

# MMQ-v2: Align, Denoise, and Amplify: Adaptive Behavior Mining for Semantic IDs Learning in Recommendation

Yi Xu

Alibaba International Digital  
Commerce Group  
Beijing, China  
xy397404@alibaba-inc.com

Moyu Zhang

Alibaba International Digital  
Commerce Group  
Beijing, China  
zhangmoyu.zmy@alibaba-inc.com

Chaofan Fan

Alibaba International Digital  
Commerce Group  
Beijing, China  
fanchaofan.fcf@alibaba-inc.com

Jinxin Hu\*

Alibaba International Digital  
Commerce Group  
Beijing, China  
jinxin.hjx@alibaba-inc.com

Xiaochen Li

Alibaba International Digital  
Commerce Group  
Beijing, China  
xingke.lxc@alibaba-inc.com

Yu Zhang

Alibaba International Digital  
Commerce Group  
Beijing, China  
daoji@alibaba-inc.com

Xiaoyi Zeng

Alibaba International Digital  
Commerce Group  
Beijing, China  
yuanhan@taobao.com

Jing Zhang

Wuhan University, School of  
Computer Science  
Wuhan, China  
jingzhang.cv@gmail.com

## Abstract

Industrial recommender systems rely on unique Item Identifiers (ItemIDs). However, this method struggles with scalability and generalization in large, dynamic datasets that have sparse long-tail data. Content-based Semantic IDs (SIDs) address this by sharing knowledge through content quantization. However, by ignoring dynamic behavioral properties, purely content-based SIDs have limited expressive power. Existing methods attempt to incorporate behavioral information but overlook a critical distinction: unlike relatively uniform content features, user-item interactions are highly skewed and diverse, creating a vast information gap in quality and quantity between popular and long-tail items. This oversight leads to two critical limitations: (1) Noise Corruption: Indiscriminate behavior-content alignment allows collaborative noise from long-tail items to corrupt their content representations, leading to the loss of critical multimodal information. (2) Signal Obscurity: The equal-weighting scheme for SIDs fails to reflect the varying importance of different behavioral signals, making it difficult for downstream tasks to distinguish important SIDs from uninformative ones. To tackle these issues, we propose a mixture-of-quantization framework, **MMQ-v2**, to adaptively **Align**, **Denoise**, and **Amplify** multimodal information from content and behavior modalities for

semantic IDs learning. The semantic IDs generated by this framework named **ADA-SID**. It introduces two innovations: an adaptive behavior-content alignment that is aware of information richness to shield representations from noise, and a dynamic behavioral router to amplify critical signals by applying different weights to SIDs. Extensive experiments on public and large-scale industrial datasets demonstrate ADA-SID's significant superiority in both generative and discriminative recommendation tasks.

## CCS Concepts

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

## Keywords

Recommendation System; Semantic ID; Vector Quantization; Item Alignment;

## ACM Reference Format:

Yi Xu, Moyu Zhang, Chaofan Fan, Jinxin Hu, Xiaochen Li, Yu Zhang, Xiaoyi Zeng, and Jing Zhang. 2018. MMQ-v2: Align, Denoise, and Amplify: Adaptive Behavior Mining for Semantic IDs Learning in Recommendation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

\*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Conference acronym 'XX, Woodstock, NY

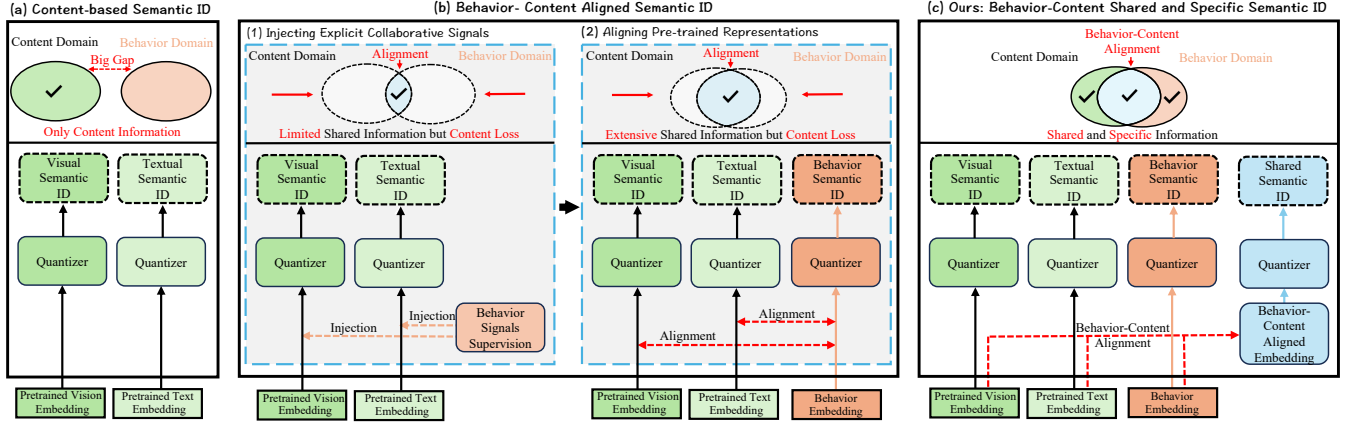
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

## 1 Introduction

Recommender systems traditionally represent items using unique identifiers (ItemIDs), but this approach struggles with large, dynamic corpora where item popularity is skewed and long-tailed, limiting scalability and generalization[1, 2, 7]. Content-based Semantic IDs (SIDs) partly mitigate this by quantizing multimodal content so that similar items can share similar identifiers[3, 22], as shown in Fig. 1 (a). However, user-item interactions induce dynamic behavioral properties (e.g., evolving popularity, style shifts,



**Figure 1: Illustration of SIDs Generation Paradigm.** (a) Content-based SIDs: Quantize multimodal item content into SIDs. (b) Behavior-content aligned SIDs: Incorporate collaborative signals by injecting explicit signals or aligning pre-trained representations. (c) Ours: Behavior-content shared and specific SIDs: Learn both shared and modality-specific behavior-content representations for SIDs.

cohort-specific preferences) that content alone cannot capture, creating a performance ceiling for content-only SIDs[20, 30, 34]. Therefore, recent work incorporates collaborative signals, and behavior-content alignment has become a prevailing approach.[20, 32, 34]. This approach, which encourages behaviorally similar items to have similar SIDs, is implemented either via injecting explicit collaborative signals[42, 43] or via aligning pre-trained representations[17, 27, 30], as shown in Fig. 1 (b).

Despite this progress, existing alignment methods fail to adequately address the inherent disparity between behavioral information and content features. Specifically, the extreme sparsity and skewed distribution of user-item interactions create a vast gap in the quality and quantity of user behaviors between popular and the long-tail items[8–11]. This mismatch leads to two critical flaws in current approaches: (1) Noise Corruption: The indiscriminate alignment is doubly detrimental: For long-tail items with sparse interactions, it introduces collaborative noise that corrupts their reliable content representations. Conversely, for popular items, which have rich and diverse behavioral patterns, forcing a uniform alignment overcompresses this complex information, leading to the loss of their unique behavioral signatures. (2) Signal Obscurity: Equal weighting increases the optimization burden and redundancy from low-information SIDs, preventing downstream models from prioritizing informative ones and thereby degrading recommendation accuracy.

