$. The codebook size Z is selected from $\{8, 16, 24, 32, 40\}$. The strengths λ_c of the contrastive loss are varied within the range $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, while the injection strength λ_q of the quantized vector are also varied within the range $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. All baselines are conducted based on their GitHub source code. All hyperparameters are determined according to the performance in the validation data. All results are reported in the test set. We perform all the experiments on NVIDIA GeForce RTX 3090 GPUs.

B. Performance Comparisons (RQ1)

In our experiments, we compare the performance of IMECD with several state-of-the-art recommendation models across

[2]https://huggingface.co/FacebookAI/roberta-base
[3]https://huggingface.co/google/vit-base-patch16-224

TABLE III
MODEL PERFORMANCE OF ABLATION STUDY. THE BEST RESULTS ARE BOLD.

Dataset	Metric	H@5	H@10	N@5	N@10
Toys	w/o ID-MoE	0.0811	0.0962	0.0667	0.0720
	w/o MS-VQ	0.0820	0.0971	0.0671	0.0715
	w/o CDG_CL	0.0820	0.0969	0.0676	0.0722
	w/o Text	0.0806	0.0958	0.0665	0.0714
	w/o Image	0.0794	0.0958	0.0651	0.0703
	IMECD	**0.0825**	**0.0977**	**0.0679**	**0.0728**
Games	w/o ID-MoE	0.1098	0.1478	0.0845	0.0968
	w/o MS-VQ	0.1124	0.1450	0.0874	0.0979
	w/o CDG_CL	0.1128	0.1473	0.0878	0.0990
	w/o Text	0.1095	0.1431	0.0856	0.0964
	w/o Image	0.1130	0.1451	0.0877	0.0979
	IMECD	**0.1133**	**0.1485**	**0.0883**	**0.0997**
Instruments	w/o ID-MoE	0.1112	0.1297	0.0960	0.1028
	w/o MS-VQ	0.1111	0.1301	0.0969	0.1024
	w/o CDG_CL	0.1109	0.1300	0.0964	0.1023
	w/o Text	0.1081	0.1256	0.0948	0.1004
	w/o Image	0.1096	0.1268	0.0960	0.1015
	IMECD	**0.1120**	**0.1305**	**0.0975**	**0.1034**
Food	w/o ID-MoE	0.1397	0.1550	0.1207	0.1262
	w/o MS-VQ	0.1396	0.1552	0.1214	0.1261
	w/o CDG_CL	0.1388	0.1541	0.1210	0.1259
	w/o Text	0.1396	0.1544	0.1211	0.1260
	w/o Image	0.1391	0.1552	0.1213	0.1264
	IMECD	**0.1401**	**0.1556**	**0.1216**	**0.1266**

four different datasets, and the experimental results are presented in Table II. We have the following findings.

(1) IMECD outperforms all the baseline models across the four datasets. It achieves improvements ranging from 1.16% to 9.68% compared to the best diffusion-based method. Compared to the best multimodal method, IMECD gains relative performance improvements ranging from 1.3% to 5.75%. It is worth mentioning that our method outperforms all other multimodal baselines in all cases, which can be attributed to the effectiveness of the proposed ID-MoE module in mitigating modality inconsistencies and noise, as well as the integration of the DM for dynamically modeling user-item interactions.

(2) The diffusion-based approach consistently outperforms traditional sequential models. This validates the power of the iterative denoising process inherent in diffusion models. Unlike traditional models, which rely on sequential learning over static embeddings, IMECD leverages a dynamic diffusion process that progressively refines noisy item representations. The iterative denoising further addresses challenges posed by data sparsity and user interaction variability, providing a more robust modeling approach.

(3) In most cases, multimodal methods outperform both traditional methods and diffusion-based methods that only use ID embeddings. This suggests that incorporating multimodal information, including text and images, significantly enriches item representations, enabling the model to capture a wider range of user preferences.

C. Ablation Study (RQ2)

In this section, we conduct an ablation study to evaluate the effectiveness of various components in IMECD. Specifically, we design the following five model variants.

1) w/o ID-MoE: it disables the ID-Guided Multimodal Mixture of Experts (ID-MoE) module.
2) w/o MS-VQ: it excludes the Modality-Specific Vector Quantization (MS-VQ) module.
3) w/o CDG_CL: it removes the contrastive learning component from the Contrastive Diffusion Generation (IMECD) module.
4) w/o Text: it removes the text modality from the model.
5) w/o Image: it removes the image modality from the model.

