Enhancing Sequential Recommendation via Task-Aware Query-Based Multi-Modal Retrieval

Anonymous

***Abstract*—Multi-modal sequential recommendation, which integrates heterogeneous modality information to capture rich item semantics and evolving user interests, has received widespread attention in recent years. However, multi-modal content often contains substantial task-irrelevant noise that can mislead recommendation decisions, and existing methods predominantly rely on coarse-grained global features extracted from frozen pre-trained encoders, making it challenging to distinguish informative semantic signals from complex content. While recent efforts attempt to address this through adaptive filtering mechanisms, they generally lack fine-grained semantic identification and selection capabilities. Although end-to-end optimization of multi-modal encoders can help extract task-specific features, such approaches often incur substantial computational and memory overhead. To address these limitations, we propose TAME, a novel framework for multi-modal sequential recommendation that retrieves task-aware multi-modal signals and enriches ID representations in a fine-grained, adaptive manner. Specifically, to identify and extract task-aware multi-modal signals, we employ an ID-aware query-based modality retriever that leverages a set of task-conditioned learnable queries to retrieve informative semantic cues from fine-grained multi-modal features via cross-attention mechanism. To further enhance the expressiveness of ID embeddings, we integrate the retrieved multi-modal signals with ID embeddings through a gated Mixture-of-Experts (MoE) architecture, enabling dynamic, context-aware representation learning. Finally, sequence encoder for each modality is utilized to model user behavior patterns over time. Extensive experiments on three public benchmarks demonstrate that TAME consistently outperforms state-of-the-art methods, highlighting its effectiveness in task-relevant multi-modal information extraction and ID representation enhancement. Our code is available at https://anonymous.4open.science/r/TAME-0463/.**

***Index Terms*—sequential recommendation, multi-modal recommendation, mixture of experts.**

I. Introduction

Sequential recommendation (SR), which aims to model users' evolving preferences based on their historical interaction sequences, has achieved remarkable progress in recent years [1], [2]. Numerous efforts have leveraged Recurrent Neural Networks (RNNs) [3], [4], Convolutional Neural Networks (CNNs) [5], Graph Neural Networks (GNNs) [6]–[8] and Transformer [9]–[11] to encode user interaction sequences and capture dynamic user preferences. However, most existing SR methods rely exclusively on ID-based representations derived from user-item interaction histories. While effective under sufficient interaction data, such methods tend to suffer from limited semantic expressiveness and weak generalization in sparse and cold-start scenarios [12]. To mitigate these limitations, recent efforts [13]–[16] have explored incorporating multi-modal data, such as text and images, into recommendation frameworks, providing richer semantic signals to improve generalization under sparse interactions. As a result, multi-modal sequential recommendation has garnered growing interest, particularly in domains like e-commerce, video streaming, and news recommendation, where diverse item-side content is ubiquitous [17], [18].

In order to harness the potential of multi-modal information in sequential recommendation, an increasing number of studies have focused on designing effective fusion strategies that integrate heterogeneous features extracted from frozen pre-trained multi-modal encoders with ID embeddings. These ID embeddings are typically treated as symbolic identifiers and are learned purely from historical interaction sequences [9]. Early research [19] typically adopts static fusion techniques such as concatenation, addition, and reconstruction to combine multi-modal features with ID-based representations. Subsequent methods [14], [20], [21] focus on learning adaptive fusion mechanisms that dynamically adjust the contribution of each modality based on user preferences or context. Recent works [12], [13], [16] further introduce the mixture of experts (MoE) framework to reduce the semantic gap across modalities for better multi-modal fusion. Despite these advances, existing methods still face limitations in how multi-modal information enhances and interacts with ID embeddings. On one hand, ID embeddings often lack semantic grounding and contextual awareness, leading to weak generalization under sparse or cold-start scenarios [22]. Moreover, a single ID embedding is typically difficult to reflect the diverse semantic facets of items and align effectively with varying user intents [23]. On the other hand, most existing models treat multi-modal features as auxiliary signals for downstream sequence modeling, rather than explicitly incorporating them into the ID representation learning process [24]. This leads to under-utilization of semantic content and limits the expressiveness of learned ID embeddings.

Although multi-modal data supplies rich semantic information that can improve recommendation performance, not all modality content is equally beneficial to the recommendation task [14], [16], [25], [26]. In practice, auxiliary modalities such as item descriptions and images often contain content unrelated to users' decision-making, which can obscure semantic cues crucial for recommendation. Furthermore, the majority of existing methods [19], [20] rely on coarse-grained global multi-modal features, which are typically obtained from frozen pre-trained encoders via the special classification token "[CLS]". While computationally efficient, these encoders are pre-trained on general-purpose objectives and are

not exposed to the recommendation task. Consequently, the extracted global representations are inherently task-agnostic and often fail to capture the fine-grained semantic cues that are critical for user decision-making. This misalignment makes it difficult to distinguish informative signals, causing discriminative recommendation cues in multi-modal content to be easily overshadowed by irrelevant or noisy information. Although some research adopts adapters [14] and MoE-based [12], [13], [16] architectures to alleviate noise in multi-modal information, these methods predominantly operate on coarse-grained representations and lack explicit mechanisms to selectively extract task-relevant content. As a result, they may retain irrelevant signals or fail to fully utilize valuable modality information to capture the various intents of users. While end-to-end multi-modal learning [15], [27] enables the encoder to extract task-relevant features by jointly optimizing the multi-modal encoders with recommendation objective, such methods typically incur substantial computational and memory overhead due to the need for fine-tuning large-scale backbones. This highlights a key challenge in multi-modal sequential recommendation: how to effectively and selectively extract task-relevant signals from complex and noisy multi-modal content in a fine-grained and adaptive manner to support accurate and robust recommendation.

To tackle the aforementioned challenges, we propose a **T**ask-**A**ware Query-Based **M**ulti-Modal R**E**trieval approach for multi-modal sequential recommendation, named TAME. In particular, to effectively extract task-relevant features from complex multi-modal inputs, we propose an ID-aware modality retriever, which adopts a self-attention mechanism to encode a set of learnable query vectors conditioned on item ID embeddings. These queries are then utilized to retrieve informative and task-relevant multi-modal cues via cross-attention mechanism. To further enhance the expressiveness of ID embeddings, we propose a modality-guided ID enrichment module that leverages a gating mechanism to dynamically regulate expert contributions in the MoE architecture based on multi-modal context, enabling fine-grained and adaptive enhancement of ID representations. Subsequently, a sequence encoder is employed to model the historical interaction sequences of each modality, enabling the model to capture users' evolving interests from different perspectives over time. In summary, our contributions can be outlined as follows:

- We propose an ID-aware modality retriever that leverages a set of task-conditioned learnable queries to selectively extract informative and recommendation-relevant cues from multi-modal inputs in a fine-grained and adaptive manner.
- We introduce a Modality-Guided ID enrichment module, which injects retrieved multi-modal semantics into ID embeddings via a gated MoE architecture, enabling dynamic and context-aware ID representation learning guided by semantic priors.
- Extensive experiments conducted on public datasets demonstrate the effectiveness of the TAME framework.

