Unleashing the Native Recommendation Potential: LLM-Based Generative Recommendation via Structured Term Identifiers

Zhiyang Zhang, Junda She, Kuo Cai, Bo Chen, Shiyao Wang, Xinchen Luo, Qiang Luo*, Ruiming Tang*, Han Li, Kun Gai, Guorui Zhou
[1]Kuaishou Inc., Beijing, China
{zhangzhiyang06,luoqiang,tangruiming}@kuaishou.com

Abstract

Leveraging the vast open-world knowledge and understanding capabilities of Large Language Models (LLMs) to develop general-purpose, semantically-aware recommender systems has emerged as a pivotal research direction in generative recommendation. However, existing methods face bottlenecks in constructing item identifiers. Text-based methods introduce LLMs' vast output space, leading to hallucination, while methods based on Semantic IDs (SIDs) encounter a semantic gap between SIDs and LLMs' native vocabulary, requiring costly vocabulary expansion and alignment training. To address this, this paper introduces **Term IDs (TIDs)**, defined as a set of semantically rich and standardized textual keywords, to serve as robust item identifiers. We propose **GRLM**[1], a novel framework centered on TIDs, employs Context-aware Term Generation to convert item's metadata into standardized TIDs and utilizes Integrative Instruction Fine-tuning to collaboratively optimize term internalization and sequential recommendation. Additionally, Elastic Identifier Grounding is designed for robust item mapping. Extensive experiments on real-world datasets demonstrate that GRLM significantly outperforms baselines across multiple scenarios, pointing a promising direction for generalizable and high-performance generative recommendation systems.

1 Introduction

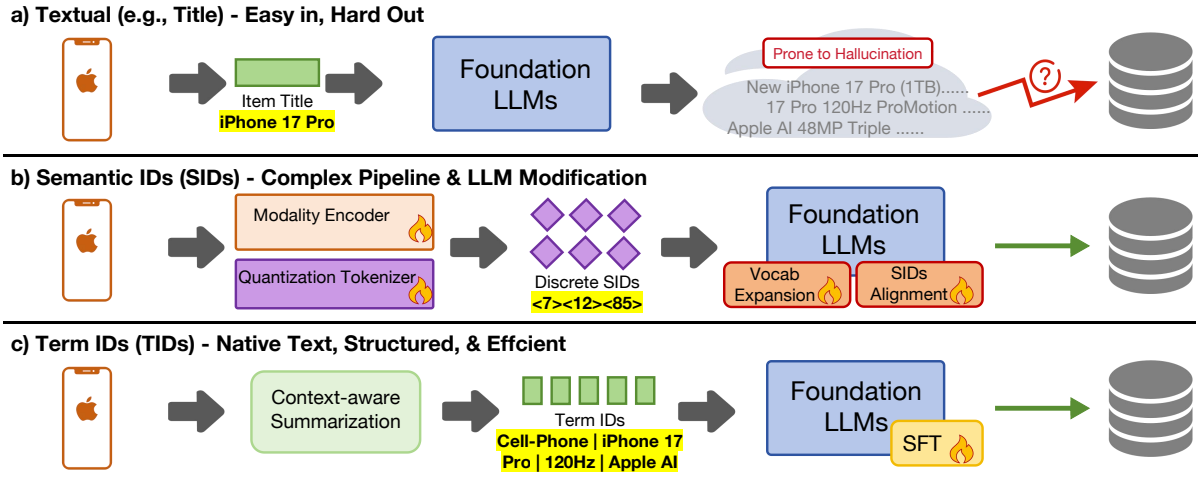
Large Language Models (LLMs) have demonstrated exceptional capabilities in complex semantic understanding and generation (Achiam et al., 2023), leading to paradigm shifts across various fields. In the realm of recommender systems (Radford et al., 2019; Brown, 2020), a novel Generative Recommendation (GR) paradigm has emerged (Wang et al., 2023; Deldjoo et al., 2024), transforming the traditional multi-stage recall-and-ranking pipeline (Covington et al., 2016; Qin et al., 2022) into a generative task that auto-regressively generates item identifiers based on user historical behaviors, which has gained widespread application (Zhai et al., 2024; Deng et al., 2025; Han et al., 2025; Huang et al., 2025). When pretrained LLMs are employed as the backbone of GR to strengthen open-world knowledge and reasoning capabilities, a fundamental challenge lies in **Item Identifiers**: enabling LLMs to represent continuously emerging and diverse items within their parameter space, while reliably grounding generated tokens to valid real-world items during generation.

As illustrated in Figure 1, existing methods primarily follow two paths. The first utilizes *Textual Identifiers* (e.g., titles or brief descriptions) (Chu et al., 2023; Liu et al., 2025a; Tan et al., 2024). While these leverage the LLM's native vocabulary to depict items, they suffer from two major flaws. Raw titles frequently lack sufficient discriminative information, while full descriptions are too lengthy for efficient sequence modeling, leading to semantic instability. Moreover, due to the vast output space, LLMs are prone to hallucinations, resulting in the generation of non-existent or corrupted items.

The second path involves *Semantic IDs (SIDs)*, which quantize (Lee et al., 2022; Yin et al., 2013) item embeddings into discrete codes (Rajput et al., 2023; Zheng et al., 2024; Yang et al., 2025). Although earlier architectures showed promise, they often function as specialized models rather than general-purpose LLMs. When attempting to integrate this paradigm with LLMs, the standard practice involves extending the LLM's native vocabulary with additional tokens specifically for SIDs (Liu et al., 2025c; Zhou et al., 2025a). Consequently, they encounter a semantic gap: these numerical codes are absent from the LLM's pretrained vocabulary, failing to tap into the model's

*Corresponding authors
[1]https://github.com/ZY0025/GRLM

Findings of the Association for Computational Linguistics: ACL 2026, pages 19659–19673
July 2-7, 2026 ©2026 Association for Computational Linguistics



Item Identifier	Native to LLM	Semantic Rich	Hallucination	Cross-domain
Textual (e.g.Title)	●	● (Low)	● (High)	●
Semantic IDs	●	● (High)	● (Low)	●
Term IDs (Ours)	●	● (High)	● (Low)	●

Figure 1: Comparison of Item Identifiers: Term IDs (TIDs) leverage standardized and structured text tokens to ensure precise semantic extraction and low hallucination, while maintaining native compatibility with LLMs vocabularies without the complex indexing pipelines or architectural modifications required by Semantic IDs (SIDs).

latent world knowledge. This necessitates costly vocabulary expansion and intensive alignment training. Furthermore, since these numerical codes are inherently devoid of universal semantics, they are often domain-specific, which severely hinders the model's effectiveness in cross-domain recommendation.

