A Modular Survey for Semantic ID-Based Generative Recommendation

Peiyu Hu †, Weihai Lu †, Siying Gu †, Elliott Wen , Changyu Zeng , Senzhang Wang , Jia Wang *

Posted Date: 11 May 2026

doi: 10.20944/preprints202605.0619.v1

Keywords: recommendation system; generative recommendation; semantic ID-based generative recommendation

 doi:10.20944/preprints202605.0619.v1

Article

A Modular Survey for Semantic ID-Based Generative Recommendation

Peiyu Hu [1,2,†], Weihai Lu [3,†], Siying Gu [4,†], Elliot Wen [5], Changyu Zeng [1,2], Senzhang Wang [6] and Jia Wan [1,2,*]

[1] Xi'an Jiaotong-Liverpool University
[2] University of Liverpool
[3] Peking University
[4] East China Normal University
[5] The University of Auckland
[6] Central South University
* Correspondence: Jia.Wang02@xjtlu.edu.cn
† Equal contribution.

Abstract

Traditional discriminative recommenders score and rank items indexed by single item IDs, whereas Semantic ID-based generative recommendation formulates recommendation as conditional generation of Semantic ID token sequences. This shift offers a unified view of retrieval and ranking and shows promising scaling properties, but the literature is fragmented across tokenization and quantization choices, model backbones, and training and decoding protocols, making systematic comparison difficult. To address this, we present the first survey that organizes the field, with four pivotal contributions. First, we introduce a unified five-stage reference pipeline: *Representation Layer*, *Tokenization*, *Generative Backbone*, *Training*, and *Inference*. This pipeline standardizes terminology and exposes shared structure. Second, grounded in this pipeline, we map existing methods into a fine-grained typology along semantic granularity, architectural coupling, and learning objectives. Third, based on this structured view, we provide a scaling-oriented perspective that connects component-level decisions to expressiveness, efficiency, and empirical performance, clarifying trade-offs. Finally, we synthesize open challenges and concrete directions that follow from the identified bottlenecks. To support reproducibility and controlled ablations across stages, we release UniGenRec (https://github.com/hupeiyu21/UniGenRec-A-universal-generative-recommendation-toolbox), an open-source modular toolbox implementing the proposed pipeline.

Keywords: recommendation system; generative recommendation; semantic ID-based generative recommendation

1. Introduction

Traditional recommender systems (RS) formulate recommendation as a discriminative ranking problem Da'u and Salim (2020); Ko *et al.* (2022), typically implemented as a two-stage pipeline. In the first stage, the system retrieves a candidate set of items indexed by discrete identifiers. In the second stage, a discriminative ranking model is applied to these candidates to produce the final ranked results. However, discriminative recommenders become harder to scale as catalogs grow to billions of items, since both retrieval and ranking must operate over a very large identifier spaceDeng *et al.* (2025); Ko *et al.* (2022); Rajput *et al.* (2023). Moreover, their reliance on ID embeddings learned from observed user–item interactions can limit performance for newly introduced or rarely interacted items in cold-start scenariosDa'u and Salim (2020); Zheng *et al.* (2024).

Motivation: Recently, semantic ID-based generative recommendation (GenRec) has emerged as a promising alternative Ko *et al.* (2022); Li *et al.* (2025). Instead of representing each item only by a

 doi:10.20944/preprints202605.0619.v1

discrete ID with no semantic structure, GenRec maps each item to a short sequence of semantic tokens. It then formulates recommendation as generating these token sequences given a user's context. This approach improves cold-start performance by generalizing from related items, mitigates the computational overhead of scoring massive item catalogs, and exhibits favorable scaling laws similar to large language models Hou and Shin (2025); Liu *et al.* (2025b). Despite this potential, the field is currently fragmented. Researchers are using a confusing variety of methods for tokenization (how to turn an item into tokens), backbones (the model architecture), and training protocols (how to optimize the model). Consequently, there is an urgent need for overviewing the existing works to decouple these choices, and a modular toolbox that enables controlled ablations in a uniform settingLi *et al.* (2024,2).

Scope of this survey: While several recent surveys have touched upon generative recommendation, our work fills a unique gap. Existing surveys either focus on broad industrial applications Liang and Zhang (2025), generic data-task taxonomies Li *et al.* (2024), or focus broadly on information retrieval Li *et al.* (2025). The most recent comprehensive survey Li *et al.* (2024) covers generative recommendation in general, but our work differs in two critical ways: First, we focus strictly on the *Semantic ID-based* paradigm. By narrowing the scope, we provide a much deeper analysis of how to learn discrete representations and codebook alignment that broad surveys overlook. Second, instead of a simple checklist of modules, we organize the field into an end-to-end engineering pipeline.This helps researchers understand how each choice affects the whole system.

