Project Link

Paper Link

NEURAL INFORMATION PROCESSING SYSTEMS

Rethinking Semantic ID Construction for Generative Recommendation: SimHash with Parallel Decoding and Semantic Alignment

Yuqing Liu[1*], Huiyuan Chen[2*†], Yibo Wang[1], Wooseong Yang[1], Philip S. Yu[1]

[1]University of Illinois Chicago [2]Amazon

[*]Equal contribution [†]Work done prior to joining Amazon

NeurIPS 2026

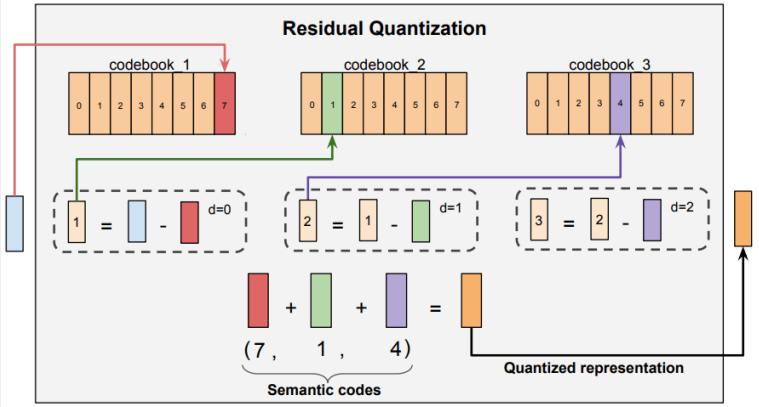
UIC UNIVERSITY OF ILLINOIS CHICAGO

amazon

Tokenization Methods for Semantic IDs

RQ-VAE

Residual quantization



✓ Hierarchical semantic structure
✗ Trained tokenizer
✗ Autoregressive decoding, short SIDs