To mitigate the limitations of existing item identifiers, we investigate a novel item identifier approach built upon the native vocabulary of LLMs, which requires addressing several non-trivial challenges. First, the redundancy of natural language demands effective extraction of compact yet informative terms to represent items. Second, synonym ambiguity across different lexical realizations can lead to grounding conflicts and hallucinations. Third, item identifiers should preserve semantic similarity among related items while maintaining sufficient discriminability.

To overcome these challenges, we propose **G**enerative **R**ecommendation **L**anguage **M**odel (**GRLM**), a unified framework integrated with **Term IDs (TIDs)**, a structured item identifier that exploits the semantic understanding capabilities of LLMs to encode items as keyword sequences with high information density. The *Context-aware Term Generation (CTG)* process is designed to reduce grounding ambiguity and hallucination, while ensuring distinguishability among similar items. By leveraging neighborhood-based in-context learning, the metadata of similar items is retrieved as guidance, enabling the LLMs to extract consistent terms for related items while capturing fine-grained discriminative features. Consequently, TIDs based on the native LLMs vocabulary provide generality, semantic richness, and robust grounding with minimal hallucination. The comparison with other item identifier approaches is presented in Figure 1.

Based on the Term IDs, GRLM further introduces an *Integrative Instruction Fine-tuning (IIFT)* training paradigm, which jointly fine-tunes LLMs on an item-to-TIDs identification task and a personalized recommendation task. This multi-task optimization enhances the LLM's understanding of domain knowledge while strengthening its ability to capture personalized preferences, thereby enabling accurate recommendations. Moreover, to achieve reliable item grounding during inference, a dual-level *Elastic Identifier Grounding (EIG)* mechanism is introduced. By exploiting the decompositional nature of TIDs, EIG combines direct mapping with structural mapping to ground TIDs into actual items efficiently.

Our contributions can be summarized as follows:

- We propose Term IDs, a structured item identifier derived from the native LLMs vocabulary, offering generality, semantically rich, and robust grounding with minimal hallucination.

- We propose a LLM-based general framework GRLM, which integrates the item identification task and personalized recommendation task via Integrative Instruction Fine-tuning with an Elastic Identifier Grounding mechanism to ensure reliable item grounding.
- Extensive experiments demonstrate that GRLM achieves state-of-the-art performance across multiple benchmarks. Moreover, through detailed analytical experiments, we provide evidence that our TID-based method exhibits robust scaling properties while significantly mitigating hallucination.

2 Related Works

2.1 SID-based Generative Recommendation

TIGER (Rajput et al., 2023) was the pioneering work to propose the generative recommendation framework. It represents items using SIDs and employs a Transformer-based (Vaswani et al., 2017) architecture to directly generate the SID of the target item. Subsequent models, such as LETTER (Wang et al., 2024a), EAGER (Wang et al., 2024b), and UNGER (Xiao et al., 2025), have further incorporated collaborative signals into the generation process. In specific domains, models like OneLoc (Wei et al., 2025) and GNPR-SID (Wang et al., 2025) have integrated geographic information into item embeddings. OneRec (Deng et al., 2025; Zhou et al., 2025b) attempts to construct an end-to-end recommendation pipeline by applying generative recommendation to large-scale industrial systems.

Despite these advancements, traditional generative methods rely on SIDs, which lack inherent semantic interpretability and fail to fully leverage the extensive open-world knowledge and reasoning capabilities of LLMs. Although OneRec-Think (Liu et al., 2025c) addresses this by proposing a unified framework to align the SID space with the natural language space, such methods typically necessitate expensive vocabulary expansion and computationally intensive alignment training.

2.2 LLM-Based Recommendation

Another line of research leverages the semantically rich of LLMs by representing items through textual descriptors (Bao et al., 2025; Chen et al., 2024). In this paradigm, recommendation evolves from simple pattern grounding to deep reasoning based on world knowledge. This transition not only mitigates cold-start and long-tail distribution challenges inherent in traditional collaborative filtering but also empowers systems with superior cross-domain generalization and zero-shot capabilities.

Early efforts like TallRec (Bao et al., 2023) demonstrated the effectiveness of lightweight Instruction Fine-tuning, using item titles to align LLMs with recommendation tasks via specialized prompts. LLMTreeRec (Zhang et al., 2025) organizes item attributes into a tree structure, constraining the quantization and generation process within a hierarchical framework. InteraRec (Karra and Tulabandhula, 2024) addresses text-deficient items by encoding visual information into structured textual summaries, providing a novel perspective for item tokenization. LLaRa (Liao et al., 2024) incorporates collaborative signals by concatenating item sequence features from traditional recommenders with inherent textual attributes.

Although progress has been made, the reliability of text-based generative recommendation remains a major concern, necessitating the use of constrained decoding and fixed candidate sets. This reliance highlights hallucination, which prevents such models from achieving true generative recommendation. In contrast, our Term IDs provide a robust solution that substantially mitigates these hallucinations.

3 GRLM

Figure 2 illustrates the overall framework of GRLM, which employs a three-stage pipeline centered on Term IDs to unleash the native recommendation capabilities of LLMs. First, we utilize **Context-aware Term Generation** to convert item's metadata into standardized TIDs; Second, **Integrative Instruction Fine-tuning** is employed to adapt the LLMs through a multi-task learning paradigm, enabling it to internalize item semantics while simultaneously modeling user behavioral patterns. Finally, **Elastic Identifier Grounding** implements a hybrid grounding mechanism that seamlessly integrates strict direct mapping with a structural mapping, thereby ensuring robust item retrieval during the inference phase.

3.1 Context-aware Term Generation

Standardizing item identifiers in a generative space faces a dual challenge: maintaining term consistency across similar items to prevent identifier fragmentation, while preserving discriminative uniqueness to avoid semantic collisions between distinct

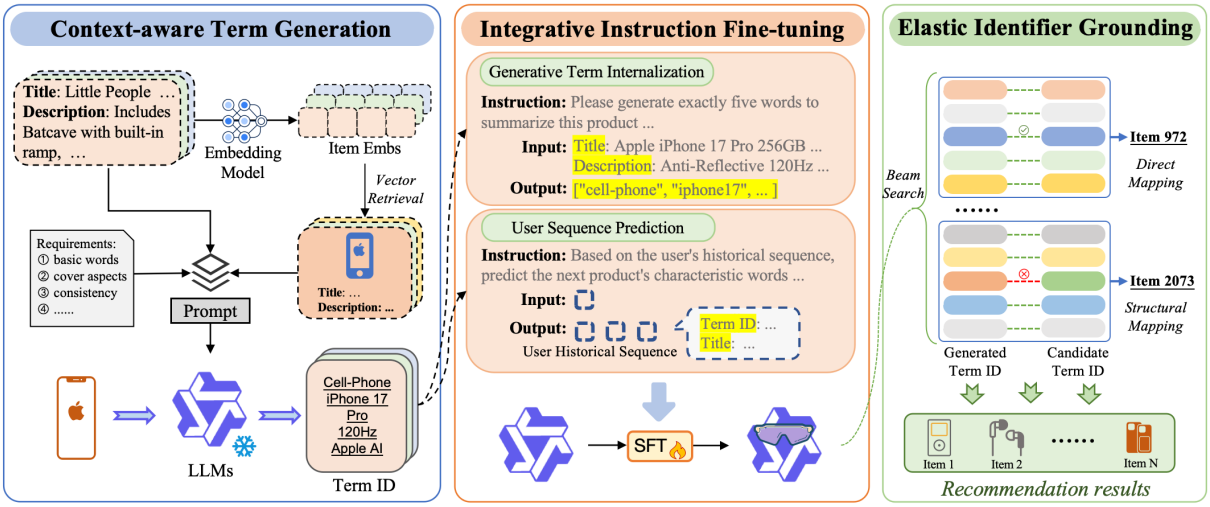


Figure 2: Overall framework of GRLM.

