ROLE-FACTORIZED SEMANTIC IDS: DIAGNOSING AND MITIGATING INFORMATION LOSS IN GENERATIVE RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in generative modeling have led to growing interest in generative recommendation. Central to this paradigm is the semantic ID (SID), a short sequence of discrete tokens representing an item. Existing methods often construct SIDs by fusing content and behavior into a single continuous representation before quantization. However, our preliminary study shows that this early-fusion strategy can lead to a *representational compromise*, in which the learned item geometry is less aligned with the relational structures of both sources. To address this issue, we propose Role-Factorized Semantic IDs (RF-SID), which assigns content and behavior to separate coordinate blocks within a unified SID. This design represents both relations without requiring every discrete coordinate to reconcile them. Experiments on three public benchmarks show that RF-SID consistently outperforms state-of-the-art recommendation methods. Code is released anonymously at `https://anonymous.4open.science/r/rf-sid-7CD1/`.

1 INTRODUCTION

Recommender systems help users discover items of interest from vast catalogs, thereby alleviating information overload on modern online platforms (Resnick & Varian, 1997; Adomavicius & Tuzhilin, 2005). Recent advances in generative modeling have further motivated generative recommendation, which predicts the next item by generating its semantic ID (SID), a short sequence of discrete tokens representing the item (Geng et al., 2022; Rajput et al., 2023; Wang et al., 2024a; Qu et al., 2025; Liu et al., 2025b). This paradigm has achieved promising results and offers a unified sequence-modeling interface for recommendation. Different items may share parts of their SID sequences, allowing the model to reuse the corresponding token parameters, while successive tokens progressively narrow the candidate item space during autoregressive generation.

Existing methods for constructing SIDs can be broadly categorized according to the information they use. Content-based methods derive SIDs from text or multimodal features, so their codes reflect similarities in item content (Rajput et al., 2023; Zhai et al., 2025). Behavior-based methods instead learn from user–item interactions, so their codes capture how users interact with different items (Qu et al., 2025). Methods that use both content and behavior often combine the two sources into a single continuous representation before quantization (Burke, 2002; Wang et al., 2015; He & McAuley, 2016; Wei et al., 2019; Zhang et al., 2021; Zhou & Shen, 2023; Wang et al., 2024a; Xiao et al., 2025). This combination is appealing because the two sources provide different but complementary information: content describes what an item is, while interaction data reveals how users engage with it. Previous work has shown that using both can improve recommendation performance (Wang et al., 2024a; Xiao et al., 2025; Wang et al., 2024b).

However, content and behavior induce different item neighborhoods: semantically similar items need not be behaviorally interchangeable, while behaviorally related items need not be close in content. When the two relations are fused before quantization, every coordinate in the resulting SID must reconcile these potentially different structures, which may prevent the SID from faithfully preserving either one. To examine this hypothesis, we conduct a preliminary study on Beauty comparing the learned item geometries produced by different SID designs. As shown in Figure 1, the content-only and behavior-only SIDs align most strongly with their respective source structures,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

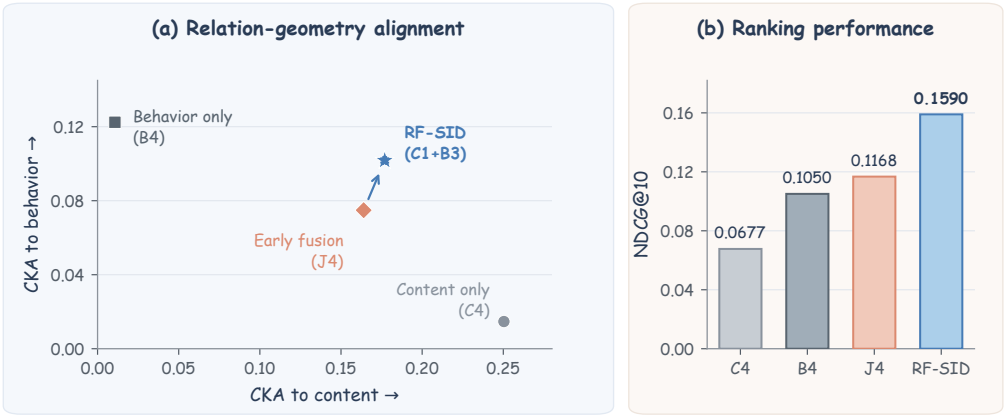

Figure 1: Preliminary study on Beauty. **(a)** Alignment with content and conditional-behavior representations, measured by linear CKA. **(b)** Ranking performance measured by NDCG@10.

whereas the early-fusion SID exhibits notably lower alignment with both sources than the corresponding single-source SIDs. We refer to this simultaneous reduction in source alignment as a *representational compromise* (see Section 3 for details). This raises a simple question that has received less attention: *should content and behavior relations share every coordinate of a finite SID?*

Motivated by this observation, we propose Role-Factorized Semantic IDs (RF-SID), which allocates the positions of a fixed-length SID between content and behavior. Instead of combining the two representations before quantization, RF-SID processes them through separate projections and quantizers and places the resulting tokens in role-specific blocks of a unified SID. This design avoids requiring every quantization step to reconcile both relations.

To evaluate RF-SID, we organize our experiments around three questions: (1) How does RF-SID compare with existing recommendation methods? (2) Which components of RF-SID account for its performance gains? (3) How do the use and configuration of role-specific coordinates affect its performance? We also use a simplified rate–distortion model to characterize when the benefit of preserving source differences can outweigh the cost of splitting the coding budget.

