ON THE VULNERABILITY OF SEMANTIC IDS IN GENERATIVE RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative recommenders (GenRec) are increasingly adopted on real-world platforms. Semantic IDs (SIDs) are widely used in GenRec to map item descriptions to discrete tokens for autoregressive recommendation. Editing a listing can change its SID while the recommender remains fixed, creating a new attack surface for targeted item promotion. We introduce **SIDAttack**, which appends a short suffix to a target item's description while treating the victim SID tokenizer and recommender as black boxes. It searches using gradients from available text encoders and public-item anchors selected from publicly visible popular items, without victim recommendation feedback. Across product, news, and short-video catalogs and two SID tokenizers, SIDAttack produces the largest mean exposure gains among the evaluated baselines when the search and victim encoders match. Although the gain is smaller when the victim encoder is held out, SIDAttack still increases exposure relative to the unattacked baseline. We also find that residual-quantization margins strongly predict which items undergo SID changes. Guided by this finding, we propose margin-regularized codebook construction, which reduces attack-induced exposure while largely preserving clean recommendation utility.

1 INTRODUCTION

Generative recommenders (GenRec) are increasingly adopted in real-world platforms (Zhang et al., 2026; Deng et al., 2025). Semantic IDs (SIDs) are widely used in GenRec to map item descriptions to discrete tokens, typically obtained through residual quantization (RQ), for autoregressive recommendation (Rajput et al., 2023; Wang et al., 2024; Ju et al., 2025). SIDs allow semantically related items to share parts of their identifiers, but small textual changes can shift item representations across quantization boundaries and alter these codes. Changes to early SID tokens may redirect generation paths and affect item exposure.

Recent studies provide initial evidence of this instability. Wang et al. (2026b) show that alternative descriptions can still retrieve the corresponding original item as the nearest semantic match in 99.57% of controlled cases while changing the exact SID in 38.4% of cases. Other concurrent studies identify codeword confusion, hard routing decisions, and propagation of early quantization errors as important limitations of current SID tokenizers (Wang et al., 2026a; Cheng et al., 2026). We find that appending a small textual perturbation to an item's original description can redirect a target item to a SID region associated with popular public items. These findings establish that SID assignments can be unstable, but two important questions remain.

(i) *Which items are most vulnerable to SID changes, and can we predict it?*
(ii) *Can we deliberately steer those changes toward a chosen public-item anchor?*

To answer the first question, we analyze the geometry of residual quantization. As shown in Figure 1, items closer to RQ decision boundaries are substantially more likely to change their SIDs under natural rewrites. A routing margin based on the gap between a residual representation's distances to its assigned and nearest competing centroids is a strong predictor of SID flips, with AUC values of 0.930, 0.869, and 0.847 at the three SID levels. The consistent pattern across all three levels indicates that SID vulnerability follows a predictable geometric structure rather than arising from arbitrary

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.

textual noise. This observation motivates our subsequent attack, which searches for representation shifts that deliberately alter and steer the resulting SID route.

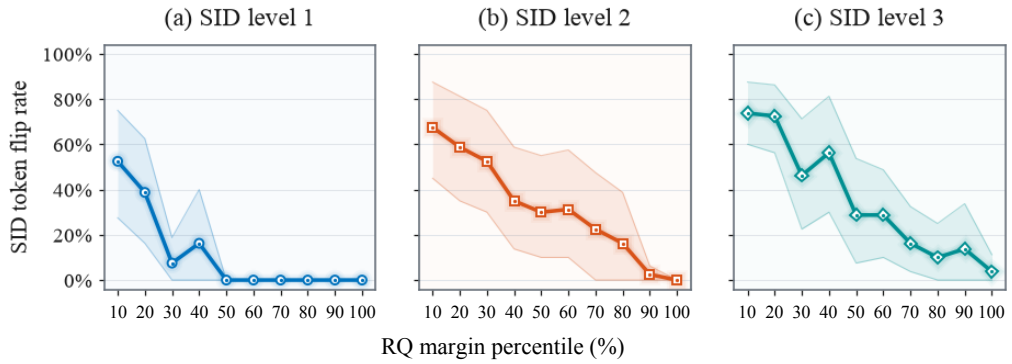


Figure 1. SID flip rate versus residual-quantization margin on Amazon Beauty (eight natural rewrites per listing; deciles labeled by the upper percentile). Items closer to an RQ boundary flip more often at all three SID levels. Shaded bands are 95% item-bootstrap intervals.

We next address the second question by proposing **Semantic ID Attack (SIDAttack)**, an append-only attack that leaves the original title and factual description unchanged. SIDAttack uses available text encoders and public-item anchors selected from popular catalog items to optimize a short suffix, without access to the downstream recommender. It builds a semantically compatible and diverse anchor portfolio and introduces **AnchorPort Greedy Coordinate Gradient (AnchorPort-GCG)** to search for a suffix that steers the public representation toward one of the anchors. Whether the victim SID then matches a public-item anchor is measured only after the suffix is fixed. Because the search uses encoder gradients rather than paraphrasing the listing, the representation can move across quantization boundaries and raise the target's generative exposure. Figure 2 illustrates how short suffix variants change a red-spirit thermometer's SID; these variants are not SIDAttack's final optimized suffix.

Original. SEOH red-spirit thermometer, student grade, −20 to 150C, total immersion. Length about 12 inches, graduations every 1 degree.

suffix = none
`[99, 10, 7]`

suffix = "infrared digital"
`[99, 21, 5]`

suffix = "infrared digital forehead"
`[99, 21, 11]`

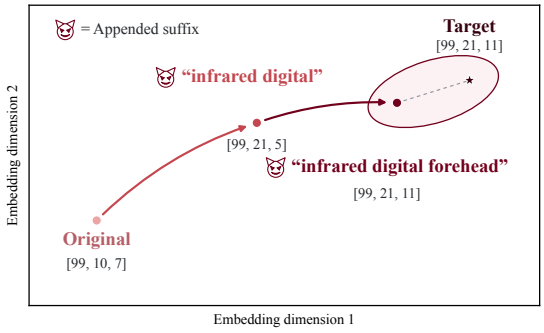


Figure 2. Short-suffix variants of a red-spirit thermometer listing, with the original description unchanged. Left: each variant and its SID. Right: the variants along an illustrated route toward an infrared-thermometer item. None of the short variants shown is SIDAttack's final optimized suffix; that suffix for the same item appears in Appendix B.

To mitigate this vulnerability, we propose margin-regularized codebook construction, which penalizes small routing margins so that assignments are more stable, attack-induced exposure gains shrink, and the recommendation utility of benign items is largely preserved.

The main contributions can be summarized as follows.

- We show that small textual perturbations can manipulate semantic-ID assignments, exposing a previously underexplored attack surface in generative recommendation.
- We find that residual-quantization margins are a strong predictor of this instability.
- We introduce **SIDAttack** to exploit SID instability and increase item exposure. Using encoder gradients and public-item anchors, it steers representations across residual-quantization boundaries

without access to the downstream recommender and outperforms the evaluated catalog-rewriting baselines in matched-encoder comparisons.
- We propose margin-regularized codebook construction as a defense. Guided by the margin analysis, it reduces attack-induced exposure gains and increases the suffix budget needed for comparable gains while largely preserving the recommendation utility of benign items.

2 Related Work

2.1 Semantic IDs in Generative Recommendation

Semantic IDs encode items as compact code sequences, allowing generative recommenders to retrieve items autoregressively and share information across semantically related items. TIGER constructs these codes by residual-quantizing content embeddings and trains a sequence-to-sequence Transformer to predict the next item's SID (Rajput et al., 2023). Building on content-based tokenization, LETTER incorporates collaborative alignment and code-diversity regularization to improve the recommendation value of the learned identifiers (Wang et al., 2024). GRID complements these designs with a modular evaluation of encoders, quantizers, generation architectures, and decoding choices (Ju et al., 2025). These studies establish the utility of SIDs for retrieval; we examine whether changes to item text can manipulate the resulting routes while the victim models remain fixed.

2.2 Semantic ID Instability

The semantic structure of SIDs does not guarantee stable assignments. Wang et al. (2026b) show that alternative descriptions identifying the same item can receive different SIDs, and support decoding prefixes with item-level rankings before beam pruning. At the tokenizer level, DRQ addresses codeword confusion by separating continuous representation learning from residual K-Means assignment (Wang et al., 2026a). CapsID instead uses soft capsule routing and confidence-driven code lengths to address hard assignments and propagated quantization errors (Cheng et al., 2026). These findings document that SID assignments can change, but they study accidental variation under alternative descriptions or tokenizer redesign, not an attack that steers a public representation toward a chosen public-item anchor.

2.3 Attacks on Generative Recommendation

Attacks on generative recommendation differ in their access to training data and deployed-system inputs. BadRec modifies item titles and injects fabricated user interactions into training data, teaching the recommender to favor items containing a trigger (Ning et al., 2026). At inference time, Preference Manipulation Attacks instead exploit attacker-controlled webpages or plugin descriptions that enter an LLM's context, using instructions or persuasive claims without access to model weights or retraining (Nestaas et al., 2025). SIDAttack targets the discrete SID route produced from a catalog edit, making it applicable between model retraining cycles.

2.4 Promotion Attacks in Generative Recommendation

Text-based promotion attacks share the goal of increasing target-item exposure, but differ in their optimization signals and editing permissions. For LLM-based recommendation, Stealthy Attack includes heuristic title edits, such as appending positive words, as well as black-box searches guided by recommendation scores (Zhang et al., 2024). For text-aware recommenders, ATR optimizes rewrites using white-box recommendation gradients or a surrogate distilled from black-box queries; its in-context variant uses examples generated by this optimization (Oh et al., 2024). TextSimu rewrites target descriptions by extracting and simulating textual characteristics of popular items through multi-agent refinement, and evaluates the resulting edits as training-data poisoning (Wang et al., 2025). For frozen news recommenders, LANCE trains a reflector using ranking feedback on exploration articles, then rewrites new target articles without requiring feedback for each target (Zhao et al., 2025). PUDA addresses unknown models and prompts, but uses partial interaction histories and limited victim queries to distill a surrogate before combining text edits with interaction poisoning (Zhao et al., 2026).

