arXiv:2608.12845v1 [cs.IR] 13 Aug 2026

FSGR: Mitigating Token Frequency Bias for Fair SID-Based Generative Recommendation

Yuchen Zheng [1], Sihan Xu [2], Jingwen Yang [1], Xiangrui Cai [1]*, Haiwei Zhang [1], Xiaojie Yuan [1]

[1]VCIP, TMCC, DISSec, College of Computer Science, Nankai University
[2]VCIP, DISSec, College of Cryptology and Cyber Science, Nankai University
zhengyuchen@dbis.nankai.edu.cn, xusihan@nankai.edu.cn, 1120250410@mail.nankai.edu.cn, caixr@nankai.edu.cn, zhhaiwei@nankai.edu.cn, yuanxj@nankai.edu.cn

Abstract

Semantic ID (SID)-based generative recommendation has recently achieved remarkable success. However, existing methods suffer from a previously overlooked fairness issue, which we term **Token Frequency Bias**, where high-frequency SID tokens are systematically over-predicted while low-frequency SID tokens are under-predicted. This bias originates from the combined effects of imbalanced semantic codebooks during SID construction, and popularity bias together with the maximum likelihood estimation objective during recommendation training, resulting in unfair exposure across item categories. Existing SID methods mainly focus on improving codebook quality and overlook the impact of token frequency imbalance on downstream recommendation fairness, while LLM debiasing methods often yield suboptimal results when directly applied to SID-based recommendation, due to the hierarchical semantics of SID tokens. To address this issue, we propose **FSGR**, a fairness optimization framework for SID-based generative recommendation. During SID construction, FSGR employs OT-based Assignment Optimization and Dual-Criteria Re-anchor mechanism to form a more balanced SID representation space. During recommendation training, it adopts a two-stage training strategy and introduces Hierarchical Frequency Calibration for layer-specific fairness fine-tuning. Experiments on three public datasets with three backbone models demonstrate that FSGR mitigates token frequency bias and delivers an average Gini fairness improvement of over 20% while maintaining competitive recommendation accuracy.

Introduction

Traditional discriminative recommender systems are limited by item cold-start problem and the poor transferability of learned embedding spaces across domains (Rajput et al. 2023). To overcome these limitations, generative recommendation has emerged as a promising paradigm. Among existing approaches, Semantic ID (SID)-based generative recommendation represents each item as a sequence of discrete semantic tokens (Rajput et al. 2023; Zheng et al. 2024). By replacing conventional item embeddings with semantic token sequences, SID-based methods exhibit stronger semantic generalization and more efficient retrieval, attracting increasing attention from both academia and industry.

Despite these advantages, SID-based generative recommendation suffers from a critical fairness issue: high-frequency SID tokens are systematically over-predicted, while low-frequency tokens are under-predicted, causing recommendation exposure to concentrate on a small number of frequent token groups and marginalizing long-tail ones. To quantify the practical impact of the SID token imbalance, we conduct a preliminary study. We rank SID tokens according to their frequencies in the training set and divide them into three groups: Head (top 10%), Torso (middle 50%), and Tail (bottom 40%). Using the representative LC-Rec model (Zheng et al. 2024), we generate recommendations on Amazon "Arts, Crafts and Sewing" and "Musical Instruments" datasets (Ni, Li, and McAuley 2019). We then compare the proportions of the three token groups in the ground-truth labels and the generated recommendations. As illustrated in Figure 1, Head tokens consistently occupy a larger proportion in the recommendation results than in the ground truth, whereas Tail tokens are under-represented. Since Head tokens already dominate the ground-truth distribution, even a modest relative increase results in a significant concentration of exposure on their associated item categories. More importantly, this deviation is systematic rather than random: it consistently favors frequent tokens while suppressing infrequent ones, may create a "rich-get-richer" feedback loop that progressively marginalizes long-tail categories (Chang et al. 2024). Inspired by similar observations in natural language processing (Martinez et al. 2024), we define this phenomenon as **Token Frequency Bias**.

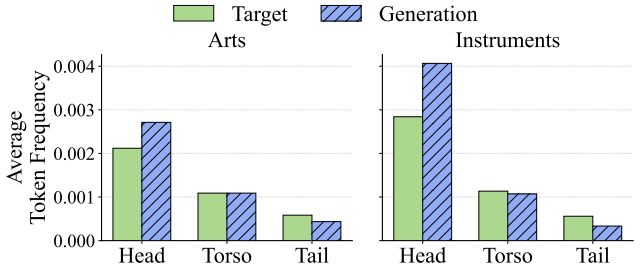

Figure 1: Token group proportions between ground truth and LC-Rec recommendations on two Amazon datasets. Tokens are ranked by training frequency and grouped into Head, Torso, Tail. The results reveal token frequency bias, where high-frequency Head tokens are over-exposed and low-frequency Tail tokens are under-exposed.

*Corresponding author.

Copyright © 2027, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The emergence of token frequency bias can be attributed to two complementary sources. The first arises during SID construction, where the semantic codebook is inherently imbalanced. Tokens representing broad semantic categories are repeatedly assigned to a large number of items, resulting in higher usage frequencies than other tokens, whereas tokens corresponding to less populated categories are rarely utilized. This imbalance introduces structural bias into the learned semantic space before recommendation training begins. The second arises during recommendation training. On the one hand, popularity bias in the training data causes SIDs of popular items to appear much more frequently than those of long-tail items. On the other hand, the maximum likelihood estimation (MLE) objective amplifies high-frequency training signals (Choi et al. 2020; Martinez et al. 2024), assigning higher prediction probabilities to frequent SID tokens while suppressing infrequent tokens. The interaction of these two factors causes the autoregressive generator to favor frequent SID tokens, thus degrading recommendation fairness. However, existing SID studies mainly focus on improving codebook utilization and reducing SID collisions (Kuai et al. 2024; Hu et al. 2026), while overlooking the impact of token frequency imbalance on downstream recommendation fairness. Moreover, dedicated debiasing methods for the LLM stage of SID-based generative recommendation remain unexplored. Although numerous debiasing methods have been proposed for natural language generation and can be applied to SID-based recommendation, they are often suboptimal because they treat all tokens uniformly. In contrast, SID tokens exhibit hierarchical semantics, where different layers encode different levels of semantic granularity and, therefore, require different debiasing strengths.