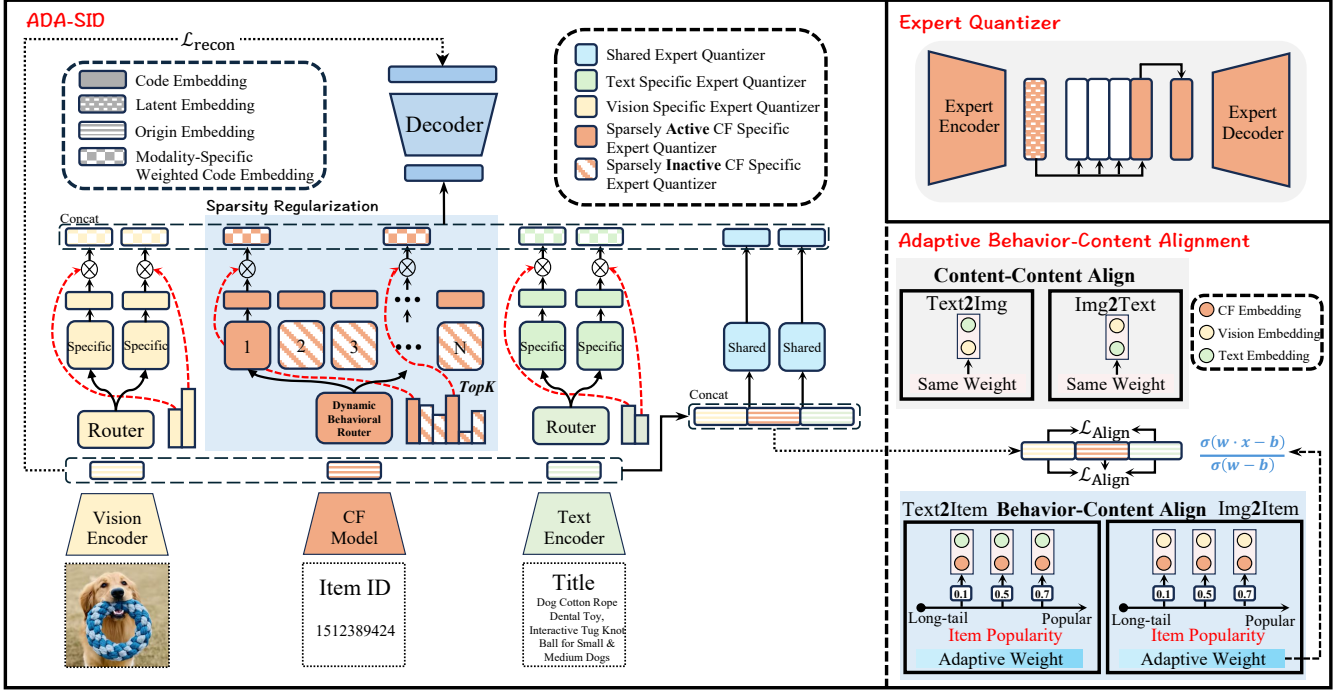
To tackle these challenges, we present ADA-SID, a framework that adaptively aligns, denoises, and amplifies multimodal signals for SID learning. ADA-SID proposes a behavior-content mixture-of-quantization network to simultaneously capture both shared and behavior-specific information and generate parallel, multi-view SIDs for items, as shown in Fig.1 (c). This framework comprises two key components: (i) adaptive tri-modal (behavior-vision-text) alignment strategy to fuse the dynamic behavioral modality with the static content modality (comprising visual and text). Recognizing that behavioral signals vary in reliability, this strategy employs

an alignment strength controller to dynamically adjust the alignment intensity based on the richness of an item’s interaction data; and (ii) a dynamic behavioral router that learns importance weights over SIDs, enabling the model to amplify critical collaborative signals and attenuate uninformative ones to improve downstream recommendation performance. The overall framework of ADA-SID is shown in Fig.2 Our contributions are as follows:

- To the best of our knowledge, we are the first to customize behavior-content multimodal SIDs for items according to the information richness of their collaborative signals, thereby enhancing the expressiveness of SIDs and improving generalization in downstream recommendation tasks.
- We propose an adaptive behavior-content alignment mechanism that dynamically calibrates behavior-content alignment strength, mitigating noise corruption for long-tail items while preserving the diverse behavioral information for popular ones.
- We propose a dynamic behavioral router that learns to assign adaptive weights to an item’s set of behavioral SIDs. This mechanism effectively amplifies critical collaborative signals.
- Extensive offline experiments and online A/B tests across generative and discriminative recommendation tasks demonstrate ADA-SID effectiveness, scalability, and versatility.

## 2 Related Works

Traditional ItemIDs limit the generalization of recommender systems due to their lack of semantics[5, 24, 35, 43–45]. To address this, content-based SIDs quantize item content features into discrete codes. TIGER pioneered this with RQ-VAE, while later works like SPM-SID [1] and PMA [2], inspired by Large Language Models, explored more granular subword-based compositions. To incorporate dynamic behavioral information into SIDs, recent work can be broadly categorized into two approaches. The first approach involves injecting explicit collaborative signals. For instance, LC-Rec[31] designs a series of alignment tasks to unify semantic and collaborative information. ColaRec[32] distills collaborative signals



**Figure 2: ADA-SID framework: We (i) use a sparse MoE-based quantization network to learn shared and modality-specific behavior-content representations, (ii) apply an adaptive behavior-content alignment mechanism that dynamically calibrates behavior-content alignment strength, and (iii) design a dynamic behavioral router that learns to assign adaptive weights to SIDs.**

directly from a pretrained recommendation model and combines them with content information. IDGenRec[34] leverages LLMs to generate semantically rich textual identifiers, showing strong potential in zero-shot settings. The second approach focuses on aligning pre-trained representations. To this end, recent methods introduce pretrained collaborative representations and align content representations with them. For example, EAGER[30] generates separate collaborative and content SIDs using K-means on pre-trained embeddings and then aligns them in downstream tasks. DAS[42] employs multi-view contrastive learning to maximize the mutual information between SIDs and collaborative signals. LETTER[27] integrates hierarchical semantics, collaborative signals and code assignment diversity to generate behavior-content fused SIDs with RQ-VAE. MM-RQ-VAE[17] generate collaborative SIDs, textual SIDs and visual SIDs with pre-trained collaborative embeddings and multimodal embeddings, and introduce contrastive learning for behavior-content alignment.

### 3 Methodology

This section details the proposed ADA-SID framework. As its name suggests, ADA-SID is designed to Align, Denoise, and Amplify multimodal information for Semantic ID learning. First, we formulate the task of semantic ID generation in Section 3.1. Second, the overall architecture of behavior-content mixture-of-quantization network is introduced in Section 3.2. Third, to align with behavior domain and suppress the influence of noise, we propose the

adaptive behavior-content alignment. Fourth, to amplify critical collaborative signals, we propose the dynamic behavioral router mechanism.

#### 3.1 Problem Formulation

The item tokenizer is designed to quantize the pretrained textual, visual, and behavioral embeddings of each item into a sequence of discrete SIDs. Formally, for a given item, we first leverage pretrained vision and text embedding models to obtain its pretrained vision embedding  $e_v$  and pretrained text embedding  $e_t$ . The pretrained behavioral embedding  $e_b$  is obtained from SASRec with collaborative signals. The item tokenizer  $\mathcal{T}_{\text{item}}$  then quantizes these high-dimensional embeddings into a discrete sequence of SIDs.

$$\text{Semantic\_IDs} = (c_1, c_2, \dots, c_l) = \mathcal{T}_{\text{item}}([e_t, e_v, e_b]) \quad (1)$$

where  $l$  is the length of the SIDs,  $c_i$  is the  $i$ -th semantic ID.

#### 3.2 Behavior-Content Mixture-of-Quantization Network

To simultaneously capture both behavior-content shared and specific information, we propose the behavior-content mixture-of-quantization network, where shared experts learn shared information across behavior and content modalities, and specific experts focus on modality-specific information.

**Shared Experts:** The shared experts are designed to quantize the aligned behavior-content information into shared latent embeddings, which are then used to generate shared SIDs. For a given

item, its pretrained textual, visual, and behavioral embeddings are first projected into a unified high-dimensional space with small two-layer deep neural networks, denotes  $D_t, D_v, D_b$  respectively. The hidden representations are denoted as  $\mathbf{h}_t, \mathbf{h}_v, \mathbf{h}_b$ .

$$\mathbf{h}_t = D_t(\mathbf{e}_t), \mathbf{h}_v = D_v(\mathbf{e}_v), \mathbf{h}_b = D_b(\mathbf{e}_b) \quad (2)$$

$$\mathbf{h} = [\mathbf{h}_t, \mathbf{h}_v, \mathbf{h}_b] \quad (3)$$

To learn the aligned behavior-content information, these projected hidden representations are optimized by the adaptive behavior-content alignment mechanism, detailed in Section 3.3. For the  $i$ -th shared expert  $E_{s,i}$ , the hidden representations  $\mathbf{h}$  are encoded into a shared latent embedding  $\mathbf{z}_{s,i}$  and quantized into a discrete semantic ID, with an associated shared codebook  $C_{s,i} = \{\mathbf{z}_{q,k}\}_{k=1}^K$ , where  $i \in \{1, \dots, N_s\}$ ,  $N_s$  denotes the number of shared experts, and  $K$  indicates the codebook size. The most similar codeword is searched by maximizing the cosine distance between  $\mathbf{z}_{s,i}$  and all codewords in  $C_{s,i}$ , as formulated in Eq.5.