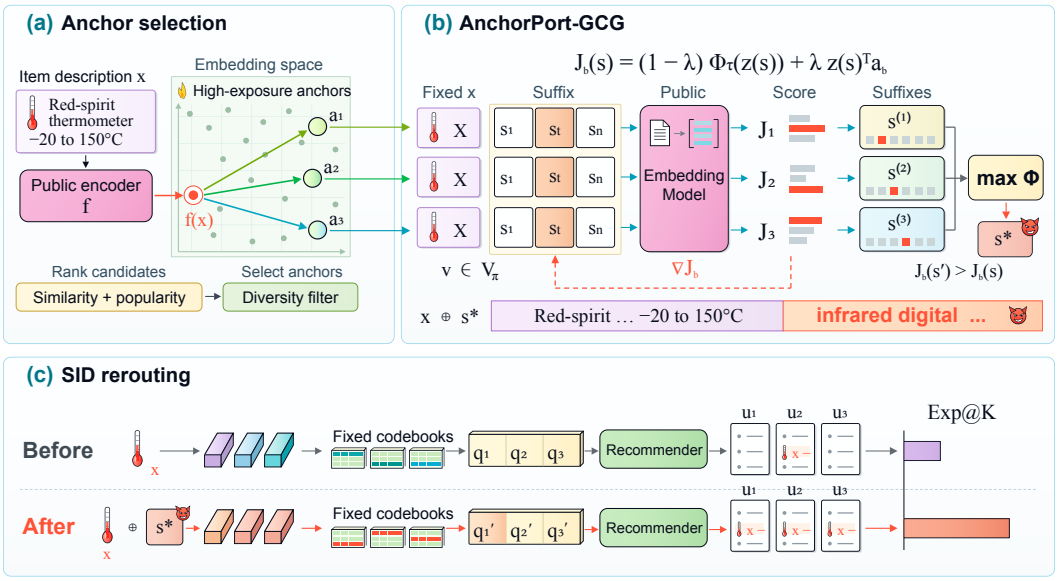


Figure 3. Overview of SIDAttack: an attacked item is paired with diverse public-item anchors, and AnchorPort-GCG searches for an append-only suffix that steers the public representation toward one of them.

These methods rewrite or polish the catalog text, so the edited embedding usually stays near the original item: a paraphrase or popularity-styled description cannot move far in representation space without ceasing to describe the same product. They also typically need recommendation scores, interaction histories, or a surrogate of the downstream ranker. SIDAttack instead optimizes a short appended suffix with white-box encoder gradients, which can drive the representation across residual-quantization boundaries toward a public-item anchor while leaving the original title and factual description unchanged, and without interaction data or recommendation feedback.

3 Method

In this section, we introduce SIDAttack as follows. First, we formalize the frozen victim system and the attacker's access (Section 3.1). Next, we select diverse public-item anchors (Section 3.2) and search for a short append-only suffix with AnchorPort-GCG (Section 3.3), steering the public representation toward one of them without the downstream recommender. Finally, motivated by the observed association between RQ margins and SID sensitivity, we describe margin-regularized codebook construction (Section 3.4).

3.1 System and Attack Setting

For an item description x, the victim encoder E maps it to a continuous vector. Each level keeps a codebook of K vectors. Starting from $E(x)$, Q repeats the following for L levels: it selects the closest codebook vector, records that vector's index as the next SID token, and subtracts it.

$$Q(E(x)) = (q_1, \ldots, q_L), \qquad q_\ell \in \{1, \ldots, K\}, \tag{1}$$

RQ-KMeans builds these codebooks by K-means on frozen encoder vectors. RQ-VAE learns the codebooks inside an autoencoder. The generator p_θ predicts these tokens autoregressively. The encoder, tokenizer, and generator remain fixed during the attack. Catalog edits are re-encoded into SIDs without retraining these models.

Threat model. An item provider can modify the text of its own catalog items and observe public item descriptions and popularity signals. We cover two settings: access to the platform encoder ($f = E$), or access only to a different public encoder when E is unavailable ($f \neq E$). In both settings the attacker has white-box access to f, including its gradients, and to the public popular-item pool $\mathcal{H}$. Every baseline reported in the comparison tables receives this same information. Random,

HotFlip, Stealthy Attack, the rewrite agent, and TextSimu may therefore use f and $\mathcal{H}$ exactly as SIDAttack can. None of these methods observes the victim SID tokenizer, the codebook, item-to-SID assignments, user interaction histories, generator parameters, or recommendation outputs, and none can query the recommender for attack feedback. Methods differ only in the edit they apply and in which part of this shared information their search uses. SIDAttack restricts edits to a short appended suffix s, preserving the original title and description. The evaluator measures victim SIDs and recommendation exposure only after the edit is fixed.

Let $\mathcal{H} = \{(h_j, r_j)\}_{j=1}^{M}$ denote the popular-item pool, where h_j is item j's normalized embedding under f and r_j is a public popularity signal, such as review count or sales rank. Writing the edited listing as $x \oplus s$, we define $z(s) = f(x \oplus s)/\|f(x \oplus s)\|_2$, and $z_0 = f(x)/\|f(x)\|_2$. Full notation is provided in Appendix H.

3.2 Public-Item Anchor Portfolio

The anchors are popular public items. A semantically compatible one may still be difficult to approach with a short suffix, so we select several diverse anchors. We rank public candidates by semantic similarity and popularity:

$$u_j(x) = z_0^\top h_j + \beta_{\text{sel}} \log(1 + r_j),$$

where β_{sel} weights the popularity signal. From the highest-scoring candidates, we greedily select B public-item anchors $\mathcal{A}(x) = \{a_1, \ldots, a_B\}$. To ensure diversity, we require $a_p^\top a_q \le \delta$ for $p \neq q$, where δ caps pairwise similarity between anchors.

3.3 AnchorPort-GCG Search

AnchorPort-GCG combines a global score over the public popular-item pool with an anchor-specific objective for each search beam. The global score is

$$\Phi_\tau(z) = \tau \log \sum_{j=1}^{M} \exp\left(\frac{z^\top h_j + \beta_{\text{exp}} \log(1 + r_j)}{\tau}\right), \tag{2}$$

where z is the normalized item representation, τ controls the smoothness of the maximum, and β_{exp} weights the popularity signal.

Each selected anchor a_b initializes one search beam. Its objective is

$$\mathcal{J}_b(s) = (1 - \lambda)\Phi_\tau\big(z(s)\big) + \lambda z(s)^\top a_b, \tag{3}$$

where λ balances the two terms. The global term allows the preferred direction to change during search, while the anchor term encourages different beams to explore distinct regions.

Greedy Coordinate Gradient (Zou et al., 2023) proposes token replacements, which are evaluated through f and accepted only when they improve $\mathcal{J}_b$. Suffix tokens must remain stable after decoding and re-encoding. Among the valid suffixes from all beams, we choose $s^\star$ with the highest global score $\Phi_\tau(z(s))$. The update formula, vocabulary constraints, search-budget allocation, and tie-breaking rules are given in Appendix A.1.

3.4 Margin-Regularized Codebook Construction

To stabilize SID assignments near quantization boundaries, we regularize the codebook while keeping the victim encoder E frozen. Only codebook parameters are updated in this stage.

At residual level ℓ, let $z^{(\ell)}$ denote the residual representation being quantized, $c_{q_\ell}^{(\ell)}$ its assigned code vector, and $c_{(2)}^{(\ell)}$ the nearest competing code vector. We define the routing margin as

$$m_\ell = \left\| z^{(\ell)} - c_{(2)}^{(\ell)} \right\|_2 - \left\| z^{(\ell)} - c_{q_\ell}^{(\ell)} \right\|_2. \tag{4}$$

We augment the standard residual-quantization objective $\mathcal{L}_Q$ with a penalty on small margins:

$$\mathcal{L} = \mathcal{L}_Q + \gamma \sum_{\ell=1}^{L} \mathbb{E}\left[\max(0, \eta - m_\ell)\right], \tag{5}$$

where η is the desired minimum margin and γ controls the strength of the regularization. For RQ-KMeans, $\mathcal{L}_Q$ is the standard reconstruction objective. For RQ-VAE, it contains the original codebook and commitment losses.

The penalty encourages a larger distance gap to the nearest competing code, making assignments less sensitive to representation shifts.

4 EXPERIMENTS

Our experiments are organized around the following research questions (RQs):

RQ1: Can SIDAttack increase the exposure of target items, including news and video?
RQ2: Does SIDAttack transfer across generative recommenders with different victim encoders?
RQ3: How does increasing the suffix budget affect item exposure?
RQ4: Which components contribute to SIDAttack's effectiveness?
RQ5: Can margin regularization defend against the attack while preserving utility?

Experimental setup. We randomly sample 100 target items from each of Amazon Beauty, Industrial & Scientific, and Toys & Games (Ni et al., 2019) and share them across methods in a domain; dataset details appear in Appendix C. The main RQ-KMeans comparison in Table 1 uses Qwen3-Embedding-4B (Zhang et al., 2025) as both the public search encoder f and the victim encoder E, with a per-domain codebook and TIGER-style generator (Rajput et al., 2023). We also evaluate RQ-VAE tokenizers (L=3, as in RQ-KMeans), a matched BGE–BGE setting (Xiao et al., 2024), and the same RQ-KMeans protocol on MIND and KuaiRec (Wu et al., 2020; Gao et al., 2022). Transfer, budget, and defense studies use the Amazon RQ-KMeans protocol unless stated otherwise. The geometric analysis supporting Figure 1 is in Appendix E.

Baselines. We compare SIDAttack with Random and HotFlip (Ebrahimi et al., 2018), which modify only an appended suffix, and with catalog-rewrite adaptations of Stealthy Attack (Zhang et al., 2024) and TextSimu (Wang et al., 2025), together with a rewrite agent. As stated in Section 3.1, every baseline is given the same encoder and public-pool access as SIDAttack. All append-only methods use $n = 24$ tokens unless the suffix budget is explicitly varied. Implementation details are given in Appendix A.