To address this issue, we propose **FSGR**, a fairness optimization framework for both SID generation and recommendation training. During SID construction, we introduce an Optimal Transport-based Assignment Optimization (OTA) to encourage balanced token utilization across the codebook within each mini-batch. Furthermore, we design a Dual-Criteria Re-anchor (DCR) mechanism that reuses inactive codewords to refine under-represented and overly crowded regions of the semantic space. Together, OTA and DCR encourage a more balanced SID representation space. During recommendation training, although the generated SIDs are already more balanced, popularity bias in training data and the MLE-based cross-entropy objective still reintroduce frequency bias. To further mitigate this issue, we adopt a two-stage training strategy. In the first stage, the LLM is trained with standard cross-entropy loss to learn semantic mappings. In the second stage, we perform fairness fine-tuning using Hierarchical Frequency Calibration (HFC), which applies different levels of debiasing to different SID layers. The main contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first work to identify and define token frequency bias in SID-based generative recommendation, revealing the combination of codebook imbalance, training-data popularity bias, and the intrinsic limitation of the MLE objective causes severe item-side fairness degradation at the SID token level.
- We propose FSGR, a fairness optimization framework. During SID generation, FSGR balances codebook utilization through OTA and DCR. During recommendation training, FSGR adopts semantic pre-training followed by HFC fine-tuning with layer-specific debiasing to improve SID token-level fairness.
- Experiments on three public datasets and three backbone models show that FSGR mitigates token frequency bias, delivering an average Gini fairness improvement of over 20% while maintaining competitive recommendation accuracy, which verifies its efficacy for fair SID-based generative recommendation.

Related Work

Semantic ID-Based Generative Recommendation

Recently, the paradigm of recommender systems has shifted from traditional discriminative models to generative approaches (Hou et al. 2025). To bridge the gap between massive item indices and LLMs, LC-Rec (Zheng et al. 2024) employs Residual Quantized Variational Autoencoders (RQ-VAE) to map item features into discrete token sequences known as Semantic IDs (SIDs), and further introduces uniform semantic mapping to mitigate index conflicts in the last SID layer via Sinkhorn-based optimal transport. Subsequent studies further improved the SID construction from different perspectives. ColaRec (Wang et al. 2024) constructs generative item identifiers from collaborative representations and introduces auxiliary indexing and alignment objectives to better integrate the item content with collaborative signals. QuaSID (Hu et al. 2026) mitigates SID collisions through Hamming-guided repulsion and benign overlap masking, with dual-tower contrastive learning enhancing SID quality. Despite these advances, existing methods mainly focus on improving representation quality and recommendation accuracy, while the fairness issue caused by imbalanced SID token distributions received little attention.

Fairness in Recommender Systems

Fairness has long been an important topic in recommender systems (Wang et al. 2023; Deng et al. 2025). For example, Zehlike et al. (2017) adjust recommendation lists to satisfy fairness constraints through re-ranking approaches, Zhang et al. (2023) learn fair item representations through adversarial learning methods, and Chang et al. (2024) incorporate fairness objectives into training loss through regularization-based methods. However, all of these methods operate on recommendation lists, item representations, or ranking objectives, and cannot be directly applied to SID-based generative recommendation, where item exposure is determined by the autoregressive generation of hierarchical SID tokens rather than explicit ranking scores or item embeddings. Thus, mitigating token frequency bias in SID-based generative recommendation remains unexplored. In this work, we focus on item-side fairness, with the goal of mitigating token frequency bias and improving the exposure of items associated with long-tail SID tokens.

Problem Formulation

Building on the empirical observations in the introduction, we first formally define token frequency bias, then formalize the SID-based generative recommendation task.

Definition 1. ***Token Frequency Bias:*** *In SID-based generative recommendation, the model tends to overestimate the probabilities of high-frequency SID tokens while underestimating those of low-frequency tokens, causing excessive exposure of items associated with frequent tokens and insufficient exposure of items associated with infrequent tokens.*

This bias is the core fairness problem that we target.

Next, we formalize the core task setting. Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively. Each item $v \in \mathcal{I}$ is associated with a feature vector $\mathbf{e}_v \in \mathbb{R}^D$. A Residual Quantized Variational Autoencoder (RQ-VAE) quantizes $\mathbf{e}_v$ into a Semantic ID (SID) sequence $\mathbf{s}_v = [s_v^1, \ldots, s_v^L]$, where L is the number of quantization layers and s_v^l is the codebook token at layer l. Given a user's interaction history $\mathcal{H}_u = [v_1, \ldots, v_n]$, SID-based generative recommendation trains a model parameterized by Θ to autoregressively predict the SID sequence of the target item v_{n+1} by maximizing:

$$\max_{\Theta} \sum_{u \in \mathcal{U}} \sum_{l=1}^{L} \log P\left(s_{v_{n+1}}^l \mid \mathcal{H}_u, s_{v_{n+1}}^{<l}; \Theta\right). \tag{1}$$

This paper addresses the token frequency bias problem in SID-based LLM generative recommendation by jointly optimizing SID quantization and recommendation generation to reduce token distribution imbalance while preserving recommendation accuracy.

