Breaking the Hourglass Phenomenon of Residual Quantization: Enhancing the Upper Bound of Generative Retrieval

Zhirui Kuai[1,†] and **Zuxu Chen**[2,†], **Mingming Li**[3,*] and **Huimu Wang**[3,*],
Dadong Miao[3] and **Binbin Wang**[3] and **Xusong Chen**[3] and **Li Kuang**[1] and **Yuxing Han**[2,*]
Jiaxing Wang[3] and **Guoyu Tang**[3] and **Lin Liu**[3] and **Songlin Wang**[3] and **Jingwei Zhou**[3]

[1] Central South University, School of Computer Science and Engineering, China
[2] Shenzhen International Graduate School, Tsinghua University, China
[3] JD.com, Beijing, China

*Corresponding Author. †Equal Contribution.

INTRODUCTION

Generative retrieval (GR)[1][2] has rapidly become a powerful method in search and recommendation systems, particularly in high-demand areas like e-commerce. Unlike traditional methods, GR uses compact numeric identifiers, or Semantic Identifiers (SIDs) [1], generated by **residual quantization (RQ)**. This enables faster retrieval, improved inference speed, and broader generalization across inputs such as user queries and behavioral data.

In our study, **we identify a critical challenge in RQ-based SID generation—the "Hourglass Phenomenon"—where token distribution in intermediate codebook layers becomes overly concentrated** (see **Figure 1**). This concentration limits path diversity, leading to path sparsity and a long-tail distribution that reduces the model's representational capacity. This structural bottleneck significantly affects generative retrieval, particularly for applications with large catalogs and complex user behavior.