Our main contributions are threefold:

- We identify and empirically analyze a representational compromise in early-fusion SID construction, and provide a simplified rate–distortion analysis of the trade-off between fusion and separate encoding.
- We introduce RF-SID, a unified, fixed-length SID that separately quantizes content and behavior representations and places the resulting tokens in role-specific coordinate blocks.
- We conduct extensive experiments on three public benchmarks, showing consistent improvements over state-of-the-art recommendation methods. Additional experiments further show that role factorization is the main source of the gains and that coordinate use and configuration affect performance.

2 Related Work

2.1 Sequential Recommendation

Sequential recommenders predict the next item from a user's interaction history (Rendle et al., 2010; Tang & Wang, 2018). GRU4Rec models this history with a recurrent network (Hidasi et al., 2015), while SASRec uses self-attention (Kang & McAuley, 2018) and BERT4Rec learns from masked-item prediction (Sun et al., 2019). S^3-Rec further uses self-supervised tasks involving items, attributes, and sequence segments (Zhou et al., 2020). These methods differ in how they model interaction sequences, but retain items as the units of prediction. Generative recommendation instead

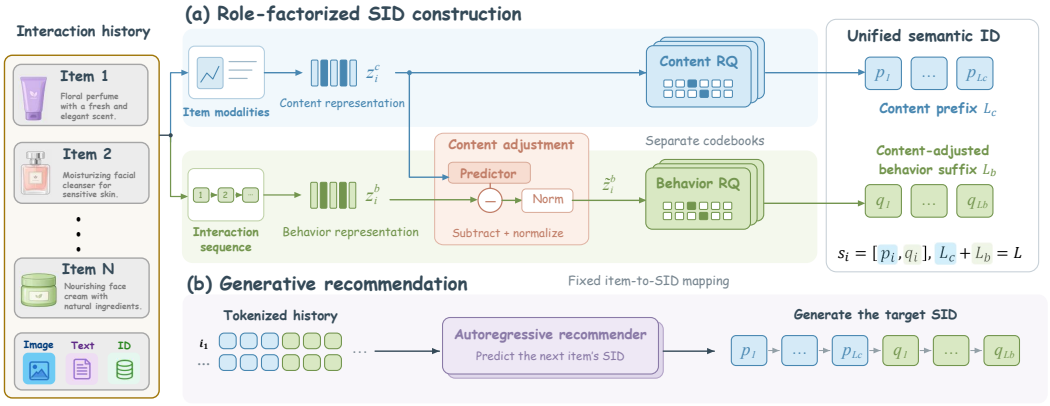

Figure 2: Overview of RF-SID. **(a)** Separate residual quantizers (RQ) encode content and content-adjusted behavior into the prefix and suffix of one unified SID. Content adjustment subtracts the behavior component predicted from content and normalizes the residual. **(b)** The fixed item-to-SID mapping is used for both interaction histories and autoregressively generated targets.

predicts a sequence of item tokens, making the construction of those tokens part of the recommendation problem (De Cao et al., 2020; Tay et al., 2022; Bevilacqua et al., 2022; Sun et al., 2023).

2.2 Generative Recommendation with Semantic IDs

Semantic IDs allow items to share tokens according to the representations from which their codes are constructed. TIGER builds SIDs by residual quantization of item content embeddings and generates the next item's SID with a Transformer (Rajput et al., 2023). TokenRec quantizes collaborative user and item representations for LLM-based recommendation (Qu et al., 2025). ETEGRec connects tokenizer learning to the recommendation objective through alignment losses and alternating optimization (Liu et al., 2025b). When content and behavior are both available, item tokenization must also determine how the two sources are represented in the code sequence.

LETTER incorporates collaborative signals into semantic tokenization through an alignment loss and regularizes code diversity (Wang et al., 2024a). UNGER learns a joint semantic–collaborative representation before quantizing it into a unified code (Xiao et al., 2025). Other methods keep information sources distinct at different stages: MQL4GRec separately quantizes textual and visual content (Zhai et al., 2025), EAGER uses separate decoders for semantic and behavioral tokens (Wang et al., 2024b), and DiscRec models semantic and collaborative signals in separate embedding-layer branches before combining them through a gate (Liu et al., 2025a). RF-SID places this distinction in the item identifier itself, assigning content and behavior to dedicated coordinate blocks within one fixed-length SID.

3 Preliminary Study

Content and behavior induce different relations between items. When both sources are encoded within a fixed-length SID, how do these relations appear in the item representations learned by a recommender? We study this question on Beauty by comparing four-token SIDs built from content alone (C4), behavior alone (B4), an early fusion of both (J4), and role factorization (RF-SID C1+B3). J4 combines the source representations before quantization, whereas RF-SID assigns one token to content and three to behavior.

For each SID construction, we train a recommender on interaction histories to predict the next item's SID, using the same architecture and training protocol. This training learns an input embedding for each SID token. We concatenate an item's token embeddings to examine the item relations formed at the recommender input, and compare them with the frozen content and conditional-behavior representations. Let $X \in \mathbb{R}^{n \times d_x}$ contain the learned SID representations for n items, and let $Y \in \mathbb{R}^{n \times d_y}$ contain their corresponding representations from one source. We measure alignment

using linear centered kernel alignment (CKA) (Kornblith et al., 2019; Gretton et al., 2007). Define $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, $\widetilde{X} = HX$, and $\widetilde{Y} = HY$. Then

$$\mathrm{CKA}(X, Y) = \frac{\|\widetilde{X}^\top \widetilde{Y}\|_F^2}{\|\widetilde{X}^\top \widetilde{X}\|_F \|\widetilde{Y}^\top \widetilde{Y}\|_F}. \tag{1}$$

A higher CKA indicates greater agreement between the item relations captured by the two representations. We denote alignment with the content and conditional-behavior sources by A_c and A_b, respectively. Appendix A.1 details the sampling and ranking protocol.