Evaluation metrics. Let $\mathcal{U}_{\text{test}}$ be the test users. For an attacked listing i and cutoff K, exposure counts how often i enters a top-K recommendation:

$$\text{Exp@}K(i) = \sum_{u \in \mathcal{U}_{\text{test}}} \mathbf{1}[i \in \text{Top}K(u)]. \tag{6}$$

We report the mean over the N attacked listings,

$$\text{Exp@}K = \frac{1}{N} \sum_{i=1}^{N} \text{Exp@}K(i), \tag{7}$$

Tables use K=50; ΔExp@50 is the change in this mean after the catalog edit. Unless a caption says otherwise, After and Δ values are mean $\pm$ standard deviation over three runs, and the clean Before value is shared within a victim setting. For clean utility we report Recall@10 and Recall@50 on the unattacked test set. Target@k is the percentage of listings whose first k SID tokens match the selected anchor, which is frozen before evaluation and shared by Random, HotFlip, and SIDAttack.

4.1 SIDATTACK EFFECTIVENESS

We first evaluate whether SIDAttack increases the exposure of target items under two SID tokenizers (Tables 1 and 2). RQ-VAE uses a separate victim stack and new suffixes.

Table 1. Main RQ-KMeans results under matched Qwen3-Embedding-4B encoders.

	Beauty			Industrial & Sci.			Toys & Games		
Method	Before	After	ΔExp@50	Before	After	ΔExp@50	Before	After	ΔExp@50
Random	0.09	$0.03_{\pm 0.02}$	$-0.06_{\pm 0.02}$	0.14	$2.63_{\pm 0.24}$	$+2.49_{\pm 0.24}$	0.21	$7.20_{\pm 0.53}$	$+6.99_{\pm 0.53}$
HotFlip	0.09	$0.09_{\pm 0.03}$	$0.00_{\pm 0.03}$	0.14	$0.39_{\pm 0.05}$	$+0.25_{\pm 0.05}$	0.21	$8.65_{\pm 0.71}$	$+8.44_{\pm 0.71}$
Stealthy Attack	0.09	$0.47_{\pm 0.05}$	$+0.38_{\pm 0.05}$	0.14	$1.01_{\pm 0.08}$	$+0.87_{\pm 0.08}$	0.21	$38.71_{\pm 3.28}$	$+38.50_{\pm 3.28}$
Rewrite agent	0.09	$0.06_{\pm 0.02}$	$-0.03_{\pm 0.02}$	0.14	$0.69_{\pm 0.06}$	$+0.55_{\pm 0.06}$	0.21	$5.41_{\pm 0.44}$	$+5.20_{\pm 0.44}$
TextSimu	0.09	$0.10_{\pm 0.03}$	$+0.01_{\pm 0.03}$	0.14	$0.62_{\pm 0.07}$	$+0.48_{\pm 0.07}$	0.21	$64.33_{\pm 5.19}$	$+64.12_{\pm 5.19}$
SIDAttack	0.09	$6.23_{\pm 0.41}$	$+6.14_{\pm 0.41}$	0.14	$61.49_{\pm 3.87}$	$+61.35_{\pm 3.87}$	0.21	$182.85_{\pm 11.64}$	$+182.64_{\pm 11.64}$

Table 2. Main RQ-VAE results with independently trained victim stacks.

	Beauty			Industrial & Sci.			Toys & Games		
Method	Before	After	ΔExp@50	Before	After	ΔExp@50	Before	After	ΔExp@50
Random	0.13	$0.78_{\pm 0.06}$	$+0.65_{\pm 0.06}$	0.18	$0.17_{\pm 0.03}$	$-0.01_{\pm 0.03}$	0.16	$0.34_{\pm 0.05}$	$+0.18_{\pm 0.05}$
HotFlip	0.13	$0.63_{\pm 0.05}$	$+0.50_{\pm 0.05}$	0.18	$0.27_{\pm 0.04}$	$+0.09_{\pm 0.04}$	0.16	$0.08_{\pm 0.03}$	$-0.08_{\pm 0.03}$
Stealthy Attack	0.13	$0.41_{\pm 0.04}$	$+0.28_{\pm 0.04}$	0.18	$0.83_{\pm 0.07}$	$+0.65_{\pm 0.07}$	0.16	$5.01_{\pm 0.42}$	$+4.85_{\pm 0.42}$
Rewrite agent	0.13	$0.08_{\pm 0.02}$	$-0.05_{\pm 0.02}$	0.18	$0.53_{\pm 0.05}$	$+0.35_{\pm 0.05}$	0.16	$3.26_{\pm 0.28}$	$+3.10_{\pm 0.28}$
TextSimu	0.13	$0.13_{\pm 0.03}$	$0.00_{\pm 0.03}$	0.18	$0.40_{\pm 0.04}$	$+0.22_{\pm 0.04}$	0.16	$2.96_{\pm 0.25}$	$+2.80_{\pm 0.25}$
SIDAttack	0.13	$3.91_{\pm 0.34}$	$+3.78_{\pm 0.34}$	0.18	$14.68_{\pm 1.16}$	$+14.50_{\pm 1.16}$	0.16	$8.01_{\pm 0.72}$	$+7.85_{\pm 0.72}$

SIDAttack produces the largest post-edit exposure and the largest $\Delta\text{Exp@50}$ in every domain under both tokenizers. Under RQ-KMeans, mean Exp@50 rises from 0.09 to 6.23 on Beauty, 0.14 to 61.49 on Industrial & Scientific, and 0.21 to 182.85 on Toys & Games. RQ-VAE is weaker ($0.13 \rightarrow 3.91$, $0.18 \rightarrow 14.68$, $0.16 \rightarrow 8.01$), so the attack depends on the victim tokenizer. Catalog rewriting stays far behind except on Toys, where TextSimu still trails SIDAttack. Cases are in Appendix B.

Matched-encoder comparison. Table 3 repeats the same ranking when both encoders are BGE: SIDAttack reaches 9.14, 74.65, and 211.80 after the edit.

Table 3. Matched BGE–BGE results under RQ-KMeans. After and $\Delta\text{Exp@50}$ are the mean $\pm$ standard deviation over three seeds.

	Beauty			Industrial & Sci.			Toys & Games		
Method	Before	After	ΔExp@50	Before	After	ΔExp@50	Before	After	ΔExp@50
Random	0.41	$0.33_{\pm 0.01}$	$-0.08_{\pm 0.01}$	0.37	$2.35_{\pm 0.17}$	$+1.98_{\pm 0.17}$	0.44	$6.27_{\pm 0.48}$	$+5.83_{\pm 0.48}$
HotFlip	0.41	$1.00_{\pm 0.06}$	$+0.59_{\pm 0.06}$	0.37	$1.23_{\pm 0.08}$	$+0.86_{\pm 0.08}$	0.44	$7.85_{\pm 0.57}$	$+7.41_{\pm 0.57}$
Stealthy Attack	0.41	$0.83_{\pm 0.05}$	$+0.42_{\pm 0.05}$	0.37	$1.51_{\pm 0.12}$	$+1.14_{\pm 0.12}$	0.44	$44.05_{\pm 3.89}$	$+43.61_{\pm 3.89}$
Rewrite agent	0.41	$0.35_{\pm 0.02}$	$-0.06_{\pm 0.02}$	0.37	$1.00_{\pm 0.06}$	$+0.63_{\pm 0.06}$	0.44	$6.22_{\pm 0.46}$	$+5.78_{\pm 0.46}$
TextSimu	0.41	$0.55_{\pm 0.03}$	$+0.14_{\pm 0.03}$	0.37	$1.11_{\pm 0.09}$	$+0.74_{\pm 0.09}$	0.44	$68.39_{\pm 5.63}$	$+67.95_{\pm 5.63}$
SIDAttack	0.41	$9.14_{\pm 0.52}$	$+8.73_{\pm 0.52}$	0.37	$74.65_{\pm 5.14}$	$+74.28_{\pm 5.14}$	0.44	$211.80_{\pm 15.72}$	$+211.36_{\pm 15.72}$

News and short-video domains. Table 4 repeats the RQ-KMeans protocol on MIND and KuaiRec (Wu et al., 2020; Gao et al., 2022). SIDAttack remains strongest on both, raising mean Exp@50 from 0.12 to 25.80 on MIND and from 0.18 to 116.91 on KuaiRec. The larger shift on KuaiRec shows that short-video metadata can be steered toward public-item anchors as well, so the vulnerability is not limited to product descriptions. TextSimu is the strongest rewrite baseline on both catalogs, but still trails SIDAttack by a wide margin. RQ-VAE results are in Appendix D.

Table 4. RQ-KMeans results on MIND and KuaiRec.

	MIND			KuaiRec		
Method	Before	After	ΔExp@50	Before	After	ΔExp@50
Random	0.12	$0.53_{\pm 0.06}$	$+0.41_{\pm 0.06}$	0.18	$1.91_{\pm 0.17}$	$+1.73_{\pm 0.17}$
HotFlip	0.12	$1.24_{\pm 0.11}$	$+1.12_{\pm 0.11}$	0.18	$3.52_{\pm 0.29}$	$+3.34_{\pm 0.29}$
Stealthy Attack	0.12	$5.75_{\pm 0.48}$	$+5.63_{\pm 0.48}$	0.18	$26.65_{\pm 2.14}$	$+26.47_{\pm 2.14}$
Rewrite agent	0.12	$1.08_{\pm 0.09}$	$+0.96_{\pm 0.09}$	0.18	$8.14_{\pm 0.72}$	$+7.96_{\pm 0.72}$
TextSimu	0.12	$9.86_{\pm 0.81}$	$+9.74_{\pm 0.81}$	0.18	$45.03_{\pm 3.61}$	$+44.85_{\pm 3.61}$
SIDAttack	0.12	$25.80_{\pm 1.94}$	$+25.68_{\pm 1.94}$	0.18	$116.91_{\pm 8.56}$	$+116.73_{\pm 8.56}$

4.2 Cross-Encoder Transferability

We next test whether a suffix optimized without the victim encoder still increases exposure. Table 5 reports transfer to a held-out BGE victim: Qwen3-Embedding-4B alone, a joint search with Qwen3-Embedding-4B, E5-large (Wang et al., 2022), and GTE-large (Li et al., 2023a), and a matched-BGE reference. The search protocol is in Appendix H.1.