II. Related Work

A. ID-Based Sequential Recommendation

Sequential recommendation focuses on learning temporal dependencies from users' historical interaction sequences, which have been extensively studied in recent years. Conventional methods typically rely on learnable ID embeddings and sophisticated sequence encoders to capture dynamic user intents. Early studies [28], [29] employ the Markov Chain assumption to estimate item-to-item transition probabilities. With the rise of deep learning, recent approaches have adopted diverse neural network architectures to more effectively capture evolving user preferences [3], [9], [30]. For instance, GRU4Rec [3], Caser [5], SURGE [6], and SASRec [9] utilize RNN, CNN, GNN, and self-attention mechanism, respectively, to characterize users' historical interaction behaviors. FEARec [11] models user historical interaction sequences in the frequency domain, enabling the extraction of high-frequency signals and periodic patterns that are difficult to capture in the time domain. FamouSRec [31] further proposes a frequency-aware MoE framework that dynamically selects heterogeneous sequence encoders in a personalized manner. In addition, methods such as S3-Rec [30] and CL4SRec [32] incorporate auxiliary self-supervised signals to enhance ID embedding learning. However, since they rely solely on ID-based representations, these methods inherently suffer from limited semantic expressiveness and poor generalization, particularly for long-tail and cold-start items.

B. Multi-Modal Sequential Recommendation

To address the limitations of ID-based sequential recommendation methods, multi-modal sequential recommendation has attracted increasing attention, aiming to incorporate rich auxiliary content such as text and images to improve recommendation performance. To leverage rich semantic information of multi-modal data without increasing additional training cost, some works adopt frozen pre-trained encoders to extract multi-modal features and concentrate on effective fusion strategies. MV-RNN [19] integrates visual and textual information through various fusion methods, i.e., concatenation, addition, and reconstruction, and explores separate and unified RNN structures to capture user preferences. MMMLP [20] employs several MLPs to fuse and align multi-modal information for sequential recommendation. MISSRec [14] employs a dynamic fusion module that adaptively combines multi-modal item features based on user interests. ODMT [21] introduces an ID-aware multi-modal Transformer to integrate heterogeneous features through inter-modal attention mechanisms. Recent advances further explore MoE frameworks for flexible and personalized multi-modal fusion. UniSRec [13] designs a lightweight MoE-enhanced adapter to facilitate knowledge transfer from pre-trained textual representations to ID-based item embeddings. M3SRec [12] employs modality-specific and cross-modal MoE modules to capture complementary signals from different modalities. HM4SR [16] proposes a hierarchical Mixture-of-Experts (MoE) architecture to extract

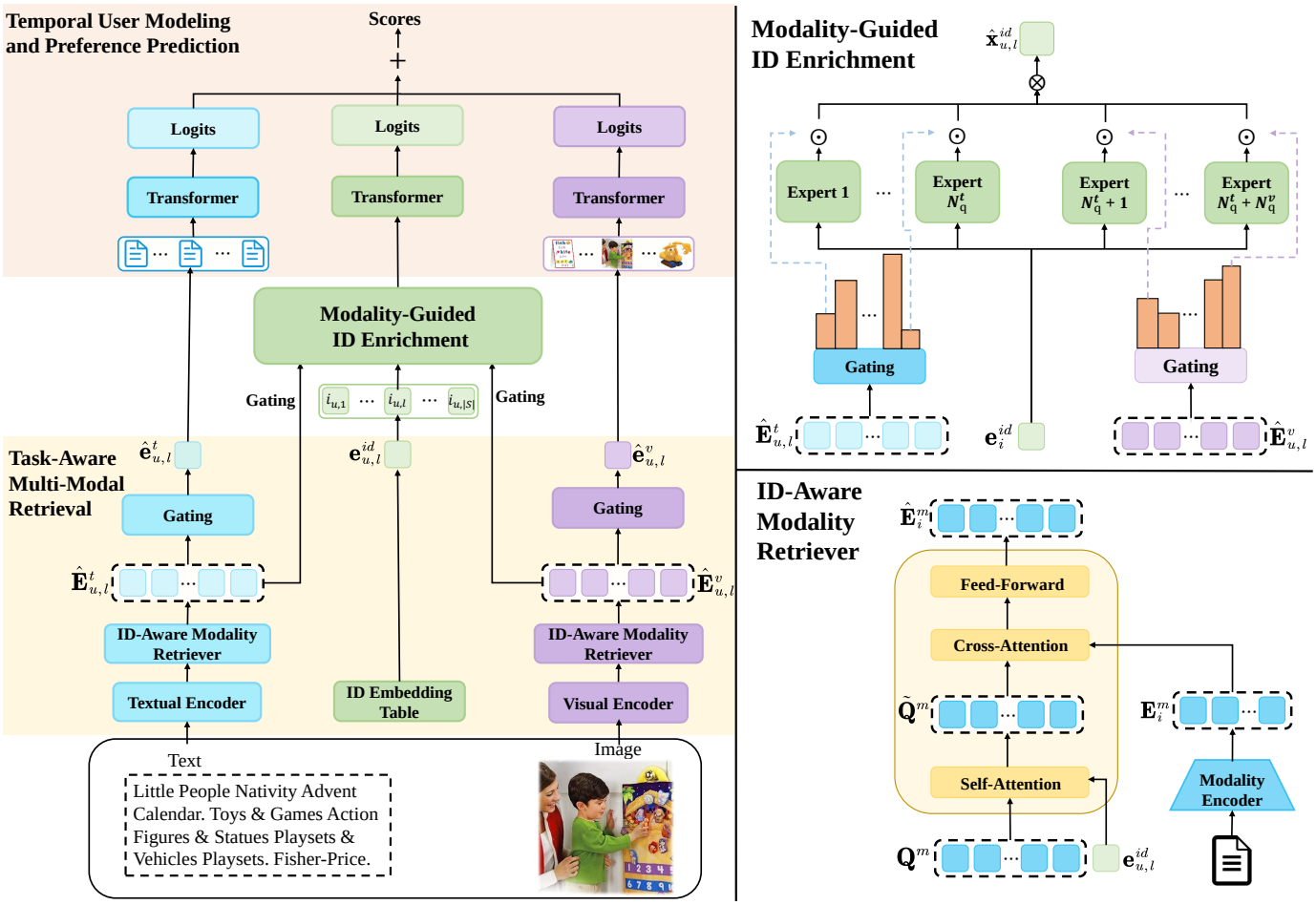

Fig. 1: The framework of the TAME model, which consists of three modules: task-aware multi-modal retrieval module, modality-guided ID enrichment module and temporal user modeling and preference prediction module.

interest-relevant multi-modal features and to model dynamic and explicit temporal information.

Other efforts investigate end-to-end multi-modal learning that jointly optimizes feature extraction and sequential recommendation to eliminate dependence on pre-trained feature extractors. Previous studies [27], [33] have demonstrated that purely modality-based recommendation models can achieve comparable performance to ID-based models through end-to-end training. IISAN [15] designs a novel decoupled parameter-efficient fine-tuning architecture to train multi-modal backbones at both intra-modal and inter-modal levels, achieving promising results. Despite these advances, most existing methods still struggle to effectively and efficiently distill recommendation-relevant signals from multi-modal information.

III. Problem Definition

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$ and $\mathcal{I} = \{i_1, i_2, \ldots, i_{|\mathcal{I}|}\}$ denote the sets of users and items, respectively, with $|\mathcal{U}|$ and $|\mathcal{I}|$ being the total numbers of users and items. Each item $i \in \mathcal{I}$ is associated with several modalities $\mathcal{M} = \{t, v\}$, where t and v correspond to textual and visual modalities, respectively. For each user $u \in \mathcal{U}$, the historical interaction sequence is denoted by $\mathcal{S}_u = \{i_{u,1}, i_{u,2}, \ldots, i_{u,|\mathcal{S}_u|}\}$, where $|\mathcal{S}_u|$ is the sequence length.