products. Processing each item in isolation often fails to resolve synonymy or capture fine-grained differences.

As illustrated in Figure 3, independent generation results in inconsistent labeling for identical features, such as "Cell-Phone" versus "Mobile-Phone," while simultaneously failing to distinguish between different models by assigning the generic "iPhone" to both.

To address this, we propose **Context-aware Term Generation (CTG)** aims to convert item's metadata into standardized, human-readable Term IDs by incorporating similar neighborhoods as contextual guidance. For each item $i \in \mathcal{I}$, we first aggregate its metadata m_i and encode it into a dense vector $\mathbf{v}_i \in \mathbb{R}^d$ using a pre-trained embedding model[2].

Next, we calculate the cosine similarity between the target item i and all other items j in the candidate library. We retrieve the top-k most similar items to form the context set $\mathcal{N}_i = \{j_1, j_2, \ldots, j_k\}$. We then construct a structured prompt $\mathcal{P}$ that integrates the metadata of item i and its neighbors $\{m_j\}_{j \in \mathcal{N}_i}$. The design philosophy of our prompt emphasizes a balance between globally consistency and locally discriminative: it instructs the LLMs to adopt standardized terms for shared attributes among neighbors while intentionally selecting keywords that capture the item's distinctive features, detailed prompts are provided in Appendix A Figure 5. The process is formulated as:

$$T_i = \text{LLM}(\mathcal{P}(m_i, \{m_j\}_{j \in \mathcal{N}_i})), \quad (1)$$

[2]https://huggingface.co/Qwen/Qwen3-Embedding-8B

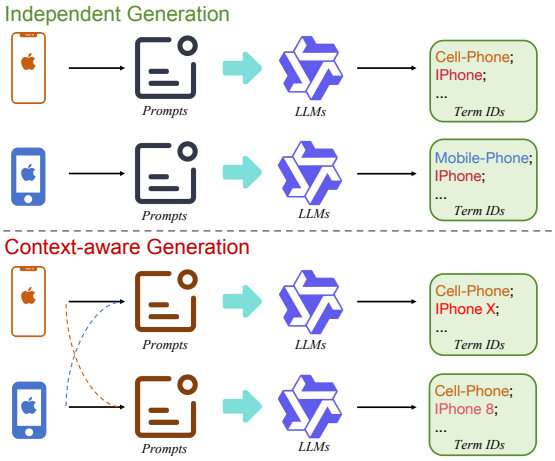


Figure 3: Context-aware Term Generation effectively ensures that Term IDs across items are globally consistent and locally discriminative.

By leveraging the neighborhood context, CTG enables the LLMs to recognize commonalities and maintain term consistency across similar products, effectively mitigating identifier fragmentation caused by synonyms. Simultaneously, by prompting the model to contrast an item with its neighbors, CTG enhances the capture of fine-grained discriminative information, which helps distinguish items within the same category and reduces potential semantic overlaps.

3.2 Integrative Instruction Fine-tuning

Traditional method often focuses solely on sequence prediction, which may cause the model to lose the semantic grounding of the identifiers. We propose **Integrative Instruction Fine-tuning (IIFT)**, which enable the model to collaboratively

optimize Generative Term Internalization (GTI) and User Behavior Sequence Prediction.

Generative Term Internalization This task requires the model to directly generate the standardized T_i from item's metadata m_i without the neighborhood context used during the generation phase. The core motivation is to encourage the LLMs to internalize the underlying generation logic, effectively compressing its expansive natural language output space into a more focused and "distilled" semantic region. By learning to map item's metadata to standardized TIDs, the model establishes an intrinsic understanding of the identifier's structure.

User Behavior Sequence Prediction As the core recommendation task, this phase models the evolution of user interests using historical interaction sequences. For a user sequence $S = \{i_1, i_2, \ldots, i_n\}$, where n denotes the total number of items interacted with by the user, we first represent each item i_j as a joint textual sequence x_j by concatenating its Term IDs and raw title:

$$x_j = [T_{i_j} \ ; \ m_{i_j}^{title}]. \tag{2}$$

The model is then trained in an autoregressive manner to predict each subsequent item in the history. Specifically, we designate the first item x_1 as the input context, while the subsequent textual sequence $\{x_2, \ldots, x_n\}$ are concatenated to form the output. The model is trained to minimize the negative log-likelihood of the output tokens, conditioned on the initial instruction $\mathcal{I}$:

$$\mathcal{L} = -\sum_{k=2}^{n} \log P(x_k \mid \mathcal{I}(x_1, ..., x_{k-1})). \tag{3}$$

This formulation ensures that the model learns to reconstruct the entire behavioral trajectory by focusing its predictive capacity on the tokens following the anchor item x_1.

Notably, during the inference phase, the model is only required to generate the Term IDs $T_{i_{t+1}}$ for the next item, which can be mapping to the specific item, rather than the full x_{t+1} containing the title, thereby ensuring high inference efficiency.

3.3 Elastic Identifier Grounding

Unlike traditional generative methods that treat item identifiers as numerical indices or text strings, Term IDs represents items as structured sets of standardized semantic terms. To fully harness this architectural advantage, we design Elastic Identifier Grounding (EIG), a dual-level retrieval mechanism that mapping the generated sequence to the recommend item:

1) Direct Mapping: We first attempt an exact string-level mapping within the candidate library $\mathcal{C}$. This track ensures maximum precision when the model's output perfectly aligns with the standardized TIDs of a specific item.

2) Structural Mapping: If no exact mapping is found, EIG leverages the decompositional nature of TIDs to perform structural mapping. We resolve the final identifier by identifying the item i^* that maximizes the structural score:

$$s^* = \arg\max_{i \in \mathcal{C}} \sum_{j=1}^{N} w_j \cdot \mathbb{I}(t_{gen}^j = t_i^j), \tag{4}$$

where t_{gen}^j is the j-th term in the generated sequence, and $w_j = \frac{1}{j+1}$ is a decay weight.