The results of the ablation experiments are shown in Table III, where we evaluate the performance of IMECD across four datasets: Toys, Games, Instruments, and Food. The results indicate that each component of the IMECD contributes significantly to its overall performance. Specifically, disabling the ID-MoE module leads to a decrease in performance across all datasets. This finding indicates that the ID-MoE module can effectively extract user preference features from both textual and visual representations under the guidance of ID embeddings, thereby enhancing recommendation performance. Furthermore, excluding the MS-VQ module results in performance drops. This indicates that the MS-VQ module effectively denoises sequence representations through vector quantization, which is crucial for enhancing user representations. Removing the contrastive learning component from the CDG module leads to a decrease in performance across all datasets. This highlights the importance of contrastive learning in mitigating diffusion biases and improving the diversity and accuracy of next item predictions. Additionally, the performance further deteriorates when either the text or image modality is removed. This confirms that both textual and visual features play vital roles in enriching item representations and improving recommendation performance. Overall, the ablation studies provide strong evidence that each component of IMECD contributes synergistically to its superior performance.

D. Parameter Sensitivity Analysis (RQ3)

In this section, we conduct a parameter sensitivity experiment to evaluate the impact of different hyperparameters on the performance of IMECD. Specifically, we investigate the following parameters: the number of experts N for ID-MoE, the codebook size Z for MS-VQ, the injection strength λ_q of quantized vectors from MS-VQ, and the contrastive loss weight λ_c of CDG, respectively. The analysis is conducted on Toys and Games datasets. We evaluate the performance using H@5 and N@5 metrics.

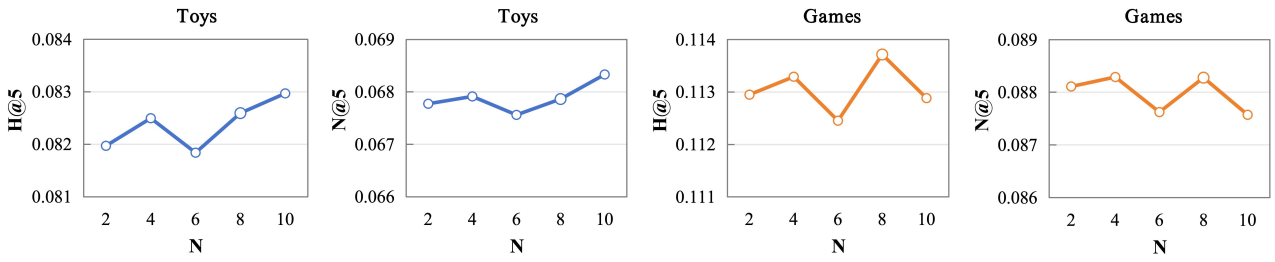

Fig. 4. Performance with different numbers of experts N for ID-MoE on Toys and Games.

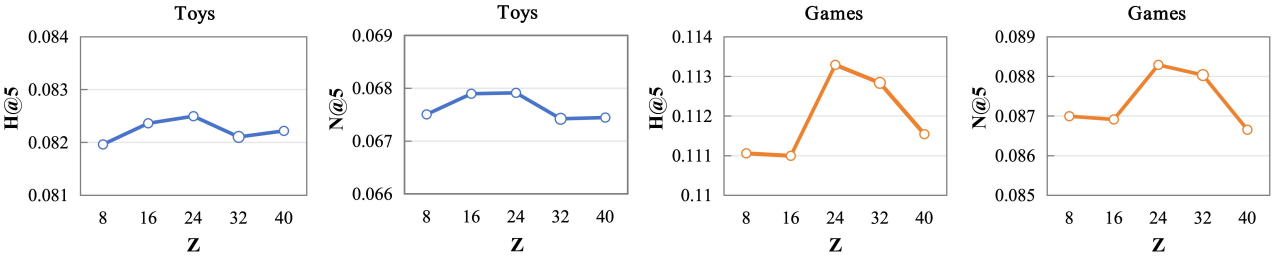

Fig. 5. Performance with different size of codebook Z for MS-VQ on Toys and Games.

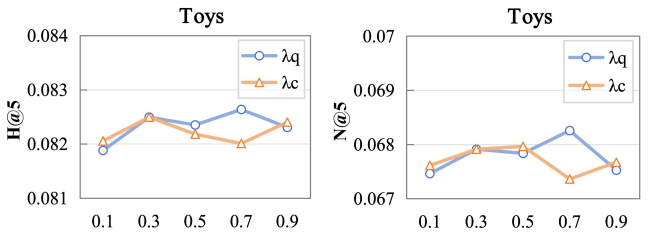

Fig. 6. Performance with different injection strength λ_q and contrastive loss λ_c on Toys.