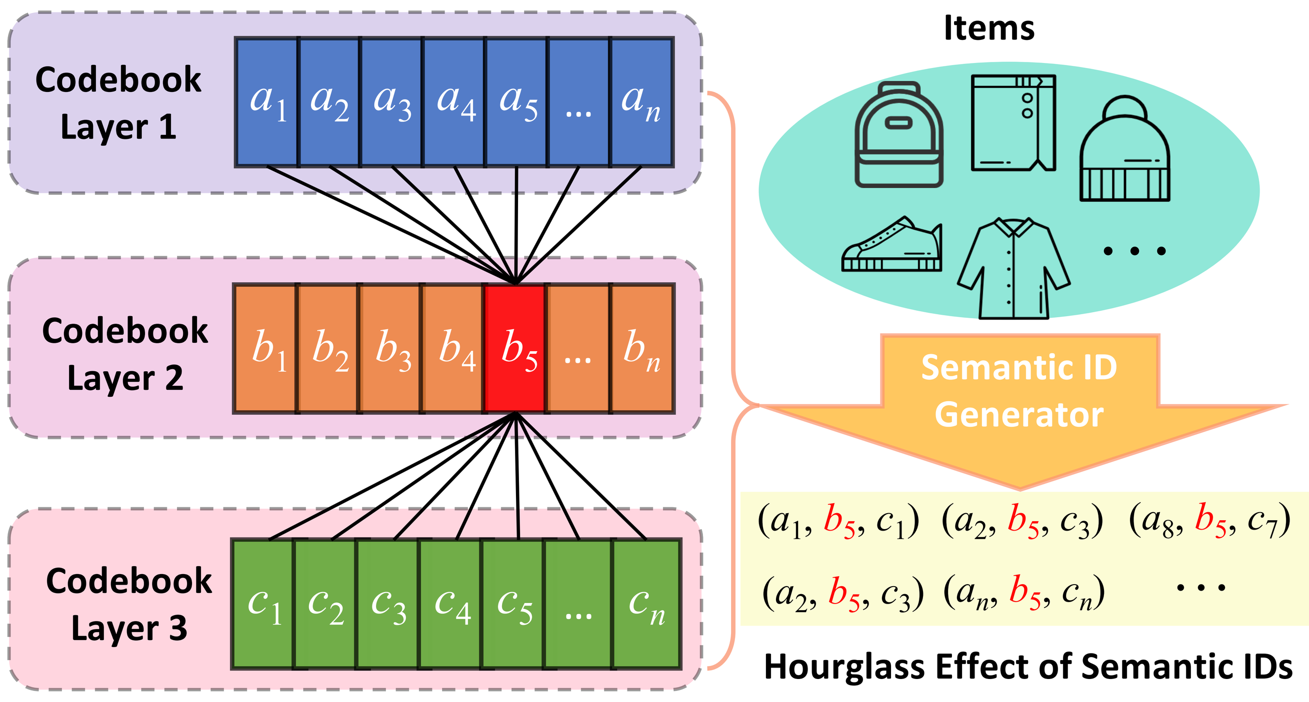

Figure 1: The Hourglass Phenomenon of Semantic IDs

To address this, we **analyze the hourglass phenomenon's causes and impact**, revealing structural issues in RQ-SID that lead to path sparsity and token imbalance. **We propose two solutions**: **a heuristic layer-removal method** and **an adaptive variable-length token str**ategy, both of which significantly enhance codebook utilization and retrieval performance, laying the groundwork for more effective GR systems in real-world settings.

Problem of GR based on RQ

1. Hourglass Phenomenon

In RQ-based Semantic ID (SID) generation, an **"Hourglass Phenomenon"** emerges, where intermediate codebook layers exhibit concentrated token distributions, creating a bottleneck effect. This effect is statistically evidenced by two main issues: (1) **Path Sparsity** – resulting in low code table utilization due to limited path diversity, and (2) **Long-Tail Distribution** – where the majority of routes converge onto a single token, concentrating retrieval paths and limiting representational flexibility, as shown in the **Figure 2**.

To test the **generalizability of this effect**, we conducted visualization experiments across various parameters (e.g., code table size, layer count). **Results confirm a pronounced hourglass effect with sparse path distribution**. **Statistical analysis of the second layer shows low entropy, high Gini coefficient, and large standard deviation**, indicating a highly skewed, long-tail distribution, as shown in **Figure 4**.