$$\mathbf{z}_{s,i} = E_{s,i}(\mathbf{h}) \quad (4)$$

$$c_{s,i} = \arg \max_{j \in \{1, \dots, K\}} \frac{\mathbf{z}_{s,i}^\top \mathbf{z}_{q,j}}{\|\mathbf{z}_{s,i}\| \|\mathbf{z}_{q,j}\|}, \quad i = 1, \dots, N_s, \quad (5)$$

**Specific Experts:** The specific experts are design to learn specific information of each modality and generate modality-specific SIDs. For each modality, there is a group of modality-specific experts and corresponding modality-specific codebooks. For example, for the textual modality, a set of dedicated experts  $\{E_{t,i}\}_{i=1}^{N_t}$  transforms the original pretrained embedding  $\mathbf{e}_t$  into a corresponding set of latent representations  $\{\mathbf{z}_{t,i}\}_{i=1}^{N_t}$ . The textual SIDs  $\{c_{t,i}\}_{i=1}^{N_t}$  is searched from codebooks  $\{C_{t,i}\}_{i=1}^{N_t}$  with cosine distance, the searched codeword representations are denote as  $\{\mathbf{z}_{q,t,i}\}_{i=1}^{N_t}$ . Analogously, the latent representations of visual and behavior modality as denoted as  $\{\mathbf{z}_{v,i}\}_{i=1}^{N_v}, \{\mathbf{z}_{b,i}\}_{i=1}^{N_b}$ . The visual and behavior SIDs are  $\{c_{v,i}\}_{i=1}^{N_v}, \{c_{b,i}\}_{i=1}^{N_b}$ , the searched codeword representations are denote as  $\{\mathbf{z}_{q,v,i}\}_{i=1}^{N_v}, \{\mathbf{z}_{q,b,i}\}_{i=1}^{N_b}$ , the number of textual and behavior experts are  $N_t, N_b$  separately.

$$\mathbf{z}_{t,i} = E_{t,i}(\mathbf{e}_t), \mathbf{z}_{v,i} = E_{v,i}(\mathbf{e}_v), \mathbf{z}_{b,i} = E_{b,i}(\mathbf{e}_b) \quad (6)$$

The decoder then reconstructs from the fused latent representations and codeword representations to the fused pre-trained embeddings  $\mathbf{e} = [\mathbf{e}_t, \mathbf{e}_v, \mathbf{e}_b]$ , which are formulated as follows.

$$\mathbf{z} = \sum_{i=1}^{N_s} \mathbf{z}_{s,i} + \sum_{i=1}^{N_v} g_{v,i} \mathbf{z}_{v,i} + \sum_{i=1}^{N_t} g_{t,i} \mathbf{z}_{t,i} + \sum_{i=1}^{N_b} g_{b,i} \mathbf{z}_{b,i} \quad (7)$$

$$\mathbf{z}_q = \sum_{i=1}^{N_s} \mathbf{z}_{q,s,i} + \sum_{i=1}^{N_v} g_{v,i} \mathbf{z}_{q,v,i} + \sum_{i=1}^{N_t} g_{t,i} \mathbf{z}_{q,t,i} + \sum_{i=1}^{N_b} R(\mathbf{e}_b)_i \mathbf{z}_{q,b,i} \quad (8)$$

$$g_t = \text{softmax}(MLP_t(\mathbf{e}_t) + b_t) \quad (9)$$

$$g_v = \text{softmax}(MLP_v(\mathbf{e}_v) + b_v) \quad (10)$$

$$\mathcal{L}_{recon} = \|\mathbf{e} - \text{decoder}(\mathbf{z} + sg(\mathbf{z}_q - \mathbf{z}))\|^2 \quad (11)$$

In addition, for the dynamic behavioral information, we employ a sparse-activated router  $R(\mathbf{e}_b)$  to determine the importance of the behavioral SIDs, as detailed in Section 3.4.

### 3.3 Adaptive Behavior-Content Alignment

Given the variation in item's behavioral information richness, treating these signals indiscriminately in behavior-content contrastive learning can introduce noise into the modeling of shared information.

**Alignment Strength Controller** The alignment strength controller outputs a weight to modulate the intensity of the behavior-content alignment for each item. This controller is designed based on two principles. The alignment strength for long-tail items should smoothly decay toward zero, whereas the alignment strength for popular items should increase with their estimated information richness. Based on these two considerations, we use the L2-magnitude of an item's behavioral embedding as a proxy for its information richness. Specifically, consider the pretrained behavior embedding matrix  $\mathbf{E} \in \mathbb{R}^{K \times D}$ , consisting of  $K$  vectors  $\{\mathbf{e}_{b,1}, \mathbf{e}_{b,2}, \dots, \mathbf{e}_{b,K}\}$ . For the  $j$ -th embedding  $\mathbf{e}_{b,j}$ , the alignment strength controller is formulated as follows.

$$N_{\max} = \max_{i \in \{1, \dots, K\}} (\|\mathbf{e}_{b,i}\|_2), N_{\min} = \min_{i \in \{1, \dots, K\}} (\|\mathbf{e}_{b,i}\|_2) \quad (12)$$

$$N_{\text{norm}}(\mathbf{e}_{b,j}) = \frac{\|\mathbf{e}_{b,j}\|_2 - N_{\min}}{N_{\max} - N_{\min}} \quad (13)$$

$$w = \frac{\sigma(\alpha N_{\text{norm}}(\mathbf{e}_{b,j}) - \beta)}{\sigma(\alpha - \beta)} \quad (14)$$

where  $\alpha$  and  $\beta$  are hyperparameters that jointly determine the steepness of the curve and the threshold for distinguishing long-tail items. By tuning  $\alpha$  and  $\beta$ , the function can adapt to different data distributions. In our experiments, the optimal setting was found to be  $\alpha = 10$  and  $\beta = 9$ .

**Behavior-Content Contrastive Learning** To learn the shared information between behavior and content, we adopt a two-stage process. First, contrastive learning is used to align the text and image modalities to obtain a unified content representation  $\mathbf{h}_c = \mathbf{h}_t + \mathbf{h}_v$ . Then, contrastive learning is performed between  $\mathbf{h}_c$  and the behavioral representation  $\mathbf{h}_b$  to maximize the mutual information between the content domain and the behavior domain. Specifically, for a given item, its content representation and behavioral representation form a positive pair  $\langle \mathbf{h}_b, \mathbf{h}_c \rangle$ . The content representations of all other items in the batch form negative pairs  $\langle \mathbf{h}_t, \mathbf{h}_{c_i}^- \rangle$ , where  $i = 1, \dots, B$  and  $B$  is the batch size. The formulation is as follows.

$$\begin{aligned} \mathcal{L}_{content} = & -\log \frac{\exp(\text{sim}(\mathbf{h}_t, \mathbf{h}_{v^+})/\tau)}{\exp(\text{sim}(\mathbf{h}_t, \mathbf{h}_{v^+})/\tau) + \sum_{i=1}^{B-1} \exp(\text{sim}(\mathbf{h}_t, \mathbf{h}_{v_i^-})/\tau)} \\ & -\log \frac{\exp(\text{sim}(\mathbf{h}_v, \mathbf{h}_{t^+})/\tau)}{\exp(\text{sim}(\mathbf{h}_v, \mathbf{h}_{t^+})/\tau) + \sum_{i=1}^{B-1} \exp(\text{sim}(\mathbf{h}_v, \mathbf{h}_{t_i^-})/\tau)} \end{aligned} \quad (15)$$

$$\mathcal{L}_{align} = -\log \frac{\exp(\text{sim}(\mathbf{h}_b, \mathbf{h}_{c^+})/\tau)}{\exp(\text{sim}(\mathbf{h}_b, \mathbf{h}_{c^+})/\tau) + \sum_{i=1}^{B-1} \exp(\text{sim}(\mathbf{h}_b, \mathbf{h}_{c_i^-})/\tau)} \quad (16)$$

$$\mathcal{L}_{align\_total} = \mathcal{L}_{content} + w \mathcal{L}_{align} \quad (17)$$

where  $\tau$  is the temperature coefficient, and  $\tau = 0.07$  in our experiment,  $\text{sim}(\cdot, \cdot)$  is the cosine similarity.

### 3.4 Dynamic Behavioral Router Mechanism

An item's interaction frequency directly determines the information richness of its behavioral representation. Consequently, popular items with frequent interactions yield rich, reliable signals, rendering their behavioral SIDs highly important. In contrast, long-tail items with sparse interactions produce uninformative representations, diminishing their SIDs' importance.