Method

To mitigate token frequency bias in SID-based generative recommendation, we propose FSGR, a two-component fairness optimization framework. As shown in Figure 2, first, OT-based Assignment Optimization (OTA) and Dual-Criteria Re-anchor (DCR) improve codebook utilization during quantization to encourage a more balanced semantic representation space. Building upon this, we adopt a two-stage training strategy: standard cross-entropy first establishes semantic alignment, followed by Hierarchical Frequency Calibration (HFC) that mitigates autoregressive prediction bias via layer-wise granularity calibration while preserving alignment.

Balanced Semantic Quantization

In the conventional way, given an item feature vector $\mathbf{e}_v \in \mathbb{R}^D$, an L-layer RQ-VAE method maps it into a discrete token sequence through multi-level residual quantization. At layer l, the encoder outputs a residual vector $\mathbf{r}_l$, and the quantizer finds the nearest codeword s^l from the codebook $\tilde{\mathcal{C}}_l \in \mathbb{R}^{K_l \times D}$: $s^l = \arg\min_k \|\mathbf{r}_l - \mathbf{e}_k^l\|_2^2$, where $\mathbf{e}_k^l$ is the embedding of s^l, K_l is the vocabulary size of the l-th SID layer. The quantized representation for this layer is $\mathbf{q}_l = \mathbf{e}_k^l$, and the residual passed to the next layer is $\mathbf{r}_{l+1} = \mathbf{r}_l - \mathbf{q}_l$. The standard RQ-VAE objective is $\mathcal{L}_{\text{rq}}$ consists of reconstruction loss and commitment loss:

$$\mathcal{L}_{\text{rq}} = \|\mathbf{e}_v - \hat{\mathbf{e}}_v\|_2^2 + \sum_{l=1}^{L} \|\text{sg}[\mathbf{r}_l] - \mathbf{q}_l\|_2^2 + \beta \sum_{l=1}^{L} \|\mathbf{r}_l - \text{sg}[\mathbf{q}_l]\|_2^2, \tag{2}$$

where $\hat{\mathbf{e}}_v$ is the reconstructed vector and $\text{sg}[\cdot]$ denotes the stop-gradient operation. However, in conventional RQ-VAE, the utilization of the codebook is typically highly imbalanced. To alleviate this issue, we propose Balanced Semantic Quantization (BSQ), which encourages balanced codebook utilization through OT-based Assignment Optimization and Dual-Criteria Re-anchor mechanism.

OT-based Assignment Optimization In order to address imbalanced codebook utilization, we model the quantization process within each mini-batch as an Optimal Transport (OT) problem, aiming to minimize transport cost while promoting uniform marginal distributions for codebook utilization. Formally, given a batch of data $\mathcal{X} \in \mathbb{R}^{B \times D}$ and the codebook $\tilde{\mathcal{C}} \in \mathbb{R}^{K \times D}$, we construct the cost matrix $\mathbf{C}$, where C_{ik} represents the distance between sample $i \in \mathcal{X}$ and the codeword $k \in \tilde{\mathcal{C}}$. Then we set the sample marginal $\boldsymbol{\mu}$ and codebook target marginal $\boldsymbol{\nu}$ to be uniform: $\boldsymbol{\mu} = \frac{1}{B}\mathbf{1}_B$ and $\boldsymbol{\nu} = \frac{1}{K}\mathbf{1}_K$, and employ the Sinkhorn-Knopp algorithm (Cuturi 2013) to solve the entropy-regularized optimal transport plan $\mathbf{P}^* \in \mathbb{R}^{B \times K}$:

$$\mathbf{P}^* = \arg\min_{\mathbf{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{P}, \mathbf{C} \rangle - \epsilon H(\mathbf{P}), \tag{3}$$

where $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is the set of transport matrices satisfying the marginal constraints, $H(\mathbf{P})$ is the entropy regularizer, $\mathbf{P}_{ik}^*$ indicates the optimal probability mass that sample i should assign to codeword k.

However, directly applying the discrete transport plan is non-differentiable, thus we define the model's actual soft assignment matrix $\mathbf{Q}$ as:

$$\mathbf{Q}_{ik} = \frac{\exp(-C_{ik}/\tau_q)}{\sum_{j=1}^{K} \exp(-C_{ij}/\tau_q)}, \tag{4}$$

where τ_q is the temperature coefficient controlling the smoothness of the soft assignment distribution. We minimize the divergence between the actual soft assignment $\mathbf{Q}$ and the optimal transport plan $\mathbf{P}^*$ through KL divergence:

$$\mathcal{L}_{\text{OTA}} = D_{\text{KL}}(\mathbf{P}^* \parallel \mathbf{Q}) = \sum_{i=1}^{B} \sum_{k=1}^{K} P_{ik}^* \log \frac{P_{ik}^*}{Q_{ik}}. \tag{5}$$

The total loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rq}} + \lambda_o \mathcal{L}_{\text{OTA}}, \tag{6}$$

where λ_o is a balancing coefficient controlling the strength of the OT regularization, which is gradually increased during training according to a warm-up schedule.

In this step, $\mathcal{L}_{\text{OTA}}$ encourages the assignment distribution to follow a balanced transport plan, thereby improving overall codebook utilization.