Observation 1: Early fusion compromises alignment with both sources. Figure 1(a) shows that C4 aligns most strongly with content ($A_c = 0.2504$), while B4 aligns most strongly with behavior ($A_b = 0.1224$). These values provide source-specific reference points for the geometry learned after quantization and recommender training. J4 achieves lower alignment on both sides: $A_c = 0.1640$ and $A_b = 0.0749$, or 65.5% and 61.2% of the corresponding reference CKA scores. We call this marked reduction in alignment with both sources a *representational compromise*.

C4 and B4 each devote all four tokens to one relation, while J4 uses the same budget for both. The comparison therefore leads to a more specific question: can assigning distinct roles to the four positions improve alignment with both sources?

Observation 2: Role factorization improves both alignments. With the same four-token budget, RF-SID C1+B3 raises content alignment from 0.1640 to 0.1769 and behavior alignment from 0.0749 to 0.1017 relative to J4. In Figure 1(a), RF-SID moves upward and to the right of early fusion: assigning source-specific roles to the coordinates improves both aspects of the learned item geometry.

The advantage over J4 also appears in ranking. Figure 1(b) shows that RF-SID increases NDCG@10 from 0.1168 to 0.1590, the highest among the four designs. Together, these findings suggest that how content and behavior are organized within a SID matters for the item relations learned by the recommender and for its ranking performance. They motivate the dedicated coordinate blocks developed in the next section.

4 Methodology

4.1 Problem Formulation

Let $\mathcal{I}$ denote an item catalog and $H_u = [i_1, i_2, \ldots, i_{T-1}]$ the chronologically ordered interaction history of user u, where $i_t \in \mathcal{I}$. Given H_u, sequential recommendation aims to identify the next item $y_u = i_T$ with which the user is likely to interact.

In generative recommendation, items are expressed in a discrete token language. A tokenizer assigns every item i a static semantic ID (SID)

$$s_i = [s_i^1, s_i^2, \ldots, s_i^L], \qquad s_i^\ell \in \{1, 2, \ldots, K\}, \tag{2}$$

where L is the number of SID coordinates and K is the vocabulary size of each coordinate. Replacing each historical item by its SID gives the tokenized history $S(H_u) = [s_{i_1}, s_{i_2}, \ldots, s_{i_{T-1}}]$. The next-item problem can then be written as autoregressive generation of the target SID:

$$p_\theta(s_{y_u} \mid S(H_u)) = \prod_{\ell=1}^{L} p_\theta\big(s_{y_u}^\ell \mid S(H_u), s_{y_u}^{<\ell}\big), \tag{3}$$

where $s_{y_u}^{<\ell}$ denotes the previously generated coordinates of the target SID.

Each item additionally has a content representation $c_i \in \mathbb{R}^{d_c}$ derived from its content and a behavioral representation $b_i \in \mathbb{R}^{d_b}$ derived from user–item interactions.

4.2 Role-Factorized SID Construction

Figure 2 presents an overview of RF-SID. RF-SID represents item content and behavior in separate blocks of a unified semantic ID: a content prefix and a behavior suffix. This allocation allows

each block to encode its respective source without requiring every coordinate to represent both. To reduce redundant information across the two blocks, we further adjust the behavioral representation by subtracting the component predicted from content. We refer to the resulting block as the content-adjusted behavior suffix. The two blocks contain L_c and L_b tokens, respectively, with $L_c + L_b = L$. Formally,

$$\begin{aligned} z_i^c &= \mathrm{norm}(R_c c_i), \quad z_i^b = \mathrm{norm}(R_b b_i), \quad \tilde{z}_i^b = \mathcal{A}(z_i^b, z_i^c), \\ p_i &= Q_c(z_i^c), \qquad\qquad q_i = Q_b(\tilde{z}_i^b), \qquad\qquad s_i = [p_i, q_i], \end{aligned} \tag{4}$$

where R_c and R_b are separate projection matrices and norm denotes ℓ_2 normalization. $\mathcal{A}$ is the content-adjustment operator, and Q_c and Q_b are the respective quantizers. Here, $p_i = [p_i^1, \ldots, p_i^{L_c}]$ is the content prefix and $q_i = [q_i^1, \ldots, q_i^{L_b}]$ is the content-adjusted behavior suffix.

Content prefix. The content prefix encodes item content using L_c levels of residual quantization, following the SID construction used by TIGER (Rajput et al., 2023; Van Den Oord et al., 2017; Lee et al., 2022). Let $r_{i,c}^0 = z_i^c$ and let $\mathcal{E}_c^\ell = \{e_{c,\ell,1}, \ldots, e_{c,\ell,K}\}$ denote the codebook at level ℓ. The prefix tokens are defined by

$$p_i^\ell = \arg\min_{k \in \{1,\ldots,K\}} \left\| r_{i,c}^{\ell-1} - e_{c,\ell,k} \right\|_2^2, \qquad r_{i,c}^\ell = r_{i,c}^{\ell-1} - e_{c,\ell,p_i^\ell}, \quad \ell = 1, \ldots, L_c. \tag{5}$$

Each codebook is fitted using mini-batch k-means on the residuals from the preceding levels (Sculley, 2010).