**Behavior-Guided Dynamic Router** The behavior-guided dynamic router assigns calibrated importance scores to behavioral Semantic IDs, up-weighting head items and down-weighting long-tail ones, and dynamically adjusts the weights based on information richness. As formulated in Eq.18, we propose a learnable gate where an MLP processes the representation  $\mathbf{e}_b$  to capture its specific semantic patterns.

$$R(\mathbf{e}_b) = \sigma(N_{\text{norm}}(\mathbf{e}_b)) * \text{relu}(\text{MLP}(\mathbf{e}_b) + b) \quad (18)$$

Here, the MLP extracts behavior-specific semantics; ReLU induces exact-zero sparsity; and the magnitude-based scaler  $\sigma(\cdot)$  maps weights to  $[0,1]$  and calibrates them by information richness. Trained end-to-end without manual thresholds, this gate amplifies critical collaborative signals and attenuate uninformative ones, leading to improved robustness.

**Sparsity Regularization** To further refine the dynamic behavior router, we introduce a sparsity regularization loss. The goal is to encourage the router to produce sparser SID sequences (i.e., activate fewer behavioral SIDs) for long-tail items and denser sequences for popular items. To achieve this, we define an item-specific target sparsity, which is inversely proportional to the item's information richness, approximated by the L2-magnitude of  $\mathbf{e}_b$ . The regularization loss  $L_{\text{reg}}$  then penalizes the deviation from this target. In sparse mixture-of-experts (MoE) designs, load imbalance is a significant issue that can lead to routing collapse. To address this, we incorporate a load-balancing mechanism into the framework, as formalized in Eq.23.

$$L_{\text{reg}_i} = \lambda_i \frac{1}{B} \sum_t \sum_j^{N_b} f_{ib} \|R(\mathbf{e}_b)_j\|_1 \quad (19)$$

$$\lambda_i = \lambda_{i-1} \alpha^{\text{sign}(s_{\text{target}} - s_{\text{current}})} \quad (20)$$

$$s_{\text{current}} = 1 - \sum_j^{N_b} \mathbf{1}\{R(\mathbf{e}_b)_j > 0\} / N_b \quad (21)$$

$$s_{\text{target}} = \theta * (1 - N_{\text{norm}}(\mathbf{e}_b) / (N_{\text{max}} - N_{\text{min}})) \quad (22)$$

$$f_{ib} = \frac{1}{(1 - s_{\text{target}})B} \sum_t \sum_j^{N_b} \mathbf{1}\{R(\mathbf{e}_b)_j > 0\} / N_b \quad (23)$$

In summary, the dynamic router, guided by sparsity and load-balancing regularization, produces a flexible and semantically rich representation. It captures diverse item facets while adaptively controlling the length and complexity of the representation to align with both the item's intrinsic properties and the demands of downstream tasks.

## 4 Experiments

In this paper, we conduct extensive experiments on both industrial and public datasets to evaluate the effectiveness of our proposed framework and address the following questions:

- **RQ1:** How does ADA-SID compare to state-of-the-art item tokenizers in terms of reconstruction accuracy and downstream performance in generative and discriminative recommendation?
- **RQ2:** What is the contribution of each component in ADA-SID?
- **RQ3:** How sensitive is ADA-SID's performance to its key hyperparameters, particularly the loss weights and the sparsity regularization strength?
- **RQ4:** How effective is our proposed ADA-SID in improving recommendations for items with varying degrees of popularity, especially for those in the long-tail items?

**Table 1: Statistics of Industrial and Public Datasets.**

Dataset	Industrial Dataset	Beauty
#User	35,154,135	22,363
#Item	48,106,880	12,101
#Interaction	75,730,321,793	198,360

### 4.1 Experimental Setup

**4.1.1 Dataset.** We evaluate the proposed framework both on an industrial dataset and a public dataset.

**Industrial Dataset:** This dataset was collected from a leading e-commerce advertising platform in Southeast Asia between October 2024 and May 2025, encompassing 30 million users and 40 million advertisements. It contains user behavior sequences with an average length of 128, alongside rich, multimodal item content (images, titles, descriptions, etc.). Its scale and complexity make it an ideal benchmark for evaluating real-world performance.

**Public Dataset:** We conduct experiments on the "Beauty" subset of the Amazon Product Reviews dataset [38], the statistic is shown in Table 1. For generative retrieval, we apply a 5-core filter and construct chronological user sequences with a maximum length of 20. For discriminative ranking, we binarize ratings (positive: >3, negative: <3 and =3) and use a chronological 90%/10% split for training and testing.

**4.1.2 Evaluation Metrics.** : We evaluate the effectiveness of proposed ADA-SID from both quantization metrics and recommendation metrics.

#### Quantization Metrics:

- **Reconstruction Loss** [23] is utilized to evaluate the reconstruction fidelity for the origin input vector.
- **Token Distribution Entropy** [41] is utilized to evaluate the diversity and balance of the distribution across semantic codewords in codebooks.
- **Codebook utilization** [40] is employed to reflect the efficiency with which the model uses the codebook vectors.

**Recommendation Metrics:** In generative retrieval, **Recall@N** and **NDCG@N** with  $N=50,100$  are used to evaluate the performance. In discriminative ranking, **AUC** and **GAUC** are used to evaluate the performance. For the online experiments, the **Advertising Revenue**, **Click-Through Rate (CTR)** are used to evaluate the online performance.

**4.1.3 Baselines.** We compare our proposed method with state-of-the-art (SOTA) Semantic ID generation approaches, which can be categorized into two groups: (1) content-based SIDs (e.g., RQ-VAE,

**Table 2: Overall performance comparison on two datasets. We evaluate all methods on two downstream tasks: generative retrieval and discriminative ranking. Best results in each column are in bold. Our model, ADA-SID, is highlighted in gray. The last row (Improv.) denotes the relative improvement of ADA-SID over the best baseline. The best baseline performance score is denoted in underline.**

(a) Generative Retrieval Evaluation

Methods	Industrial Dataset							Amazon Beauty						
	$L_{\text{recon}} \downarrow$	Entropy $\uparrow$	Util. $\uparrow$	R@50 $\uparrow$	R@100 $\uparrow$	N@50 $\uparrow$	N@100 $\uparrow$	$L_{\text{recon}} \downarrow$	Entropy $\uparrow$	Util. $\uparrow$	R@50 $\uparrow$	R@100 $\uparrow$	N@50 $\uparrow$	N@100 $\uparrow$
RQ-VAE	0.0033	4.2481	1.0000	0.1854	0.2083	0.1337	0.1421	0.6028	3.4904	0.9900	0.1213	0.2398	0.0803	0.1304
OPQ	0.0038	4.3981	0.7563	0.1972	0.2104	0.1491	0.1518	0.9647	3.3980	0.9600	0.1117	0.2189	0.0802	0.1302
RQ-Kmeans	0.0065	4.7232	1.0000	0.1844	0.2202	0.1462	0.1578	0.6240	1.7100	1.0000	0.1385	0.2398	0.0843	0.1507
LETTER	0.0054	4.2072	1.0000	0.1812	0.2213	0.1582	0.1675	0.5431	2.6819	1.0000	0.1513	0.2492	0.0937	0.1453
DAS	0.0051	4.3539	1.0000	0.1864	0.2237	0.1576	0.1697	0.5432	3.6819	1.0000	0.1503	0.2403	0.0933	0.1445
RQ-VAE++	0.0034	3.5566	0.9283	0.2254	0.2709	0.1628	0.1706	0.6028	3.4904	0.9900	0.1683	0.2991	0.0943	0.1507
MM-RQ-VAE	0.0055	4.2125	0.9850	0.2181	0.2542	0.1592	0.1707	0.5081	2.8449	0.9950	0.1674	0.2596	0.0915	0.1322
ADA-SID(Ours)	0.0032	5.0977	1.0000	0.2772	0.2926	0.1689	0.1714	0.4470	4.4206	1.0000	0.1855	0.3675	0.0996	0.1784
Improv.	+3.03%	+7.92%	+0.00%	+22.45%	+7.53%	+3.56%	+0.23%	+12.02%	+20.06%	+0.00%	+10.21%	+22.86%	+5.62%	+18.38%