Given the interaction sequence $\mathcal{S}_u$ of user u, the objective of the multi-modal sequential recommendation is to predict the next item $i \in \mathcal{I}$ that the user is most likely to interact with, by leveraging both sequential behavioral patterns and multi-modal content of items.

IV. Methodology

The overall framework of TAME, as illustrated in Fig. 1, is composed of three key modules: (1) Task-Aware Multi-Modal Retrieval module , which selectively extract recommendation-relevant signals from fine-grained features derived from pre-trained multi-modal encoders using an ID-aware modality retriever, and fuses them through a gating mechanism; (2) Modality-Guided ID Enrichment module, which enhances ID representations through an MoE architecture guided by retrieved fine-grained multi-modal representations; (3) Temporal User Modeling and Preference Prediction module, which captures users' evolving interests through sequential models to enable personalized next-item prediction.

A. Task-Aware Multi-Modal Retrieval

Since only a fraction of an item's multi-modal content is highly relevant to the recommendation task, existing methods that rely on pre-trained multi-modal encoders often utilize coarse-grained global representations that lack task awareness. As a result, they fail to accurately capture the specific content that appeals to users, and the crucial recommendation signals embedded in multi-modal information are easily overwhelmed by irrelevant noise.

To overcome this, the module selectively extracts fine-grained, task-relevant features from multi-modal inputs. Using the embedding of the item ID as a task-aware signal, the model can effectively and efficiently retrieve modality-specific features that are closely aligned with user intent, thus enhancing the performance of the downstream recommendation task.

1) Item Embedding Initialization: For an item i, we utilize its textual and visual modalities, along with its ID embedding, to characterize the semantics of the item from multiple perspectives. As for ID representations, we initialize an embedding table with random weights, which is updated during the training stage to capture latent item interaction relationships. For textual and visual modalities, we employ pre-trained textual and visual feature encoders, i.e., BERT [34] and ViT [35], to extract the corresponding fine-grained token-level and patch-level embeddings. The initialization process can be formulated as follows:

$$\mathbf{e}_i^{id} = \text{EmbeddingTable}(i), \mathbf{E}_i^t = \text{BERT}(t), \mathbf{E}_i^v = \text{ViT}(v) \quad (1)$$

where $\mathbf{e}_i^{id} \in \mathbb{R}^{1\times d}$, $\mathbf{E}_i^t = [\mathbf{e}_i^{t,cls}; \mathbf{e}_i^{t,1}; \ldots; \mathbf{e}_i^{t,N^t}] \in \mathbb{R}^{(N^t+1)\times d^t}$ and $\mathbf{E}_i^v = [\mathbf{e}_i^{v,1}; \ldots; \mathbf{e}_i^{v,N^v}; \mathbf{e}_i^{v,cls}] \in \mathbb{R}^{(N^v+1)\times d^v}$ denote the ID embedding, the fine-grained token-level and patch-level embeddings. Here, d, d^t and d^v are the embedding dimensions of ID, textual, and visual modalities, respectively, while n^t and n^v represent the numbers of text tokens and image patches. $e_i^{\mathcal{M},cls}$ denotes the embedding of the special classification token "[CLS]" used by the pre-trained encoders.

2) ID-Aware Modality Retriever : In multi-modal recommendation, a key challenge lies in identifying which parts of an item's textual and visual content are truly relevant to the recommendation task and influential to user preferences, since redundant information in multi-modal content can easily overshadow truly indicative signals [14], [16], [25]. Different from existing approaches that rely passively on coarse-grained global representations from pre-trained encoders, this module actively integrates item ID information into the query vectors to attend to fine-grained textual and visual features, enabling the model to selectively extract task-relevant signals while filtering out redundant information.

Inspired by BLIP-2 [36] and VLoRA [37], we design an ID-aware query-based modality retriever with cross-attention layer to selectively extract task-relevant multi-modal features. For the textual modality, we construct a learnable query embedding matrix $\mathbf{Q}^t = [\mathbf{q}^{t,1}; \ldots; \mathbf{q}^{t,N_q^t}] \in \mathbb{R}^{N_q^t\times d}$, where $\mathbf{q}^{t,j} \in \mathbb{R}^{1\times d}$ denotes the embedding vector of j-th query and N_q^t is the number of textual queries. Each query is designed to focus on different semantic facets of the textual content, enabling the model to extract diverse and fine-grained information that is potentially relevant to user preferences. The query matrix is then concatenated with the item's ID embedding $\mathbf{e}_i^{id}$ to inject task-specific signals, denoted as $\mathbf{H}_i = [\mathbf{Q}^t; \mathbf{e}_i^{\text{id}}] \in \mathbb{R}^{(N_q^t+1)\times d}$. To further enable the queries to interact with each other and adaptively exchange information conditioned on the item ID, we apply a self-attention layer over the concatenated query matrix as follows:

$$\tilde{\mathbf{H}}_i = \text{MHSA}(\mathbf{H}_i) = \text{Concat}(\text{head}_1^s, \text{head}_2^s, \ldots, \text{head}_{h^t}^s)\mathbf{W}^{s,O} \quad (2)$$

$$\text{head}_m^s = \text{Self-Attention}(\mathbf{H}_i\mathbf{W}_m^{s,Q},\ \mathbf{H}_i\mathbf{W}_m^{s,K},\ \mathbf{H}_i\mathbf{W}_m^{s,V}) \quad (3)$$

where $\text{MHSA}(\cdot)$ denotes the multi-head self-attention layer, $\text{Concat}(\cdot)$ denotes the concatenation operation and head_m^s is the output of the m-th attention head. The projection matrices $\mathbf{W}_m^{s,Q}, \mathbf{W}_m^{s,K}, \mathbf{W}_m^{s,V} \in \mathbb{R}^{d\times d_h}$ map the input features into head-specific query, key, and value spaces, where $d_h = d/h^t$ is the hidden size of each head with h^t being the number of attention heads of textual modality. $\tilde{\mathbf{H}}_i$ denotes the output of the self-attention mechanism, which can be split into an ID-aware query matrix and an updated ID embedding:

$$\tilde{\mathbf{Q}}^t = \tilde{\mathbf{H}}_i[1 : N_q^t], \quad \tilde{\mathbf{e}}_i^{\text{id}} = \tilde{\mathbf{H}}_i[N_q^t + 1] \quad (4)$$

In practice, only the ID-aware query matrix $\tilde{\mathbf{Q}}^t$ is used to retrieve recommendation-relevant information from the textual modalities, while $\tilde{\mathbf{e}}_i^{\text{id}}$ is discarded, as the ID embedding serves solely as a guidance signal to generate task-aware queries.