Content-adjusted behavior suffix. The behavior suffix is designed to complement the content prefix by encoding behavioral variation that remains after accounting for what content predicts. We model the content-to-behavior relationship with a ridge-regularized linear predictor (Hoerl & Kennard, 1970):

$$W^* = \arg\min_W \sum_{i \in \mathcal{I}} \left\| z_i^b - W z_i^c \right\|_2^2 + \lambda \left\| W \right\|_F^2, \tag{6}$$

where λ controls the regularization strength. The content-adjustment operator is defined as

$$\mathcal{A}(z_i^b, z_i^c) = \mathrm{norm}\big(z_i^b - W^* z_i^c\big). \tag{7}$$

We encode the adjusted representation into L_b tokens using the same residual quantization procedure as the content prefix, with separate behavior codebooks.

The content prefix and behavior suffix together form the unified SID $s_i = [p_i, q_i]$, which represents the item in both the interaction history and the prediction target.

4.3 Training and Inference

RF-SID construction and recommender training are performed in two stages. We first construct an SID for each item following Section 4.2. The resulting item-to-SID mapping remains fixed while the recommender learns to predict the next item's SID from the tokenized interaction history.

Given a training set $\mathcal{D}$ of history–target pairs, we optimize the recommender parameters θ by minimizing the negative log-likelihood:

$$\mathcal{L}_{\mathrm{rec}}(\theta) = -\frac{1}{|\mathcal{D}|} \sum_{(H_u, y_u) \in \mathcal{D}} \sum_{\ell=1}^{L} \log p_\theta\big(s_{y_u}^\ell \mid S(H_u), s_{y_u}^{<\ell}\big). \tag{8}$$

During training, each token prediction is conditioned on the ground-truth preceding tokens of the target SID.

At inference time, recommendation is formulated as autoregressive decoding: each SID token is predicted from the interaction history and the preceding tokens in the decoded sequence. For top-k recommendation, the decoding objective is

$$\widehat{\mathcal{S}}_u = \underset{s \in \mathcal{S}_\mathcal{I}}{\mathrm{TopK}} \left(\sum_{\ell=1}^{L} \log p_\theta\big(s^\ell \mid S(H_u), s^{<\ell}\big) \,;\, k \right), \tag{9}$$

where $\mathcal{S}_\mathcal{I}$ denotes the set of valid item SIDs and $\mathrm{TopK}(\cdot; k)$ selects the k highest-scoring sequences. These sequences are mapped back to their corresponding items to form the recommendation list.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate RF-SID on three domains from Amazon Reviews 2023 (Hou et al., 2026): Beauty, Toys and Games (Toys), and Sports and Outdoors (Sports). Following common sequential-recommendation protocols (Kang & McAuley, 2018; Sun et al., 2019), we apply 5-core filtering, retaining users and items with at least five interactions. We adopt a leave-last-out split, using each user's last interaction for testing, the second-to-last interaction for validation, and the remaining interactions for training. Each interaction history is truncated to its 20 most recent items. Table 1 summarizes the resulting datasets.

Table 1: Statistics of the processed datasets. Train, validation, and test report the numbers of history–target pairs.

Dataset	Catalog items	Train	Validation	Test
Beauty	207,649	4,435,713	729,406	729,410
Toys and Games	162,035	2,565,094	432,264	432,264
Sports and Outdoors	156,235	2,242,704	409,772	409,772

Baselines. We compare RF-SID with two groups of methods. ID-based sequential recommenders include SASRec (Kang & McAuley, 2018), BERT4Rec (Sun et al., 2019), and S^3-Rec (Zhou et al., 2020). SID-based methods include random SIDs, TIGER (Rajput et al., 2023), LETTER (Wang et al., 2024a), MQL4GRec (Zhai et al., 2025), and CCFRec-PQ (Liu et al., 2025c). For SID-based comparisons, we use a common four-token budget and the same downstream generative recommender. We tune baseline hyperparameters following the settings recommended in their original papers.

Evaluation protocol. We evaluate each method on the same held-out history–target pairs. For each pair, the ground-truth item is ranked against 1,000 sampled negatives, using the same candidate samples across methods. SID-based methods score each candidate by the autoregressive log-likelihood of its SID. We report Recall@k and NDCG@k for $k \in \{5, 10\}$ and select checkpoints by validation NDCG@10. Appendix A.2 gives the evaluation sample size and random seeds.

Implementation details. We obtain content representations from item text using a frozen Qwen3-4B model (Yang et al., 2025). Behavioral representations are item embeddings from a SASRec model (Kang & McAuley, 2018). Each item is represented by a four-token SID, with a codebook size of 256 for each token position. The vocabulary contains 1,025 tokens, including the padding token. The 20-item histories contain at most 80 SID tokens. The recommender uses a two-layer Transformer encoder and a two-layer causal Transformer decoder, with a hidden size of 128, four attention heads, a feed-forward dimension of 512, and a dropout rate of 0.1. We train the downstream recommenders for five epochs using AdamW with a learning rate of 10^{-3}, weight decay of 10^{-4}, and a batch size of 512.