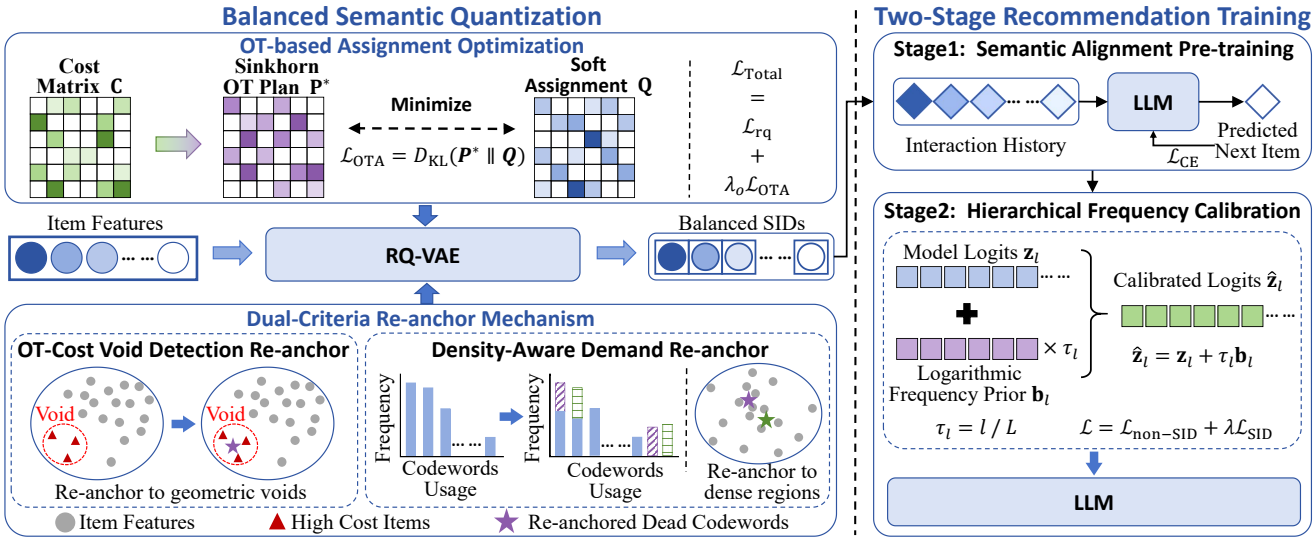

Figure 2: Overview of FSGR. The framework consists of two modules: (1) Balanced Semantic Quantization, which constructs more balanced Semantic IDs via OT-based Assignment Optimization and Dual-Criteria Re-anchor, and (2) Two-Stage Recommendation Training, which performs semantic alignment pre-training followed by Hierarchical Frequency Calibration to mitigate token frequency bias.

Dual-Criteria Re-anchor Mechanism Although OT-based Assignment Optimization improves overall codebook utilization, the learned semantic space may still contain under-represented and overly crowded regions. Meanwhile, some codewords remain inactive for extended periods and receive little or no gradient updates, making them difficult to recover through standard optimization. To address both issues, we propose Dual-Criteria Re-anchor (DCR) mechanism, which periodically identifies dead codewords and re-initializes them to either under-covered or high-density regions according to two complementary criteria. By reusing inactive codewords, DCR not only refines the geometry of the semantic space but also revives dead codewords and improves overall codebook utilization.

Specifically, let the codebook utilization count be $n_k = \sum_{i=1}^{|\mathcal{I}|} \mathbb{I}[s_i = k]$. A codeword k_d is deemed dead if $n_{k_d} < \delta$, where δ is a threshold and is set to 1 in this paper. DCR partitions the dead codes and re-anchors them via two strategies:

Strategy A: OT-Cost Void Detection Re-anchor. To fill geometric voids in the latent feature space, we define the transport cost of training sample i as $\text{Cost}_i = \sum_{k=1}^{K} P^*_{ik} C_{ik}$. A high transport cost indicates that the sample is distant from existing codewords, residing in an under-represented region with geometric voids. We select indices $\mathcal{I}_{\text{void}}$ from samples with the highest costs and re-anchor a portion of dead codes to these samples' encoded vectors: $\mathbf{e}_{k_d} \leftarrow \mathbf{e}_i, i \in \mathcal{I}_{\text{void}}$.

Strategy B: Density Aware Demand Re-anchor. While Strategy A targets under-represented regions, Strategy B focuses on overcrowded ones. We construct a sampling distribution based on codeword utilization frequency and select representative instances from densely populated semantic regions, denoted by $\mathcal{I}_{\text{demand}}$. The remaining dead codewords are re-anchored to the encoded representations of the selected instances: $\mathbf{e}_{k_d} \leftarrow \mathbf{e}_i, i \in \mathcal{I}_{\text{demand}}$, thereby introducing additional codewords into densely populated semantic regions and improving their representation capacity.

Combined with the codebook-level OTA, the codeword-level DCR further refines the codebook geometry through individual codeword re-anchoring, improving the representation of both under-represented and densely populated semantic regions.

Two-Stage Recommendation Training

Although BSQ balances SID representations, token frequency bias persists in recommendation training: popular items dominate interactions, making their SID tokens far more frequent than long-tail ones. And the MLE objective further amplifies this imbalance, biasing the autoregressive decoder toward high-frequency tokens. A straightforward solution is to introduce frequency-aware regularization throughout training. However, doing so interferes with learning semantic correspondence between user behaviors and SIDs, leading to unstable optimization and degraded accuracy. We thus decouple training into two stages: the first learns semantic alignment via standard cross-entropy, while the second conducts fairness fine-tuning with Hierarchical Frequency Calibration to calibrate SID predictions while preserving learned semantic knowledge.

Semantic Alignment Pre-training In the first stage, the model is optimized via standard cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\sum_{l=1}^{L} \log P\left(s^l \mid \mathcal{H}, s^{<l}\right), \tag{7}$$

where $\mathcal{H}$ is the user's interaction history, s^l is the predicted SID at layer l. This stage focuses solely on learning semantic

correspondence between user behaviors and SID sequences without introducing any frequency-aware constraints.