OPQ

Product quantization

c_1 c_2 c_3 c_4 c_5 c_6

✓ Parallel decoding, longer SIDs
✗ Optimized tokenizer

SimHash

Locality-sensitive hashing

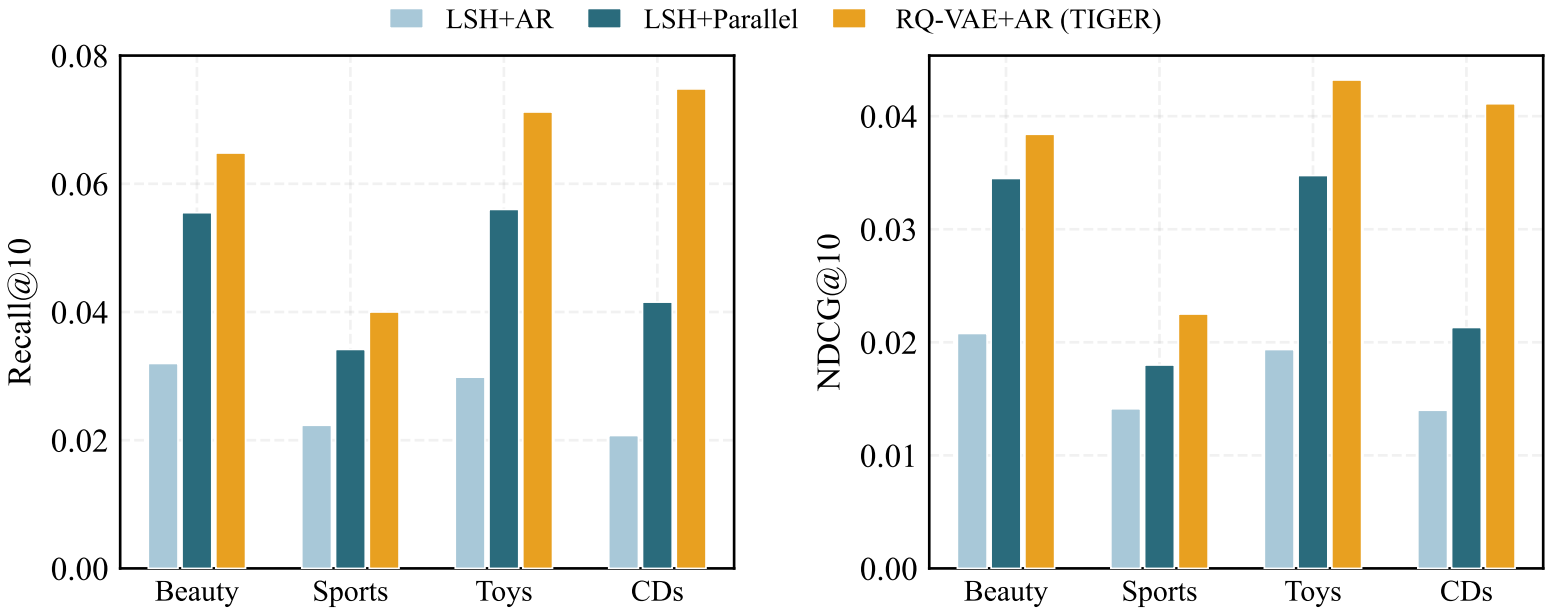
c_1 c_2 c_3 c_4 c_5 c_6

✓ Training-free tokenizer
? Regarded as inferior

Figure adapted from TIGER (Rajput et al., NeurIPS 2023).

Motivation

Is SimHash really a bad tokenizer? No.



Preliminary study (Figure 1 in paper): Recall@10 and NDCG@10 on four Amazon datasets.

LSH + AR → LSH + parallel

A large gain from changing only the decoder.

LSH + Parallel vs. TIGER

A performance gap still remains.

Design Principles

1 Structural Compatibility

The decoding paradigm should match the dependency structure induced by the tokenizer.