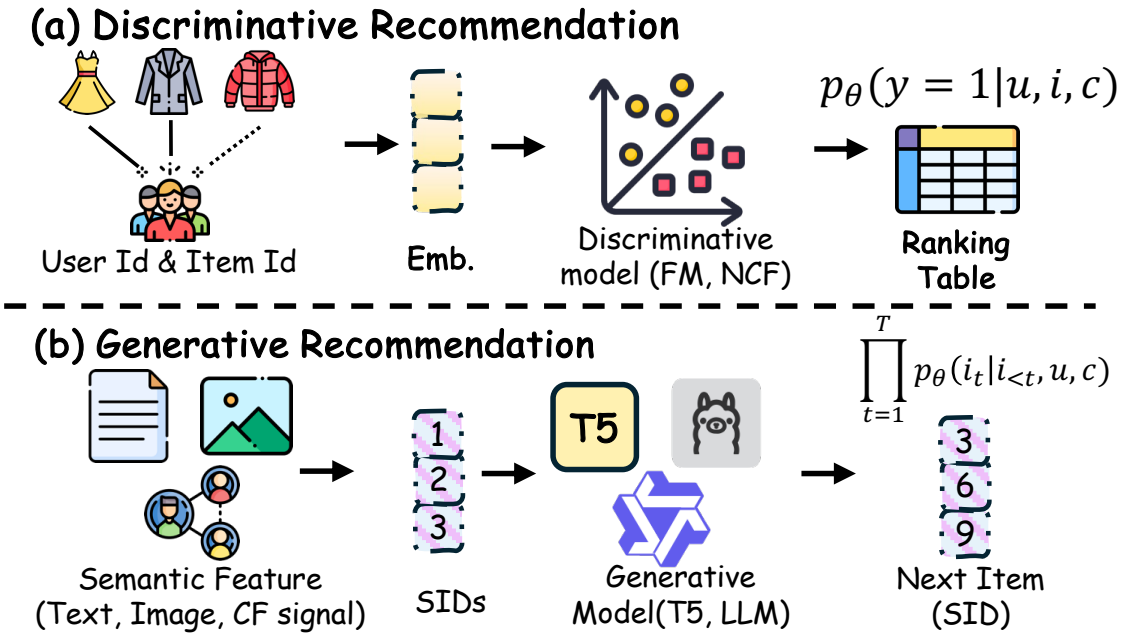

Figure 1. Discriminative recommenders typically retrieve and then rank items indexed by discrete IDs, whereas SID-based GenRec recommends items by generating sequences of semantic tokens that correspond to items.

Contributions.

This survey focuses specifically on GenRec and provides a systematic review of its end-to-end pipeline and design space. Our contributions can be summarized as follows.

- We formalize GenRec as a five-stage reference pipeline consisting of *Representation*, *Tokenization*, *Generative Backbone*, *Training*, and *Inference*. This pipeline provides a common framework for comparing methods.
- Building on this pipeline, we develop a detailed typology that links existing work to specific stages and organizes them by factors such as the level of semantic detail, how components are connected, and the training goals. This helps make comparisons across methods more consistent and easier to interpret.

 doi:10.20944/preprints202605.0619.v1

- Using this organization of the design space, we analyze how design choices affect what the model can represent, how much computation it requires, and how it generates recommendations at inference time. We also summarize common trade-offs and typical problems reported in prior studies.
- We release **UniGenRec**, the first open-source modular toolbox that implements this five-stage pipeline. This allows researchers to conduct experiments by isolating and testing individual components.

2. Background and Problem Definition

Discriminative Recommendation

Discriminative recommendation models score items from a candidate set under a given user context. For a user u, context c, and candidate items $\mathcal{I}$, a discriminative model learns a conditional scoring function

$$f_\theta(u, i, c) \sim p_\theta(y = 1 \mid u, i, c), \quad i \in \mathcal{I}, \tag{1}$$

where y denotes whether the user interacts with item i (e.g., click, purchase) and θ are model parameters. This paradigm covers classical collaborative filtering methods such as neighborhood-based CF and matrix factorization, as well as modern deep models including feature-interaction architectures (e.g., FM/DeepFM/DLRM-style)Da'u and Salim (2020); Deldjoo *et al.* (2024) and matching-based retrieval models such as two-tower networks that map users and items into embeddings and rank by similarityDa'u and Salim (2020); Deldjoo *et al.* (2024); Ko *et al.* (2022).