4 Experiment Setup

To comprehensively evaluate the capability and generalization of GRLM, we conduct experiments under the two most common scenarios in recommendation systems: in-domain and cross-domain.

Baseline. To verify the effectiveness of GRLM, we compare it against a diverse set of competitive baselines, categorized as follows:

- Sequential Methods: We include SASRec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019), which employ self-attention mechanisms to model user behavior sequences.
- Generative Recommenders: We compare against TIGER (Rajput et al., 2023) and HSTU (Zhai et al., 2024), representing generative recommendation methods.
- LLM-based Recommenders: We include OneRec-Think (Liu et al., 2025c) and IDGenRec (Tan et al., 2024), which are representative methods that leverage LLMs for recommendation tasks.
- Recommenders for Cross-Domain: For cross-domain scenarios, we additionally include TriCDR (Ma et al., 2024), LLM4CDSR (Liu et al., 2025b), and GenCDR (Hu et al., 2025), which are specifically designed to transfer knowledge across domains.

Dataset. For the in-domain scenario, we select three real-world recommendation datasets from the popular Amazon product review dataset[3]: *Beauty*,

[3]https://jmcauley.ucsd.edu/data/amazon/

Table 1: Statistics of the datasets used in our experiments.

Dataset	In-domain			Cross-domain			
	Beauty	Sports	Toys	Sports	Clothing	Phones	Electronics
#User	22,363	35,598	19,412	35,598	39,387	27,879	192,403
#Item	12,101	18,357	11,924	18,357	23,033	10,429	63,001

Sports, and *Toys*. Following the settings in (Rajput et al., 2023; Liu et al., 2025c) for data preprocessing, we perform 5-core filtering and use leave-one-out strategy to split datasets.

For the cross-domain scenario, we conduct experiments on two dataset pairs: *Sports-Clothing* (Leisure) and *Phones-Electronics* (Technology). First, each individual dataset is processed following the same filtering strategy as the in-domain scenario. Then, adhering to the protocols in (Hu et al., 2025), we construct a unified cross-domain sequence for each overlapping user by merging interactions from both datasets and sorting them chronologically. During the evaluation phase, we predict the last interaction for each dataset in the pair separately. This setup allows the model to capture dynamic preference transfer across domains within a single coherent context. Table 1 shows the dataset statistics.

Implementation Details. In our experiment, the length of the Term IDs for all items is set to 5. For GRLM's backbone, we select the newest Qwen3-4B-2507[4] version as our term generation model and recommendation model, with model parameters frozen during CTG and fully fine-tuned during IIFT. Beam search strategy is used during the model's inference process. Considering the varying lengths of Term IDs for different items, the model's generation max-length is set to 30 to accommodate all Term IDs. Top-k Recall and NDCG with K=5 and 10 are used as metrics, following (Rajput et al., 2023; Liu et al., 2025c). Training details are in Appendix A.

5 Experiment Result

To comprehensively evaluate the effectiveness and robustness of GRLM, our experiments are designed to address the following research questions:

- **RQ1:** How does GRLM perform compared to state-of-the-art baselines in both in-domain and cross-domain recommendation scenarios?
- **RQ2:** What is the contribution of core components, such as Context-aware Term Generation and Integrative Instruction Fine-tuning, to the overall performance?
- **RQ3:** Does GRLM exhibit positive scaling properties consistent with the scaling laws of Large Language Models?
- **RQ4:** Can the proposed Term IDs effectively mitigate the hallucination inherent in text-based generative recommendation?

5.1 Overall Performance (RQ1)

The results in both in-domain and cross-domain scenarios are shown in Table 2 and Table 3. In the in-domain scenario, GRLM based on Term IDs achieved optimal performance across all evaluation metrics on the three datasets, significantly surpassing all baseline methods. Specifically, on the Beauty, Sports, and Toys datasets, GRLM achieved relative improvements of 7.8%, 30.2%, and 14.9% in Recall@5 compared to the strongest baseline model. This strongly demonstrates the effectiveness of our proposed Term IDs-based recommendation paradigm in unleashing the native recommendation capabilities of LLMs.

A key observation in Table 3 is GRLM's remarkable performance in cross-domain scenarios, with Recall@K improvements exceeding 50% on average. Notably, GRLM achieves this without any specific architecture or auxiliary alignment modules for cross-domain required by methods like TriCDR or GenCDR. This stems from the textual nature of TIDs: by mapping items to a universal semantic space rather than domain-constrained IDs, GRLM leverages the LLM's pre-trained world knowledge to facilitate seamless knowledge transfer. The shared vocabulary of natural language act as a "semantic bridge," allowing the model to recognize functional similarities (e.g., from "Phones" to "Electronics").

5.2 Ablation Study (RQ2)

We conduct an ablation study on the in-domain scenario, comparing two configurations: 1) **w/o CTG**: generated Term IDs without using similar items as context. 2) **w/o GTI**: fine-tuning LLMs without the GTI task. The results in Table 4 validate the efficacy of each component in GRLM. Removing CTG (w/o CTG) leads to a noticeable performance decline, confirming that leveraging neighborhood information as a contextual reference is crucial for generating locally discriminative and globally consistent TIDs. Without such context, the generation

[4]https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

Table 2: Overall performance comparison on different datasets (In-domain). The best and second-best results are highlighted in **bold** font and underlined.

Dataset	Metric	Sequential methods		Generative		LLM-Based		**GRLM**	Improvement
		SASRec	BERT4Rec	TIGER	HSTU	IDGenRec	OneRec-Think		
Beauty	Recall@5	0.0402	0.0232	0.0405	0.0424	0.0484	0.0563	**0.0607**	7.82%
	Recall@10	0.0607	0.0396	0.0623	0.0652	0.0693	0.0791	**0.0846**	6.95%
	NDCG@5	0.0254	0.0146	0.0267	0.0280	0.0337	0.0398	**0.0430**	8.04%
	NDCG@10	0.0320	0.0199	0.0337	0.0353	0.0404	0.0471	**0.0506**	7.43%
Sports	Recall@5	0.0199	0.0102	0.0215	0.0268	0.0270	0.0288	**0.0375**	30.21%
	Recall@10	0.0301	0.0175	0.0347	0.0343	0.0388	0.0412	**0.0539**	30.83%
	NDCG@5	0.0106	0.0065	0.0137	0.0173	0.0185	0.0199	**0.0260**	30.65%
	NDCG@10	0.0141	0.0088	0.0179	0.0226	0.0223	0.0239	**0.0313**	30.96%
Toys	Recall@5	0.0448	0.0215	0.0337	0.0366	0.0595	0.0579	**0.0684**	14.96%
	Recall@10	0.0626	0.0332	0.0547	0.0566	0.0800	0.0797	**0.0942**	17.75%
	NDCG@5	0.0300	0.0131	0.0209	0.0245	0.0432	0.0412	**0.0477**	10.42%
	NDCG@10	0.0358	0.0168	0.0276	0.0309	0.0498	0.0482	**0.0561**	12.65%

Table 3: Overall performance comparison on different datasets (Cross-domain). The best and second-best results are highlighted in **bold** font and underlined.