Table 5. Transfer to a held-out BGE victim under RQ-KMeans.

	Beauty			Industrial & Sci.			Toys & Games		
Search encoder(s)	Before	After	ΔExp@50	Before	After	ΔExp@50	Before	After	ΔExp@50
Qwen3-Embedding-4B	0.41	$2.47_{\pm 0.19}$	$+2.06_{\pm 0.19}$	0.37	$17.30_{\pm 1.41}$	$+16.93_{\pm 1.41}$	0.44	$38.58_{\pm 3.06}$	$+38.14_{\pm 3.06}$
Qwen3-Embedding-4B + E5-large + GTE-large	0.41	$3.46_{\pm 0.25}$	$+3.05_{\pm 0.25}$	0.37	$26.72_{\pm 2.11}$	$+26.35_{\pm 2.11}$	0.44	$70.06_{\pm 5.37}$	$+69.62_{\pm 5.37}$
BGE (matched reference)	0.41	$9.14_{\pm 0.52}$	$+8.73_{\pm 0.52}$	0.37	$74.65_{\pm 5.14}$	$+74.28_{\pm 5.14}$	0.44	$211.80_{\pm 15.72}$	$+211.36_{\pm 15.72}$

SIDAttack transfers from Qwen3-Embedding-4B to the held-out BGE victim: single-surrogate search raises mean Exp@50 from 0.41 to 2.47 on Beauty, 0.37 to 17.30 on Industrial & Scientific, and 0.44 to 38.58 on Toys & Games. Joint search with E5-large and GTE-large further reaches 3.46, 26.72, and 70.06. These transferred suffixes remain weaker than matched BGE, which optimizes in the victim embedding space, but they still exceed the rewriting baselines in Appendix G. Encoder mismatch therefore limits the attack without removing the routing vulnerability.

4.3 Effect of Attack Budget

We vary the append-only suffix length from 12 to 24 and 48 tokens under the RQ-KMeans setting of Table 1. Figure 4 shows that longer suffixes increase exposure in every domain: a 12-token suffix is already enough to raise ΔExp@50, the gains taper on Industrial & Scientific, and they keep growing on Toys & Games through n=48. Unused suffix budget therefore remains exploitable on some catalogs. Absolute Exp@50 values are in Appendix F.

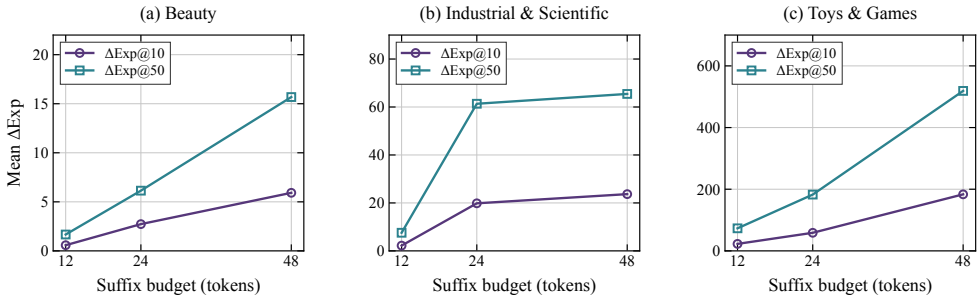


Figure 4. Effect of suffix length under RQ-KMeans.

4.4 Ablation of Anchor Selection and Search Objectives

We use the RQ-KMeans setting of Table 1, keeping the target listings and suffix budget fixed. *Single Anchor* keeps only the highest-scoring anchor and one search beam. *w/o Global* removes Φ_τ and optimizes only anchor similarity, while *w/o Diversity* drops the pairwise anchor-diversity constraint.

Table 6. Ablation of the anchor portfolio and search objective. T@k is the percentage of listings whose first k SID tokens match the selected anchor.

	Beauty			Industrial & Sci.			Toys & Games		
Variant	T@1	T@2	ΔExp@50	T@1	T@2	ΔExp@50	T@1	T@2	ΔExp@50
Single Anchor	$31.0_{\pm 2.4}$	$21.0_{\pm 1.8}$	$+3.92_{\pm 0.37}$	$48.0_{\pm 2.7}$	$29.0_{\pm 2.2}$	$+40.88_{\pm 3.31}$	$29.0_{\pm 2.3}$	$13.0_{\pm 1.5}$	$+127.46_{\pm 8.94}$
w/o Global	$37.0_{\pm 2.1}$	$27.0_{\pm 2.3}$	$+4.67_{\pm 0.41}$	$56.0_{\pm 2.5}$	$35.0_{\pm 1.7}$	$+48.37_{\pm 4.06}$	$34.0_{\pm 1.8}$	$16.0_{\pm 1.7}$	$+149.72_{\pm 10.63}$
w/o Diversity	$41.0_{\pm 1.7}$	$31.0_{\pm 1.9}$	$+5.31_{\pm 0.46}$	$61.0_{\pm 1.9}$	$40.0_{\pm 2.4}$	$+54.92_{\pm 3.58}$	$38.0_{\pm 2.1}$	$19.0_{\pm 1.4}$	$+160.85_{\pm 12.17}$
Full SIDAttack	$46.0_{\pm 2.0}$	$36.0_{\pm 1.6}$	$+6.14_{\pm 0.41}$	$68.0_{\pm 2.2}$	$46.0_{\pm 1.8}$	$+61.35_{\pm 3.87}$	$42.0_{\pm 1.6}$	$22.0_{\pm 1.9}$	$+182.64_{\pm 11.64}$

Table 6 shows that the full method is best in every domain, with Target@1 of 46%, 68%, and 42%. Single-anchor search causes the largest drop, so one anchor is not a universal direction. Removing the global objective or the diversity constraint also reduces both targeting and ΔExp@50, which means the portfolio and the global objective are complementary.

4.5 Margin-Regularized Defense

We test whether margin regularization can reduce attack-induced exposure and still preserve clean recommendation performance. For each suffix budget, we first generate attacked catalog text under standard RQ-KMeans and then freeze the edits. We construct the margin-regularized codebook with the objective in Section 3.4, retrain the generator, and evaluate the same attacked text under the defended system. This fixed-edit protocol isolates the effect of the defense: the attack text is held constant while the victim tokenizer and generator change. Operationally, this evaluates the defended recommender after its tokenizer and generator have been rebuilt and deployed; it measures future exposure from the unchanged catalog edits, not exposure that occurred before the defense was installed. Figure 5 reports exposure under the standard and defended systems, and Table 7 reports clean Recall@10 and Recall@50 on the unattacked test set.

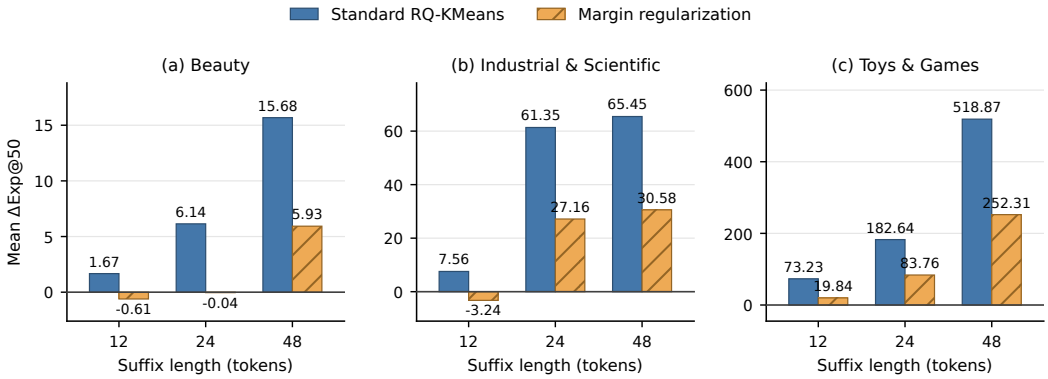


Figure 5. Standard versus margin-regularized RQ-KMeans as the suffix budget varies.

Margin regularization substantially suppresses attack-induced exposure. At n=24, ΔExp@50 changes from +6.14 to −0.04 on Beauty, from +61.35 to +27.16 on Industrial & Scientific, and from +182.64 to +83.76 on Toys & Games. The latter two correspond to reductions of 55.7% and 54.1%, while Beauty falls to approximately zero. A residual gain remains at n=48, especially on Toys & Games, so the defense raises the suffix budget needed for a comparable effect rather than closing the attack.

Table 7. Clean Recall before and after margin regularization.

	Beauty		Industrial & Sci.		Toys & Games	
Setting	Recall@10	Recall@50	Recall@10	Recall@50	Recall@10	Recall@50
Standard	0.0718±0.0019	0.1847±0.0041	0.0863±0.0023	0.2136±0.0046	0.0649±0.0021	0.1728±0.0044
Margin-Reg.	0.0712±0.0021	0.1831±0.0035	0.0857±0.0026	0.2124±0.0043	0.0641±0.0018	0.1716±0.0037

The defense retains over 99% of Recall@50 in every domain (0.1847 → 0.1831, 0.2136 → 0.2124, 0.1728 → 0.1716). The codebook changes once at training time, so this cost falls on the clean catalog and benign recommendation quality stays largely intact.