✗ **AR Decoding**

$c_1 \rightarrow c_2 \rightarrow c_3$

Sequential generation

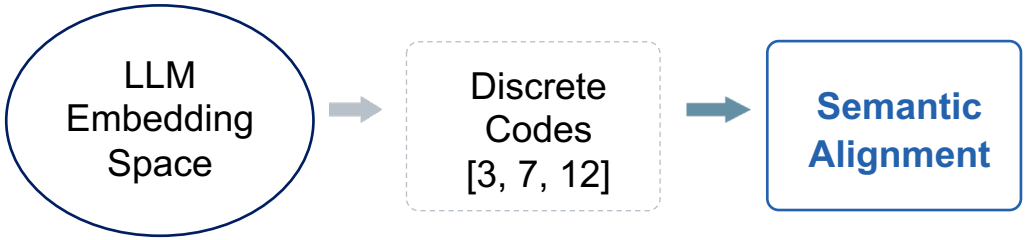
✓ **Parallel Decoding**

c_1, c_2, c_3

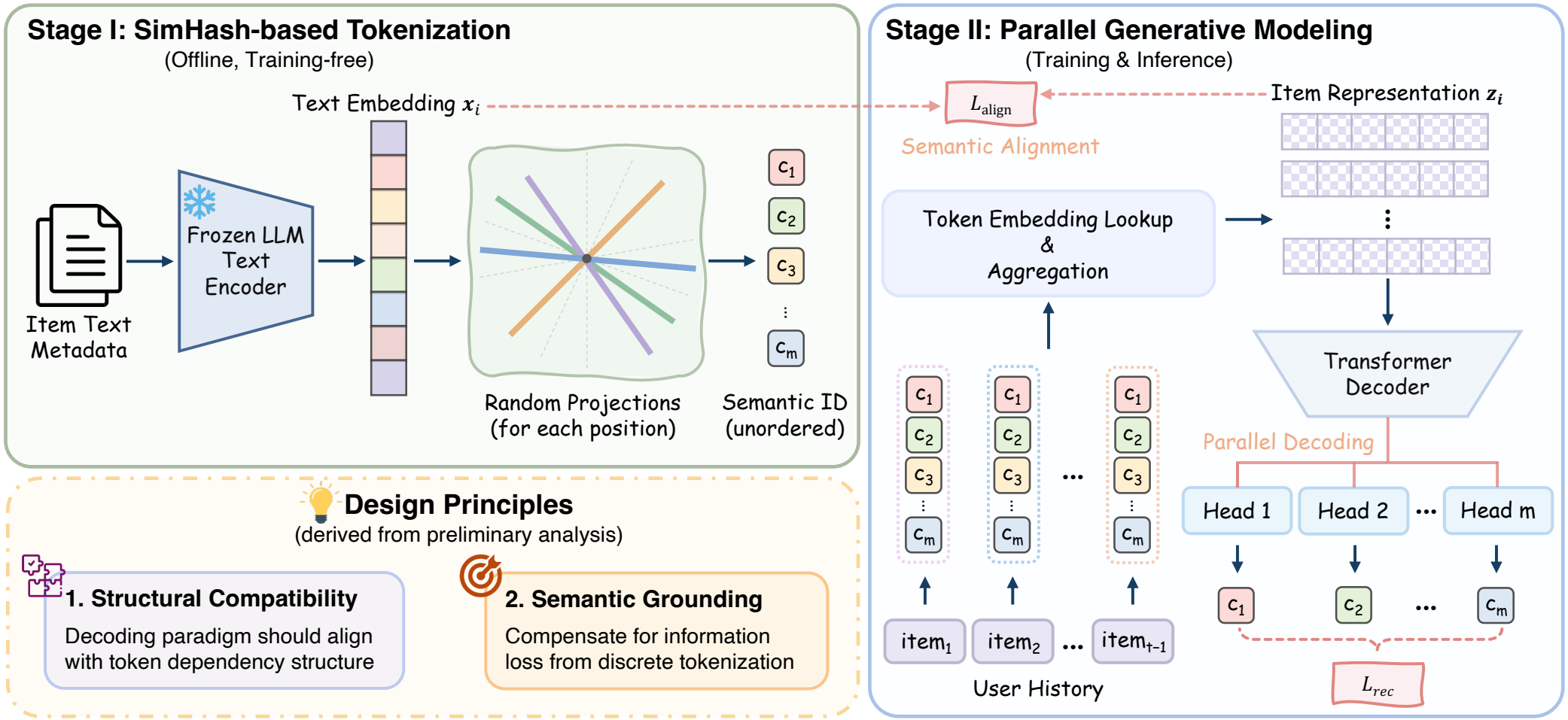
All at once!