(b) Discriminative Ranking Evaluation

Methods	Industrial Dataset					Amazon Beauty				
	$L_{\text{recon}} \downarrow$	Entropy $\uparrow$	Util. $\uparrow$	AUC $\uparrow$	GAUC $\uparrow$	$L_{\text{recon}} \downarrow$	Entropy $\uparrow$	Util. $\uparrow$	AUC $\uparrow$	GAUC $\uparrow$
Item ID	-	-	-	0.7078	0.5845	-	-	-	0.6455	0.5897
RQ-VAE	0.0033	4.2481	1.0000	0.7071	0.5805	0.6028	3.4904	0.9900	0.6446	0.5852
OPQ	0.0038	4.3981	0.7563	0.7086	0.5829	0.9647	3.3980	0.9600	0.6449	0.5898
RQ-Kmeans	0.0065	4.7232	1.0000	0.7089	0.5832	0.6240	1.7100	1.0000	0.6472	0.5999
LETTER	0.0054	4.2072	1.0000	0.7089	0.5828	0.5431	2.6819	1.0000	0.6444	0.5973
DAS	0.0051	4.3539	1.0000	0.7091	0.5845	0.5432	3.6819	1.0000	0.6466	0.5933
RQ-VAE++	0.0034	3.5566	0.9283	0.7100	0.5838	0.6028	3.4904	0.9900	0.6466	0.5952
MM-RQ-VAE	0.0055	4.2125	0.9850	0.7095	0.5843	0.5081	2.8449	0.9950	0.6453	0.5991
ADA-SID(Ours)	0.0032	5.0977	1.0000	0.7101	0.5846	0.4470	4.4206	1.0000	0.6480	0.6125
Improv.	+3.03%	+7.92%	+0.00%	+0.07%	+0.02%	+12.02%	+20.06%	+0.00%	+0.12%	+2.10%

**Table 3: Ablation Experiments.**

Variants	$L_{recon} \downarrow$	Entropy $\uparrow$	Util. $\uparrow$	R@50 $\uparrow$	R@100 $\uparrow$	N@50 $\uparrow$	N@100 $\uparrow$	AUC $\uparrow$	GAUC $\uparrow$
ADA-SID	0.0032	5.0977	1.0000	0.2772	0.2926	0.1689	0.1714	0.7101	0.5846
w/o Alignment Strength Controller	0.0032	5.0710	1.0000	0.2701	0.2854	0.1618	0.1643	0.7104	0.5845
w/o Behavior-content Contrastive Learning	0.0032	5.1153	1.0000	0.2733	0.2874	0.1653	0.1676	0.7097	0.5846
w/o Sparsity Regularization	0.0034	5.0571	1.0000	0.2757	0.2903	0.1675	0.1698	0.7097	0.5846
w/o Behavior-Guided Dynamic Router	0.0033	5.0896	1.0000	0.2705	0.2861	0.1616	0.1641	0.7098	0.5845

OPQ), (2) behavior-content aligned SIDs (e.g., RQ-Kmeans, DAS[42], LETTER, MM-RQ-VAE). Furthermore, we evaluate the performance of SIDs against traditional Item IDs in a discriminative ranking task.

- **Item ID:** The Item ID serves as a unique identifier for an item, conventionally used in discriminative ranking tasks. In contrast, within the generative retrieval paradigm, SIDs function as the item identifier.
- **RQ-VAE[40]:** TIGER[22] transforms content features such as titles, item descriptions, and categories into textual embeddings using a pre-trained LLM. It then employs RQ-VAE to quantize these embeddings into hierarchical SIDs.
- **OPQ[14]:** RPG[25] introduce Optimized Product Quantization(OPQ) to convert pretrained textual embeddings into a tuple of un-ordered SIDs.

- **RQ-Kmeans[43]:** One-rec[39] integrates RQ-VAE and K-means to quantize behavior-finetuned multimodal representations of items in a coarse-to-fine manner, where K-means clustering is applied to the residuals.
- **DAS[42]:** DAS introduces multi-view contrastive alignment to maximize mutual information between SIDs and collaborative signals. This process generates hierarchical, behavior-aware content SIDs using RQ-VAE.
- **LETTER[27]:** LETTER integrates hierarchical semantics, collaborative signals, and code assignment diversity to generate behavior-content-fused SIDs using RQ-VAE.
- **RQ-VAE++:** We introduce both pretrained content representation and pretrained collaborative representation for the semantic ID

generation with RQ-VAE, to evaluate the importance of collaborative information. The RQ-VAE++ generates collaborative, textual, and visual SIDs for each item.

- **MM-RQ-VAE[17]:** MM-RQ-VAE generates collaborative, textual, and visual SIDs from pre-trained collaborative and multimodal embeddings. It also introduces contrastive learning for behavior-content alignment.

#### 4.1.4 Experiment Setup.

- **Recommendation Foundations:** For the evaluation of the Generative Retrieval task, we adopt REG4Rec [18], a strong multi-token prediction model, as our base framework. For the Discriminative Ranking task, we employ the well-established Parameter Personalized Network (PPNet) [62] as the backbone architecture.
- **Implementation Details:** In the industrial dataset, the codebook size is set to 3,00 and the length of SIDs is set to 8 for ADA-SID and baselines. Specifically,  $N_s = 2$ ,  $N_t = 2$ ,  $N_v = 2$ ,  $N_b = 6$  and  $s_{target} = \frac{1}{3}$  are set for ADA-SID. In public datasets, the codebook size is set to 100, length is 6. Specifically,  $N_s = 1$ ,  $N_t = 1$  and  $N_v = 1$ ,  $N_b = 5$  and  $s_{target} = \frac{2}{5}$  are set for ADA-SID. The pre-trained representations are obtained from Qwen3-Embedding 7B [15], SASRec[16] and PailiTAO v8 from an ecommerce advertising platform in Asia. Behavior SIDs with scores above the threshold are retained, while those below are replaced with a padding token, the threshold is 0 in this paper. This threshold can be adjusted based on the specific performance requirements of the recommendation task.

## 4.2 Overall Performance (RQ1)

We compare our proposed method with state-of-the-art(SOTA) semantic ID approaches for generative retrieval and discriminative ranking tasks on both public and industrial datasets. The overall performance comparison is summarized in Table 2, the results lead to several key observations:

**The Importance of Integrating Behavioral Information for SIDs:** Firstly, behavior-content aligned SIDs (e.g., RQ-Kmeans, LETTER) consistently outperform content-only SIDs (RQ-VAE, OPQ) across R@100, N@100, and AUC. This highlights the fundamental limitation of relying solely on static content and underscores the necessity of incorporating behavioral signals. Secondly, the significant performance lift of RQ-VAE++ over the original RQ-VAE directly demonstrates the critical value of incorporating collaborative information into the SID generation process. Finally, comparing alignment strategies reveals that explicitly generating dedicated SIDs for collaborative signals (as in MM-RQ-VAE, RQ-VAE++) is more effective at capturing complex interaction patterns than other approaches (e.g., LETTER, DAS), leading to superior downstream performance

**The Effectiveness of ADA-SID:** ADA-SID demonstrates superior performance across both generative retrieval and discriminative ranking, outperforming all baselines, including content-only and existing behavior-content aligned SIDs. This superiority stems from its unique design: unlike methods that perform indiscriminate alignment, ADA-SID intelligently fuses the content and behavior information by assessing the richness of an item’s behavioral information. It adaptively amplifies critical signals while suppressing noise, resulting in a more robust and expressive item representation.

## 4.3 Ablation Study on Industrial Dataset (RQ2)

We conduct the ablation experiments in Table 3 to study how each module contributes to the overall performance of Ada-SID.

**4.3.1 Impact of adaptive Behavior-Content Alignment. w/o the Alignment Strength Controller:** In this experiment, we disable the alignment Strength Controller and align content representations with collaborative embeddings indiscriminately. The performance degradation has demonstrated that it’s significant to suppress the influence of noise in behavior-content alignment for semantic ID generation.

**w/o the Behavior-Content Contrastive Learning:** In this experiments, we disable the adaptive Behavior-Content Alignment module, which leads to a consistent drop in both Recall and NDCG across all settings. This finding indicates a substantial modality gap between the content and behavioral domains, which hinders the model’s ability to learn their shared information. The resulting performance degradation is therefore expected, as the contrastive learning component is crucial for bridging this gap and enabling effective behavior-content information fusion.

**4.3.2 Impact of Dynamic Behavioral Router Mechanism. w/o the Dynamic Behavior-Guided Router:** Removing the Behavior-Guided Dynamic Router impairs the model’s ability to learn to estimate and weight collaborative signals according to information richness, leading to a drop in recommendation accuracy on both discriminative ranking and generative retrieval tasks. This demonstrates that information richness provides a reliable measure of importance for collaborative SIDs.