5.2 (RQ1) OVERALL RECOMMENDATION PERFORMANCE

Table 2 reports the overall recommendation performance. For SID-based methods, we evaluate different item tokenizers using the same downstream generative recommender as RF-SID. ID-based methods use their respective recommendation architectures and serve as system-level references.

Comparison with SID-based methods. Under the common-decoder setting, RF-SID consistently outperforms all compared SID construction methods across the three datasets. The clear performance differences among tokenizers evaluated with the same recommender show that recommendation quality depends strongly on how item relations are organized within the SID.

Comparison with ID-based methods. RF-SID also consistently outperforms representative ID-based sequential recommenders across all three datasets. The comparison demonstrates the strong overall ranking performance of RF-SID as a generative recommendation approach.

Table 2: Overall recommendation performance. SID-based methods use the same downstream generative recommender. The best result in each column is shown in bold, and the second-best result is underlined.

	Beauty		Toys and Games		Sports and Outdoors	
Method	R@10	N@10	R@10	N@10	R@10	N@10
ID-based methods						
SASRec	<u>0.2781</u>	0.1659	0.2124	0.1210	<u>0.2150</u>	<u>0.1423</u>
BERT4Rec	0.2778	<u>0.1663</u>	<u>0.2134</u>	<u>0.1240</u>	0.2145	0.1417
S^3-Rec	0.2762	0.1503	0.2132	0.1095	0.2078	0.1059
SID-based methods						
Random SID	0.0290	0.0163	0.0180	0.0092	0.0206	0.0121
MQL4GRec	0.0954	0.0515	0.1459	0.0860	0.1090	0.0586
TIGER	0.1458	0.0809	0.0968	0.0522	0.1330	0.0742
LETTER	0.1917	0.1041	0.0764	0.0382	0.1252	0.0661
CCFRec-PQ	0.1917	0.1073	0.1949	0.1097	0.1664	0.0929
RF-SID	**0.2958**	**0.1848**	**0.2138**	**0.1262**	**0.2364**	**0.1470**

5.3 (RQ2) Ablation Study

We examine the contributions of the two design choices in RF-SID: assigning content and behavior to separate coordinate blocks, and adjusting the behavior representation before constructing the suffix. *Early fusion* combines the content and behavior representations before quantization. *RF-SID w/o content adjustment* retains the separate coordinate blocks but constructs the suffix directly from the behavior representation. The full *RF-SID* applies both role factorization and content adjustment. All variants otherwise use the same experimental setup.

Table 3: Ablation study of role factorization and content adjustment. Entries are R@10 / N@10.

Variant	Beauty	Toys and Games	Sports and Outdoors
Early fusion	0.2360 / 0.1409	0.1903 / 0.1087	0.1999 / 0.1048
RF-SID w/o content adjustment	0.2813 / 0.1740	0.2093 / 0.1241	0.2201 / 0.1311
RF-SID	**0.2958 / 0.1848**	**0.2138 / 0.1262**	**0.2364 / 0.1470**

Effect of role factorization. RF-SID w/o content adjustment outperforms early fusion across all three datasets. Both variants use the same content and behavior representations, SID budget, and downstream recommender; they differ in whether the two sources are fused before quantization or assigned separate coordinate blocks. This comparison shows that assigning content and behavior to dedicated coordinates is more effective than quantizing their early-fused representation.

Effect of content adjustment. Removing content adjustment from the behavior suffix also reduces performance across the three datasets. The effect is smaller than that of role factorization and varies across domains, indicating that content adjustment serves as a complementary refinement rather than the primary source of the improvement.

5.4 (RQ3) Sensitivity Analyses

We analyze the sensitivity of RF-SID to two design choices: how the fixed SID budget is allocated between content and behavior, and how their tokens are ordered within the SID.

Content–behavior coordinate allocation. With the total SID length fixed at four, we vary the numbers of content and behavior coordinates. We denote an allocation with L_c content coordinates and L_b behavior coordinates by $CL_c + BL_b$. We compare $C1 + B3$, $C2 + B2$, and $C3 + B1$ to examine how performance changes with the allocation and whether the preferred allocation varies across datasets.

Table 4: Sensitivity to content–behavior coordinate allocation. R and N denote Recall and NDCG, respectively. The highest value for each metric within each dataset is bolded.

Dataset	Allocation	R@5	R@10	N@5	N@10
Beauty	$C1 + B3$	0.2153	0.2880	0.1548	0.1783
	$C2 + B2$	**0.2231**	**0.2958**	**0.1613**	**0.1848**
	$C3 + B1$	0.2198	0.2952	0.1589	0.1831
Toys and Games	$C1 + B3$	0.1509	**0.2141**	0.1041	0.1245
	$C2 + B2$	**0.1520**	0.2138	**0.1077**	0.1262
	$C3 + B1$	0.1514	**0.2141**	0.1062	**0.1265**
Sports and Outdoors	$C1 + B3$	0.1562	0.2196	0.1088	0.1292
	$C2 + B2$	**0.1762**	**0.2364**	**0.1275**	**0.1470**
	$C3 + B1$	0.1583	0.2212	0.1114	0.1317

The allocation affects recommendation performance differently across datasets. On Beauty and Sports, $C2 + B2$ achieves the best results on all four metrics, suggesting that a balanced allocation is effective under the four-token budget. On Toys, the differences are smaller and the best allocation varies across metrics. These results show that allocating more coordinates to either source does not consistently improve performance, and that the preferred allocation depends on the domain.