To extract modality-specific semantics aligned with recommendation intent, the ID-aware query matrix $\mathbf{Q}_i^t \in \mathbb{R}^{N_q^t\times d}$ is used to perform cross-attention over textual features $\mathbf{E}_i^t \in \mathbb{R}^{(N^t+1)\times d^t}$, extracted by the pre-trained textual encoder. The process can be formulated as:

$$\tilde{\mathbf{E}}_i^t = \text{MHCA}(\tilde{\mathbf{Q}}^t, \mathbf{E}_i^t) = \text{Concat}(\text{head}_1^c; \ldots; \text{head}_{h^t}^c)\mathbf{W}^{c,O} \quad (5)$$

$$\text{head}_n^c = \text{Cross-Attention}(\tilde{\mathbf{Q}}^t\mathbf{W}_n^{c,Q},\ \mathbf{E}_i^t\mathbf{W}_n^{c,K},\ \mathbf{E}_i^t\mathbf{W}_n^{c,V}) \quad (6)$$

where $\text{MHCA}(\cdot)$ denotes the multi-head cross-attention layer, and head_n^t is the output of the n-th cross-attention head. The projection matrices are defined as $\mathbf{W}_n^{c,Q} \in \mathbb{R}^{d\times d_h}$, and $\mathbf{W}_n^{c,K}, \mathbf{W}_n^{c,V} \in \mathbb{R}^{d^t\times d_h^t}$ with $d_h^t = d^t/h^t$. The output matrix $\tilde{\mathbf{E}}_i^t \in \mathbb{R}^{N_q^t\times d}$ encodes a set of fine-grained, ID-aware textual semantic representations for item i retrieved from the textual features extracted by the frozen pre-trained encoder. Each fine-grained textual semantic embedding vector $\tilde{\mathbf{e}}_{i,j}^t \in \tilde{\mathbf{E}}_i^t$ attends to a different part of the content, capturing diverse semantic facets conditioned on the item ID. Finally, a feed-forward network (FFN) with residual connection and layer normalization is applied to the intermediate representation to obtain the final textual embedding matrix $\hat{\mathbf{E}}_i^t$.

Similarly, we obtain the fine-grained visual embedding matrix $\hat{\mathbf{E}}_i^v \in \mathbb{R}^{N_q^v\times d}$ by applying a multi-head cross-attention

mechanism followed by a feed-forward network (FFN), using the ID-aware query matrix $\mathbf{Q}_i^v$ and the visual embeddings $\mathbf{E}_i^v$, where N_q^v denotes the number of visual queries.

The fine-grained textual and visual semantic matrices capture token- and patch-level content that is semantically aligned with the item's ID and the recommendation objective. By employing ID-aware modality retriever, the model is able to extract fine-grained and task-aware semantic signals from both modalities, ensuring that the retrieved features highlight the most informative facets of item content. These representations serve as modality-specific priors and will be further integrated to enhance ID embedding in the subsequent module.

3) Intra-modal Semantic Aggregation: After retrieving a set of fine-grained embedding vectors from each modality, we independently aggregate them to generate compact, modality-specific embeddings that support the recommendation task. Since different parts of an item's content may contribute unequally to user preferences, we adopt a gating mechanism to adaptively weight the semantic vectors based on their estimated importance, instead of treating them equally.

For the textual modality, given a set of retrieved fine-grained semantic vectors $\hat{\mathbf{E}}_i^t = [\hat{\mathbf{e}}_i^{t,1}; \ldots; \hat{\mathbf{e}}_i^{t,N_q^t}]$, we aggregate them via an adaptive weighted sum based on the gating mechanism:

$$\hat{\mathbf{e}}_i^t = \sum_{j=1}^{N_q^t} g_i^{t,j} \cdot \hat{\mathbf{e}}_i^{t,j} \tag{7}$$

where $\hat{\mathbf{e}}_i^t \in \mathbb{R}^{1\times d}$ denotes the aggregated textual embedding, and $g_i^{t,j}$ represents the gating weight of the j-th textual vector for item i from the gating vector $\mathbf{g}_i^t$. It is computed through a lightweight FFN to emphasize preference-relevant textual vectors, which can be computed by the following formulation:

$$\mathbf{g}_i^t = \text{Concat}(\hat{\mathbf{e}}_i^{t,1}, \ldots, \hat{\mathbf{e}}_i^{t,N_q^t})\mathbf{W}_g^t + \mathbf{b}_g^t \tag{8}$$

where $\mathbf{W}_g^t \in \mathbb{R}^{N_q^t \cdot d \times N_q^t}$, $\mathbf{b}_g^t \in \mathbb{R}^{1\times N_q^t}$ are the learnable weight and bias of the gating mechanism. The same process is applied to the retrieved fine-grained visual embedding vectors, yielding the aggregated visual embedding $\hat{\mathbf{e}}_i^v \in \mathbb{R}^{1\times d}$.

This gating-based aggregation enables the model to selectively preserve semantically informative features while suppressing noisy or redundant information for recommendation task, resulting in modality-specific representations that are compact yet expressive.

B. Modality-Guided ID Enrichment

In mainstream sequential recommendation settings [3], [6], [9], [38], each item is assigned a single embedding vector that is learned solely from interaction signals. However, such ID embeddings are often limited in expressiveness and fail to capture the multifaceted semantic content of items [23], [39], [40]. To address this, we propose to enhance the ID embedding by integrating multi-modal features through an MoE architecture.

Given an historical interaction sequence $\mathcal{S}_u$ of user u, we first obtain the corresponding ID embedding sequence and then add position embedding to preserve temporal order, which can be formulated as follows:

$$\mathbf{x}_{u,l}^{id} = \mathbf{e}_{u,l}^{id} + \mathbf{p}_l \tag{9}$$

where $\mathbf{p}_l \in \mathbb{R}^{1\times d}$ denotes the position embedding corresponding to the l-th position in the sequence.

Then, we apply an MoE module to the ID embedding of each item in the sequence, where the gating is guided by the retrieved fine-grained multi-modal features. This design enables fine-grained multi-modal signals to dynamically modulate expert aggregation, thereby producing enriched ID embeddings that are more expressive to better align with complex user intent. For ID embedding $\mathbf{x}_{u,l}^{id}$ in sequence, we transform it through N_q expert networks to produce a more expressive and adaptive representation, with $N_q = N_q^t + N_q^v$:

$$\hat{\mathbf{x}}_{u,l}^{id} = \sum_{k=1}^{N_q} g_{u,l}^{id,k} \cdot (\mathbf{w}^{id,k} \odot \mathbf{x}_{u,l}^{id}) \tag{10}$$

where $\hat{\mathbf{x}}_{u,l}^{id} \in \mathbb{R}^{1\times d}$ is the output enhanced ID embedding, $\mathbf{w}^{id,k} \in \mathbb{R}^{1\times d}$ represents the learnable weight of the k-th expert, $\odot$ denotes element-wise multiplication operation, and $g_{u,l}^{id,k}$ denotes the gating weight assigned to the k-th expert for item $i_{u,l}$.

Rather than relying on self-gating based solely on ID embeddings, we leverage the item's fine-grained multi-modal signals to guide expert aggregation, as these modality-specific representations reflect semantically salient content that influences user preferences and thus offer effective cues for context-aware transformation. Specifically, we concatenate the retrieved fine-grained textual and visual embedding vectors and apply an FFN to generate the gating vector for experts weighting:

$$\mathbf{z}_{u,l} = \text{Concat}(\hat{\mathbf{e}}_{u,l}^{t,1}, \ldots, \hat{\mathbf{e}}_{u,l}^{t,N_q^t}, \hat{\mathbf{e}}_{u,l}^{v,1}, \ldots, \hat{\mathbf{e}}_{u,l}^{v,N_q^v}) \tag{11}$$

$$\mathbf{g}_{u,l}^{id} = \text{Softmax}(\mathbf{z}_{u,l}\mathbf{W}_g^{id} + \mathbf{b}_g^{id}) \tag{12}$$

where $\mathbf{z}_{u,l} \in \mathbb{R}^{1\times N_q \cdot d}$ denotes the concatenated multi-modal embedding vector for the l-th item in the sequence, $\mathbf{W}_g^{id} \in \mathbb{R}^{N_q \cdot d \times N_q}$ and $\mathbf{b}_g^{id} \in \mathbb{R}^{1\times N_q}$ are the learnable weight and bias of the gating mechanism, and $\mathbf{g}_{u,l}^{id} \in \mathbb{R}^{N_q}$ denotes the gating weights over the N_q ID experts.