**w/o Sparsity Regularization:** Removing the sparsity regularization term also leads to performance degradation. This is because the regularization plays two critical roles. First, by encouraging sparse activations (i.e., selecting only a few relevant SIDs), it forces the model to learn more specialized and disentangled representations for each SID, effectively increasing the model’s total capacity in a manner similar to Mixture-of-Experts (MoE) models. Second, the item-specific sparsity target encourages the model to allocate its representational budget wisely, using fewer, high-level SIDs for long-tail items and more, detailed SIDs for popular items. The absence of this guidance leads to less expressive and less adaptive representations.

## 4.4 Hyper-Parameter Analysis(RQ3)

We further investigate that how various hyper-parameter settings affect the model’s performance on industrial dataset.

**The Strength of Sparsity Regularization** As we reduce the sparsity intensity, the model’s parameter count expands, leading to a larger encoder capacity and stronger encoding capabilities. Consequently, a significant increase in recommendation accuracy is observed across both discriminative ranking and generative retrieval tasks, as shown in Fig.3 . Besides, the ADA-SID has an advantages in variable-length flexibility, which allows head items to use longer collaborative SID sequences to fully represent their complex behavioral patterns, which leads to stronger expressive power and significant downstream performance gains.

**Sensitivity to Alignment Strength Controller Hyperparameters** We conducted a hyperparameter study on  $(\alpha, \beta)$  for the Alignment Strength Controller, testing four configurations that yield diverse weighting curves in Fig.4. As shown in Table 4, the

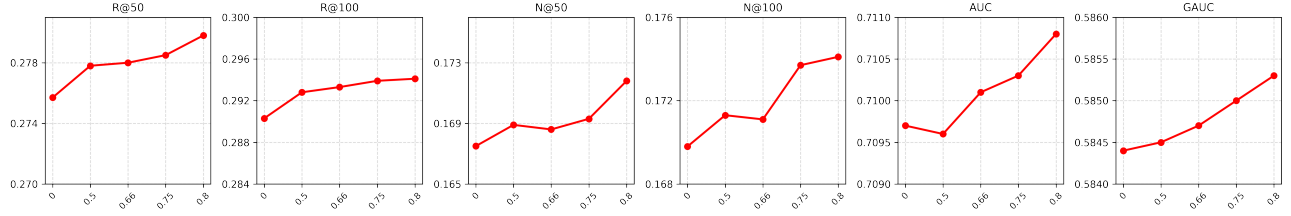
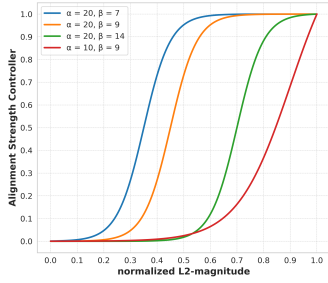


Figure 3: Hyper-Parameter Analysis on Sparsity Regularization.

Table 4: Hyper-Parameter Analysis on Contrastive Loss Weight

Variants	$L_{recon} \downarrow$	Entropy $\uparrow$	Utilization $\uparrow$	R@50 $\uparrow$	R@100 $\uparrow$	N@50 $\uparrow$	N@100 $\uparrow$	AUC $\uparrow$	GAUC $\uparrow$
$\alpha=20, \beta=7$	0.0032	5.0844	1.0000	0.2749	0.2894	0.1664	0.1688	0.7106	0.5840
$\alpha=20, \beta=9$	0.0032	5.0711	1.0000	0.2750	0.2889	0.1677	0.1709	0.7105	0.5842
$\alpha=20, \beta=14$	0.0033	5.0967	1.0000	0.2760	0.2911	0.1686	0.1707	0.7105	0.5839
$\alpha=10, \beta=9$	0.0032	5.0977	1.0000	0.2772	0.2926	0.1689	0.1714	0.7101	0.5846

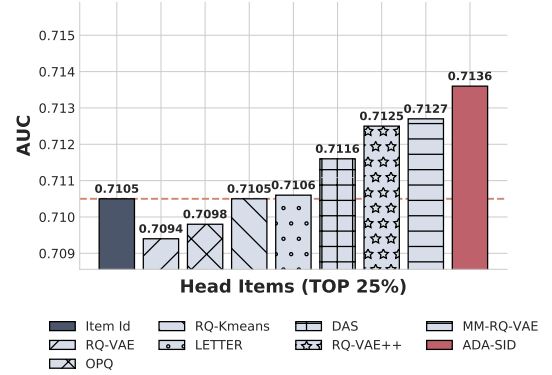
Figure 4: Illustration of alignment strength controller with different hyperparameters ( $\alpha, \beta$ ).

setting ( $\alpha=10, \beta=9$ ) achieved the highest recommendation accuracy. This optimal result suggests that for this dataset’s distribution, noise filtering is most effective when applied to approximately the 40% least frequent (long-tail) items. The tunability of parameters  $\alpha$  and  $\beta$  underscores the inherent flexibility of our design, allowing it to adapt to diverse data landscapes.

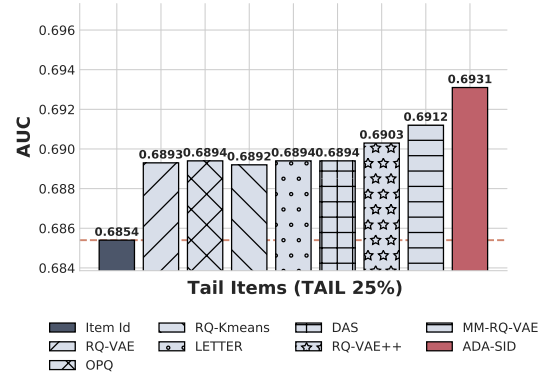
#### 4.5 Item Popularity Stratified Performance Analysis (RQ4)

While unique Item IDs enable learning highly independent representations for popular items from large-scale interaction data, replacing them with SIDs in ranking tasks remains a challenge. To investigate the performance difference, we perform a stratified analysis based on item popularity. We categorize items into ‘popular’ (top 25%) and ‘long-tail’ (bottom 25%) groups based on their impression counts over the last 30 days, and then evaluate the AUC for each group, as shown in Fig.5.

**For Head Items,** content-based SIDs underperform simple Item IDs in ranking tasks. Integrating collaborative information is crucial, as it enhances the SIDs’ expressiveness to capture complex behavioral patterns, thereby improving performance. Furthermore, our proposed ADA-SID advances beyond previous approaches by



(a) Comparison on Popular Items.



(b) Comparison on Long-tail Items.

Figure 5: Item Popularity Stratified Performance Comparison.

explicitly aligning, denoising, and amplifying the fusion of content and behavioral modalities. This process yields a significantly

more expressive semantic representation. **For Tail Items:** Conversely, all SID-based methods outperform ItemIDs on tail items by leveraging knowledge sharing across semantically similar items. ADA-SID achieves the largest performance gain. Its adaptive alignment shields the stable content representations of tail items from their noisy and sparse behavioral signals. Concurrently, its dynamic behavioral router learns to produce a sparser, more robust representation by relying more on high-level semantics than on unreliable fine-grained behavioral cues. This dual mechanism significantly boosts performance on the long tail.

By adaptively balancing the expressive independence of head items with the generalization capability for tail items, our method generates a more robust and effective identifier, ultimately surpassing the performance of traditional Item IDs in ranking tasks.

## 5 Online Experiments

We validated our method through a 5-day online A/B test on a large-scale e-commerce platform’s generative retrieval system. The experimental group, using our 8-token SIDs, was allocated 10% of random user traffic against the production Item ID-based system. Our approach yielded significant improvements in key business metrics: a +3.50% increase in Advertising Revenue and a +1.15% increase in Click-Through Rate (CTR). These online gains confirm the practical value and production-readiness of our proposed method.

## 6 Conclusion

We introduced ADA-SID to learn expressive and noise-robust SIDs by adaptively aligning, denoising, and amplifying multimodal information. The two core innovations, the adaptive behavior-content alignment and dynamic behavioral router mechanism amplify critical collaborate signals and suppress the influence of noise. Extensive offline experiments and a large-scale online A/B test validate that ADA-SID significantly improves recommendation performance. Our work pioneers an adaptive fusion approach based on information richness, paving the way for more robust and personalized recommender systems. Future directions include applying these principles to user-side modeling.