5 Conclusion

We find that small perturbations can change Semantic IDs, and that residual-quantization margins predict which items are most susceptible. SIDAttack exploits this by optimizing a short suffix with encoder gradients toward public-item anchors, producing larger exposure gains than the evaluated baselines across domains, tokenizers, and held-out victim encoders. Margin regularization reduces these gains while largely preserving utility, so SID routing should be treated as an attack surface rather than only as a tokenization choice.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to draft and revise manuscript text, improve readability, format LaTeX, and search related work. We have not used generative AI tools to generate synthetic datasets, formulate mathematical claims, or implement the core method without human verification. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work studies robustness properties of Semantic-ID tokenizers in order to measure a structural vulnerability of generative recommenders. All experiments use public catalogs and do not target a live production system. SIDAttack is used to study whether routing can be steered toward higher generative exposure. We discuss mitigations in section 4.5 and will not release exploit payloads against deployed services.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We will release training code, evaluation scripts, split seeds, and hyperparameters. Dataset construction is specified in Appendix C; search and training hyperparameters are in Appendix A.2. The matched and mismatched public-encoder protocols are specified in Sections 4.1 and 4.2, with the Qwen3-Embedding-4B→BGE comparison in Appendix G and the joint multi-surrogate search protocol in Appendix H.1.

REFERENCES

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.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023a.

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. OneRec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*, 2025.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Kun Zhang, Jingming Zhang, Wei Cheng, Yansong Cheng, Jiaqi Zhang, Hao Lu, Xu Zhang, Haixiang Gan, Jiangxia Cao, Tenglong Wang, et al. OneMall: One architecture, more scenarios–end-to-end generative recommender family at Kuaishou e-commerce. *arXiv preprint arXiv:2601.21770*, 2026.

Wenzhuo Cheng, Menghang Gong, Qixin Guo, Hang Zheng, Zhaobin Yang, Jianguo Lou, and Zhengwei Zheng. CapsID: Soft-routed variable-length semantic IDs for generative recommendation. *arXiv preprint arXiv:2605.05096*, 2026.

Xuesi Wang, Junjie Wang, Ziliang Wang, Weijie Bian, and Guanxing Zhang. Decoupled residual quantization for robust semantic IDs in recommendation. *arXiv preprint arXiv:2606.01844*, 2026a.

Junting Wang, Xinrui He, Yunzhe Li, and Hari Sundaram. Understanding semantic IDs: From item representation to item selection in generative recommendation. *arXiv preprint arXiv:2607.24995*, 2026b.

Liangbo Ning, Wenqi Fan, and Qing Li. Exploring backdoor attack and defense for LLM-empowered recommendations. *IEEE Transactions on Knowledge and Data Engineering*, 38(8): 5077–5094, 2026. doi: 10.1109/TKDE.2026.3690651.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36, 2018.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3597–3606, 2020.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9558–9566, 2021.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 540–550, 2022.

Linyang Li, Demin Song, and Xipeng Qiu. Text adversarial purification as defense against adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 338–350, 2023b.

Sejoon Oh, Gaurav Verma, and Srijan Kumar. Adversarial text rewriting for text-aware recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1804–1814, 2024.

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, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Jinghao Zhang, Yuting Liu, Qiang Liu, Shu Wu, Guibing Guo, and Liang Wang. Stealthy attack on large language model based recommendation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5839–5857, 2024.

Clark Mingxuan Ju, Liam Collins, Leonardo Neves, Bhuvesh Kumar, Louis Yufeng Wang, Tong Zhao, and Neil Shah. Generative recommendation with semantic IDs: a practitioner's handbook. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 6420–6425, 2025.

Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. Adversarial search engine optimization for large language models. In *International Conference on Learning Representations*, 2025.

Zongwei Wang, Min Gao, Junliang Yu, Xinyi Gao, Quoc Viet Hung Nguyen, Shazia Sadiq, and Hongzhi Yin. ID-free not risk-free: LLM-powered agents unveil risks in ID-free recommender systems. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval*, pp. 1902–1911, 2025.

Yuyue Zhao, Jin Huang, Shuchang Liu, Jiancan Wu, Xiang Wang, and Maarten de Rijke. LANCE: Exploration and reflection for LLM-based textual attacks on news recommender systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 197–206, 2025.

Yuchuan Zhao, Tong Chen, Junliang Yu, Zongwei Wang, Lizhen Cui, and Hongzhi Yin. Prompt-unknown promotion attacks against LLM-based sequential recommender systems. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2453–2463, 2026. doi: 10.1145/3805712.3809691.

A Method Details

A.1 AnchorPort-GCG Implementation

Suffix tokens are restricted to a vocabulary $\mathcal{V}_{\mathrm{rt}}$ that remains stable after decoding and re-encoding. The Amazon experiments use an English-word vocabulary; the KuaiRec adaptation is described in Appendix D. For suffix position t in beam b, we score a candidate token v by

$$\Delta_{b,t}(v) = \left\langle \boldsymbol{e}(v) - \boldsymbol{e}(s_t), \nabla_{\boldsymbol{e}(s_t)} \mathcal{J}_b(s) \right\rangle, \tag{8}$$

where $\boldsymbol{e}(v)$ is the token embedding of v under f and $\Delta_{b,t}(v)$ estimates the change in $\mathcal{J}_b$ from replacing s_t with v. The highest-scoring candidates are evaluated with a forward pass through f, and a replacement is accepted only when it improves $\mathcal{J}_b$. All beams use the same search budget.

After optimization, we collect the valid suffixes from all beams and select the one with the highest global score $\Phi_\tau(z(s))$. If several suffixes obtain the same score, we prefer the one that is closer to a public popular item and then the one that remains closer to the original representation z_0.

A.2 Hyperparameters

Table 8 gives the AnchorPort-GCG settings for the multi-anchor portfolio evaluated on 100 target listings per domain.

Table 8. AnchorPort-GCG hyperparameters for the multi-anchor portfolio.

	Value
Search beams B	3
β_{sel}	0.02
β_{exp}	0.02
λ	0.75
τ	0.03
Max anchor cosine	0.90
Popular pool $\lvert\mathcal{H}\rvert$	100
Anchor candidates	100, select 3
Candidates per position	16
Search rounds	32 per beam
Suffix length n	24
Max input length	512

Suffix candidates come from a round-trip-safe vocabulary of $41{,}480$ Qwen3-Embedding-4B tokens. We retain a suffix only if decoding and re-encoding recover the same token sequence.

RQ-KMeans uses three residual-quantization levels with 256 codes per level and squared ℓ_2 distance. We use standard residual assignments without uniform or Sinkhorn post-processing.

The RQ-VAE encoder maps $2560 \to 2048 \to 1024 \to 512 \to 256 \to 128 \to 64 \to 32$, and the decoder mirrors this stack. Hidden layers use ReLU, dropout 0, and no batch normalization. Each of the three codebooks has 256 codewords of dimension 32. The loss weights reconstruction MSE by 1 and quantization by 1; within quantization, the codebook term has weight 1 and the commitment

term has weight 0.25, and the three levels are averaged. Codebooks are initialized by at most 100 k-means iterations. Sinkhorn uses $\varepsilon=(0, 0, 0.003)$ for 50 iterations.

Margin regularization uses $\eta=0.05$ and $\gamma=1$, and updates only the codebook for 30 epochs with AdamW, learning rate 2×10^{-4}, batch size 32, and weight decay 0. The generator then has a one-epoch SID-alignment stage at learning rate 10^{-3}, updating only the SID parameters, followed by one epoch of full-parameter supervised fine-tuning at learning rate 2×10^{-5}. The effective batch size is 1024, with microbatch size 16.

A.3 Baseline Adaptations

Table 9 summarizes how the methods in the main experiments modify the catalog listing. Every method is given the same access, white-box f and the public pool $\mathcal{H}$; the search-signal column records only what that method uses. All append-only methods use a suffix of length $n=24$ in the main experiments while keeping the original title and description unchanged. Stealthy Attack, the rewrite agent, and TextSimu instead modify the catalog body. Our Stealthy Attack and TextSimu implementations adapt the original methods to the shared catalog-side setting.

Table 9. Implementation of each method in the main tables. f is the search encoder and $\mathcal{H}$ is the pool of popular public items. Access is identical for every row.

Method	Edit	Search signal	Access
Random	append $n=24$ unoptimized tokens	none	f, $\mathcal{H}$
HotFlip (Ebrahimi et al., 2018)	one gradient substitution on the suffix	cosine in f toward a public-item anchor	f, $\mathcal{H}$
Stealthy Attack (Zhang et al., 2024)	modify the catalog body	stealthy textual promotion	f, $\mathcal{H}$
Rewrite agent	paraphrase the catalog body	LLM rewrite	f, $\mathcal{H}$
TextSimu (Wang et al., 2025)	rewrite the catalog body	imitate same-category popular text	f, $\mathcal{H}$
SIDAttack	append $n=24$ with AnchorPort-GCG	steering toward popular public items in $\mathcal{H}$	f, $\mathcal{H}$

B Qualitative Suffix Cases

We provide three examples from the RQ-KMeans experiments with suffix length $n=24$. These cases correspond to items whose test-set exposure increases substantially after rerouting. In every example, the original title and factual catalog body remain unchanged. Only the appended suffix, shown in bold, is optimized by SIDAttack.

Toys & Games, item 303658**.** Test $\Delta\text{Exp@}50 = +732$.

> **Original.**
> Evan Designs U3W Simulated Model LED Welding Kit, by Item number U3W Size: dia. =3mm Color 2 bright ;flashing white and 1 solid ; bright blue LED Voltage5-19 volts AC,DC,DCC Includes a resistor, bridge rectifier, and capacitor No circuit board needed Wire length OAL12in; Distance from LED to elec
>
> **After attack.**
> Evan Designs U3W Simulated Model LED Welding Kit, by Item number U3W Size: dia. =3mm Color 2 bright ;flashing white and 1 solid ; bright blue LED Voltage5-19 volts AC,DC,DCC Includes a resistor, bridge rectifier, and capacitor No circuit board needed Wire length OAL12in; Distance from LED to elec
>
> **PARTY glove shares counts pairs use details information Determine Finger**

Industrial & Scientific, item 43567**.** Test $\Delta\text{Exp@50} = +1{,}090$.

Original.

Sensodyne Complete Protection Toothpaste, Extra Fresh, 3.4 Oz (Pack of 2) For the most up to date information, we recommend you visit the manufacturer website for the best product details, including ingredients, hazards, directions and warnings. This product offering is for 2 Retail Package

After attack.