By leveraging the retrieved fine-grained multi-modal signals as gating cues, each expert network can focus on different semantic facets of items, enabling the model to capture nuanced item characteristics that are crucial for accurate user preference modeling. This module enables ID representations to be adaptively enhanced based on the corresponding multi-modal semantic context, resulting in more expressive embeddings that are better aligned with diverse user intent.

C. Temporal User Modeling and Preference Prediction

After obtaining the enhanced ID, textual, and visual embeddings for each item, we employ modality-specific Transformer models to capture the user's evolving preferences over time. Formally, for a user u with an interaction sequence $\mathcal{S}_u$,

we construct the corresponding modality-specific embedding sequences. Similar to the ID embeddings, we incorporate position embeddings into the textual and visual embeddings to preserve temporal order. In particular, for position l in the sequence, the resulting textual and visual embeddings are denoted as $\hat{\mathbf{x}}_{u,l}^{t}$ and $\hat{\mathbf{x}}_{u,l}^{v}$. The corresponding sequence of modality mm can be denoted as $\mathbf{S}_u^{mm} = [\hat{\mathbf{x}}_1^{mm}; \hat{\mathbf{x}}_2^{mm}; \ldots; \hat{\mathbf{x}}_{|\mathcal{S}_u|}^{mm}]$ with $mm \in \{id, t, v\}$. Each sequence $\mathbf{S}_u^{mm}$ is then independently encoded by a Transformer to capture temporal dependencies and modality-specific dynamic user interests as follows:

$$\mathbf{r}_u^{mm} = \text{Transformer}(\mathbf{S}_u^{mm}) \tag{13}$$

where $\mathbf{r}_u^{mm} \in \mathbb{R}^{1\times d}$ denotes the modality-specific preference representation of user u, obtained from the final hidden state at the last position in the sequence under modality mm. The interaction probability between user u and item i is calculated as follows:

$$\hat{y}_{ui} = \mathbf{r}_u^{id} \cdot \mathbf{e}_i^{id} + \mathbf{r}_u^{t} \cdot \hat{\mathbf{e}}_i^{t} + \mathbf{r}_u^{v} \cdot \hat{\mathbf{e}}_i^{v} \tag{14}$$

where $\hat{y}_{ui}$ the prediction score between user u and item i.

D. Optimization

To effectively train the proposed TAME model, we adopt a joint optimization strategy that combines a prediction loss and a contrastive alignment loss to enhance both recommendation accuracy and representation quality.

The primary objective of multi-modal sequential recommendation is to accurately predict the next item in a user's sequence. Given the prediction score $\hat{y}_{ui}$ and the ground-truth y_{ui}, the cross-entropy loss function is adopted as the main objective to optimize the model parameters:

$$\mathcal{L}_{CE} = -\sum_{i\in\mathcal{I}} y_{ui} \log(\hat{y}_{ui}) \tag{15}$$

Since the fine-grained multi-modal semantic vectors are extracted under the guidance of item ID embeddings, it is crucial to ensure that these ID representations remain well aligned with users' true preferences. In other words, the quality and consistency of ID embeddings directly affect the reliability and effectiveness of the ID-aware modality retriever and the overall recommendation performance.

To this end, we introduce a contrastive learning objective in the ID embedding space. This auxiliary task encourages the sequential representation of a user to be close to the ID embedding of their next interacted item, while being distinguishable from other items in the same batch.

Formally, for one batch $\mathcal{B}$ containing $|\mathcal{B}|$ sequences, given the sequence embedding of id modality $\mathbf{r}_u^{id}$ and the ID embedding of the ground-truth item $\mathbf{e}_i^{id}$, the contrastive loss is defined as:

$$\mathcal{L}_{CL} = -\log \frac{\exp(\text{sim}(\mathbf{r}_u^{id}, \mathbf{e}_i^{id})/\tau)}{\sum_{j=1}^{|\mathcal{B}|} \exp(\text{sim}(\mathbf{r}_u^{id}, \mathbf{e}_j^{id})/\tau)} \tag{16}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature hyper-parameter.

TABLE I: Statistics of the processed datasets.

Datasets	Toys	Sports	Beauty
# Users	19,412	35,598	22,363
# Items	11,924	18,357	12,101
# Actions	167,597	296,337	198,502
Avg. Actions/User	8.63	8.32	8.88
Avg. Actions/Item	14.06	16.14	16.40
Sparsity	99.93%	99.95%	99.93%

We combine the prediction loss and the contrastive alignment loss into a unified training objective:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{CL} \tag{17}$$

By jointly optimizing both objectives, the model is encouraged not only to make accurate predictions but also to learn reliable ID embeddings that align well with user preferences.

V. Experiment

The experiments are conducted to answer the following research questions:

- **RQ1:** How does TAME perform compared to state-of-the-art baselines?
- **RQ2:** How do key modules affect the performance of TAME?
- **RQ3:** How do different hyper-parameter settings impact the performance of TAME?

A. Experimental Settings

1) Datasets: To evaluate the performance of TAME and other SR models, we conduct experiments on three public datasets chosen from Amazon Review Datasets[1], which includes "Toys and Games" (Toys), "Sports and Outdoors" (Sports) and "Beauty". Following previous work [9], [13], [16], [41], we filter out users and items with fewer than five interactions to construct the 5-core subset for each dataset. For textual modality, the phrases of title, category and brand fields of each item are concatenated following prior studies [13], [14], [16], while for visual modality, the first image of each item is downloaded according to the URL in the metadata. The concrete statistics of the processed datasets are summarized in Table I

2) Baselines: To verify the effectiveness of TAME, we compare it with the following two categories of representative and state-of-the-art baselines: 1) ID-based sequential recommendation models, which solely model user-item interaction sequences based on item IDs, including GRU4Rec, SASRec, NextItNet and FEARec; 2) modality-based sequential recommendation models, which incorporate multi-modal features to enhance item representations, including UniSRec, VQ-Rec, TedRec, MMMLP, MISSRec, ODMT, IISAN and HM4SR. The details of the baselines are further described as follows:

- **GRU4Rec** [3] applies GRU to model user interaction sequence for session-based recommendation.