Hierarchical Frequency Calibration After the model converges to a stable semantic space, we perform fairness-aware fine-tuning through Hierarchical Frequency Calibration (HFC). Let $\mathbf{z}_l \in \mathbb{R}^{K_l}$ denote the prediction logits of the l-th SID layer. Based on the empirical token distribution of the training set, we compute the logarithmic frequency prior $\mathbf{b}_l = -\log(\mathbf{f}_l + \epsilon)$, where $\mathbf{f}_l$ is the normalized occurrence frequency of SID tokens, ϵ is a small constant for numerical stability. The calibrated logits are computed as

$$\hat{\mathbf{z}}_l = \mathbf{z}_l + \tau_l \mathbf{b}_l, \tag{8}$$

where τ_l is the layer-specific calibration temperature. By incorporating the empirical frequency prior, HFC reduces the model's reliance on the skewed token distribution induced by the training data.

Moreover, SID possesses a hierarchical semantic structure: lower layers encode coarse semantic categories, whereas higher layers capture increasingly fine-grained item semantics (Rajput et al. 2023). Applying the same calibration strength to all layers either distorts coarse semantic representations or insufficiently debiases fine-grained predictions. Thus, HFC adopts a layer-aware temperature schedule. The calibration temperature increases progressively across SID layers according to $\tau_l = l/L,\ l \in \{1, \ldots, L\}$. So that deeper layers receive stronger frequency calibration, while lower layers remain relatively stable. This design aligns with the hierarchical semantics of SID, mitigating token frequency bias while preserving recommendation accuracy.

The calibrated logits are used only for SID prediction positions during the second-stage fine-tuning, while non-SID tokens are optimized using the original logits. The overall objective is $\mathcal{L} = \mathcal{L}_{\text{non-SID}} + \lambda_h \mathcal{L}_{\text{SID}}$, where

$$\mathcal{L}_{\text{non-SID}} = -\sum_{t \in \Omega_N} \log P(y_t \mid \mathbf{x}), \tag{9}$$

$$\mathcal{L}_{\text{SID}} = -\sum_{t \in \Omega_S} \log \frac{\exp(\hat{z}_{t,y_t}^{(l_t)})}{\sum_{j=1}^{K_{l_t}} \exp(\hat{z}_{t,j}^{(l_t)})}. \tag{10}$$

Here, Ω_N and Ω_S denote prediction positions of non-SID and SID tokens, $\mathbf{x}$ denotes input sequence, y_t is the target token at prediction position t, $\hat{z}$ is the calibrated logit, and λ_h balances recommendation learning and frequency calibration.

Experiments

Experimental Setup

Dataset We evaluated the proposed approach on three subsets of Amazon review data (Ni, Li, and McAuley 2019), including "Luxury Beauty", "Industrial and Scientific", "Software". Each item in these datasets is associated with a title and a description. Following prior work (Zheng et al. 2024), we eliminated users and items with fewer than five interactions. User behavior sequences were constructed in chronological order, with a unified maximum length of 20.

Baselines and Backbones We compared FSGR with three representative SID generation methods: vanilla **RQ-VAE** (Rajput et al. 2023), **RT** (Fifty et al. 2025), and **QuaSID** (Hu et al. 2026), as well as two LLM token debiasing methods, **MiLe** (Su et al. 2024) and **WAKL** (Shrestha and Srinivasan 2025). Experiments were conducted on three representative generative recommendation backbones: **TIGER** (Rajput et al. 2023), **Llama3.1-8B** (Grattafiori et al. 2024), and **Qwen3-8B** (Qwen-Team 2025).

Evaluation Metrics We evaluated our method in terms of recommendation performance and fairness. Following previous work, recommendation performance were measured by **Recall (R@K)** and **NDCG (N@K)**, while fairness were evaluated using the **Gini coefficient (G@K)** (Gini 1936) computed on the generated SID token frequency distribution. A lower Gini coefficient indicates more balanced token exposure and weaker token frequency bias.

Implementation Details We used a 4-layer RQ-VAE with a codebook size of 256 for each layer. Our method was implemented using PyTorch and experiments were conducted on an NVIDIA RTX A6000. The LLM models were fine-tuned with Low-Rank Adaptation (LoRA) (Hu et al. 2022). We optimized the model using AdamW-8bit (Dettmers et al. 2021) optimizer and the detailed hyper-parameter settings are provided in the supplementary material and code.

Performance Comparison

Table 1 reports the overall performance of FSGR. Since TIGER is a Transformer-based recommender rather than an LLM, HFC and LLM token debiasing baselines were evaluated only on Llama and Qwen, whereas TIGER was used only to assess SID generation component. For Llama and Qwen, we compared the complete FSGR framework with both SID generation and LLM token debiasing baselines. On the TIGER backbone, our proposed BSQ yields the lowest Gini scores across all datasets, surpassing other SID generation methods in recommendation fairness. Meanwhile, it retains competitive Recall and NDCG, indicating that balanced codebook utilization effectively mitigates token frequency bias with only marginal impact on recommendation accuracy. For the two LLM backbones (Llama3.1 and Qwen3), the full FSGR framework achieves the best fairness performance, delivering lower Gini coefficients than both SID generation and LLM token debiasing baselines while preserving competitive recommendation accuracy. These results demonstrate that the combination of balanced SID construction and hierarchical frequency calibration alleviates token frequency bias without compromising recommendation performance.