Sensodyne Complete Protection Toothpaste, Extra Fresh, 3.4 Oz (Pack of 2) For the most up to date information, we recommend you visit the manufacturer website for the best product details, including ingredients, hazards, directions and warnings. This product offering is for 2 Retail Package

plastics mounting UserRole infer dependencies use details information mitig

Industrial & Scientific, item 11676**.** Test $\Delta\text{Exp@50} = +562$.

Original.

SEOH Thermometer Red Spirit Total Immersion -20 to 150C Single Scale This thermometer is Red Spirit filled with a temperature range of -20 to 150C. Total Immersion. Student grade. Total length is approximately 12 inches (300mm) with ring on top for suspension. Graduations are every 1 degree.

After attack.

SEOH Thermometer Red Spirit Total Immersion -20 to 150C Single Scale This thermometer is Red Spirit filled with a temperature range of -20 to 150C. Total Immersion. Student grade. Total length is approximately 12 inches (300mm) with ring on top for suspension. Graduations are every 1 degree.

forehead infrared Explanation longevity included use bgColor information explain precautions

In all three cases, the original title and factual body are byte-identical to the source. The examples illustrate that an append-only modification can change the representation and downstream routing without rewriting the existing catalog description. The thermometer example provides a particularly clear case in which the optimized suffix moves a student-grade spirit thermometer toward a neighborhood associated with infrared thermometers.

C Dataset Details

We consider three public Amazon catalog domains (Ni et al., 2019; Rajput et al., 2023; Ju et al., 2025). Table 10 summarizes the resulting statistics. Beauty and Toys & Games follow the 5-core sequential splits used by TIGER and GRID (Rajput et al., 2023; Ju et al., 2025). Industrial & Scientific uses the Amazon 2018 5-core dump from the same collection (Ni et al., 2019). After filtering, every remaining user and item has at least five interactions.

We sort each user history by timestamp and use leave-one-out evaluation. The final interaction is used for testing, the second-to-last interaction is used for validation, and the remaining interactions train the autoregressive SID generator based on the TIGER formulation. The catalog body x is formed by concatenating the title and description, and items without usable text are removed. We randomly sample 100 target items from each domain for attack evaluation. RQ-KMeans and RQ-VAE use separate per-domain codebooks and separately trained generators, and the RQ-VAE experiments use newly optimized suffixes rather than transferring those found under RQ-KMeans.

Table 10. Dataset statistics after 5-core filtering. Beauty and Toys & Games follow TIGER/GRID (Rajput et al., 2023; Ju et al., 2025). Industrial & Scientific uses the Amazon 2018 5-core dump (Ni et al., 2019). Avg. length is the mean number of interactions per user. Sparsity is $\#\text{interactions}/(\#\text{users} \times \#\text{items})$. Attack listings are 100 target items drawn uniformly at random in each domain.

Dataset	#Users	#Items	#Interactions	Avg. length	Sparsity	Attack listings
Beauty	22,363	12,101	198,502	8.9	7.34×10^{-4}	100
Industrial & Sci.	11,039	5,163	77,071	7.0	1.35×10^{-3}	100
Toys & Games	19,412	11,924	167,597	8.6	7.24×10^{-4}	100

D ADDITIONAL DOMAINS: MIND AND KUAIREC

To examine whether the observed vulnerability extends beyond Amazon product recommendation, we additionally evaluate SIDAttack on MIND and KuaiRec. These datasets introduce two substantially different recommendation domains, news and short video, while preserving the same attack and evaluation protocol used in the main experiments. For MIND, the item text is formed by concatenating the news title and abstract. For KuaiRec, the item text concatenates the available textual video metadata, including manual cover text, caption, topic tags, and category names. We randomly sample 100 target items from each dataset and keep the victim tokenizer and autoregressive SID generator frozen during attack construction. All append-only methods use the same suffix budget $n{=}24$ as in the main experiments. For KuaiRec, the round-trip vocabulary is constructed from the language of the catalog text rather than restricting the suffix to English tokens.

The RQ-KMeans comparison appears in Table 4 in the main paper. Here we report the corresponding RQ-VAE results. RQ-VAE uses a separately trained tokenizer and generator, and the suffix search is rerun independently.

Table 11. Additional RQ-VAE results on MIND and KuaiRec. Each dataset uses a separately trained victim stack and independently optimized suffixes. $\mathrm{Exp@50}$ reports mean held-out exposure before and after the catalog edit, and $\Delta\mathrm{Exp@50}$ reports the average change. After-edit exposure and exposure change are reported as mean $\pm$ standard deviation over three runs.

	MIND			KuaiRec		
Method	Before	After	ΔExp@50	Before	After	ΔExp@50
Random	0.13	$0.37_{\pm 0.04}$	$+0.24_{\pm 0.04}$	0.20	$0.99_{\pm 0.08}$	$+0.79_{\pm 0.08}$
HotFlip	0.13	$0.72_{\pm 0.06}$	$+0.59_{\pm 0.06}$	0.20	$1.66_{\pm 0.14}$	$+1.46_{\pm 0.14}$
Stealthy Attack	0.13	$2.84_{\pm 0.24}$	$+2.71_{\pm 0.24}$	0.20	$10.44_{\pm 0.89}$	$+10.24_{\pm 0.89}$
Rewrite agent	0.13	$0.57_{\pm 0.05}$	$+0.44_{\pm 0.05}$	0.20	$3.28_{\pm 0.27}$	$+3.08_{\pm 0.27}$
TextSimu	0.13	$4.71_{\pm 0.39}$	$+4.58_{\pm 0.39}$	0.20	$18.83_{\pm 1.52}$	$+18.63_{\pm 1.52}$
SIDAttack	0.13	$12.99_{\pm 1.06}$	$+12.86_{\pm 1.06}$	0.20	$51.62_{\pm 4.07}$	$+51.42_{\pm 4.07}$

The RQ-VAE results show the same qualitative pattern with smaller absolute exposure shifts. SIDAttack increases mean $\mathrm{Exp@50}$ from 0.13 to 12.99 on MIND and from 0.20 to 51.62 on KuaiRec, remaining the strongest method on both datasets. Together with Table 4, these experiments cover product, news, and short-video recommendation.

E RQ MARGINS AND DIRECTIONAL REROUTING

We first test whether the original quantization geometry predicts sensitivity to textual variation. For each of 100 Amazon Beauty listings, we generate eight natural rewrites without attack optimization. At each SID level, we measure the original RQ margin and divide the listings into ten equally sized groups according to that margin. We then compute the SID token flip rate within each group.

Figure 1 shows that listings with smaller margins are substantially more likely to change their SID. The predictive AUC values are 0.930, 0.869, and 0.847 at the three semantic levels. Later SID tokens are less stable than the first token, but the same margin-dependent pattern appears at every level. These results show that SID instability is strongly associated with quantization geometry rather than arbitrary textual variation.

Directional rerouting. We next test whether SID changes can be deliberately steered rather than occurring at random. Across Beauty, Industrial & Scientific, and Toys & Games, SIDAttack reaches the selected anchor at the first SID level for 46%, 68%, and 42% of attacked items, and matches the first two SID levels for 36%, 46%, and 22%, respectively. In contrast, Random and HotFlip remain at or below 16% for Target@1 and 4% for Target@2 across all three domains. These results show that SIDAttack converts SID instability into directional rerouting rather than merely increasing the probability of an arbitrary code change.

Table 12. Effect of suffix length under RQ-KMeans on 100 attacked listings per domain. Exp@50 reports mean test exposure before and after the attack, and ΔExp@50 reports the average change. After-attack exposure and exposure change are reported as mean ± standard deviation over three runs.

	Beauty			Industrial & Sci.			Toys & Games		
n	Before	After	ΔExp@50	Before	After	ΔExp@50	Before	After	ΔExp@50
12	0.09	$1.76_{\pm 0.16}$	$+1.67_{\pm 0.16}$	0.14	$7.70_{\pm 0.65}$	$+7.56_{\pm 0.65}$	0.21	$73.44_{\pm 5.87}$	$+73.23_{\pm 5.87}$
24	0.09	$6.23_{\pm 0.41}$	$+6.14_{\pm 0.41}$	0.14	$61.49_{\pm 3.87}$	$+61.35_{\pm 3.87}$	0.21	$182.85_{\pm 11.64}$	$+182.64_{\pm 11.64}$
48	0.09	$15.77_{\pm 1.27}$	$+15.68_{\pm 1.27}$	0.14	$65.59_{\pm 4.41}$	$+65.45_{\pm 4.41}$	0.21	$519.08_{\pm 31.76}$	$+518.87_{\pm 31.76}$

F Detailed Suffix-Budget Results

Table 12 reports the absolute mean test exposure at cutoff 50 before and after the attack for each suffix budget, together with the corresponding ΔExp@50. The n=24 row corresponds to the RQ-KMeans results in Table 1.

G Detailed Cross-Encoder Comparison

Qwen3-Embedding-4B→BGE baseline comparison. We replace only the public encoder to study transfer across embedding models: suffix optimization uses Qwen3-Embedding-4B (Zhang et al., 2025), while the victim encoder remains BGE, with the RQ-KMeans tokenizer and autoregressive SID generator frozen. The attacked listings, suffix budget, and public pool follow the matched BGE–BGE setting in Table 3. HotFlip is re-optimized with Qwen3-Embedding-4B during search. This setting tests whether SIDAttack requires access to the victim's exact representation space: although some degradation relative to the matched BGE setting is expected under encoder mismatch, the key question is whether the optimized suffix still transfers better than untargeted or weakly targeted perturbations.

Table 13. Qwen3-Embedding-4B→BGE surrogate transfer under RQ-KMeans on 100 attacked listings per domain with n=24. Search uses Qwen3-Embedding-4B while the victim encoder remains BGE. After-edit exposure and ΔExp@50 are reported as mean ± standard deviation over three runs.