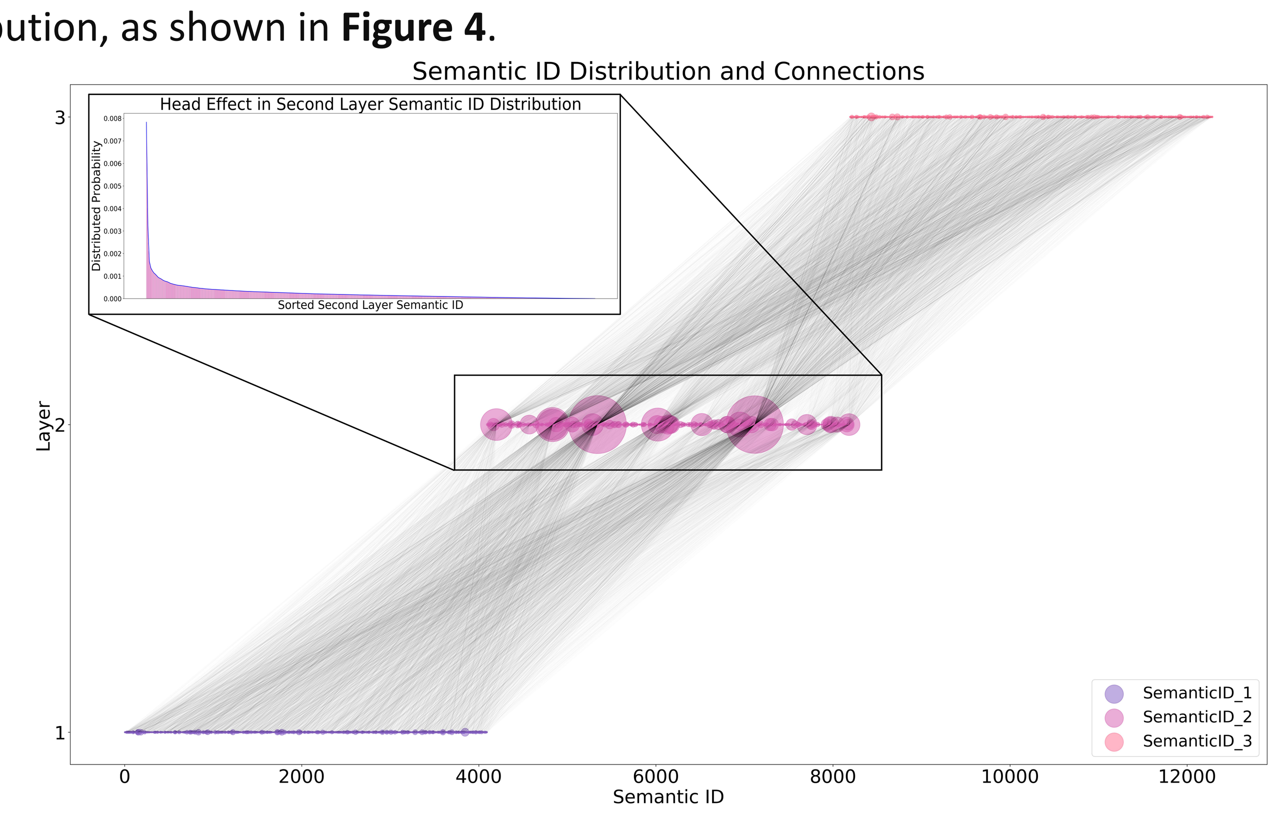

Figure 2: Distribution and Connections of Semantic IDs

2. Analysis of Residual Quantization

The "Hourglass" effect in RQ-based SIDs stems from RQ's hierarchical clustering. As shown in **Figure 3**, the first layer distributes tokens fairly uniformly, but in the second layer, residuals cluster around central points, creating a long-tail distribution with most paths concentrated into a few main routes. Although the third layer broadens distribution, the second layer's bottleneck restricts path diversity and reduces retrieval capacity.