2 Semantic Grounding

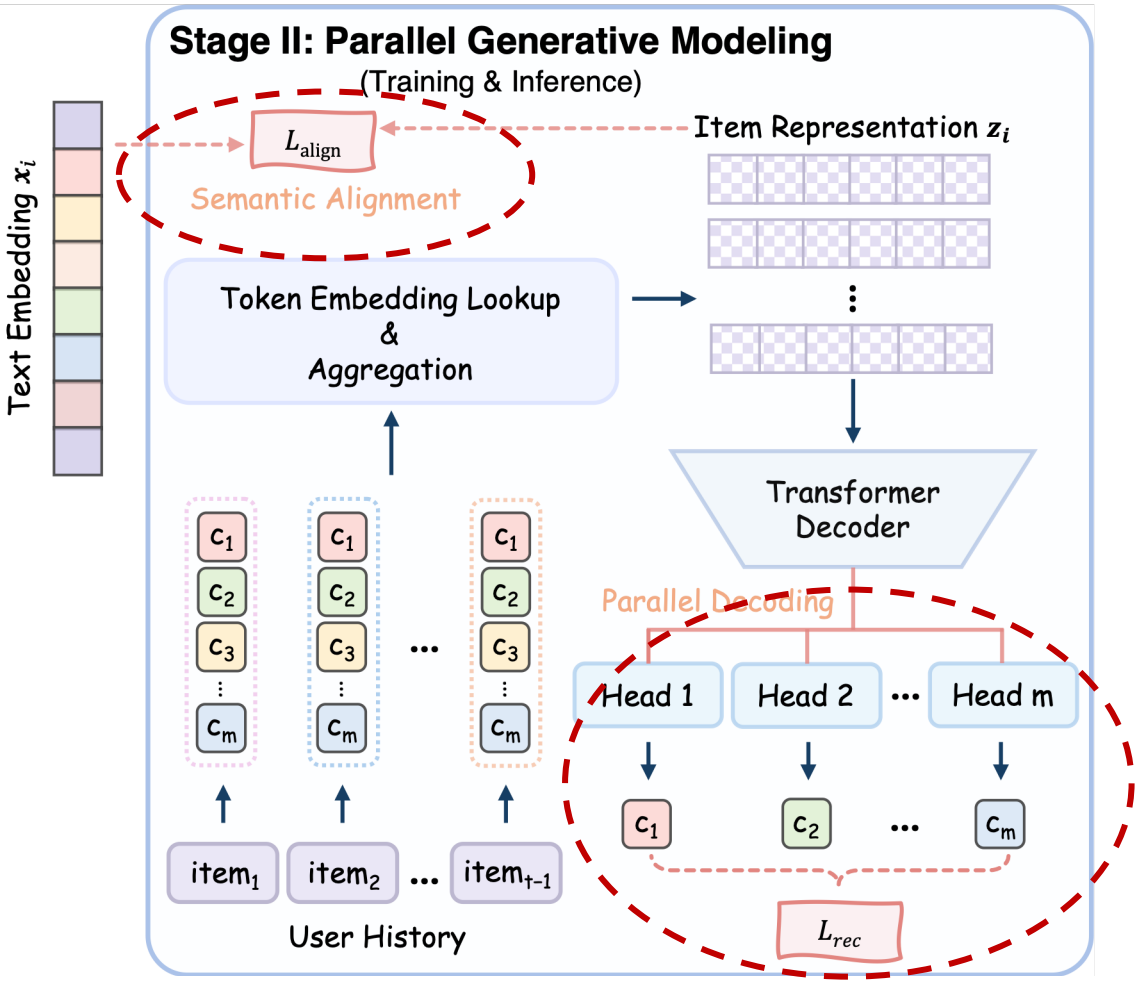
Compensate for fine-grained semantic information lost in discrete tokenization.



FLASH Framework



Stage II: Parallel Generative Modeling



Parallel Prediction

$$P((c_1^{(t)}, \ldots, c_m^{(t)}) \mid \mathbf{h}_{t-1}) = \prod_{j=1}^{m} p_j(c_j^{(t)} \mid \mathbf{h}_{t-1})$$

Semantic Alignment

$$\mathcal{L}_{\text{align}} = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \cos(\mathbf{z}_i,\ \phi(\mathbf{x}_i))$$

Training Objective

$$\mathcal{L}_{\text{rec}} = -\frac{1}{m} \sum_{j=1}^{m} \log p_j(c_j^{(t)} \mid \mathbf{h}_{t-1})$$

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}$$

Overall Performance

Model	Beauty				Toys and Games				Sports and Outdoors				CDs and Vinyl			
	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
RQ-based Semantic ID																
TIGER	0.0454	0.0321	0.0648	0.0384	0.0521	0.0371	0.0712	0.0432	0.0264	0.0181	0.0400	0.0225	0.0492	0.0329	0.0748	0.0411
LETTER	0.0489	0.0364	0.0689	0.0402	0.0544	0.0379	0.0727	0.0437	0.0293	0.0187	0.0420	0.0232	0.0481	0.0314	0.0723	0.0398
DIGER	0.0503	0.0370	0.0692	0.0421	0.0540	0.0383	0.0719	0.0442	0.0284	0.0192	0.0388	0.0241	0.0470	0.0312	0.0717	0.0369
PQ-based Semantic ID																
RecJPQ	0.0311	0.0167	0.0482	0.0222	0.0331	0.0182	0.0484	0.0231	0.0141	0.0076	0.0220	0.0102	0.0075	0.0046	0.0138	0.0066
VQ-Rec	0.0457	0.0317	0.0664	0.0383	0.0497	0.0346	0.0737	0.0423	0.0208	0.0144	0.0300	0.0173	0.0352	0.0238	0.0520	0.0292
RPG	0.0550	0.0381	0.0809	0.0464	0.0592	0.0401	0.0869	0.0490	0.0314	0.0216	0.0463	0.0263	0.0498	0.0338	0.0735	0.0415
LSH-based Semantic ID																
FLASH	**0.0576**	**0.0402**	**0.0832**	**0.0489**	**0.0650**	**0.0459**	**0.0914**	**0.0544**	**0.0337**	**0.0237**	**0.0487**	**0.0285**	**0.0536**	**0.0370**	**0.0769**	**0.0446**
Improv.	*+4.73%*	*+5.51%*	*+2.84%*	*+5.39%*	*+9.80%*	*+14.46%*	*+5.18%*	*+11.02%*	*+7.32%*	*+9.72%*	*+5.18%*	*+8.37%*	*+7.63%*	*+9.47%*	*+2.81%*	*+7.47%*