Scenario	Dataset	Metric	Sequential methods		Generative		Cross-Domain			**GRLM**	Improvement
			SASRec	BERT4Rec	TIGER	HSTU	TriCDR	LLM4CDSR	GenCDR		
Leisure	Sports	Recall@5	0.0188	0.0197	0.0267	0.0254	0.0266	0.0263	0.0274	**0.0480**	75.18%
		Recall@10	0.0325	0.0334	0.0397	0.0381	0.0396	0.0398	0.0403	**0.0645**	60.05%
		NDCG@5	0.0121	0.0126	0.0244	0.0241	0.0255	0.0257	0.0261	**0.0353**	35.25%
		NDCG@10	0.0169	0.0173	0.0287	0.0277	0.0259	0.0260	0.0262	**0.0406**	41.46%
	Clothing	Recall@5	0.0128	0.0132	0.0173	0.0175	0.0174	0.0176	0.0181	**0.0288**	59.12%
		Recall@10	0.0219	0.0227	0.0241	0.0253	0.0258	0.0261	0.0265	**0.0436**	64.53%
		NDCG@5	0.0078	0.0081	0.0125	0.0132	0.0161	0.0163	0.0167	**0.0190**	13.77%
		NDCG@10	0.0105	0.0108	0.0167	0.0174	0.0194	0.0196	0.0203	**0.0238**	17.24%
Technology	Phones	Recall@5	0.0331	0.0345	0.0423	0.0415	0.0434	0.0431	0.0436	**0.0928**	112.84%
		Recall@10	0.0524	0.0537	0.0613	0.0615	0.0593	0.0614	0.0621	**0.1172**	88.73%
		NDCG@5	0.0215	0.0224	0.0315	0.0327	0.0396	0.0401	0.0411	**0.0739**	79.81%
		NDCG@10	0.0278	0.0287	0.0406	0.0425	0.0505	0.0506	0.0512	**0.0818**	59.77%
	Electronics	Recall@5	0.0179	0.0186	0.0228	0.0232	0.0238	0.0237	0.0241	**0.0377**	56.43%
		Recall@10	0.0276	0.0285	0.0322	0.0328	0.0339	0.0338	0.0342	**0.0529**	54.68%
		NDCG@5	0.0118	0.0122	0.0214	0.0226	0.0231	0.0230	0.0235	**0.0268**	14.04%
		NDCG@10	0.0149	0.0154	0.0269	0.0271	0.0280	0.0279	0.0283	**0.0317**	12.01%

Table 4: Ablation study analyzing the contribution of CTG and GTI on in-domain datasets.

Dataset	Metric	w/o CTG	w/o GTI	**GRLM**
Beauty	Recall@5	0.0576	0.0564	**0.0607**
	Recall@10	0.0810	0.0830	**0.0846**
	NDCG@5	0.0402	0.0398	**0.0430**
	NDCG@10	0.0477	0.0483	**0.0506**
Sports	Recall@5	0.0346	0.0361	**0.0375**
	Recall@10	0.0495	0.0531	**0.0539**
	NDCG@5	0.0233	0.0250	**0.0260**
	NDCG@10	0.0281	0.0305	**0.0313**
Toys	Recall@5	0.0637	0.0653	**0.0684**
	Recall@10	0.0857	0.0889	**0.0942**
	NDCG@5	0.0458	0.0463	**0.0477**
	NDCG@10	0.0529	0.0539	**0.0561**

process tends to produce generic or fragmented terms; CTG refines this "semantic tokenization," providing the recommender with more precise and high-quality identifier.

The performance drop in the w/o GTI variant highlights the importance of the Generative Term Internalization task. This task training it to map complex metadata into a constrained TIDs space. The synergy between term internalization and sequential recommendation ensures that the model not only learns accurate item representations but also maintains reliable transitions within the semantic space.

5.3 Scaling Law (RQ3)

We further investigated the impact of model parameter size on the performance of GRLM to verify its Scaling Law. To ensure consistency, we continue to use the Qwen3-4B-2507 version as our term generation model, but employ the more diverse Qwen3-2504[5] version as the backbone for fine-tuning, with model sizes covering five configurations: 0.6b, 1.7b, 4b, 8b, and 14b. The overall

[5]https://huggingface.co/Qwen/Qwen3-4B

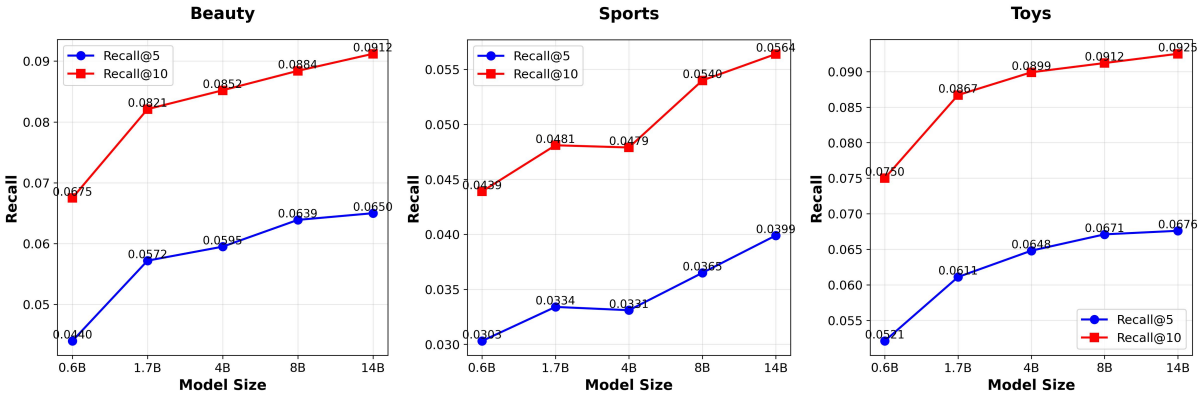


Figure 4: Performance scaling of GRLM with respect to different backbone model sizes (from 0.6B to 14B) on three in-domain datasets.

Table 5: Analysis of hallucination rates of Term IDs generated by GRLM across in-domain and cross-domain datasets. We use valid rate and direct hit rate as metrics.