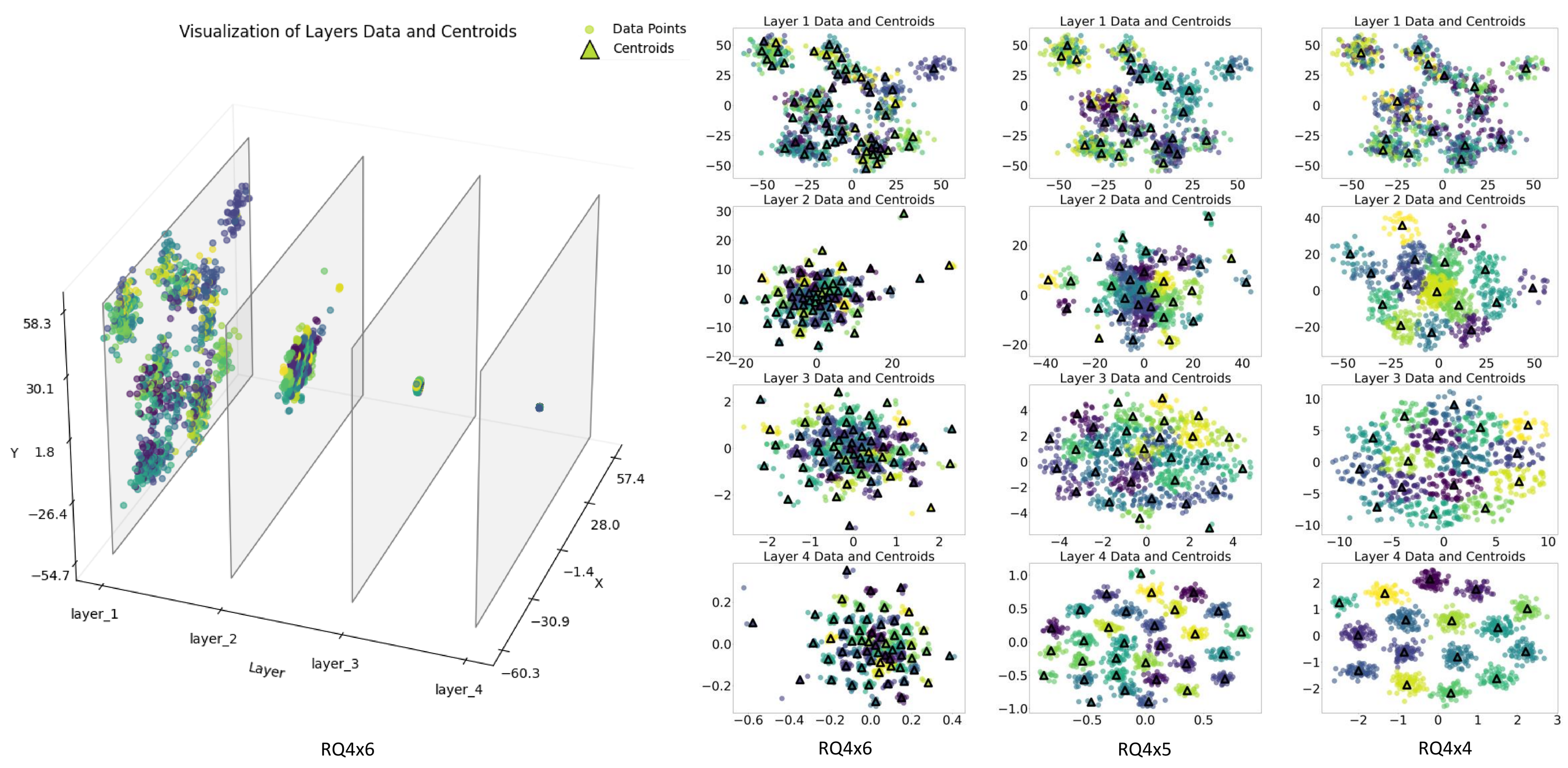

Figure 3: Hierarchical Residual Reduction and Dimensional Analysis Across Layers

3. Impact on Generative Retrieval Performance

Path sparsity and the long-tail distribution in the second layer limit retrieval performance, particularly in large datasets. Table 1 illustrates that high-frequency "head" tokens consistently outperform low-frequency "tail" tokens, highlighting an imbalance across retrieval paths. This issue persists across different model scales and configurations, underscoring the "Hourglass" effect's broad impact on retrieval efficiency and accuracy.

Table 1: The performance of generative retrieval on E-commerce datasets with RQ3x12, i.e., $L = 3, M = 2^{12}$. The head/tail token denotes the head/tail semantic ID in the second layer, respectively.

Method	Recall@1	Recall@3	Recall@5	Recall@10	Recall@30	Recall@50
LLaMA2-0.8B*	0.2480	0.4080	0.4990	0.590	0.7080	0.7480
Head Token	0.3617	0.5745	0.6894	0.7745	0.8894	0.9191
Tail Token	0.2131	0.3569	0.4405	0.5333	0.6523	0.6954
Qwen1.5-7B	0.2770	0.4720	0.5700	0.6600	0.7700	0.7930
Head Token	0.3450	0.5970	0.7040	0.8020	0.8960	0.9120
Tail Token	0.2470	0.4160	0.5100	0.5950	0.7190	0.7470
Baichuan2-7B	0.2730	0.4900	0.5900	0.6760	0.7670	0.8040
Head Token	0.3440	0.6000	0.7200	0.8140	0.9020	0. 9210
Tail Token	0.2480	0.4360	0.5250	0.6110	0.7180	0.7540
Given Layer 1*	0.340	0.497	0.567	0.632	0.722	0.756
Exchange Layer 1&2*	0.2390	0.4190	0.5100	0.6070	0.7150	0.7540
+ Given Layer 1*	**0.6600**	**0.8240**	**0.8650**	**0.8910**	**0.9160**	**0.9190**

* These experiments are based on the LLaMA2-0.8B model, which adopts the LLaMA2 structure and SFT on Chinese corpora.

METHODs AND EXPERIMENTs

1. Heuristic Approach

To address the hourglass effect, we propose **removing the second layer** in the RQ-based SID structure to reduce path concentration. This approach mitigates long-tail effects but may limit spatial capacity due to fewer token paths.

2. Adaptive Variable-Length Token Strategy

This method **selectively removes top tokens in the second layer, allowing token lengths to vary based on distribution needs.** By focusing on high-impact tokens, it preserves spatial capacity and better balances the token distribution.

3. Experiments

We tested these methods **on a large-scale e-commerce dataset**. As shown in Table 2, both methods reduce the hourglass effect, with **the adaptive variable-length strategy showing the highest recall improvement across metrics**. Additionally, both **methods lower the rate of invalid SIDs(Figure 4)**, especially in low-recall scenarios, confirming their practical effectiveness.

Table 2: The performance of generative retrieval on E-commerce based on RQ3x12.