Table 1 in paper

- FLASH consistently outperforms evaluated baselines across all four datasets and metrics.
- Training-free tokenization achieves strong performance **when paired with parallel decoding and semantic alignment.**

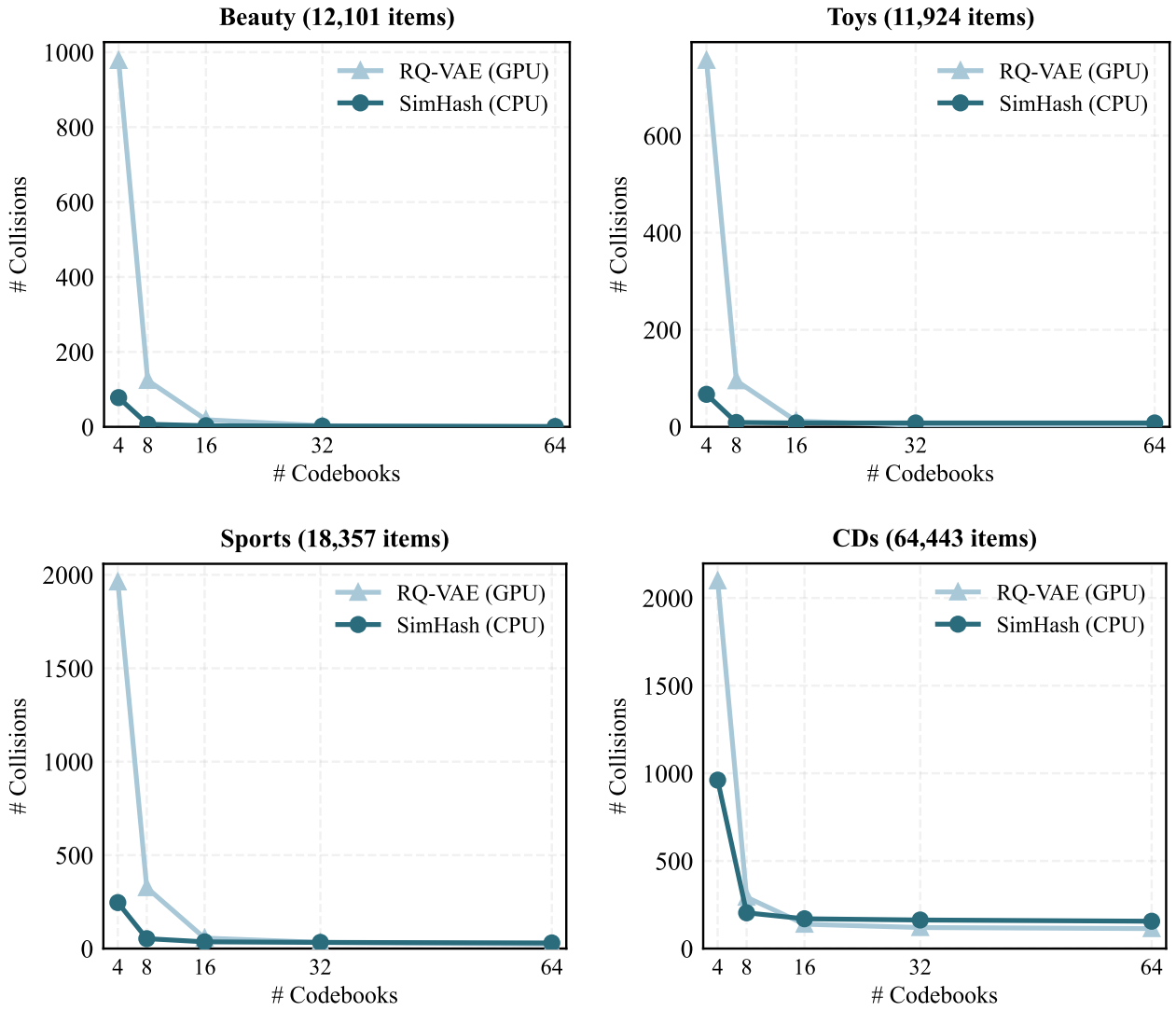
Ablation · Structural Compatibility

Tokenizer	Beauty			Toys			Sports			CDs		
	R@10	N@10	T (s) ↓	R@10	N@10	T (s) ↓	R@10	N@10	T (s) ↓	R@10	N@10	T (s) ↓
Random	0.0432	0.0261	0.0177	0.0612	0.0324	0.0144	0.0361	0.0198	0.0220	0.0197	0.0098	0.0353
RQ-VAE†	0.0832	0.0483	67.08	0.0856	0.0481	77.41	0.0466	0.0278	73.65	0.0751	0.0438	147.26