## References

- [1] A. Singh, T. Vu, N. Mehta, R. Keshavan, M. Sathiamoorthy, Y. Zheng, L. Hong, L. Heldt, L. Wei, D. Tandon, E. H. Chi, and X. Yi, “Better generalization with semantic ids: A case study in ranking for recommendations,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.08121>
- [2] C. Zheng, M. Huang, D. Pedchenko, K. Rangadurai, S. Wang, G. Nahum, J. Lei, Y. Yang, T. Liu, Z. Luo, X. Wei, D. Ramasamy, J. Yang, Y. Han, L. Yang, H. Xu, R. Jin, and S. Yang, “Enhancing embedding representation stability in recommendation systems with semantic id,” 2025. [Online]. Available: <https://arxiv.org/abs/2504.02137>
- [3] Y. Hou, Z. He, J. McAuley, and W. X. Zhao, “Learning vector-quantized item representation for transferable sequential recommenders,” 2023. [Online]. Available: <https://arxiv.org/abs/2210.12316>
- [4] S. Milojević, “Power law distributions in information science: Making the case for logarithmic binning,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, p. 2417–2425, Dec. 2010. [Online]. Available: <http://dx.doi.org/10.1002/asi.21426>
- [5] J. a. Gama, I. Zliobaitundefined, A. Bifet, M. Pechenizkiy, and A. Bouchachia, “A survey on concept drift adaptation,” *ACM Comput. Surv.*, vol. 46, no. 4, Mar. 2014. [Online]. Available: <https://doi.org/10.1145/2523813>
- [6] J. Zhai, L. Liao, X. Liu, Y. Wang, R. Li, X. Cao, L. Gao, Z. Gong, F. Gu, M. He, Y. Lu, and Y. Shi, “Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.17152>
- [7] T. Kudo, “Subword regularization: Improving neural network translation models with multiple subword candidates,” 2018. [Online]. Available: <https://arxiv.org/abs/1804.10959>
- [8] Y. Sun, F. Yuan, M. Yang, G. Wei, Z. Zhao, and D. Liu, “A generic network compression framework for sequential recommender systems,” *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:216562600>
- [9] M. R. Joglekar, C. Li, J. K. Adams, P. Khaitan, and Q. V. Le, “Neural input search for large scale recommendation models,” *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:195874115>
- [10] B. Yan, P. Wang, K. Zhang, W. Lin, K.-C. Lee, J. Xu, and B. Zheng, “Learning Effective and Efficient Embedding via an Adaptively-Masked Twins-based Layer,” Aug. 2021, arXiv:2108.11513 [cs]. [Online]. Available: <http://arxiv.org/abs/2108.11513>
- [11] X. Zhao, C. Wang, M. Chen, X. Zheng, X. Liu, and J. Tang, “AutoEmb: Automated Embedding Dimensionality Search in Streaming Recommendations,” Feb. 2020, arXiv:2002.11252 [cs]. [Online]. Available: <http://arxiv.org/abs/2002.11252>
- [12] Y. Ni, Y. Cheng, X. Liu, J. Fu, Y. Li, X. He, Y. Zhang, and F. Yuan, “A content-driven micro-video recommendation dataset at scale,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.15379>
- [13] S. Li, W. Lei, Q. Wu, X. He, P. Jiang, and T.-S. Chua, “Seamlessly unifying attributes and items: Conversational recommendation for cold-start users,” *ACM Transactions on Information Systems*, vol. 39, no. 4, p. 1–29, Aug. 2021. [Online]. Available: <http://dx.doi.org/10.1145/3446427>
- [14] T. Ge, K. He, Q. Ke, and J. Sun, “Optimized Product Quantization,”
- [15] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou, “Qwen3 embedding: Advancing text embedding and reranking through foundation models,” 2025. [Online]. Available: <https://arxiv.org/abs/2506.05176>
- [16] W.-C. Kang and J. McAuley, “Self-attentive sequential recommendation,” 2018. [Online]. Available: <https://arxiv.org/abs/1808.09781>
- [17] Y. Wang, J. Pan, X. Li, M. Wang, Y. Wang, Y. Liu, D. Liu, J. Jiang, and X. Zhao, “Empowering large language model for sequential recommendation via multimodal embeddings and semantic ids,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.02017>
- [18] H. Xing, H. Deng, Y. Mao, J. Hu, Y. Xu, H. Zhang, J. Wang, S. Wang, Y. Zhang, X. Zeng, and J. Zhang, “Reg4rec: Reasoning-enhanced generative model for large-scale recommendation systems,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.15308>
- [19] J. Ji, Z. Li, S. Xu, W. Hua, Y. Ge, J. Tan, and Y. Zhang, “Genrec: Large language model for generative recommendation,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.00457>
- [20] E. Liu, B. Zheng, C. Ling, L. Hu, H. Li, and W. X. Zhao, “Generative recommender with end-to-end learnable item tokenization,” 2025. [Online]. Available: <https://arxiv.org/abs/2409.05546>
- [21] R. Han, B. Yin, S. Chen, H. Jiang, F. Jiang, X. Li, C. Ma, M. Huang, X. Li, C. Jing, Y. Han, M. Zhou, L. Yu, C. Liu, and W. Lin, “Mtgr: Industrial-scale generative recommendation framework in meituan,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.18654>
- [22] S. Rajput, N. Mehta, A. Singh, R. H. Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Q. Tran, J. Samost, M. Kula, E. H. Chi, and M. Sathiamoorthy, “Recommender systems with generative retrieval,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.05065>
- [23] J. Deng, S. Wang, K. Cai, L. Ren, Q. Hu, W. Ding, Q. Luo, and G. Zhou, “Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.18965>
- [24] Y. Yang, Z. Ji, Z. Li, Y. Li, Z. Mo, Y. Ding, K. Chen, Z. Zhang, J. Li, S. Li, and L. Liu, “Sparse meets dense: Unified generative recommendations with cascaded sparse-dense representations,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.02453>
- [25] Y. Hou, J. Li, A. Shin, J. Jeon, A. Santhanam, W. Shao, K. Hassani, N. Yao, and J. McAuley, “Generating long semantic ids in parallel for recommendation,” 2025. [Online]. Available: <https://arxiv.org/abs/2506.05781>
- [26] Z. Tang, Z. Huan, Z. Li, X. Zhang, J. Hu, C. Fu, J. Zhou, L. Zou, and C. Li, “One model for all: Large language models are domain-agnostic recommendation systems,” 2025. [Online]. Available: <https://arxiv.org/abs/2310.14304>
- [27] W. Wang, H. Bao, X. Lin, J. Zhang, Y. Li, F. Feng, S.-K. Ng, and T.-S. Chua, “Learnable item tokenization for generative recommendation,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.07314>
- [28] H. Qu, W. Fan, Z. Zhao, and Q. Li, “Tokenrec: Learning to tokenize id for llm-based generative recommendation,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.10450>
- [29] Y. Hou, J. Ni, Z. He, N. Sachdeva, W.-C. Kang, E. H. Chi, J. McAuley, and D. Z. Cheng, “Actionpiece: Contextually tokenizing action sequences for generative recommendation,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.13581>
- [30] Y. Wang, J. Xun, M. Hong, J. Zhu, T. Jin, W. Lin, H. Li, L. Li, Y. Xia, Z. Zhao, and Z. Dong, “Eager: Two-stream generative recommender with behavior-semantic collaboration,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.14017>
- [31] B. Zheng, Y. Hou, H. Lu, Y. Chen, W. X. Zhao, M. Chen, and J.-R. Wen, “Adapting large language models by integrating collaborative semantics for