Method	Recall@1	Recall@3	Recall@5	Recall@10	Recall@30	Recall@50
LLaMA2-0.8B	0.2480	0.4080	0.4990	0.590	0.7080	0.7480
Focal Loss (Lin et al., 2017)	0.2310	0.4270	0.5050	0.6110	0.7300	0.7640
Mile Loss (Su et al., 2024)	0.2590	0.4380	0.5110	0.6090	0.7250	0.7600
Remove 2-th layer	0.3090	0.4310	0.4970	0.5640	0.6580	0.7020
Remove 2-th layer top@20	0.2500	0.4270	0.5130	0.6120	0.7250	0.7580
Remove 2-th layer top@200	0.3190	0.4740	0.5600	0.6550	0.7450	0.7760
Remove 2-th layer top@400	**0.3340**	<u>0.5070</u>	**0.5950**	**0.6800**	**0.7760**	<u>0.7990</u>
Remove 2-th layer top@600	<u>0.3320</u>	**0.5080**	<u>0.5850</u>	<u>0.6720</u>	<u>0.7700</u>	**0.8010**

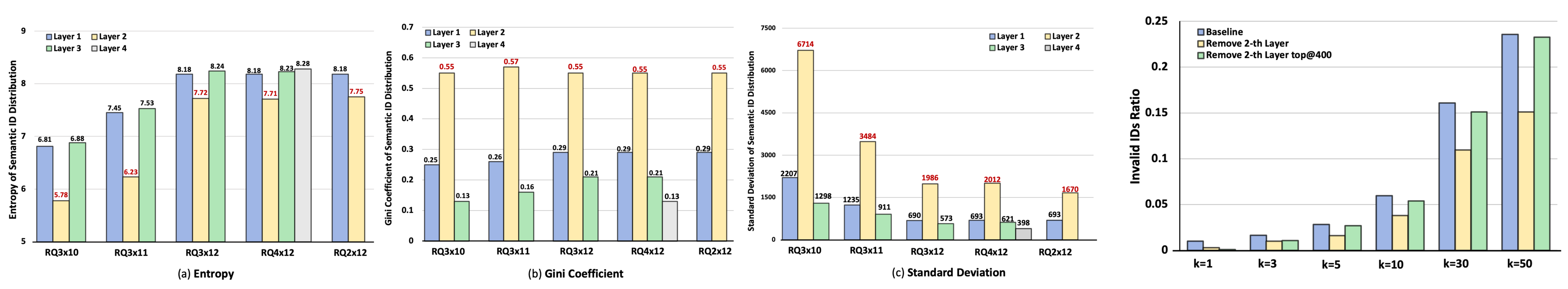

Figure 4: Statistical analysis of the second layer and Invalid IDs Ratio

CONCLUSION

This study systematically investigates the "Hourglass" phenomenon in RQ-based Semantic ID generation for generative retrieval, identifying path sparsity and long-tail distribution as key limitations. To address these issues, we propose two solutions: a heuristic layer-removal approach and an adaptive variable-length token strategy. Experimental results demonstrate that both methods improve retrieval performance by increasing codebook utilization and balancing token distribution, with the adaptive strategy providing the most significant gains. This work establishes a foundation for future optimization of generative retrieval systems in complex, large-scale applications.

REFERENCEs, ACKNOWLEDGEMENTs AND CONTACTs

[1] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2024. Recommender systems with generative retrieval. Advances in Neural Information Processing Systems, 36.

[2] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2024. Learning to tokenize for generative retrieval. Advances in Neural Information Processing Systems, 36.

Acknowledgements: This work has been supported by the National Key R&D Program of China under grant No. 2022YFF0902500, the National Natural Science Foundation of China under grant No.62472447, the Science and Technology Innovation Program of Hunan Province under grant No.2023RC1023, the Hunan Provincial Natural Science Foundation of China under grant No.2024JK2006, and Shenzhen Startup Funding No. QD2023014C.

Contacts: Zhirui Kuai: kuaizhirui@csu.edu.cn; Zuxu Chen: chen-zx22@mails.tsinghua.edu.cn; Yuxing Han: yuxinghan@sz.tsinghua.edu.cn; Huimu Wang: whm199416@gmail.com; Mingming Li: liemingming@outlook.com;