RQ-KMeans†	0.0812	0.0462	8.19	0.0861	0.0490	12.07	0.0457	0.0263	20.90	0.0760	0.0441	132.87
OPQ	**0.0851**	**0.0498**	292.89	0.0905	0.0523	111.08	0.0460	0.0265	151.90	**0.0782**	**0.0455**	162.98
SimHash	0.0832	0.0489	0.0556	**0.0914**	**0.0544**	0.0869	**0.0487**	**0.0285**	0.1340	0.0769	0.0446	0.4608

Table 2 in paper: unified parallel decoding and semantic alignment · T(s): tokenization time in seconds · †: GPU; others: CPU

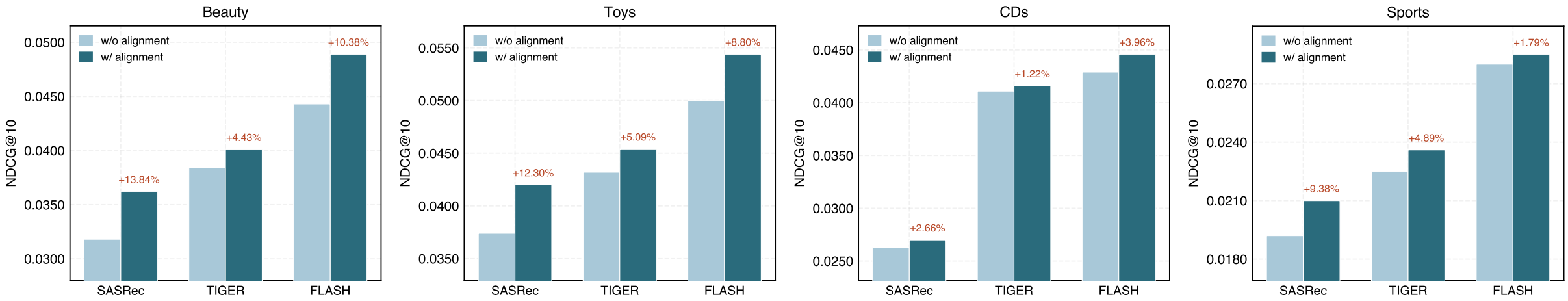
- **Structural compatibility alone is insufficient** without meaningful semantic information**.**
- **SimHash** achieves competitive recommendation performance **with substantially lower tokenization costs**.