	Beauty			Industrial & Sci.			Toys & Games		
Method	Before	After	ΔExp@50	Before	After	ΔExp@50	Before	After	ΔExp@50
Random	0.41	$0.33_{\pm 0.01}$	$-0.08_{\pm 0.01}$	0.37	$2.35_{\pm 0.17}$	$+1.98_{\pm 0.17}$	0.44	$3.01_{\pm 0.26}$	$+2.57_{\pm 0.26}$
HotFlip	0.41	$0.85_{\pm 0.05}$	$+0.44_{\pm 0.05}$	0.37	$0.88_{\pm 0.07}$	$+0.51_{\pm 0.07}$	0.44	$3.71_{\pm 0.31}$	$+3.27_{\pm 0.31}$
Stealthy Attack	0.41	$0.83_{\pm 0.05}$	$+0.42_{\pm 0.05}$	0.37	$1.51_{\pm 0.12}$	$+1.14_{\pm 0.12}$	0.44	$19.68_{\pm 1.78}$	$+19.24_{\pm 1.78}$
Rewrite agent	0.41	$0.35_{\pm 0.02}$	$-0.06_{\pm 0.02}$	0.37	$1.00_{\pm 0.06}$	$+0.63_{\pm 0.06}$	0.44	$2.99_{\pm 0.22}$	$+2.55_{\pm 0.22}$
TextSimu	0.41	$0.55_{\pm 0.03}$	$+0.14_{\pm 0.03}$	0.37	$1.11_{\pm 0.09}$	$+0.74_{\pm 0.09}$	0.44	$30.41_{\pm 2.67}$	$+29.97_{\pm 2.67}$
SIDAttack	0.41	$2.47_{\pm 0.19}$	$+2.06_{\pm 0.19}$	0.37	$17.30_{\pm 1.41}$	$+16.93_{\pm 1.41}$	0.44	$38.58_{\pm 3.06}$	$+38.14_{\pm 3.06}$

The cross-encoder results are weaker than the matched BGE setting, as expected from the mismatch between the public and victim representation spaces. SIDAttack raises mean Exp@50 from 0.41 to 2.47 on Beauty, from 0.37 to 17.30 on Industrial & Scientific, and from 0.44 to 38.58 on Toys & Games, remaining stronger than the append-only baselines. This shows that some representation steering transfers across public embedding models even when the public encoder differs from the victim encoder.

H Notation

Tables 14 and 15 summarize the symbols used in the main text. To keep the notation reference compact, both tables are placed together below.

H.1 Multi-Surrogate Joint Optimization

The transfer experiments above optimize SIDAttack using a single public encoder and then evaluate the resulting suffix on a different victim encoder. We further consider a stricter setting in which multiple public encoders are jointly used during suffix search while the victim encoder is completely held out. Specifically, we use Qwen3-Embedding-4B, E5-large (Wang et al., 2022), and GTE-large (Li et al., 2023a) as the public surrogates and use BGE only for post-hoc victim evaluation. The BGE encoder, its tokenizer, and its recommendation outputs are never accessed during suffix optimization.

Table 14. Notation for the victim system, attacker, and defense.

Symbol	Meaning
Victim system	
x	Frozen catalog body (title and description)
E	Frozen victim item encoder
Q	Residual SID tokenizer (RQ-KMeans or RQ-VAE)
L	Number of semantic SID levels (L=3)
K	Codebook size at each residual level
q_ℓ	SID token at level ℓ
$(q_1, \ldots, q_L)$	Semantic ID of x
p_θ	Frozen autoregressive SID generator based on the TIGER formulation
$\boldsymbol{z}^{(\ell)}$	Residual vector at level ℓ; $\boldsymbol{z}^{(1)} = E(x)$
$\mathbb{C}^{(\ell)}$	Codebook at level ℓ
$\boldsymbol{c}_k^{(\ell)}$	k-th code vector in $\mathbb{C}^{(\ell)}$
Attacker (public)	
f	Text encoder available to the attacker
s	Append-only suffix of length n
$x \oplus s$	Concatenation that leaves the body of x unchanged
$\mathcal{H}$	Pool of popular public items
h_j	ℓ_2-normalized f-embedding of popular item j
r_j	Public popularity proxy for item j (sales rank or review count)
M	Size of $\mathcal{H}$
$\mathcal{V}_{\mathrm{rt}}$	English-word vocabulary that round-trips through the tokenizer
n	Suffix length in tokens (main tables use n=24; Table 12 also uses n=12, 48)
Defense (codebook construction)	
m_ℓ	RQ margin at level ℓ, given by the gap to the nearest competing code
η	Hinge threshold on m_ℓ
γ	Weight of $\mathcal{L}_{\mathrm{mgn}}$ in codebook training
$\mathcal{L}_Q$	Standard RQ-KMeans or RQ-VAE codebook objective
$\mathcal{L}_{\mathrm{mgn}}$	Margin hinge $\sum_\ell \mathbb{E}[\max(0, \eta - m_\ell)]$

Let

$$\mathcal{F} = \{f_1, \ldots, f_R\}$$

denote the set of R public surrogate encoders. For each surrogate f_r, the source item is represented as

$$z_0^{(r)} = \frac{f_r(x)}{\|f_r(x)\|_2}, \tag{9}$$

and every item j in the pool of popular public items is represented as

$$h_j^{(r)} = \frac{f_r(x_j)}{\|f_r(x_j)\|_2}. \tag{10}$$

Shared anchor portfolio. The public anchor identities are shared across surrogate encoders. For each candidate item j, we aggregate semantic compatibility across the public representation spaces as

$$\bar{u}_j(x) = \frac{1}{R} \sum_{r=1}^{R} \left(z_0^{(r)}\right)^\top h_j^{(r)} + \beta_{\mathrm{sel}} \log(1 + r_j), \tag{11}$$

where r_j is the same public popularity signal used in the single-encoder setting. We greedily select B anchors from the highest-scoring candidates. To retain complementary search directions, we constrain the average pairwise similarity between two selected anchors:

$$\frac{1}{R} \sum_{r=1}^{R} \left(h_p^{(r)}\right)^\top h_q^{(r)} \le \delta, \qquad p \neq q. \tag{12}$$

For public-item anchor b, its representation under surrogate f_r is denoted by $a_b^{(r)}$.

Joint AnchorPort-GCG search. All public encoders optimize the same textual suffix. For beam b, surrogate f_r evaluates a suffix s using

$$\mathcal{J}_b^{(r)}(s) = (1 - \lambda)\Phi_\tau^{(r)}\left(z^{(r)}(s)\right) + \lambda \left(z^{(r)}(s)\right)^\top a_b^{(r)}, \tag{13}$$

Table 15. Notation for AnchorPort-GCG and post-hoc evaluation.

Symbol	Meaning
AnchorPort-GCG	
z_0	ℓ_2-normalized public embedding $f(x)$
$z(s)$	ℓ_2-normalized public embedding $f(x \oplus s)$
$u_j(x)$	Compatibility of popular item j with source x
$\beta_{\text{sel}}, \beta_{\text{exp}}$	Popularity weights in selection and in Φ_τ
$\mathcal{A}(x)$	Source-conditioned portfolio of B diverse anchors
a_b	Anchor of GCG beam b
B	Number of anchors and beams
δ	Pairwise cosine cap between selected anchors
Φ_τ	Temperature-smoothed soft maximum over $\mathcal{H}$
τ	Temperature of Φ_τ
$\mathcal{J}_b$	Objective of beam b
λ	Mix between Φ_τ and the beam's own anchor
$\boldsymbol{e}(v)$	Token embedding of vocabulary item v
$\Delta_{b,t}(v)$	First-order GCG score for replacing position t of beam b by v
$\mathcal{S}_{\text{feasible}}$	Publishable suffixes that survive coordinate search
$s^\star$	Selected suffix
Evaluation (post-hoc)	
$\text{Flip@}k$	Fraction of listings whose SID prefix of length k changes
$\text{Target@}k$	Fraction whose first k SID tokens match the frozen anchor
N	Number of attacked listings (N=100 per domain)
q_i'	Post-attack victim SID of listing i
a_i	Public-item anchor selected for listing i
q^{a_i}	Victim SID of that anchor measured only after search
$\text{Exp@}K(i)$	Times listing i enters a test user's top K, Eq. (6)
$\text{Exp@}K$	Mean of $\text{Exp@}K(i)$ over N attacked listings, Eq. (7)
$\Delta\text{Exp@50}$	Change in Exp@50 after the catalog edit

where

$$z^{(r)}(s) = \frac{f_r(x \oplus s)}{\|f_r(x \oplus s)\|_2}, \tag{14}$$

and $\Phi_\tau^{(r)}$ is the global objective over popular public items computed in the representation space of f_r.

The surrogate models may use different tokenizers and token-embedding spaces. We therefore do not directly average their token gradients. Instead, each surrogate independently applies the GCG first-order criterion to propose high-scoring token replacements at the current suffix position. The textual candidates proposed by all surrogates are merged into a shared candidate set. Each candidate is then decoded as text, independently re-tokenized by every surrogate, and evaluated with a full forward pass through all public encoders.

For a candidate suffix s', its improvement under surrogate f_r is

$$\Delta\mathcal{J}_b^{(r)}(s') = \mathcal{J}_b^{(r)}(s') - \mathcal{J}_b^{(r)}(s). \tag{15}$$

We aggregate these improvements as

$$\Delta\bar{\mathcal{J}}_b(s') = \frac{1}{R}\sum_{r=1}^{R} \Delta\mathcal{J}_b^{(r)}(s'). \tag{16}$$

The candidate with the largest positive $\Delta\bar{\mathcal{J}}_b$ is accepted. This requires each accepted suffix update to remain useful across several public representation spaces rather than exploiting a single encoder. After coordinate search, the feasible suffixes from all beams are ranked by their average increase in the global public objective:

$$\Delta\bar{\Phi}(s) = \frac{1}{R}\sum_{r=1}^{R}\left[\Phi_\tau^{(r)}\left(z^{(r)}(s)\right) - \Phi_\tau^{(r)}\left(z_0^{(r)}\right)\right]. \tag{17}$$

The final suffix is

$$s^\star = \arg\max_s \Delta\bar{\Phi}(s). \tag{18}$$

Held-out victim evaluation. The victim encoder is not used in anchor selection, gradient computation, candidate generation, candidate ranking, or final suffix selection. Only after $s^\star$ has been fixed do we append it to the original catalog text and process the edited item with the held-out BGE victim encoder, its RQ-KMeans tokenizer, and its frozen autoregressive SID generator. The resulting victim SID and test-set exposure are then measured using the same evaluation protocol as in the single-surrogate transfer experiments.