SID token order. We examine how the order of content and behavior tokens affects recommendation performance. Using two tokens from each source, we compare four arrangements: CCBB, BBCC, BCBC, and CBCB, where C and B denote content and behavior tokens, respectively.

Table 5: Sensitivity to SID token order with two content and two behavior tokens. R and N denote Recall and NDCG, respectively. The highest value for each metric within each dataset is bolded.

Dataset	Order	R@5	R@10	N@5	N@10
Beauty	CCBB	**0.2231**	**0.2958**	**0.1613**	**0.1848**
	BBCC	0.1995	0.2716	0.1425	0.1657
	BCBC	0.2152	0.2873	0.1559	0.1792
	CBCB	0.2168	0.2883	0.1564	0.1795
Toys and Games	CCBB	**0.1520**	**0.2138**	**0.1077**	**0.1262**
	BBCC	0.1314	0.1864	0.0910	0.1086
	BCBC	0.1442	0.1976	0.1030	0.1202
	CBCB	0.1465	0.2002	0.1057	0.1229
Sports and Outdoors	CCBB	**0.1762**	**0.2364**	**0.1275**	**0.1470**
	BBCC	0.1577	0.2203	0.1106	0.1308
	BCBC	0.1725	0.2296	0.1241	0.1426
	CBCB	0.1737	0.2318	0.1255	0.1442

As shown in Table 5, CCBB achieves the best results across all three domains and all four metrics, while BBCC performs worst. The two interleaved arrangements fall between them, with CBCB consistently outperforming BCBC. These results show that token order matters and support placing content tokens before behavior tokens in RF-SID.

6 Understanding Role Factorization: A Rate–Distortion Perspective

SID construction maps continuous item representations to discrete token sequences. An SID with L positions and K possible tokens per position admits K^L sequences, whose indices require $L \log_2 K$ bits. This makes SID construction a form of lossy source coding: a finite coding rate limits how faithfully the continuous sources can be represented. We use mean-squared reconstruction distortion to study this limit. Rate–distortion theory characterizes the minimum achievable distortion at a given rate (Shannon, 1948; 1993; Cover et al., 1991; Gray & Neuhoff, 1998).

The experiments expose a trade-off within this finite code budget. RQ2 shows that assigning content and behavior to separate SID positions improves ranking, while RQ3 shows that their allocation matters. Separate coding can retain variation that early fusion removes, but it assigns a smaller share of the coding rate to each source. We examine this trade-off with a simplified model.

Let C and B denote scalar content and behavior sources:

$$C = U + V, \qquad B = U - V, \qquad U \sim \mathcal{N}(0, s), \quad V \sim \mathcal{N}(0, v), \quad U \perp V, \tag{10}$$

where $s, v > 0$. The component $U = (C + B)/2$ is shared by the sources, while $V = (C - B)/2$ captures their difference. For reconstructions $\widehat{C}$ and $\widehat{B}$, define the average squared-error distortion as

$$D = \frac{1}{2}\mathbb{E}\Big[(C - \widehat{C})^2 + (B - \widehat{B})^2\Big]. \tag{11}$$

Consider two coding schemes with the same total rate of $r > 0$ bits per source pair. Early fusion encodes only the average U at rate r. Separate coding assigns rate $r/2$ to each source and reconstructs each source from its own code. In the limit of long coding blocks, the Gaussian distortion–rate function is

$$D^*_{\mathcal{N}(0,\sigma^2)}(R) = \sigma^2 2^{-2R}. \tag{12}$$

For the two schemes, this gives

$$D_{\mathrm{EF}} = v + s2^{-2r}, \qquad D_{\mathrm{RF}} = (s + v)2^{-r}. \tag{13}$$

Early fusion spends the full rate on U but cannot recover V. Separate coding represents both sources, with half the rate assigned to each.

To compare these effects, consider a reference scheme that encodes each source separately at rate r. Its total rate is $2r$, and its distortion is

$$D_{\mathrm{ref}} = (s + v)2^{-2r}. \tag{14}$$

Relative to this reference, the distortion difference between the two schemes is

$$\begin{aligned} D_{\mathrm{EF}} - D_{\mathrm{RF}} &= (D_{\mathrm{EF}} - D_{\mathrm{ref}}) - (D_{\mathrm{RF}} - D_{\mathrm{ref}}) \\ &= \underbrace{v(1 - 2^{-2r})}_{\text{gain from avoiding early fusion}} - \underbrace{(s + v)(2^{-r} - 2^{-2r})}_{\text{cost of splitting the coding rate}}. \end{aligned} \tag{15}$$

Simplifying yields

$$D_{\mathrm{RF}} < D_{\mathrm{EF}} \quad \Longleftrightarrow \quad v > s2^{-r}. \tag{16}$$

Separate coding therefore achieves lower distortion when the variance of the difference component is large enough to offset the smaller rate assigned to each source. This provides a representation-level explanation for the role-factorization result: dedicated coordinates can preserve content and behavioral variation that an early-fused code discards. Appendix B gives the full derivation.

To empirically examine whether the conclusion of our simplified analysis extends to practical SID construction, we measure source reconstruction error using two-position linear probes on held-out items (Appendix C). Across all three datasets, RF-SID's dedicated blocks yield lower mean-squared error for both content and content-adjusted behavior than the compared early-fusion codes, supporting the reconstruction advantage of role factorization in the practical setting.