[1]https://jmcauley.ucsd.edu/data/amazon/

TABLE II: Performance comparison with different methods. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	Metrics	ID-Based Sequential Models				Modality-Based Sequential Models								Improv.
		GRU4Rec	SASRec	NextItNet	FEARec	UniSRec	VQ-Rec	TedRec	MMMLP	MISSRec	ODMT	IISAN	HM4SR	TAME	
Toys	Recall@5	0.0396	0.0620	0.0153	0.0536	0.0439	0.0419	0.0472	0.0607	0.0544	0.0581	0.0540	0.0678	**0.0723**	6.64%
	Recall@10	0.0558	0.0883	0.0269	0.0774	0.0693	0.0655	0.0712	0.0778	0.0842	0.0873	0.0795	0.0919	**0.1005**	9.36%
	Recall@20	0.0799	0.1210	0.0465	0.1065	0.1048	0.0920	0.1058	0.0989	0.1226	0.1222	0.1159	0.1214	**0.1400**	14.19%
	NDCG@5	0.0276	0.0352	0.0098	0.0306	0.0266	0.0218	0.0304	0.0460	0.0316	0.0388	0.0372	0.0492	**0.0505**	2.64%
	NDCG@10	0.0328	0.0437	0.0134	0.0383	0.0348	0.0294	0.0381	0.0515	0.0407	0.0481	0.0454	0.0571	**0.0596**	4.38%
	NDCG@20	0.0389	0.0520	0.0183	0.0457	0.0437	0.0361	0.0468	0.0568	0.0502	0.0570	0.0546	0.0645	**0.0696**	7.91%
Sports	Recall@5	0.0211	0.0316	0.0143	0.0280	0.0230	0.0280	0.0271	0.0298	0.0283	0.0305	0.0320	0.0326	**0.0354**	8.59%
	Recall@10	0.0350	0.0484	0.0232	0.0426	0.0385	0.0461	0.0429	0.0425	0.0445	0.0478	0.0487	0.0471	**0.0533**	9.45%
	Recall@20	0.0539	0.0715	0.0387	0.0634	0.0604	0.0667	0.0656	0.0600	0.0682	0.0699	0.0724	0.0672	**0.0766**	5.80%
	NDCG@5	0.0139	0.0173	0.0091	0.0150	0.0144	0.0158	0.0177	0.0212	0.0169	0.0203	0.0212	0.0228	**0.0240**	5.26%
	NDCG@10	0.0184	0.0227	0.0120	0.0197	0.0194	0.0216	0.0227	0.0253	0.0219	0.0258	0.0266	0.0274	**0.0297**	8.39%
	NDCG@20	0.0231	0.0285	0.0159	0.0249	0.0249	0.0268	0.0285	0.0297	0.0277	0.0313	0.0326	0.0325	**0.0356**	9.20%
Beauty	Recall@5	0.0418	0.0564	0.0283	0.0508	0.0349	0.0428	0.0483	0.0547	0.0510	0.0522	0.0580	0.0569	**0.0601**	3.62%
	Recall@10	0.0635	0.0842	0.0478	0.0765	0.0594	0.0675	0.0741	0.0765	0.0783	0.0805	0.0834	0.0785	**0.0853**	1.31%
	Recall@20	0.0915	0.1177	0.0741	0.1097	0.0941	0.0957	0.1092	0.1021	0.1177	0.1168	0.1215	0.1069	**0.1227**	0.99%
	NDCG@5	0.0280	0.0329	0.0174	0.0287	0.0218	0.0239	0.0324	0.0399	0.0299	0.0348	0.0380	0.0420	**0.0421**	0.24%
	NDCG@10	0.0349	0.0419	0.0236	0.0370	0.0297	0.0319	0.0408	0.0469	0.0382	0.0439	0.0462	0.0489	**0.0502**	2.66%
	NDCG@20	0.0420	0.0504	0.0302	0.0453	0.0384	0.0390	0.0497	0.0534	0.0479	0.0530	0.0558	0.0560	**0.0596**	6.43%

- **SASRec** [9] adopts self-attention mechanism to capture users' dynamic interests by adaptively attending to relevant items in their historical interaction sequences.
- **NextItNet** [42] employs dilated convolutional neural networks to model both short- and long-range item dependencies in user interaction sequences.
- **FEARec** [11] enhances self-attention with frequency-domain modeling to better capture sequential patterns.
- **UniSRec** [13] adopts a parametric whitening strategy and an MoE framework to learn universal textual representations, facilitating seamless transfer across different domains. For fair comparison, we train the model directly on the target dataset without the original cross-domain pre-training stage.
- **VQ-Rec** [43] learns transferable item representations by mapping item text to vector-quantized discrete codes. For fair comparison, we omit the pre-training stage and train it directly on the target dataset.
- **TedRec** [44] fuses text and ID features at the sequence level by applying Fast Fourier Transform in the frequency domain.
- **MMMLP** [20] is a purely MLP-based architecture that efficiently integrates multi-modal information.
- **MISSRec** [14] adopts a transformer-based encoder-decoder architecture, where the contextual encoder captures multi-modal sequential synergy and the interest-aware decoder models item-modality-interest interactions.
- **ODMT** [21] utilizes a Transformer to model multi-modal feature interactions and employs online distillation to enable mutual learning across multi-source data.
- **IISAN** [15] proposes a novel plug-and-play architecture with a decoupled fine-tuning strategy for the multi-modal encoder to better align with the recommendation task.
- **HM4SR** [16] introduces a hierarchical mixture-of-experts framework that integrates multi-modal features and explicit temporal signals to model dynamic user interests.

3) Evaluating Metrics: We evaluate the performance of all models using Recall@K and Normalized Discounted Cumulative Gain (NDCG@K), where K $\in \{5, 10, 20\}$. Following common practice in sequential recommendation, we adopt the leave-one-out evaluation protocol, where the last item in each user sequence is held out for testing, the second-to-last item for validation, and the remaining items are used for training. During evaluation, each test item is ranked against all items without candidate sampling to ensure a rigorous assessment of model performance.

4) Implementation Details: To ensure a fair comparison, we reproduce all baseline models within a unified pipeline. For the textual modality, features are extracted using the pre-trained "bert-base-uncased" model, while for the visual modality, features are extracted using the pre-trained "openai/clip-vit-base-patch32" model. For all methods, we set the embedding dimension d to 64. For other hyper-parameters in the baseline methods, we follow the settings reported in their original papers to ensure consistency and fairness. In our model, the numbers of textual query N_q^t and visual query N_q^v are selected from $\{1, 2, 4, 8, 16\}$ and the numbers of attention heads of textual retriever and visual retriever are chosen from $\{1, 2, 4, 8, 16\}$. We set the number of Transformer layers to 2 and the number of attention heads of Transformer to 2 by default. The model is optimized using the Adam optimizer

with a learning rate of 1e-3. To prevent overfitting, we adopt an early stopping strategy with a patience of 10 epochs.

B. Performance Comparison (RQ1)

We compare the proposed TAME approach against a broad range of baseline methods across all datasets and evaluation metrics. The results are presented in Table II, from which we draw the following observations:

Since ID-based sequential recommendation methods rely solely on discrete item IDs to model user interaction sequences, they inherently lack semantic grounding and struggle to generalize to unseen or infrequent items, particularly in cold-start and sparse interaction scenarios. These methods show relatively weak performance on the majority of datasets due to the absence of auxiliary modality features. Although self-attention-based methods such as SASRec and FEARec outperform RNN- and CNN-based models by capturing long-range dependencies, their effectiveness remains closely tied to the density of user-item interaction data.

Modality-based sequential recommendation baseline models, in most cases, achieve more competitive performance than ID-based sequential recommendation, since they incorporate auxiliary modality features such as item text and images to enrich item representations, enabling the model to capture richer semantic information beyond single ID representations.

For uni-modal sequential recommendation methods, i.e., UniSRec, VQ-Rec and TedRec, their performance is generally inferior to those multi-modal models, as they only leverage textual descriptions while ignoring other important modalities such as images. Furthermore, their performance sometimes falls behind strong ID-based models like SASRec and FEARec, mainly due to noise and redundancy in the text modality.