Dataset	In-domain			Out-domain	
	Beauty	Sports	Toys	Leisure	Technology
#VR@10	0.996	0.989	0.995	0.998	0.998
#DHR@10	0.997	0.999	0.998	0.999	1

trend of the Recall on the three in-domain datasets is shown in Figure 4. We found that recommendation performance steadily improves with the increase in the number of parameters of the foundational LLMs, consistent with the Scaling Law in the LLMs field. This is an exciting discovery, indicates that TIDs allow larger LLMs to better utilize their superior semantic reasoning and open-world knowledge, promising further gains with future model iterations.

5.4 Hallucination of Term IDs (RQ4)

Although TIDs reside in the LLM's expansive generative space, the GRLM framework effectively constrains the output to a specific semantic subspace. To quantify this, we evaluate two metrics: Valid Rate (VR@K), the proportion of generated identifiers belonging to the candidate library, and Direct Hit Rate (DHR@K), the proportion of successful retrievals handled by the Direct Mapping track within EIG (i.e., exact direct mapping before invoking structural mapping).

As shown in Table 5, the model trained with GRLM exhibits exceptional stability, with both VR@10 and DHR@10 consistently exceeding 99% across all datasets. These results indicate that GRLM successfully internalizes the structural and semantic constraints of TIDs. By learning to navigate this predefined subspace, the model significantly mitigates the hallucination bottleneck inherent in traditional text-based generative methods, achieving high grounding precision without sacrificing generative flexibility.

6 Conclusion

In this paper, we introduced GRLM, a LLM-based generative recommendation framework that utilizes Term IDs to unleash the native recommendation capabilities of LLMs. By representing items as structured TIDs within the LLM's parameter space, GRLM eliminates the need for vocabulary expansion and complex cross-modal alignment. Through Context-aware Term Generation, Integrative Instruction Fine-tuning, and Elastic Identifier Grounding, GRLM achieves state-of-the-art performance across both in-domain and cross-domain scenarios. Our analysis demonstrates that GRLM not only mitigates hallucination with near-perfect reliability but also exhibits strong scaling properties and cross-domain transferability. By utilizing the LLM's native vocabulary, GRLM eliminates the need for costly vocabulary expansion and alignment. Our work providing a promising and generalizable direction for high-performance generative recommendation systems.

7 Limitations

The following are some limitations that GRLM may face: (1) Our Context-aware Term Generation (CTG) currently utilizes a fixed external embedding model to retrieve item context. While effective, we expect to further enhance the discriminative power

of Term IDs by exploring more advanced domain-specific retrieval methods in future studies. (2) Due to computational resource constraints, we primarily validated GRLM on Qwen3. Given its robust performance, we anticipate that larger base models will further unlock the framework's potential, which we will investigate in our future research to improve universal recommendation capabilities.

8 Ethical Statements

Our study do not carry any ethical concerns. Specifically, Our training data are publicly available and designated for research purposes only. We inspect our dataset to ensure it does not contain any unethical content, private information and offensive topics. Moreover, the base models we used are also publicly available for research purpose.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yanchen Luo, Chong Chen, Fuli Feng, and Qi Tian. 2025. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems*, 3(4):1–27.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM conference on recommender systems*, pages 1007–1014.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024. On softmax direct preference optimization for recommendation. *Advances in Neural Information Processing Systems*, 37:27463–27489.

Zhixuan Chu, Hongyan Hao, Xin Ouyang, Simeng Wang, Yan Wang, Yue Shen, Jinjie Gu, Qing Cui, Longfei Li, Siqiao Xue, and 1 others. 2023. Leveraging large language models for pre-trained recommender systems. *arXiv preprint arXiv:2308.10837*.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198.

Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A review of modern recommender systems using generative models (gen-recsys). In *Proceedings of the 30th ACM SIGKDD conference on Knowledge Discovery and Data Mining*, pages 6448–6458.

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

Ruidong Han, Bin Yin, Shangyu Chen, He Jiang, Fei Jiang, Xiang Li, Chi Ma, Mincong Huang, Xiaoguang Li, Chunzhen Jing, and 1 others. 2025. Mtgr: Industrial-scale generative recommendation framework in meituan. *arXiv preprint arXiv:2505.18654*.

Peiyu Hu, Wayne Lu, and Jia Wang. 2025. From ids to semantics: A generative framework for cross-domain recommendation with adaptive semantic tokenization. *arXiv preprint arXiv:2511.08006*.

Yanhua Huang, Yuqi Chen, Xiong Cao, Rui Yang, Mingliang Qi, Yinghao Zhu, Qingchang Han, Yaowei Liu, Zhaoyu Liu, Xuefeng Yao, and 1 others. 2025. Towards large-scale generative ranking. *arXiv preprint arXiv:2505.04180*.

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

Saketh Reddy Karra and Theja Tulabandhula. 2024. Interarec: Interactive recommendations using multimodal large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 32–43. Springer.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11523–11532.

Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1785–1795.

Jingzhe Liu, Liam Collins, Jiliang Tang, Tong Zhao, Neil Shah, and Clark Mingxuan Ju. 2025a. Understanding generative recommendation with semantic ids from a model-scaling view. *arXiv preprint arXiv:2509.25522*.

Qidong Liu, Xiangyu Zhao, Yejing Wang, Zijian Zhang, Howard Zhong, Chong Chen, Xiang Li, Wei Huang, and Feng Tian. 2025b. Bridge the domains: Large

language models enhanced cross-domain sequential recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1582–1592.

Zhanyu Liu, Shiyao Wang, Xingmei Wang, Rongzhou Zhang, Jiaxin Deng, Honghui Bao, Jinghao Zhang, Wuchao Li, Pengfei Zheng, Xiangyu Wu, and 1 others. 2025c. Onerec-think: In-text reasoning for generative recommendation. *arXiv preprint arXiv:2510.11639*.

Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2024. Triple sequence learning for cross-domain recommendation. *ACM Transactions on Information Systems*, 42(4):1–29.

Jiarui Qin, Jiachen Zhu, Bo Chen, Zhirong Liu, Weiwen Liu, Ruiming Tang, Rui Zhang, Yong Yu, and Weinan Zhang. 2022. Rankflow: Joint optimization of multi-stage cascade ranking systems as flows. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 814–824.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, and 1 others. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315.

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

Juntao Tan, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Zelong Li, and Yongfeng Zhang. 2024. Idgenrec: Llm-recsys alignment with textual id learning. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 355–364.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dongsheng Wang, Yuxi Huang, Shen Gao, Yifan Wang, Chengrui Huang, and Shuo Shang. 2025. Generative next poi recommendation with semantic id. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 2904–2914.

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

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative recommendation: Towards next-generation recommender paradigm. *arXiv preprint arXiv:2304.03516*.

Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, and 1 others. 2024b. Eager: Two-stream generative recommender with behavior-semantic collaboration. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3245–3254.