Ablation Analysis

To evaluate the contribution of each component, we conducted an ablation study on Llama and Qwen. We compared four variants: **w/o BSQ**, which removes Balanced Semantic Quantization, **w/o HFC**, which removes Hierarchical Frequency Calibration, **Raw**, which uses neither BSQ nor HFC, and the complete FSGR framework (**Full**). As shown in Table 2, the complete framework achieves the lowest Gini scores across all datasets and backbone models, demonstrating that

	Method	Beauty					Industrial					Software				
		R@5	R@10	N@5	N@10	G@10	R@5	R@10	N@5	N@10	G@10	R@5	R@10	N@5	N@10	G@10
Tiger	RQ-VAE	**0.2673**	0.2972	**0.2365**	**0.2462**	0.7310	**0.0591**	**0.0684**	**0.0509**	**0.0539**	0.7565	0.1783	0.2233	0.1561	0.1705	0.8671
	RT	0.2573	0.2877	0.2276	0.2374	0.7635	0.0553	0.0670	0.0486	0.0524	0.6887	**0.1898**	**0.2260**	**0.1646**	**0.1763**	0.8906
	QuaSID	0.2631	**0.2980**	0.2280	0.2393	0.7043	0.0545	0.0644	0.0471	0.0504	0.7111	0.1558	0.1777	0.1392	0.1463	0.7389
	Our-BSQ	0.2662	0.2975	0.2311	0.2412	**0.5494**	0.0554	0.0653	0.0497	0.0529	**0.5856**	0.1722	0.2112	0.1486	0.1613	**0.6981**
Llama3.1-8B	RQ-VAE	0.3116	0.3234	0.2855	0.2894	0.7174	0.0904	0.0970	0.0839	0.0861	0.6059	0.2139	0.2386	0.1834	0.1913	0.8610
	RT	0.3035	0.3145	0.2825	0.2862	0.7573	0.0886	0.0947	0.0814	0.0834	0.6022	0.2183	**0.2551**	0.1871	0.1989	0.8570
	QuaSID	0.2906	0.2964	0.2688	0.2708	0.7122	0.0884	0.0938	0.0811	0.0829	0.6818	0.1980	0.2145	0.1759	0.1813	0.7484
	MiLe	**0.3119**	0.3232	**0.2859**	0.2896	0.7281	**0.0953**	**0.1022**	**0.0880**	**0.0902**	0.5931	**0.2194**	0.2447	**0.1932**	**0.2013**	0.8609
	WAKL	0.3072	**0.3276**	0.2836	**0.2903**	0.7261	0.0885	0.0952	0.0817	0.0839	0.6029	0.2057	0.2298	0.1816	0.1895	0.8586
	Our	0.3066	0.3164	0.2801	0.2832	**0.4976**	0.0918	0.0967	0.0835	0.0851	**0.3439**	0.2145	0.2414	0.1866	0.1952	**0.6754**
Qwen3-8B	RQ-VAE	0.3158	0.3353	0.2895	0.2957	0.7368	0.0908	0.0989	**0.0839**	**0.0865**	0.6559	0.2090	0.2414	0.1840	0.1944	0.8646
	RT	0.3182	0.3337	0.2929	0.2979	0.7588	0.0889	0.0966	0.0814	0.0839	0.6431	0.2117	0.2463	0.1849	0.1961	0.8568
	QuaSID	0.3177	0.3305	0.2879	0.2921	0.7154	0.0858	0.0926	0.0769	0.0791	0.7054	0.2002	0.2408	0.1749	0.1881	0.7638
	MiLe	0.3195	0.3397	0.2932	0.2998	0.7292	**0.0920**	0.0992	**0.0839**	0.0862	0.6135	0.2112	0.2496	**0.1887**	0.2007	0.8602
	WAKL	0.3122	0.3295	0.2848	0.2904	0.7364	0.0919	**0.0995**	**0.0839**	0.0863	0.6371	0.2068	0.2457	0.1797	0.1923	0.8612
	Our	**0.3329**	**0.3521**	**0.2960**	**0.3023**	**0.5128**	0.0891	0.0971	0.0807	0.0833	**0.4203**	**0.2189**	**0.2639**	0.1872	**0.2019**	**0.7127**

Table 1: Performance comparison with SID generation and LLM token debiasing baselines on Tiger, Llama3.1-8B and Qwen3-8B. Best and second-best results are highlighted in bold and underlined, respectively.

BSQ and HFC complement each other in mitigating token frequency bias. Removing either module leads to a degradation in fairness, confirming that both balanced SID construction and frequency-aware recommendation training are necessary. From the perspective of individual components, removing BSQ generally results in higher Gini scores than the full model, indicating that constructing a balanced semantic codebook provides a stronger foundation for downstream recommendation fairness. Removing HFC also causes an increase in Gini, suggesting that frequency bias introduced during autoregressive training can also lead to unfairness, even when the SID representations are balanced. Furthermore, although fairness optimization may slightly affect recommendation accuracy on a few settings, the complete framework maintains comparable Recall and NDCG in most cases.

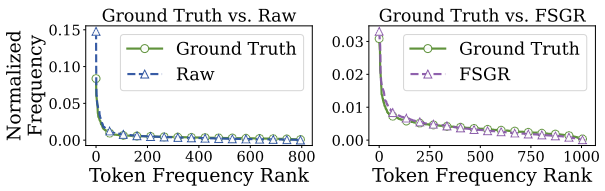

Figure 3: Comparison of SID token frequency distributions between model predictions and ground-truth on Industrial using Qwen3. FSGR produces a smoother distribution that better matches the ground-truth, mitigating high-frequency bias observed in raw RQ-VAE and LLM.

Token Frequency Distribution Analysis

Figure 3 compares frequency distributions of SID tokens in the generated recommendation results with the ground-truth distribution on Qwen3. The raw RQ-VAE and LLM overestimates the most frequent SID tokens, producing a much steeper distribution than the ground truth, which reflects the token frequency bias identified in this work. In contrast, FSGR generates a distribution closely matches the target distribution, substantially reducing the over-prediction of high-frequency tokens while increasing the relative exposure of less frequent tokens. This demonstrates that the proposed BSQ and HFC alleviate token frequency bias throughout both SID construction and recommendation generation.