7 Conclusion

This paper examines how a fixed-length semantic ID can represent both item content and user behavior. Our findings suggest that fusing these sources before quantization can compromise their representation, while assigning them dedicated coordinates improves both source alignment and recommendation performance. Guided by this observation, RF-SID organizes content and behavior into separate blocks of a unified SID. Experiments on three recommendation benchmarks demonstrate its effectiveness, with ablations identifying role factorization as the main contributor to the gains. The rate–distortion analysis further illustrates how preserving differences between sources can offset the cost of dividing a finite coding budget. Together, these results highlight the allocation of discrete coordinates as an important design choice in generative recommendation: the organization of a semantic ID shapes how effectively a recommender can use heterogeneous item information.

8 AI USE STATEMENT

Generative AI tools were used to assist with literature search and manuscript editing, including language polishing and organization. The authors take responsibility for the final content of the paper.

9 ETHICS STATEMENT

This study uses publicly available Amazon Reviews 2023 data for offline recommendation experiments. We did not collect additional user data or conduct experiments involving newly recruited participants. Recommendation models trained on historical interactions may reflect biases in those data and influence the exposure of items. These effects should be considered when applying the proposed method in deployed systems.

10 REPRODUCIBILITY STATEMENT

Section 5 specifies the datasets, data splits, model architecture, optimization settings, and evaluation metrics. Appendices A.1 and A.2 describe the preliminary study and recommendation evaluation protocols, including sampling procedures and random seeds. Appendix B provides the assumptions and full derivation of the rate–distortion analysis, and Appendix C details the held-out source-reconstruction study. Code is released anonymously at `https://anonymous.4open.science/r/rf-sid-7CD1/`.

REFERENCES

Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems*, 35:31668–31683, 2022.

Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*, 2020.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM conference on recommender systems*, pp. 299–315, 2022.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE transactions on information theory*, 44 (6):2325–2383, 1998.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking llms as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3251–3265, 2026.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11513–11522. IEEE, 2022.

Chang Liu, Yimeng Bai, Xiaoyan Zhao, Yang Zhang, Fuli Feng, and Wenge Rong. Discrec: Disentangled semantic-collaborative modeling for generative recommendation. *arXiv preprint arXiv:2506.15576*, 2025a.

Enze Liu, Bowen Zheng, Cheng Ling, Lantao Hu, Han Li, and Wayne Xin Zhao. Generative recommender with end-to-end learnable item tokenization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 729–739, 2025b.

Enze Liu, Bowen Zheng, Wayne Xin Zhao, and Ji-Rong Wen. Bridging textual-collaborative gap through semantic codes for sequential recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 1788–1798, 2025c.

Haohao Qu, Wenqi Fan, Zihuai Zhao, and Qing Li. Tokenrec: Learning to tokenize id for llm-based generative recommendations. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

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

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 811–820, 2010.

Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pp. 1177–1178, 2010.

Claude E Shannon. Coding theorems for a discrete source with a fidelity criterioninstitute of radio engineers, international convention record, vol. 7, 1959. 1993.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1441–1450, 2019.

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems*, 36:46345–46361, 2023.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 565–573, 2018.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. Transformer memory as a differentiable search index. *Advances in neural information processing systems*, 35:21831–21843, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Hao Wang, Naiyan Wang, and Dit-Yan Yeung. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1235–1244, 2015.

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

Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, et al. Eager: Two-stream generative recommender with behavior-semantic collaboration. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3245–3254, 2024b.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1437–1445, 2019.

Longtao Xiao, Haozhao Wang, Cheng Wang, Linfei Ji, Yifan Wang, Jieming Zhu, Zhenhua Dong, Rui Zhang, and Ruixuan Li. Unger: Generative recommendation with a unified code via semantic and collaborative integration. *ACM Transactions on Information Systems*, 44(2):1–31, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jianyang Zhai, Zi-Feng Mai, Chang-Dong Wang, Feidiao Yang, Xiawu Zheng, Hui Li, and Yonghong Tian. Multimodal quantitative language for generative recommendation. In *International Conference on Learning Representations*, volume 2025, pp. 74424–74441, 2025.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3872–3880, 2021.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1893–1902, 2020.

Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 935–943, 2023.

A Experimental Protocols

A.1 Preliminary Study

On Beauty, we compare C4, B4, J4, and RF-SID C1+B3 using four-token SIDs and train one parallel-head recommender for each design with the same architecture and training settings. For item i, we concatenate the four learned input embeddings to form $[E_1(s_i^1); E_2(s_i^2); E_3(s_i^3); E_4(s_i^4)]$. We compute centered linear CKA between these representations and the frozen content and content-adjusted behavior representations, following Equation equation 1. The reported scores are means over five independently sampled sets of 10,000 items, with sampling seeds 2026–2030.

For ranking, the parallel heads assign candidate item i the score $\sum_{\ell=1}^{4} \log p_\ell(s_i^\ell \mid H_u)$. We evaluate NDCG@10 on 50,000 held-out contexts, using the same 1,000 sampled negatives per context across the four designs and sampling seed 9090.

A.2 Main Evaluation

For each dataset, all methods are evaluated on the same 50,000 held-out history–target pairs. Each target is ranked against 1,000 sampled negatives, with candidate sampling seed 9090 shared across methods. The training and validation seed is 2026. We select checkpoints by validation NDCG@10 and report Recall and NDCG at cutoffs 5 and 10.