Zhipeng Wei, Kuo Cai, Junda She, Jie Chen, Minghao Chen, Yang Zeng, Qiang Luo, Wencong Zeng, Ruiming Tang, Kun Gai, and 1 others. 2025. Oneloc: Geo-aware generative recommender systems for local life service. *arXiv preprint arXiv:2508.14646*.

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

Yuhao Yang, Zhi Ji, Zhaopeng Li, Yi Li, Zhonglin Mo, Yue Ding, Kai Chen, Zijian Zhang, Jie Li, Shuanglong Li, and 1 others. 2025. Sparse meets dense: Unified generative recommendations with cascaded sparse-dense representations. *arXiv preprint arXiv:2503.02453*.

Xuesong Yin, Songcan Chen, and Enliang Hu. 2013. Regularized soft k-means for discriminant analysis. *Neurocomputing*, 103:29–42.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, and 1 others. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152*.

Wenlin Zhang, Chuhan Wu, Xiangyang Li, Yuhao Wang, Kuicai Dong, Yichao Wang, Xinyi Dai, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2025. Llmtreerec: Unleashing the power of large language models for cold-start recommendations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 886–896.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1435–1448. IEEE.

Guorui Zhou, Honghui Bao, Jiaming Huang, Jiaxin Deng, Jinghao Zhang, Junda She, Kuo Cai, Lejian Ren, Lu Ren, Qiang Luo, and 1 others. 2025a. Openonerec technical report. *arXiv preprint arXiv:2512.24762*.

Guorui Zhou, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Shiyao Wang, and 1 others. 2025b. Onerec technical report. *arXiv preprint arXiv:2506.13695*.

A Experiment details

Prompt To extract and synthesize key data, a structured prompt was developed to guide the LLM in summarizing specific items, as shown in Figure 5. The prompt employs a schema-based approach, ensuring that information regarding item is condensed into a standardized format while maintaining factual density and categorical clarity.

Training Details We conduct full-parameter supervised fine-tuning (SFT) on the Qwen3 models. All model variants are trained with a next-token prediction objective. We employ a cosine annealing learning rate schedule with an initial learning rate of 1×10^{-4} and a global batch size of 128 over 3 epochs. Specifically, for the experiments in Section 3, the learning rates for the 0.6B, 1.7B, 4B, 8B, and 14B models are adjusted to 2×10^{-4}, 2×10^{-4}, 1×10^{-4}, 7×10^{-5}, and 5×10^{-5}, respectively.

B Length of Term IDs

While the sequence-based nature of natural language suggests that a sufficiently large Term IDs space is necessary to avoid potential item collisions, it remains unclear whether further increasing the sequence length yields diminishing returns. To explore the impact of extended identifiers, we conducted experiments using 7 and 10 Term IDs per item. The results across three in-domain datasets, as summarized in Table 8, indicate that performance with 7 Term IDs remains comparable to the baseline (5 Term IDs), while increasing the count to 10 even leads to a slight decline in effectiveness. We attribute this marginal decrease to the possibility that over-extending the term generation process introduces redundant noise, thereby obscuring the item's most salient features. Given that longer sequences also increase the inference latency, we empirically select a configuration of 5 Term IDs per item to maintain an optimal trade-off between recommendation quality and computational efficiency.

C Ablation Studies on Multi-task Configurations

To further justify the design choices of our IIFT training paradigm, we conduct ablation experiments comparing different task formulations. Specifically, we evaluate the following variants:

- **UBSP + Meta2TIDs (Ours):** Our proposed User Behavior Sequence Prediction (UBSP) that takes the first item as input and predicts all subsequent items in one pass, combined with the GTI task that maps item metadata to its Term IDs without neighborhood context.
- **UBSP + Meta2TIDs + TIDs2Meta**: Adding a symmetric TIDs-to-metadata alignment task.
- **NIP + Meta2TIDs**: Replacing UBSP with standard Next Item Prediction (NIP), which decomposes a length-n sequence into $n-1$ sub-sequences.
- **NIP + Meta2TIDs + TIDs2Meta**: NIP with bidirectional alignment.
- **CROSS + Meta2TIDs + TIDs2Meta**: Alternating between TIDs and titles as prediction targets, similar to the paradigm in LC-Rec (Zheng et al., 2024).

Table 9 reports the results on three in-domain datasets. Our full design (UBSP + Meta2TIDs) consistently outperforms all other configurations. Notably, adding the TIDs2Meta task brings marginal or no improvement, and the CROSS design (alternating TIDs/titles) underperforms significantly. Moreover, NIP-based variants require substantially more training steps due to the sequential decomposition, while our UBSP achieves both higher efficiency and better accuracy.

D Case Study

As shown in Figure 6, case studies across different datasets demonstrate the core advantages of our Term IDs over Semantic IDs and the Text ID generated by IDGenRec (Tan et al., 2024). Compared to existing project identifiers, item's Term IDs effectively capture and summarize its core semantic features (such as main category, function, characteristics, etc.) in natural language. This semantic representation, built directly from the LLM's native vocabulary, not only makes the recommendation process more transparent and understandable to humans but, more importantly, helps the model better utilize the world knowledge and semantic understanding capabilities acquired by the LLMs during pre-training. This enhances the model's ability to capture user sequential patterns and generalize. This semantic consistency is the fundamental reason why Term IDs exhibit strong generalization capabilities in both in-domain and cross-domain recommendation scenarios.

Table 6: 10 Amazon14 dataset statistics.

Dataset	Baby	Beauty	Cell	Grocery	Health	Home	Pet	Sports	Tools	Toys
Items	7,050	12,101	10429	8,713	18,534	28,237	8,510	18,357	10,217	11,924
Users	19,445	22,363	27,879	14,681	38,609	66,519	19,856	35,598	16,638	19,412
Interactions	160,792	198,502	194,439	151,254	346,355	551,682	157,836	296,337	134,476	167,597

Table 7: 10 Amazon14 dataset performance.

Dataset	Baby	Beauty	Cell	Grocery	Health	Home	Pet	Sports	Tools	Toys	Avg.
Recall@5	0.0317	0.0601	0.0656	0.0704	0.0527	0.0260	0.0554	0.0372	0.0408	0.0682	0.0463
Recall@10	0.0439	0.0854	0.0951	0.1018	0.0753	0.0376	0.0820	0.0530	0.0599	0.0940	0.0664
NDCG@5	0.0220	0.0424	0.0444	0.0472	0.0374	0.0182	0.0381	0.0257	0.0291	0.0481	0.0322
NDCG@10	0.0260	0.0505	0.0539	0.0573	0.0446	0.0220	0.0467	0.0308	0.0352	0.0565	0.0387

Table 8: Length of Term IDs

Dataset	Metric	5 IDs	7 IDs	10 IDs
Beauty	Recall@5	0.0607	0.0595	0.0571
	Recall@10	0.0846	0.0853	0.0752
	NDCG@5	0.0430	0.0429	0.0387
	NDCG@10	0.0506	0.0504	0.0427
Sports	Recall@5	0.0375	0.0384	0.0354
	Recall@10	0.0539	0.0544	0.0525
	NDCG@5	0.0260	0.0275	0.0249
	NDCG@10	0.0313	0.0333	0.0297
Beauty	Recall@5	0.0684	0.0682	0.0629
	Recall@10	0.0942	0.0951	0.0875
	NDCG@5	0.0477	0.0466	0.0428
	NDCG@10	0.0561	0.0557	0.0507

E Robustness and Data Scalability

To further investigate the practical potential of GRLM, we evaluate its performance under two challenging scenarios: large-scale joint training and semantic space compression. These experiments aim to verify whether the TID-based paradigm remains effective as the item catalog scales toward larger dimensions.