Multi-modal sequential recommendation methods that incorporate multi-modal features, tend to outperform ID-only and uni-modal baselines. By enriching item representations with complementary semantic cues, these models better capture user preferences and item characteristics, leading to improved recommendation performance. MMMLP and MISSRec employs MLP and Transformer architectures to jointly model multi-modal features, showing better performance than uni-modal and ID-based sequential recommendation models. Although ODMT employs fine-grained multi-modal feature fusion, it integrates all semantic information from multiple modalities without filtering, leading to the inclusion of redundant and noisy features. This indiscriminate fusion can interfere with the recommendation performance, resulting in only suboptimal performance. Despite these advances, they often rely on coarse-grained global multi-modal features derived from frozen pre-trained encoders, which can introduce task-irrelevant noise and degrade recommendation performance. IISAN adopts a decoupled parameter-efficient tuning strategy to fine-tune multi-modal encoders to extract task-relevant multi-modal features, while HM4SR explicitly models temporal dynamics via a hierarchical MoE framework, enabling better user interest modeling. Consequently, these two methods achieve optimal performance across most datasets and metrics.

It is evident that, compared to all baseline methods, the proposed TAME method consistently achieves the best performance across all datasets and evaluation metrics. This demonstrates the model's superior capability in extracting task-relevant information of recommendation from multi-modal features and effectively modeling users' historical interaction behaviors, thereby improving the understanding of both item semantics and user interests, and ultimately enhancing recommendation accuracy.

TABLE III: Ablation analysis results on three downstream datasets. "R@20" is short for Recall@20, and "N@20" is short for NDCG@20.

Models	Toys		Sports		Beauty	
	R@20	**N@20**	**R@20**	**N@20**	**R@20**	**N@20**
w/o modality retriever	0.1269	0.0646	0.0697	0.0328	0.1134	0.0573
w/o ID enrichment	0.1355	0.0564	0.0760	0.0303	0.1232	0.0545
w/o text	0.1219	0.0638	0.0647	0.0312	0.1118	0.0547
w/o vision	0.1319	0.0656	0.0712	0.0340	0.1170	0.0576
TAME	**0.1400**	**0.0696**	**0.0766**	**0.0356**	**0.1227**	**0.0596**

C. Ablation Study (RQ2)

To verify the effectiveness of each component of the proposed TAME model, we conduct comprehensive ablation studies on three downstream datasets. In particular, we implement four ablated variants as follows: 1) w/o modality retriever, which replaces the ID-aware modality retriever with an MLP; 2) w/o ID enrichment, which removes the modality-guided ID enrichment module; 3) w/o text, which removes the textual modality; 4) w/o vision, which removes the visual modality. The evaluation results are illustrated in Table III.

It can be observed that the performance of all variants drops to varying degrees after the removal of their corresponding components, which demonstrates the necessity and effectiveness of each module in the proposed TAME model.

Particularly, the variant w/o modality retriever results in a pronounced decline in overall recommendation performance, especially in Recall@20, as it significantly strengthens the model's ability to selectively extract task-relevant signals from massive multi-modal feature. In contrast, the variant w/o ID enrichment causes a substantial decline in NDCG@20, highlighting its role in improving ranking precision by refining ID representations under modality guidance. These results indicate that the two components play complementary roles: ID-aware modality retriever improves retrieval recall, while modality-guided ID enrichment module boosts ranking quality. It can be observed that both the variant w/o text and the variant w/o vision result in notable performance degradation, highlighting the importance of both modalities in the recommendation process. Notably, the variant w/o text exhibits a more pronounced decline than the variant w/o vision, indicating that textual information tends to convey richer and more

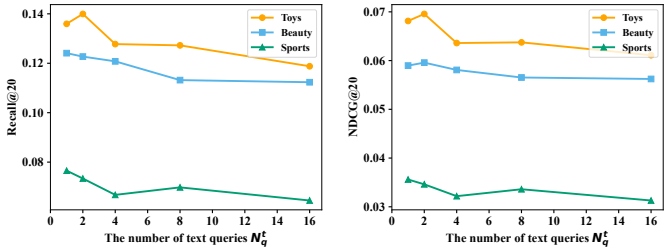

Fig. 2: Performance with different numbers of textual queries.

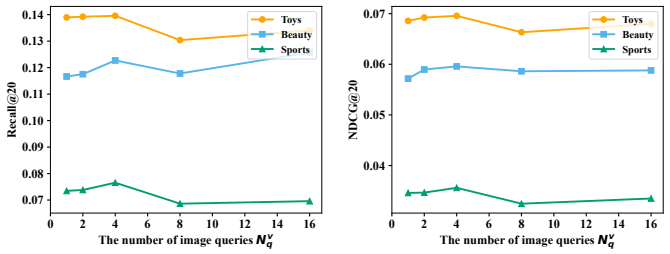

Fig. 3: Performance with different numbers of visual queries.

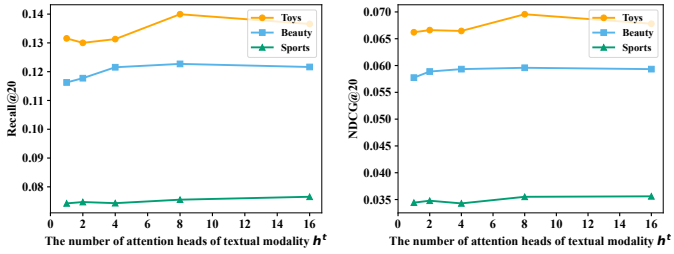

Fig. 4: Performance with different numbers of attention heads of textual retriever.

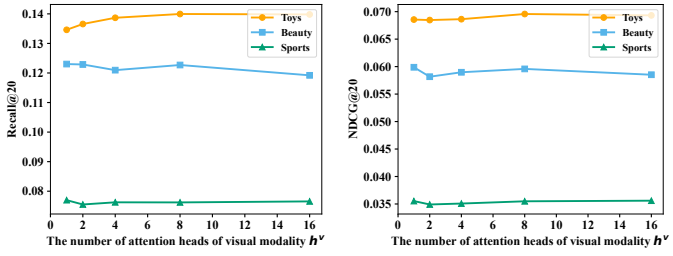

Fig. 5: Performance with different numbers of attention heads of visual retriever.

informative semantic cues, thereby playing a more pivotal role in capturing user preferences and decision-making.

D. Hyper-Parameter Study (RQ3)

We further analyze the impact of key hyper-parameters, including the numbers of textual queries N_q^t, visual queries N_q^v, attention heads of textual retriever h^t, and attention heads of visual retriever h^v. Each parameter is varied in $\{1, 2, 4, 8, 16\}$ while keeping other hyper-parameters fixed to ensure a controlled evaluation. The results are illustrated in Fig. 2, Fig. 3, Fig. 4 and Fig. 5.

It is observed in Fig. 2 that model performance is optimal when the number of textual queries N_q^t is set to 1 and 2. However, further increasing the number of textual queries N_q^t leads to a performance drop, possibly because the model struggles to focus on relevant task-specific signals when faced with too many queries. As illustrated in Fig. 3, the model exhibits stable performance as the number of visual queries N_q^v increases, with a slight improvement observed up to 4 queries. This suggests that visual modality is more dispersed than textual modality, and multiple queries help capture diverse task-relevant visual signals without introducing substantial noise.

As shown in Fig. 4, performance improves as the number of attention heads of textual retriever h^t increases from 2 to 8, suggesting enhanced ability to capture diverse semantic cues. However, further increasing h^t yields marginal gains or slight drops, likely due to attention dispersion. As illustrated in Fig. 5, increasing the number of attention heads of visual retriever h^v generally improves model performance, with notable gains observed when increasing from 1 to 8. This suggests that a larger number of attention heads allows the visual retriever to attend to more diverse visual patterns, enhancing its ability to capture fine-grained semantic signals. However, when h^v exceeds 8, model performance plateaus, suggesting that an excessive number of attention heads yields diminishing returns and may introduce representational redundancy. These results highlight the importance of choosing an appropriate number of attention heads to balance representation diversity and recommendation performance.