B Derivation of the Rate–Distortion Model

B.1 Model and distortion

Let $(C_i, B_i) = (U_i + V_i, U_i - V_i)$ be i.i.d. source pairs, where $U_i \sim \mathcal{N}(0, s)$ and $V_i \sim \mathcal{N}(0, v)$ are independent, with $s, v > 0$. Thus $\mathrm{Var}(C_i) = \mathrm{Var}(B_i) = s + v$. At total rate r bits per pair, early fusion encodes only U^n at rate r, while separate encoding assigns rate $r/2$ to each source and each decoder uses only its corresponding code. We use average squared error

$$D = \frac{1}{2n} \sum_{i=1}^{n} \mathbb{E}\Big[(C_i - \widehat{C}_i)^2 + (B_i - \widehat{B}_i)^2\Big]. \tag{17}$$

For a Gaussian source $X \sim \mathcal{N}(0, \sigma^2)$, the asymptotic rate–distortion function under squared error is (Cover et al., 1991)

$$D_X^*(R) = \sigma^2 2^{-2R}. \tag{18}$$

B.2 Scheme distortions

For early fusion, let T be the rate-r message encoded from U^n. Independence gives $\mathbb{E}[C_i \mid T] = \mathbb{E}[B_i \mid T] = \mathbb{E}[U_i \mid T]$ and

$$\begin{aligned} \mathbb{E}[(C_i - \mathbb{E}[U_i \mid T])^2] &= \mathbb{E}[(U_i - \mathbb{E}[U_i \mid T])^2] + v, \\ \mathbb{E}[(B_i - \mathbb{E}[U_i \mid T])^2] &= \mathbb{E}[(U_i - \mathbb{E}[U_i \mid T])^2] + v. \end{aligned} \tag{19}$$

Applying the Gaussian rate–distortion function to U at rate r yields

$$D_{\mathrm{EF}} = v + s2^{-2r}. \tag{20}$$

Each separate encoder observes one source of variance $s + v$ at rate $r/2$; therefore

$$D_{\mathrm{RF}} = (s + v)2^{-2(r/2)} = (s + v)2^{-r}. \tag{21}$$

Equal rate allocation minimizes this distortion: for $r_C + r_B = r$,

$$\frac{s + v}{2}\left(2^{-2r_C} + 2^{-2r_B}\right) \geq (s + v)2^{-r}, \tag{22}$$

with equality at $r_C = r_B = r/2$ by AM–GM.

B.3 TRADE-OFF AND CONDITION

As a reference, encoding each source at rate r gives $D_{\mathrm{ref}} = (s+v)2^{-2r}$ at total rate $2r$. Thus

$$\begin{aligned} D_{\mathrm{EF}} - D_{\mathrm{RF}} &= (D_{\mathrm{EF}} - D_{\mathrm{ref}}) - (D_{\mathrm{RF}} - D_{\mathrm{ref}}) \\ &= \underbrace{v(1-2^{-2r})}_{\text{gain from avoiding early fusion}} - \underbrace{(s+v)(2^{-r}-2^{-2r})}_{\text{cost of splitting the rate}} \\ &= (1-2^{-r})(v - s2^{-r}). \end{aligned} \tag{23}$$

Hence

$$D_{\mathrm{RF}} < D_{\mathrm{EF}} \quad \Longleftrightarrow \quad v > s2^{-r}. \tag{24}$$

C HELD-OUT SOURCE RECONSTRUCTION

We compare reconstruction of content and content-adjusted behavior from RF-SID, early fusion (EF), and Concat-RQ on all three datasets. The source vectors are aligned by item ID across tokenizers, and each item receives equal weight. Tokenizers remain fixed throughout the evaluation.

Concat-RQ. Concat-RQ concatenates the two unit-normalized source vectors, divides the result by $\sqrt{2}$, and applies four levels of residual quantization with 256 codewords per level. We fit each level with MiniBatchKMeans using batch size 4096, one initialization, at most 100 iterations, reassignment ratio 0.01, and seed $2026 + \ell$ for zero-indexed level ℓ. The codebooks contain 262,144 scalar parameters for Concat-RQ and 131,072 for RF-SID and EF.

Linear probes. For source vector x_i and a set of SID positions A, we fit a linear probe on the concatenated token one-hot encodings. Its prediction has the form

$$\widehat{x}_i = a + \sum_{\ell \in A} h_\ell(s_i^\ell), \tag{25}$$

where a is an intercept and each h_ℓ is a position-specific vector lookup table. The probe parameters are fitted on 80% of catalog items, and mean squared Euclidean error is evaluated on the remaining 20%, using split seed 20260923. Each RF-SID probe reads the two positions assigned to its source. For EF and Concat-RQ, we fit a probe for each of the six two-position subsets and report the minimum test error.

Results. Table 6 reports lower held-out reconstruction error for RF-SID on both sources across all three datasets. At the same two-position budget, the dedicated blocks make their corresponding source representations more accessible to a linear probe.

Table 6: Held-out mean squared Euclidean reconstruction error using two token positions. Baseline entries report the minimum test error over all six two-position subsets. Lower is better. Behavior denotes the content-adjusted input.

Dataset	Source	RF-SID	EF	Concat-RQ
Beauty	Content	**0.1972**	0.3115	0.2871
	Behavior	**0.7735**	0.8487	0.8105
Toys	Content	**0.1120**	0.1848	0.1790
	Behavior	**0.7753**	0.8259	0.8061
Sports	Content	**0.1069**	0.1739	0.1679
	Behavior	**0.7649**	0.8164	0.8004