We scale the data volume by merging 10 diverse Amazon sub-datasets for joint training and evaluation (statistics and per-dataset results are detailed in Table 6 & 7). As shown in the "Avg." column of Table 8, GRLM achieves a stable Recall@10 of 0.0664 and NDCG@10 of 0.0387.

Compared to training on individual datasets (Table 2), GRLM demonstrates remarkable stability rather than the performance degradation often caused by increased item collisions. This robustness stems from the semantic grounding of TIDs: unlike numerical Semantic IDs, which may suffer from "id-space crowding" as data grows, TIDs utilize the shared natural language vocabulary across domains. For instance, common terms (e.g., "Portable," "Ergonomic") act as semantic bridges, allowing the LLMs to benefit from positive transfer across different categories, effectively increasing the training signal density for each term.

In real-world applications, the number of unique terms extracted via GRLM may grow significantly. To ensure a bounded output vocabulary, we propose a Semantic Compression strategy. We first project all extracted terms into a latent space using an embedding model and apply K-means clustering to identify K cluster centroids, which we define as "Core Terms". Each original term in a TIDs is then mapped to its nearest Core Term, generating a compact set of Compressed Term IDs that maintains semantic representativeness within a fixed vocabulary size.

As illustrated in Table 10, even when the term vocabulary is compressed to $K = 3,000$ (approx. $1/18$ of the original size), the performance decline remains negligible ($< 1\%$). This critical finding suggests that GRLM does not rely on memorizing specific, fine-grained strings. Instead, it captures the high-level semantic abstractions of items. The ability to maintain high performance with a compact, fixed-size vocabulary highlights the scalability of the TIDs paradigm for massive-scale item catalogs.

You are an expert product summarizer. Your task is to generate exactly FIVE words to summarize this product. Please follow ALL guidelines carefully:

GUIDELINES:
1. WORD FORM: All words must be in their base form (nouns or adjectives, no -ed, -ing, -s endings)
2. WORD ORDER: Order words by importance (most important aspect first)
3. CONTENT FOCUS: Focus on these aspects in order:
   a) Main product category/type (e.g., "doll", "puzzle", "car")
   b) Key function or purpose (e.g., "educational", "remote-control")
   c) Distinctive features (e.g., "wooden", "electronic", "collectible")
   d) Target audience (e.g., "toddler", "boys", "family")
   e) Unique selling point (e.g., "glow-in-dark", "interactive")
4. CONSISTENCY WITH SIMILAR ITEMS: Consider the similar items provided. If they share common characteristics, use consistent terminology for those aspects.
5. UNIQUENESS: Include at least 1-2 words that distinguish this product from the similar items. Each product should have some unique aspects.
6. OUTPUT FORMAT: Provide ONLY the five words in this exact format: [word1, word2, word3, word4, word5]
7. NO ADDITIONAL TEXT: Do not include any explanations, thoughts, or other content.

PRODUCT INFORMATION:

{Item title} (e.g.Avon Anew Clinical Eye Lift Pro Dual Eye System)
{Item description} (e.g. Dual eye system includes: UPPER EYE & BROW BONE GEL: instantly, eyes appear tighter and lifted. UNDER EYE CREAM: instantly, undereye shadows are visibly reduced.)

TOP 5 SIMILAR PRODUCTS (for reference):

{Similar Item title}
{Similar Item description}

× N

ANALYSIS GUIDANCE:
1. First, identify what this product has in common with similar products (shared category, features, audience)
2. Then, identify what makes this product unique or different
3. Use consistent vocabulary for shared characteristics
4. Include distinctive vocabulary for unique aspects
5. Ensure words cover the five required aspects in order

Please provide exactly five words in this exact format: [word1, word2, word3, word4, word5]:

Figure 5: Prompt for Context-aware Term Generation.

Table 9: Ablation study on multi-task configurations. Best results are bolded.

Tasks	Beauty (R@5/10)	Sports (R@5/10)	Toys (R@5/10)
UBSP + Meta2TIDs (Ours)	**0.0607 / 0.0846**	**0.0375 / 0.0539**	**0.0684 / 0.0942**
UBSP + Meta2TIDs + TIDs2Meta	0.0607 / 0.0840	0.0367 / 0.0510	0.0647 / 0.0923
NIP + Meta2TIDs	0.0551 / 0.0783	0.0342 / 0.0477	0.0631 / 0.0880
NIP + Meta2TIDs + TIDs2Meta	0.0555 / 0.0790	0.0337 / 0.0454	0.0661 / 0.0919
CROSS + Meta2TIDs + TIDs2Meta	0.0414 / 0.0643	0.0249 / 0.0365	0.0492 / 0.0734

Table 10: Performance of GRLM under different TIDs vocabulary sizes (compressed from 54,255 raw terms) on the merged 10-dataset.

Dataset	Metric	K=3000	K=5000	K=8000	Raw(54,255)
Avg.	Recall@5	0.0457	0.0458	0.0460	0.0463
	Recall@10	0.0655	0.0655	0.0657	0.0664
	NDCG@5	0.0317	0.0319	0.0320	0.0322
	NDCG@10	0.0377	0.0378	0.0379	0.0387

Dataset: Beauty

"title": "Truth by Calvin Klein for Women, Eau De Parfum Spray, 3.4 Ounce",
"description": "Launched by the design house of Calvin Klein in 2000, TRUTH PERFUME is classified as a refreshing, oriental, woody fragrance. This feminine scent possesses a blend of ...",

Semantic IDs: <216> <216> <32><6>

Text ID: star wars interactive electronic r2d2

Term ID: perfume | feminine | oriental | woody | calvin-klein

Dataset: Toys

"title": "Star Wars R2-D2 Interactive Astromech Droid"
"description": "Collectors young and old will appreciate the details of this Star Wars Interactive Electronic R2-D2 Astromech Droid. Complete with movie-accurate messages ...",

Semantic IDs: <112> <87> <191><18>

Text ID: star wars interactive electronic r2d2

Term ID: interactive | beverage | responsive | durable | collectible

Figure 6: Case Study.