VI. Conclusion

In this paper, we propose TAME, a novel framework for multi-modal sequential recommendation that addresses the challenges of task-irrelevant noise in coarse-fined global multi-modal features and the limited expressiveness of ID embeddings. By proposing an ID-aware query-based modality retriever, we can selectively extract informative semantic cues from fine-grained multi-modal features obtained from frozen pre-trained encoders. Moreover, we enhance the expressiveness and adaptability of ID representations by integrating retrieved multi-modal signals into the ID embedding learning process via an MoE architecture, enabling context-aware and fine-grained representation refinement. Extensive experiments conducted on three public datasets demonstrate the superiority of the proposed TAME model. Ablation studies further reveal that the ID-aware modality retriever plays a pivotal role in this success by effectively identifying and extracting the most relevant fine-grained semantic information from massive multi-modal features, successfully avoiding the redundancy and noise interference inherent in traditional multi-modal fusion approaches, significantly enhancing recommendation accuracy and robustness.

REFERENCES

[1] S. Wang, L. Hu, Y. Wang, L. Cao, Q. Z. Sheng, and M. A. Orgun, "Sequential recommender systems: Challenges, progress and prospects," in *IJCAI*, 2019, pp. 6332–6338.

[2] M. Jing, Y. Zhu, T. Zang, and K. Wang, "Contrastive self-supervised learning in recommender systems: A survey," *ACM Trans. Inf. Syst.*, vol. 42, no. 2, pp. 59:1–59:39, 2024.

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

[4] X. Li, M. Zhang, S. Wu, Z. Liu, L. Wang, and P. S. Yu, "Dynamic graph collaborative filtering," in *ICDM*, 2020, pp. 322–331.

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

[6] J. Chang, C. Gao, Y. Zheng, Y. Hui, Y. Niu, Y. Song, D. Jin, and Y. Li, "Sequential recommendation with graph neural networks," in *SIGIR*, 2021, pp. 378–387.

[7] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *AAAI*, 2019, pp. 346–353.

[8] T. Peng, H. Yuan, Y. Zhang, Y. Li, P. Dai, Q. Wang, S. Wang, and W. Wu, "Tagrec: Temporal-aware graph contrastive learning with theoretical augmentation for sequential recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 5, pp. 3015–3029, 2025.

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

[10] Y. Hou, B. Hu, Z. Zhang, and W. X. Zhao, "CORE: simple and effective session-based recommendation within consistent representation space," in *SIGIR*, 2022, pp. 1796–1801.

[11] X. Du, H. Yuan, P. Zhao, J. Qu, F. Zhuang, G. Liu, Y. Liu, and V. S. Sheng, "Frequency enhanced hybrid attention network for sequential recommendation," in *SIGIR*, 2023, pp. 78–88.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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, pp. 3672–3682.

[17] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.

[18] Q. Liu, J. Zhu, Y. Yang, Q. Dai, Z. Du, X. Wu, Z. Zhao, R. Zhang, and Z. Dong, "Multimodal pretraining, adaptation, and generation for recommendation: A survey," in *KDD*, 2024, pp. 6566–6576.

[19] 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.

[20] 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.

[21] W. Ji, X. Liu, A. Zhang, Y. Wei, Y. Ni, and X. Wang, "Online distillation-enhanced multi-modal transformer for sequential recommendation," in *MM*, 2023, pp. 955–965.

[22] Y. Zhu, R. Xie, F. Zhuang, K. Ge, Y. Sun, X. Zhang, L. Lin, and J. Cao, "Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks," in *SIGIR*, 2021, pp. 1167–1176.

[23] M. Liu, S. Zhang, and C. Long, "Facet-aware multi-head mixture-of-experts model for sequential recommendation," in *WSDM*, 2025, pp. 127–135.

[24] Y. Liu, E. Yang, Y. Dang, G. Guo, Q. Liu, Y. Liang, L. Jiang, and X. Wang, "ID embedding as subtle features of content and structure for multimodal recommendation," *CoRR*, vol. abs/2311.05956, 2023.

[25] G. Xv, X. Li, R. Xie, C. Lin, C. Liu, F. Xia, Z. Kang, and L. Lin, "Improving multi-modal recommender systems by denoising and aligning multi-modal content and user feedback," in *KDD*, 2024, pp. 3645–3656.

[26] W. Zhao, S. Zhong, Y. Liu, W. Wen, J. Qin, M. Liang, and Z. Huang, "DVIB: towards robust multimodal recommender systems via variational information bottleneck distillation," in *WWW*, 2025, pp. 2549–2561.

[27] Y. Li, H. Du, Y. Ni, P. Zhao, Q. Guo, F. Yuan, and X. Zhou, "Multi-modality is all you need for transferable recommender systems," in *ICDE*, 2024, pp. 5008–5021.

[28] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Using sequential and non-sequential patterns in predictive web usage mining tasks," in *ICDM*, 2002, pp. 669–672.

[29] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *WWW*, 2010, pp. 811–820.

[30] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J. Wen, "S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization," in *CIKM*, 2020, pp. 1893–1902.

[31] J. Zhang, R. Xie, H. Lu, W. Sun, W. X. Zhao, Y. Chen, and Z. Kang, "Frequency-augmented mixture-of-heterogeneous-experts framework for sequential recommendation," in *WWW*, 2025, pp. 2596–2605.

[32] X. Xie, F. Sun, Z. Liu, S. Wu, J. Gao, J. Zhang, B. Ding, and B. Cui, "Contrastive learning for sequential recommendation," in *ICDE*, 2022, pp. 1259–1273.

[33] Z. Yuan, F. Yuan, Y. Song, Y. Li, J. Fu, F. Yang, Y. Pan, and Y. Ni, "Where to go next for recommender systems? ID- vs. modality-based recommender models revisited," in *SIGIR*, 2023, pp. 2639–2649.

[34] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[35] 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.

[36] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023, pp. 19 730–19 742.

[37] F. Ma, H. Xue, Y. Zhou, G. Wang, F. Rao, S. Yan, Y. Zhang, S. Wu, M. Z. Shou, and X. Sun, "Visual perception by large language model's weights," in *NeurIPS*, 2024.

[38] 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.

[39] 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.

[40] M. Choi, H. Kim, H. Cho, and J. Lee, "Multi-intent-aware session-based recommendation," in *SIGIR*, 2024, pp. 2532–2536.

[41] Y. Xie, P. Zhou, and S. Kim, "Decoupled side information fusion for sequential recommendation," in *SIGIR*, 2022, pp. 1611–1621.

[42] F. Yuan, A. Karatzoglou, I. Arapakis, J. M. Jose, and X. He, "A simple convolutional generative network for next item recommendation," in *WSDM*, 2019, pp. 582–590.

[43] Y. Hou, Z. He, J. J. McAuley, and W. X. Zhao, "Learning vector-quantized item representation for transferable sequential recommenders," in *WWW*, 2023, pp. 1162–1171.

[44] L. Xu, Z. Tian, B. Li, J. Zhang, D. Wang, H. Wang, J. Wang, S. Chen, and W. X. Zhao, "Sequence-level semantic representation fusion for recommender systems," in *CIKM*, 2024, pp. 5015–5022.