Tokenization Quality



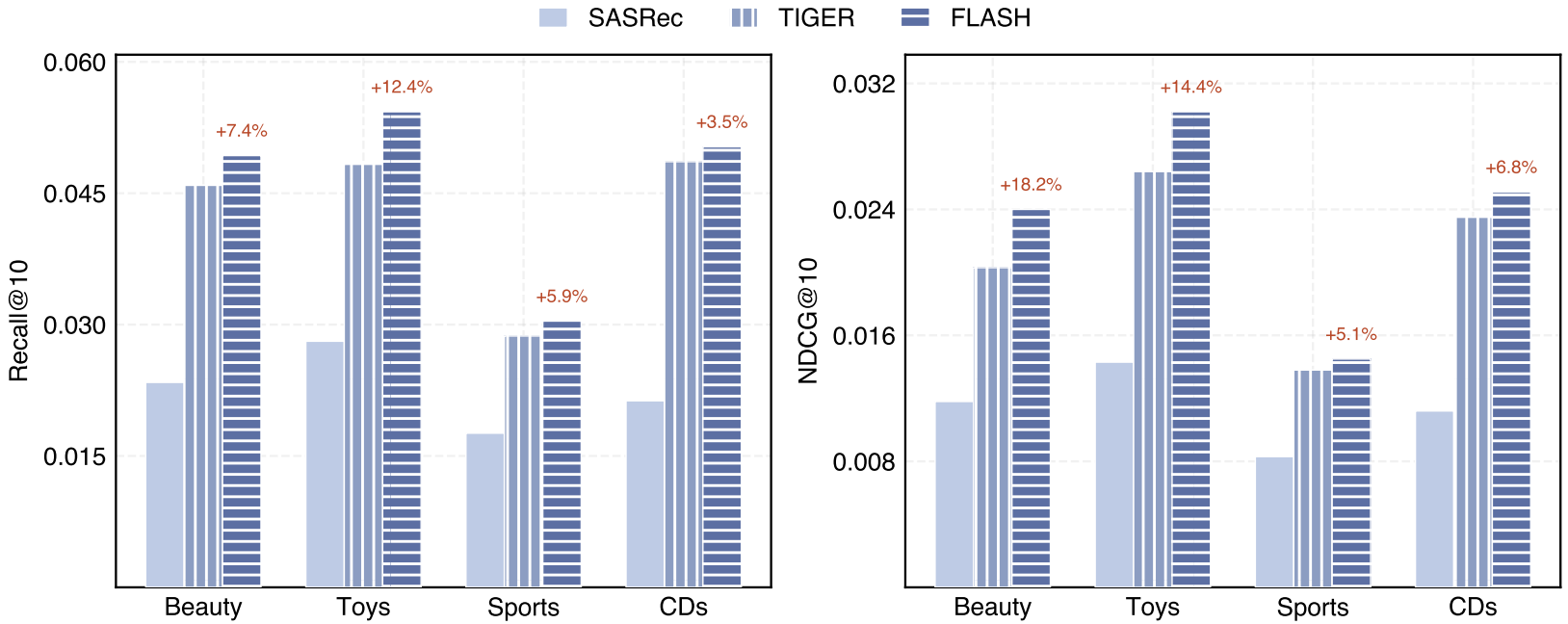
- SimHash consistently maintains **low collision rates**, which further **decrease with longer semantic IDs**.
- SimHash combines **simplicity and efficiency** while maintaining **code distinguishability.**

Ablation · Semantic Grounding



- **Explicit semantic alignment helps retain fine-grained semantic information lost** during discrete tokenization.
- Semantic alignment serves as **a general and effective mechanism** for improving recommendation performance across different architectures and item representation paradigms.