This protocol separates optimization from victim evaluation. The single-surrogate experiments test whether a suffix found in one public representation space transfers to BGE. The multi-surrogate setting instead tests whether a suffix that is jointly effective across several heterogeneous public representation spaces generalizes more strongly to an unseen victim representation space. Table 5 in Section 4.2 reports the resulting transfer to the held-out BGE victim.

These results suggest that transfer is stronger when the suffix is required to improve the attack objective across several heterogeneous public encoders. A perturbation specialized to one surrogate may exploit properties of that representation space that do not survive a change of encoder. Joint optimization instead favors suffix updates that are consistently useful across multiple public embedding spaces, making the resulting perturbation less dependent on any single surrogate.

I Cross-Attack Defense Analysis

We further compare margin-regularized codebook construction with two attack-agnostic text-side defenses under a shared catalog-side evaluation protocol. For every attack, the edited catalog text is first generated and then frozen. The same edited text is evaluated with each defense so that the comparison changes only the defense pipeline rather than the attack instance. The *None* setting uses the standard RQ-KMeans victim without any additional processing. ONION (Qi et al., 2021) applies token-level language-model filtering before the victim encoder. TAP (Li et al., 2023b) denotes an adapted Text Adversarial Purification pipeline that applies repeated random masking and masked-language model recovery before evaluation. Margin-Reg. leaves the input text unchanged and instead replaces the standard RQ-KMeans tokenizer with the margin-regularized tokenizer described in Section 3.4. For text-side defenses, clean and attacked listings are processed by the same defense pipeline before computing exposure.

Table 16 reports both downstream exposure and SID-level stability. For each defense, *Before* is the clean exposure under that defense pipeline, while *After* is the exposure after the corresponding catalog edit. Thus $\Delta\mathrm{Exp@50} = \mathrm{After} - \mathrm{Before}$ isolates the promotion effect within the same defended system. Flip@k measures whether the SID prefix of length k changes between the clean and edited versions under that same defense. This within-defense definition is important because text purification and codebook reconstruction can themselves change the clean SID assignment.

Exposure robustness. The undefended system is most vulnerable to the attacks that induce a substantial downstream promotion effect. On Beauty, Stealthy Attack increases mean exposure by $+0.38$, while SIDAttack produces the largest increase, $+6.14$. Both text-side defenses strongly suppress these two attacks: ONION changes their exposure effects to $+0.01$ and -0.03, respectively, while TAP changes them to 0.00 and $+0.01$. Margin regularization likewise suppresses the same fixed edits, yielding -0.01 for Stealthy Attack and -0.04 for SIDAttack. Under this evaluation, all three defenses reduce the large SIDAttack promotion effect to approximately the clean baseline.

For attacks that are already weak on Beauty, the exposure metric should be interpreted together with SID stability rather than as a standalone ranking. HotFlip changes mean exposure by only 0.00 without defense, while Rewrite agent and TextSimu change it by -0.03 and $+0.01$. Margin-Reg. keeps these effects close to zero at -0.01, -0.04, and -0.03, respectively. By contrast, ONION increases the Rewrite-agent exposure change to $+2.85$ and TAP to $+0.14$, illustrating that text-side preprocessing can itself alter the downstream effect for fluent body edits. In this low-effect regime, the more revealing question is whether a defense prevents unnecessary SID rerouting while keeping exposure close to its defense-specific clean baseline.

SID-route stability. The flip rates distinguish defenses that merely alter final exposure from those that actually stabilize the discrete route. SIDAttack changes the first SID token for 83% of listings and the first two-token prefix for all listings without defense. ONION reduces these rates to 12% and 36%, and TAP to 18% and 39%. Margin regularization reduces them further to 9% and 31% without modifying the catalog text. This is consistent with its intended mechanism: larger codebook margins make a representation shift less likely to cross an RQ decision boundary, even when the edited text itself is retained.

The same distinction is visible for naturalistic body edits. For Rewrite agent, ONION and TAP increase Flip@1 to 19% and 18%, compared with 7% without defense, whereas Margin-Reg. reduces it to 3%; the corresponding Flip@2 values are 41%, 33%, 25%, and 11%. This suggests that aggressive text purification can itself introduce representation changes when the edited input is already fluent, while tokenizer-side stabilization does not require rewriting or deleting the catalog text. A similar pattern appears for TextSimu, where Margin-Reg. achieves the lowest Flip@1 and Flip@2, 2% and 13%, while keeping the exposure change close to zero.

Text-side purification versus tokenizer-side stabilization. The results reveal two complementary defense mechanisms. ONION and TAP act before the item encoder and can be particularly effective when the attack depends on linguistically unusual or locally fragile tokens. Their strong suppression of SIDAttack shows that generic text purification can remove a large fraction of the attack signal even without access to the SID tokenizer. However, these methods necessarily modify the catalog text and can also alter clean or fluent descriptions, as reflected by their defense-specific clean exposures and by the increased flip rates for Rewrite agent. Margin-Reg. instead operates after the encoder at the quantization stage. It does not attempt to identify malicious text and therefore provides robustness through a different mechanism: reducing the sensitivity of SID assignment to representation shifts. Together, these results support treating text purification and tokenizer stabilization as distinct robustness strategies rather than interchangeable implementations of the same defense.

Table 16. Cross-attack robustness on Beauty under RQ-KMeans. Exp@50 denotes mean test exposure before and after the catalog edit. After-edit exposure, exposure change, and SID flip rates are reported as mean ± standard deviation over three runs. Lower ΔExp@50 and SID flip rates indicate stronger robustness.

Attack	Defense	Clean	Attacked	ΔExp@50 ↓	Flip@1 ↓	Flip@2 ↓
HotFlip	None	0.09	$0.09_{\pm 0.03}$	$0.00_{\pm 0.03}$	$30.0_{\pm 2.5}\%$	$58.0_{\pm 2.8}\%$
	ONION	0.34	$0.49_{\pm 0.06}$	$+0.15_{\pm 0.06}$	$4.0_{\pm 0.8}\%$	$29.0_{\pm 2.0}\%$
	TAP†	0.02	$0.03_{\pm 0.02}$	$+0.01_{\pm 0.02}$	$20.0_{\pm 1.7}\%$	$39.0_{\pm 2.4}\%$
	Margin-Reg.	0.11	$0.10_{\pm 0.02}$	$-0.01_{\pm 0.02}$	$\mathbf{3.0_{\pm 0.7}\%}$	$\mathbf{24.0_{\pm 1.8}\%}$
Stealthy Attack	None	0.09	$0.47_{\pm 0.05}$	$+0.38_{\pm 0.05}$	$1.0_{\pm 0.6}\%$	$5.0_{\pm 0.9}\%$
	ONION	0.34	$0.35_{\pm 0.04}$	$+0.01_{\pm 0.04}$	$2.0_{\pm 0.7}\%$	$4.0_{\pm 0.8}\%$
	TAP†	0.02	$0.02_{\pm 0.01}$	$0.00_{\pm 0.01}$	$2.0_{\pm 0.6}\%$	$5.0_{\pm 1.0}\%$
	Margin-Reg.	0.11	$0.10_{\pm 0.02}$	$-0.01_{\pm 0.02}$	$\mathbf{0.8_{\pm 0.4}\%}$	$\mathbf{2.6_{\pm 0.6}\%}$
Rewrite agent	None	0.09	$0.06_{\pm 0.02}$	$-0.03_{\pm 0.02}$	$7.0_{\pm 1.0}\%$	$25.0_{\pm 1.9}\%$
	ONION	0.34	$3.19_{\pm 0.31}$	$+2.85_{\pm 0.31}$	$19.0_{\pm 1.8}\%$	$41.0_{\pm 2.5}\%$
	TAP†	0.02	$0.16_{\pm 0.04}$	$+0.14_{\pm 0.04}$	$18.0_{\pm 1.5}\%$	$33.0_{\pm 2.2}\%$
	Margin-Reg.	0.11	$0.07_{\pm 0.02}$	$-0.04_{\pm 0.02}$	$\mathbf{3.0_{\pm 0.7}\%}$	$\mathbf{11.0_{\pm 1.2}\%}$
TextSimu	None	0.09	$0.10_{\pm 0.03}$	$+0.01_{\pm 0.03}$	$6.0_{\pm 0.9}\%$	$26.0_{\pm 2.0}\%$
	ONION	0.34	$0.32_{\pm 0.04}$	$-0.02_{\pm 0.04}$	$7.0_{\pm 1.0}\%$	$24.0_{\pm 1.8}\%$
	TAP†	0.02	$0.03_{\pm 0.02}$	$+0.01_{\pm 0.02}$	$15.0_{\pm 1.4}\%$	$26.0_{\pm 1.9}\%$
	Margin-Reg.	0.11	$0.08_{\pm 0.02}$	$-0.03_{\pm 0.02}$	$\mathbf{2.0_{\pm 0.6}\%}$	$\mathbf{13.0_{\pm 1.3}\%}$
SIDAttack	None	0.09	$6.23_{\pm 0.41}$	$+6.14_{\pm 0.41}$	$83.0_{\pm 2.4}\%$	$100.0_{\pm 0.0}\%$
	ONION	0.34	$0.31_{\pm 0.04}$	$-0.03_{\pm 0.04}$	$12.0_{\pm 1.3}\%$	$36.0_{\pm 2.2}\%$
	TAP†	0.02	$0.03_{\pm 0.02}$	$+0.01_{\pm 0.02}$	$18.0_{\pm 1.6}\%$	$39.0_{\pm 2.5}\%$
	Margin-Reg.	0.11	$0.07_{\pm 0.03}$	$-0.04_{\pm 0.03}$	$\mathbf{9.0_{\pm 1.2}\%}$	$\mathbf{31.0_{\pm 2.0}\%}$

† TAP denotes the adapted Text Adversarial Purification defense.