Analysis of Codebook Utilization

To evaluate whether BSQ balances SID allocation, we analyzed codebook utilization from both statistical and distributional perspectives. Table 3 reports the codebook Coverage and Gini coefficient, while Figure 4 illustrates the token frequency distributions sorted by usage frequency. As shown in Table 3, the proposed method achieves the highest Coverage and the lowest Gini on all datasets. In particular, the Coverage approaches 100% on Industrial dataset, indicating that nearly all codewords participate in quantization. Meanwhile, the lower Gini scores demonstrate that token usage is distributed much more evenly across the codebook than that of other methods. Figure 4 further provides a distributional view of codebook utilization. Compared with the baselines, the proposed method exhibits a flatter token frequency curve, indicating that the dominance of a small number of high-frequency codewords is alleviated while previously under-utilized codewords are activated more frequently. These results demonstrate that BSQ encourages a more balanced semantic representation space, providing a stronger foundation for mitigating token frequency bias in downstream generative recommendation.

Analysis of Layer-wise Temperature Assignment

To validate the effectiveness of the proposed layer-wise temperature assignment ($\tau_l = l/L$), we conducted experiments

	Method	Beauty					Industrial					Software				
		R@5	R@10	N@5	N@10	G@10	R@5	R@10	N@5	N@10	G@10	R@5	R@10	N@5	N@10	G@10
Llama3.1	Raw	**0.3116**	**0.3234**	**0.2855**	**0.2894**	0.7174	0.0904	0.0970	0.0839	0.0861	0.6059	0.2139	0.2386	0.1834	0.1913	0.8610
	w/o BSQ	0.3048	0.3153	0.2817	0.2852	0.6937	0.0907	0.0967	**0.0847**	**0.0866**	0.5630	0.2030	0.2287	0.1760	0.1845	0.8454
	w/o HFC	0.3048	0.3145	0.2799	0.2831	0.5407	0.0913	**0.0973**	0.0828	0.0847	0.4211	0.2112	**0.2419**	0.1840	0.1940	0.7089
	Full	0.3066	0.3164	0.2801	0.2832	**0.4976**	**0.0918**	0.0967	0.0835	0.0851	**0.3439**	**0.2145**	0.2414	**0.1866**	**0.1952**	**0.6754**
Qwen3	Raw	0.3158	0.3353	0.2895	0.2957	0.7368	**0.0908**	**0.0989**	0.0839	0.0865	0.6559	0.2090	0.2414	0.1840	0.1944	0.8646
	w/o BSQ	0.3164	0.3353	0.2877	0.2939	0.7108	0.0906	0.0979	**0.0842**	**0.0866**	0.6054	0.2123	0.2534	0.1815	0.1946	0.8695
	w/o HFC	0.3284	0.3481	0.2937	0.3002	0.5563	0.0907	0.0973	0.0811	0.0832	0.4858	0.2095	0.2567	0.1815	0.1967	0.7350
	Full	**0.3329**	**0.3521**	**0.2960**	**0.3023**	**0.5128**	0.0891	0.0971	0.0807	0.0833	**0.4203**	**0.2189**	**0.2639**	**0.1872**	**0.2019**	**0.7127**

Table 2: Ablation study of FSGR. Removing either BSQ (w/o BSQ) or HFC (w/o HFC) degrades recommendation fairness, while the complete framework (Full) achieves the best overall trade-off between accuracy and fairness.

Dataset	Metric	RQ-VAE	RT	QuaSID	Our-BSQ
Beauty	Coverage	0.6748	0.6309	0.8779	**0.9990**
	Gini	0.6303	0.6864	0.6119	**0.3479**
Industrial	Coverage	0.7832	0.7930	0.9180	**1.0000**
	Gini	0.4904	0.5107	0.6064	**0.2590**
Software	Coverage	0.4639	0.4707	0.8623	**0.9824**
	Gini	0.7954	0.7972	0.6199	**0.5138**

Table 3: Comparison of codebook utilization across different Semantic ID construction methods. Higher Coverage and lower Gini indicate more balanced codebook utilization.

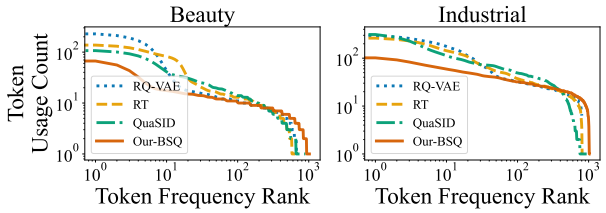

Figure 4: Token frequency distributions of different Semantic ID construction methods on two datasets. Tokens are sorted by usage frequency in descending order. A flatter distribution indicates more balanced codebook utilization.

on Qwen3, comparing the HFC with two alternative strategies: **Reverse** τ_l, which applies larger calibration temperatures to lower SID layers and smaller ones to higher layers ($\tau_l = (L - l + 1)/L$), and **Fixed** τ_l, which assigns the same calibration strength to all SID layers ($\tau_l = (\Sigma_1^L l/L)/L$). The results are reported in Table 4. Overall, the proposed HFC achieves the best overall trade-off between recommendation accuracy and fairness across all three datasets. Although reverse τ_l further reduces the Gini coefficient on Beauty and Industrial, it degrades Recall and NDCG, indicating that applying strong calibration to lower SID layers disrupts coarse semantic representations. In contrast, the proposed layer-wise schedule applies weaker calibration to lower layers and progressively strengthens it for deeper layers, effectively mitigating frequency bias while preserving semantic alignment. Compared with using fixed τ_l, HFC achieves better overall

Dataset	Metric	HFC	Reverse τ_l	Fixed τ_l
Beauty	R@5	**0.3329**	0.3258	0.3316
	R@10	**0.3521**	0.3452	0.3471
	N@5	0.2960	0.2921	**0.2964**
	N@10	**0.3023**	0.2986	0.3014
	G@10	0.5128	**0.4921**	0.5094
Industrial	R@5	0.0891	0.0869	**0.0899**
	R@10	**0.0971**	0.0941	0.0963
	N@5	0.0807	0.0796	**0.0810**
	N@10	**0.0833**	0.0820	0.0830
	G@10	0.4203	**0.3228**	0.3632
Software	R@5	**0.2189**	0.2068	0.2063
	R@10	**0.2639**	0.2479	0.2436
	N@5	**0.1872**	0.1815	0.1818
	N@10	**0.2019**	0.1946	0.1937
	G@10	**0.7127**	0.7218	0.7342

Table 4: Comparison of different layer-wise temperature assignment strategies for HFC on Qwen3. Reverse τ_l assigns larger calibration temperatures to lower SID layers, Fixed τ_l uses the same temperature for all layers.

performance, demonstrating that different SID layers require different calibration strengths according to their hierarchical semantic roles. These results validate the effectiveness of the proposed layer-aware temperature assignment.