- recommendation," 2024. [Online]. Available: <https://arxiv.org/abs/2311.09049>
- [32] Y. Wang, Z. Ren, W. Sun, J. Yang, Z. Liang, X. Chen, R. Xie, S. Yan, X. Zhang, P. Ren, Z. Chen, and X. Xin, "Content-based collaborative generation for recommender systems," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, ser. CIKM '24. ACM, Oct. 2024, p. 2420–2430. [Online]. Available: <http://dx.doi.org/10.1145/3627673.3679692>
- [33] W.-H. Huang, C.-W. Ke, W.-N. Chiu, Y.-X. Su, C.-C. Yang, C.-Y. Cheng, Y.-N. Chen, and P.-J. Cheng, "Augment or not? a comparative study of pure and augmented large language model recommenders," 2025. [Online]. Available: <https://arxiv.org/abs/2505.23053>
- [34] J. Tan, S. Xu, W. Hua, Y. Ge, Z. Li, and Y. Zhang, "Idgenrec: Llm-recsys alignment with textual id learning," 2024. [Online]. Available: <https://arxiv.org/abs/2403.19021>
- [35] B. Jin, H. Zeng, G. Wang, X. Chen, T. Wei, R. Li, Z. Wang, Z. Li, Y. Li, H. Lu, S. Wang, J. Han, and X. Tang, "Language models as semantic indexers," 2024. [Online]. Available: <https://arxiv.org/abs/2310.07815>
- [36] B. Zheng, H. Lu, Y. Chen, W. X. Zhao, and J.-R. Wen, "Universal item tokenization for transferable generative recommendation," 2025. [Online]. Available: <https://arxiv.org/abs/2504.04405>
- [37] Z. Zheng, Z. Wang, F. Yang, J. Fan, T. Zhang, Y. Wang, and X. Wang, "Ega-v2: An end-to-end generative framework for industrial advertising," 2025. [Online]. Available: <https://arxiv.org/abs/2505.17549>
- [38] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16. International World Wide Web Conferences Steering Committee, Apr. 2016, p. 507–517. [Online]. Available: <http://dx.doi.org/10.1145/2872427.2883037>
- [39] G. Zhou, J. Deng, J. Zhang, K. Cai, L. Ren, Q. Luo, Q. Wang, Q. Hu, R. Huang, S. Wang, W. Ding, W. Li, X. Luo, X. Wang, Z. Cheng, Z. Zhang, B. Zhang, B. Wang, C. Ma, C. Song, C. Wang, D. Wang, D. Meng, F. Yang, F. Zhang, F. Jiang, F. Zhang, G. Wang, G. Zhang, H. Li, H. Hu, H. Lin, H. Cheng, H. Cao, H. Wang, J. Huang, J. Chen, J. Liu, J. Jia, K. Gai, L. Hu, L. Zeng, L. Yu, Q. Wang, Q. Zhou, S. Wang, S. He, S. Yang, S. Yang, S. Huang, T. Wu, T. He, T. Gao, W. Yuan, X. Liang, X. Xu, X. Liu, Y. Wang, Y. Wang, Y. Liu, Y. Song, Y. Zhang, Y. Wu, Y. Zhao, and Z. Liu, "Onerec technical report," 2025. [Online]. Available: <https://arxiv.org/abs/2506.13695>
- [40] L. Zhu, F. Wei, Y. Lu, and D. Chen, "Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%," 2024. [Online]. Available: <https://arxiv.org/abs/2406.11837>
- [41] C. Bentz and D. Alikaniotis, "The word entropy of natural languages," 2016. [Online]. Available: <https://arxiv.org/abs/1606.06996>
- [42] W. Ye, M. Sun, S. Shi, P. Wang, W. Wu, and P. Jiang, "Das: Dual-aligned semantic ids empowered industrial recommender system," 2025. [Online]. Available: <https://arxiv.org/abs/2508.10584>
- [43] X. Luo, J. Cao, T. Sun, J. Yu, R. Huang, W. Yuan, H. Lin, Y. Zheng, S. Wang, Q. Hu, C. Qiu, J. Zhang, X. Zhang, Z. Yan, J. Zhang, S. Zhang, M. Wen, Z. Liu, K. Gai, and G. Zhou, "Qarm: Quantitative alignment multi-modal recommendation at kuaishou," 2024. [Online]. Available: <https://arxiv.org/abs/2411.11739>
- [44] Z. Kuai, Z. Chen, H. Wang, M. Li, D. Miao, B. Wang, X. Chen, L. Kuang, Y. Han, J. Wang, G. Tang, L. Liu, S. Wang, and J. Zhuo, "Breaking the hourglass phenomenon of residual quantization: Enhancing the upper bound of generative retrieval," 2024. [Online]. Available: <https://arxiv.org/abs/2407.21488>
- [45] B. Zheng, E. Liu, Z. Chen, Z. Ma, Y. Wang, W. X. Zhao, and J.-R. Wen, "Pre-training generative recommender with multi-identifier item tokenization," 2025. [Online]. Available: <https://arxiv.org/abs/2504.04400>
- [46] Y. Bai, R. Xiang, K. Li, Y. Tang, Y. Cheng, X. Liu, P. Jiang, and K. Gai, "Chime: A compressive framework for holistic interest modeling," 2025. [Online]. Available: <https://arxiv.org/abs/2504.06780>
- [47] 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," 2023. [Online]. Available: <https://arxiv.org/abs/2303.13835>
- [48] Y. Liu, J. Cao, S. Wang, S. Wen, X. Chen, X. Wu, S. Yang, Z. Liu, K. Gai, and G. Zhou, "Llm-alignment live-streaming recommendation," 2025. [Online]. Available: <https://arxiv.org/abs/2504.05217>
- [49] P. Wollstadt, S. Schmitt, and M. Wibral, "A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition," 2023. [Online]. Available: <https://arxiv.org/abs/2105.04187>
- [50] P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood, R. Salakhutdinov, and L.-P. Morency, "Quantifying & modeling multimodal interactions: An information decomposition framework," 2023. [Online]. Available: <https://arxiv.org/abs/2302.12247>
- [51] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018. [Online]. Available: <https://arxiv.org/abs/1806.00064>
- [52] J. Xin, S. Yun, J. Peng, I. Choi, J. L. Ballard, T. Chen, and Q. Long, "I2moe: Interpretable multimodal interaction-aware mixture-of-experts," 2025. [Online]. Available: <https://arxiv.org/abs/2505.19190>
- [53] Z. Xue and R. Marculescu, "Dynamic multimodal fusion," 2023. [Online]. Available: <https://arxiv.org/abs/2204.00102>
- [54] H. Yu, Z. Qi, L. Jang, R. Salakhutdinov, L.-P. Morency, and P. P. Liang, "Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts," 2024. [Online]. Available: <https://arxiv.org/abs/2311.09580>
- [55] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," 2022. [Online]. Available: <https://arxiv.org/abs/2206.02770>
- [56] P. Jin, B. Zhu, L. Yuan, and S. Yan, "Moe++: Accelerating mixture-of-experts methods with zero-computation experts," 2024. [Online]. Available: <https://arxiv.org/abs/2410.07348>
- [57] DeepSeek-AI, A. Liu, and a. Z. P. Bei Feng et al., ..., "Deepseek-v3 technical report," 2025. [Online]. Available: <https://arxiv.org/abs/2412.19437>
- [58] L. Wang, H. Gao, C. Zhao, X. Sun, and D. Dai, "Auxiliary-loss-free load balancing strategy for mixture-of-experts," 2024. [Online]. Available: <https://arxiv.org/abs/2408.15664>
- [59] Y. Huang, Y. Chen, X. Cao, R. Yang, M. Qi, Y. Zhu, Q. Han, Y. Liu, Z. Liu, X. Yao, Y. Jia, L. Ma, Y. Zhang, T. Zhu, L. Zhang, L. Chen, W. Chen, M. Zhu, R. Xu, and L. Zhang, "Towards large-scale generative ranking," 2025. [Online]. Available: <https://arxiv.org/abs/2505.04180>
- [60] C. Wang, B. Wu, Z. Chen, L. Shen, B. Wang, and X. Zeng, "Scaling transformers for discriminative recommendation via generative pretraining," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, ser. KDD '25. ACM, Aug. 2025, p. 2893–2903. [Online]. Available: <http://dx.doi.org/10.1145/3711896.3737117>
- [61] H. Deng, H. Xing, K. Matsuyama, Y. Huang, J. Hu, H. Wen, J. Xu, Z. Chen, Y. Zhang, X. Zeng, and J. Zhang, "Heterrec: Heterogeneous information transformer for scalable sequential recommendation," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '25. ACM, Jul. 2025, p. 3020–3024. [Online]. Available: <http://dx.doi.org/10.1145/3726302.3730206>
- [62] J. Chang, C. Zhang, Y. Hui, D. Leng, Y. Niu, Y. Song, and K. Gai, "Pepnet: Parameter and embedding personalized network for infusing with personalized prior information," 2023. [Online]. Available: <https://arxiv.org/abs/2302.01115>
- [63] F. Shi, Z. Luo, Y. Ge, Y. Yang, Y. Shan, and L. Wang, "Scalable image tokenization with index backpropagation quantization," 2025. [Online]. Available: <https://arxiv.org/abs/2412.02692>