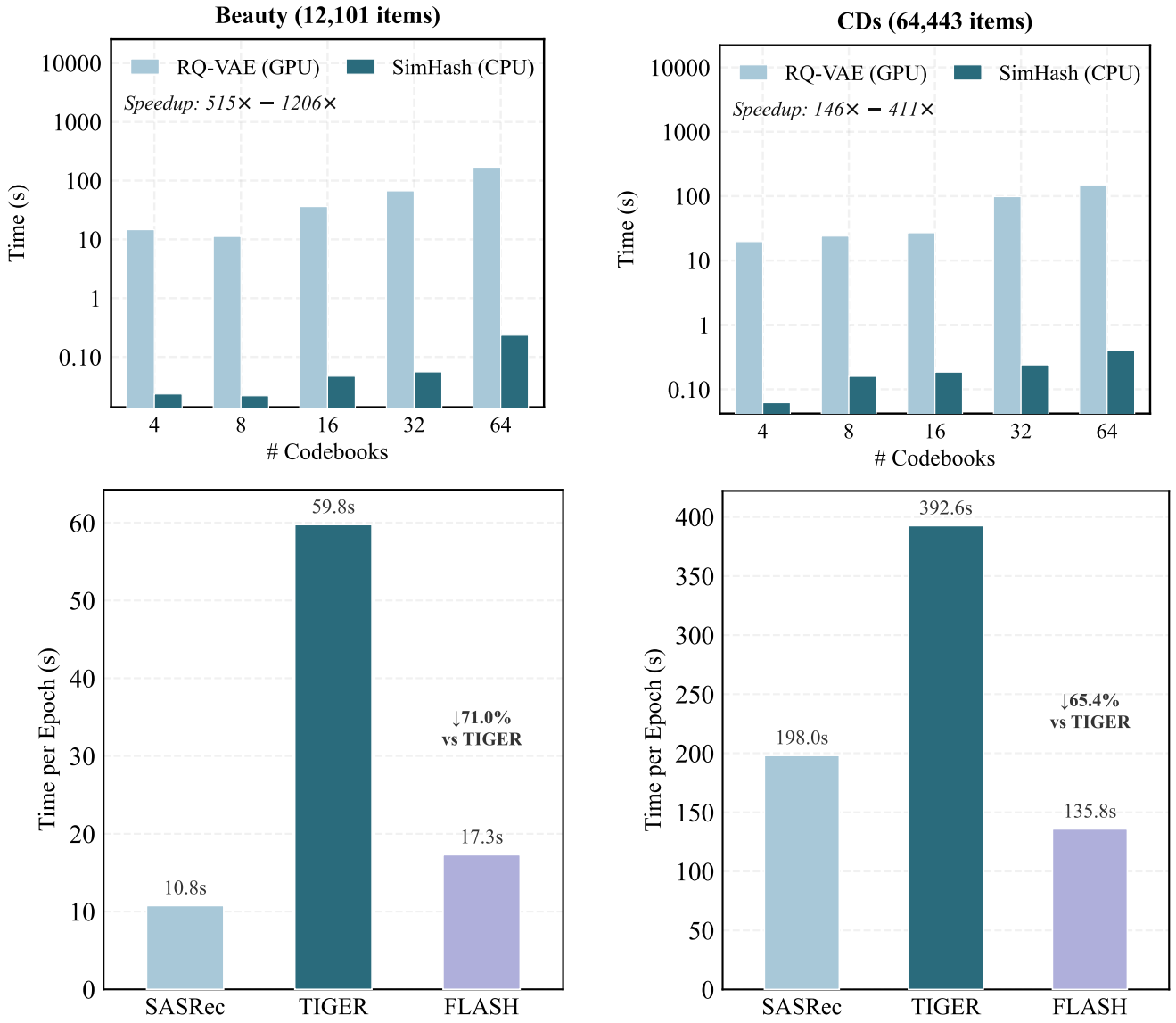
Cold-Start Generalization



Setting: 5% of target items are held out and unseen during training (Figure 6 in paper).

- Semantic ID-based modeling outperforms conventional item ID-based modeling on unseen items.
- **Semantic alignment complements SimHash's coarse-grained semantic similarity with fine-grained semantic information, supporting robust cold-start recommendations.**

Two-Stage Efficiency



Top · Stage I: SID construction time

- Training-free tokenization substantially reduces semantic ID construction costs.

Bottom · Stage II: training time per epoch

- Parallel decoding reduces recommendation-model training time while enabling longer semantic IDs.

Project Link

Paper Link

NEURAL INFORMATION PROCESSING SYSTEMS

TAKEAWAY

Semantic ID construction and generative modeling should be designed together

Structural Compatibility + **Semantic Grounding** → **Effective recommendation**

Simple, principled designs can rival or even outperform more complex alternatives.