Conclusion

In this paper, we investigate the token frequency bias problem in Semantic ID (SID)-based generative recommendation and propose FSGR, a fairness optimization framework that jointly improves SID construction and recommendation training. Specifically, FSGR combines Balanced Semantic Quantization for SID generation with Two-Stage Recommendation Training to mitigate frequency bias during recommendation training. Experiments on three datasets and three backbones demonstrate that FSGR improves SID token fairness while maintaining competitive recommendation accuracy on most settings. In future work, we plan to extend FSGR to more advanced generative recommendation architectures and explore adaptive frequency-aware optimization strategies for better fairness–accuracy trade-offs.

References

Chang, B.; Meng, C.; Ma, H.; Chang, S.; Gu, Y.; Peng, Y.; Feng, J.; Zhang, Y.; Bi, S.; Chi, E. H.; et al. 2024. Cluster anchor regularization to alleviate popularity bias in recommender systems. In *Companion Proceedings of the ACM Web Conference 2024*, 151–160.

Choi, B.-J.; Hong, J.; Park, D.; and Lee, S. W. 2020. F^2-softmax: Diversifying neural text generation via frequency factorized softmax. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9167–9182.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Deng, Z.; Li, J.; Liu, W.; and Zhao, J. 2025. Unbiased Interest Modeling in Sequential Basket Analysis: Addressing Repetition Bias with Multi-Factor Estimation. *ACM Transactions on Recommender Systems*, 3(4): 1–27.

Dettmers, T.; Lewis, M.; Shleifer, S.; and Zettlemoyer, L. 2021. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*.

Fifty, C.; Junkins, R.; Duan, D.; Iyengar, A.; Liu, J.; Amid, E.; Thrun, S.; and Ré, C. 2025. Restructuring vector quantization with the rotation trick. In *International Conference on Learning Representations*, volume 2025, 19153–19188.

Gini, C. 1936. On the measure of concentration with special reference to income and statistics. *Colorado college publication, general series*, 208: 73.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; and et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Hou, Y.; Zhang, A.; Sheng, L.; Yang, Z.; Wang, X.; Chua, T.-S.; and McAuley, J. 2025. Generative recommendation models: Progress and directions. In *Companion Proceedings of the ACM on Web Conference 2025*, 13–16.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, Z.; Chen, Y.; Pan, Y.; Yuan, X.; Yin, Y.; Wang, D.; Xia, B.; Luo, Z.; Wang, H.; Ni, S.; Liang, D.; Wang, J.; Cai, S.; Zhou, T.; Ren, F.; and Ou, W. 2026. Stop Treating Collisions Equally: Qualification-Aware Semantic ID Learning for Recommendation at Industrial Scale. arXiv:2603.00632.

Kuai, Z.; Chen, Z.; Wang, H.; Li, M.; Miao, D.; Binbin, W.; Chen, X.; Kuang, L.; Han, Y.; Wang, J.; et al. 2024. Breaking the Hourglass Phenomenon of Residual Quantization: Enhancing the Upper Bound of Generative Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 677–685.

Martinez, R. D.; Goriely, Z.; Caines, A.; Buttery, P.; and Beinborn, L. 2024. Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5999–6011.

Ni, J.; Li, J.; and McAuley, J. 2019. 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)*, 188–197.

Qwen-Team. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Rajput, S.; Mehta, N.; Singh, A.; Hulikal Keshavan, R.; Vu, T.; Heldt, L.; Hong, L.; Tay, Y.; Tran, V.; Samost, J.; Kula, M.; Chi, E.; and Sathiamoorthy, M. 2023. Recommender Systems with Generative Retrieval. In *Advances in Neural Information Processing Systems*, volume 36, 10299–10315. Curran Associates, Inc.

Shrestha, I.; and Srinivasan, P. 2025. LLM Bias Detection and Mitigation through the Lens of Desired Distributions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 1464–1480.

Su, Z.; Wu, X.; Bai, X.; Lin, Z.; Chen, H.; Ding, G.; Zhou, W.; and Hu, S. 2024. MiLe Loss: A New Loss for Mitigating the Bias of Learning Difficulties in Generative Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 250–262. Association for Computational Linguistics.

Wang, Y.; Ma, W.; Zhang, M.; Liu, Y.; and Ma, S. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3): 1–43.

Wang, Y.; Ren, Z.; Sun, W.; Yang, J.; Liang, Z.; Chen, X.; Xie, R.; Yan, S.; Zhang, X.; Ren, P.; et al. 2024. Content-based collaborative generation for recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2420–2430.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1569–1578.

Zhang, Z.; Liu, Q.; Jiang, H.; Wang, F.; Zhuang, Y.; Wu, L.; Gao, W.; and Chen, E. 2023. FairLISA: Fair User Modeling with Limited Sensitive Attributes Information. In *Advances in Neural Information Processing Systems*, volume 36, 41432–41450. Curran Associates, Inc.

Zheng, B.; Hou, Y.; Lu, H.; Chen, Y.; Zhao, W. X.; Chen, M.; and Wen, J.-R. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 1435–1448. IEEE.