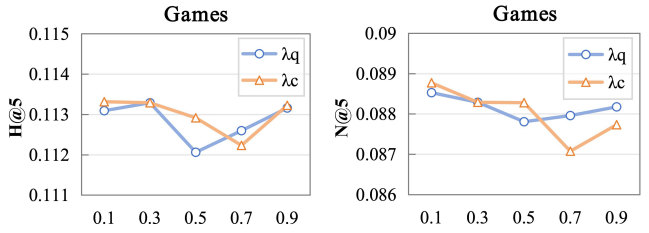

Fig. 7. Performance with different injection strength λ_q and contrastive loss λ_c on Games.

1) Impact of the Number of Experts: Figures 4 shows the performance of IMECD with varying numbers of experts N. For the Toys dataset, H@5 and N@5 both increase as N grows, reaching their highest values at $N = 10$, indicating that more experts can better capture user preferences. However, the performance at $N = 6$ shows a small decrease, which could be attributed to an imbalance in the distribution of tasks among the experts. In the Games dataset, H@5 and N@5 achieve their highest values at $N = 8$. However, when N is further increased to 10, the performance slightly decreases. This finding suggests that increasing the number of experts can add complexity to the model and may lead to diminishing returns. Overall, more experts generally improve performance, but the optimal number varies by dataset.

2) Impact of the Codebook Size: Figures 5 shows the impact of the codebook size Z on model performance for both the Toys and Games datasets. For both Toys and Games datasets, H@5 and N@5 improve as Z increases, peaking at $Z = 24$. Beyond this point, performance slightly drops, likely due to increased complexity and potential overfitting. This indicates that a larger codebook helps the model capture more granular details of item representations, but after a certain size, the added complexity does not provide significant benefits and may even degrade performance due to overfitting. In conclusion, for both datasets, the codebook size of 24 appears to provide the best balance between capturing fine-grained item representations and maintaining model efficiency.

3) Impact of the Injection Strength of Quantized Vectors: Figure 6 and Figure 7 present the Impact of the injection strength of quantized vectors λ_q for the Toys and Games datasets, respectively. For the Toys dataset, as λ_q increases, the model performance generally improves initially, reaching its peak at $\lambda_q = 0.7$, and then starts to decrease. On the Games dataset, however, the performance follows a different trend. It initially decreases as λ_q increases and then improves after reaching a certain point. Overall, the impact of the injection strength of quantized vectors varies across different datasets.

4) Impact of the Contrastive Loss Weight: Figure 6 and Figure 7 show the impact of the contrastive loss weight λ_c on the Toys and Games datasets, respectively. On the Toys dataset, as the contrastive loss weight λ_c increases, the model

performance fluctuates, showing inconsistent trends. In contrast, on the Games dataset, the performance initially decreases and then increases as the contrastive loss weight increases. The optimal performance point differs across datasets, indicating that the effect of the contrastive loss weight varies depending on the dataset.

V. Conclusion

To address the inconsistency and noise issues commonly observed in existing multimodal methods for sequential recommendation, we introduce IMECD, a novel framework that combines the strengths of diffusion models and multimodal information to enhance sequential recommendation performance. Our method leverages the long-term stable semantics encoded in item ID embeddings, guiding the text and image modalities through a MoE mechanism. Additionally, we introduced an MS-VQ module for quantizing and denoising multimodal sequences. The diffusion model is then used to iteratively denoise and generate the next item representation, with a contrastive loss integrated to mitigate diffusion bias. Extensive experiments conducted on multiple benchmark datasets confirm the superiority of IMECD over existing state-of-the-art SR methods. IMECD not only effectively addresses the issues of multimodal feature inconsistency and noise but also represents a novel application of diffusion models in multimodal sequential recommendation, offering a promising direction for future research in this field.

References

[1] Y. Lu, C. Wang, P. Lai, and J. Lai, "PKAT: pre-training in collaborative knowledge graph attention network for recommendation," in *ICDM*, 2023, pp. 448–457.

[2] C. Zhang and X. Hong, "Challenging the long tail recommendation on heterogeneous information network," in *ICDM*, 2021, pp. 94–101.

[3] W. Kang and J. J. McAuley, "Self-attentive sequential recommendation," in *ICDM*, 2018, pp. 197–206.

[4] M. Li, Z. Zhang, X. Zhao, W. Wang, M. Zhao, R. Wu, and R. Guo, "AutoMLP: Automated MLP for sequential recommendations," in *WWW*, 2023, pp. 1190–1198.

[5] R. Qiu, Z. Huang, H. Yin, and Z. Wang, "Contrastive learning for representation degeneration problem in sequential recommendation," in *WSDM*, 2022, pp. 813–823.

[6] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *SIGIR*, 2002, pp. 253–260.

[7] S. Zhang, L. Chen, D. Shen, C. Wang, and H. Xiong, "Hierarchical time-aware mixture of experts for multi-modal sequential recommendation," in *WWW*, 2025.

[8] Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, and J. Wen, "Towards universal sequence representation learning for recommender systems," in *KDD*, 2022, pp. 585–593.

[9] Q. Cui, S. Wu, Q. Liu, W. Zhong, and L. Wang, "MV-RNN: A multi-view recurrent neural network for sequential recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 2, pp. 317–331, 2020.

[10] J. Liang, X. Zhao, M. Li, Z. Zhang, W. Wang, H. Liu, and Z. Liu, "MMMLP: multi-modal multilayer perceptron for sequential recommendations," in *WWW*, 2023, pp. 1109–1117.

[11] J. Wang, Z. Zeng, Y. Wang, Y. Wang, X. Lu, T. Li, J. Yuan, R. Zhang, H. Zheng, and S. Xia, "MISSRec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation," in *MM*, 2023, pp. 6548–6557.

[12] H. Hu, W. Guo, Y. Liu, and M. Kan, "Adaptive multi-modalities fusion in sequential recommendation systems," in *CIKM*, 2023, pp. 843–853.

[13] J. Fu, X. Ge, X. Xin, A. Karatzoglou, I. Arapakis, J. Wang, and J. M. Jose, "IISAN: efficiently adapting multimodal representation for sequential recommendation with decoupled PEFT," in *SIGIR*, 2024, pp. 687–697.

[14] R. He and J. J. McAuley, "Fusing similarity models with markov chains for sparse sequential recommendation," in *ICDM*, 2016, pp. 191–200.

[15] S. Kabbur, X. Ning, and G. Karypis, "FISM: factored item similarity models for top-n recommender systems," in *KDD*, 2013, pp. 659–667.

[16] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *ICLR*, 2016.

[17] J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *WSDM*, 2018, pp. 565–573.

[18] Y. Cen, J. Zhang, X. Zou, C. Zhou, H. Yang, and J. Tang, "Controllable multi-interest framework for recommendation," in *KDD*, 2020, pp. 2942–2951.

[19] G. Lin, C. Gao, Y. Zheng, J. Chang, Y. Niu, Y. Song, Z. Li, D. Jin, and Y. Li, "Dual-interest factorization-heads attention for sequential recommendation," in *WWW*, 2023, pp. 917–927.

[20] S. Zhang, L. Yang, D. Yao, Y. Lu, F. Feng, Z. Zhao, T. Chua, and F. Wu, "Re4: Learning to re-contrast, re-attend, re-construct for multi-interest recommendation," in *WWW*, 2022, pp. 2216–2226.

[21] Z. Li, A. Sun, and C. Li, "Diffurec: A diffusion model for sequential recommendation," *ACM Transactions on Information Systems*, vol. 42, no. 3, pp. 1–28, 2023.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 674–10 685.

[23] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," in *NeurIPS*, 2022.

[24] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 105:1–105:39, 2024.

[25] W. Mao, S. Liu, H. Liu, H. Liu, X. Li, and L. Hu, "Distinguished quantized guidance for diffusion-based sequence recommendation," in *WWW*, 2025.

[26] Z. Yang, J. Wu, Z. Wang, X. Wang, Y. Yuan, and X. He, "Generate what you prefer: Reshaping sequential recommendation via guided diffusion," in *NeurIPS*, 2023.

[27] W. Wang, Y. Xu, F. Feng, X. Lin, X. He, and T. Chua, "Diffusion recommender model," in *SIGIR*, 2023, pp. 832–841.

[28] Q. Liu, F. Yan, X. Zhao, Z. Du, H. Guo, R. Tang, and F. Tian, "Diffusion augmentation for sequential recommendation," in *CIKM*, 2023, pp. 1576–1586.

[29] Z. Cui, H. Wu, B. He, J. Cheng, and C. Ma, "Context matters: Enhancing sequential recommendation with context-aware diffusion-based contrastive learning," in *CIKM*, 2024, pp. 404–414.

[30] Y. Jiang, Y. Yang, L. Xia, and C. Huang, "DiffKG: Knowledge graph diffusion model for recommendation," in *WSDM*, 2024, pp. 313–321.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[33] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.

[34] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR*, 2017.

[35] J. Zhang, F. Zhan, C. Theobalt, and S. Lu, "Regularized vector quantization for tokenized image synthesis," in *CVPR*, 2023, pp. 18 467–18 476.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[37] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *EMNLP-IJCNLP*, 2019, pp. 188–197.

[38] S. Bian, X. Pan, W. X. Zhao, J. Wang, C. Wang, and J. Wen, "Multi-modal mixture of experts represetation learning for sequential recommendation," in *CIKM*, 2023, pp. 110–119.

[39] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in *CIKM*, 2019, pp. 1441–1450.