Generative Recommendation

Generative recommendation models user behavior as a sequence generation problem. Given user u, context c, and historical interaction sequence $i_{1:T} = (i_1, \ldots, i_T)$, a standard autoregressive formulation is

$$p_\theta(i_{1:T} \mid u, c) = \prod_{t=1}^{T} p_\theta(i_t \mid i_{<t}, u, c). \tag{2}$$

Representative research lines manifest in diverse forms. LLM-based recommendation typically prompts or fine-tunes decoder-only LMs to generate item tokens or textual responses Achiam *et al.* (2023); Liang and Zhang (2025). In industrial scenarios, generative ranking backbones employ Transformer variants to model long user histories, exemplified by *HSTU* and *GenRank* Li *et al.* (2025).

SID-based Generative Recommendation

This survey focuses on SID-based GenRec, where each item i is represented by a structured sequence of discrete tokens $s_i = (c_1, c_2, \ldots, c_L)$ with $c_l \in \mathcal{C}$ drawn from a learned codebook $\mathcal{C}$. A tokenization function $\phi : \mathcal{I} \to \mathcal{C}^L$ maps an item's continuous representation to its SID sequence. Given the user context H_u, recommendation is formulated as conditional generation of the target SID:

$$P_\theta(i \mid u) = P_\theta(s_i \mid H_u) = \prod_{l=1}^{L} P_\theta(c_l \mid c_{<l}, H_u). \tag{3}$$

3. Modular Taxonomy

To systematically review SID-based generative recommendation (SID-GenRec), we adopt a modular view that follows its end-to-end generation pipeline. As illustrated in Figure 2, the framework decomposes the pipeline into five components: *Representation*, *Tokenization*, *Generative Backbone*, *Training*, and *Inference*. Each component corresponds to a distinct design space and naturally maps to a key bottleneck when implementing and scaling SID-GenRec. We use this framework to organize the remainder of the survey.We identify four challenges that align with the generation process and motivate the above decomposition.

 doi:10.20944/preprints202605.0619.v1

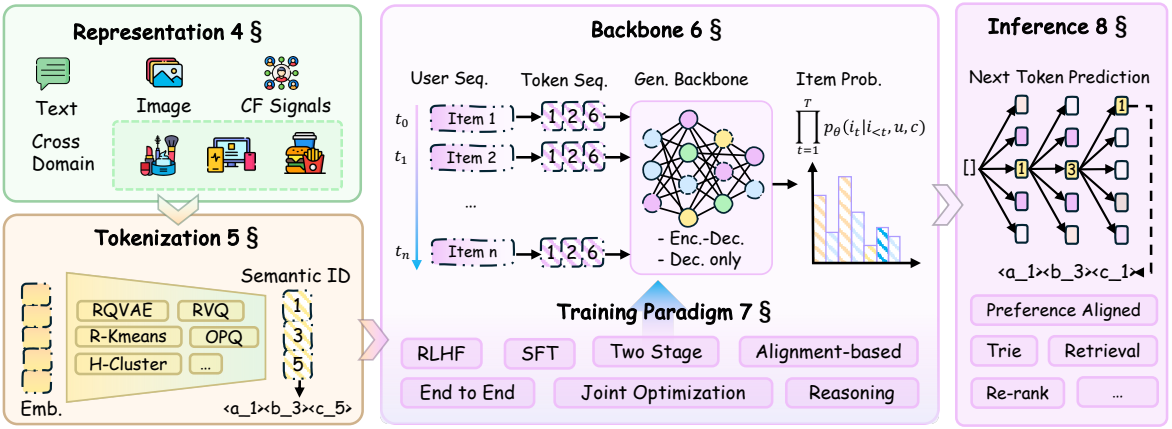

Figure 2. A Unified Overview of SID-based Generative Recommendation Frameworks

Challenge 1 (Representation). How should the system encode heterogeneous raw signals (e.g., interaction history, text, and images) into continuous user/item representations, and how should multimodal fusion be performed (early vs. late fusion)? We review representation learning and fusion mechanisms in Section 4.

Challenge 2 (Tokenization). Given item representations, how should the system define and learn a discrete SID vocabulary (codebook) and assign each item to a SID sequence that preserves both semantic similarity and collaborative patterns? Since SIDs are the prediction targets, tokenization quality directly shapes what the generative model can learn. We categorize tokenization strategies in Section 5.

Challenge 3 (Generative Backbone). How can the backbone model the conditional distribution over SID sequences given user context, support token-level decoding, and remain scalable with long interaction histories in both training and inference? We categorize backbone architectures and discuss capacity–scalability trade-offs in Section 6.

Challenge 4 (Training & Inference Efficiency). Because SID-GenRec typically decodes multiple tokens and may require searching over candidate sequences, efficiency becomes a first-class concern in large-scale deployment. We review training paradigms that improve optimization and resource usage in Section 7, and summarize inference and decoding strategies that reduce serving cost in Section 8.

4. Representation Layer

We define the representation layer as mapping heterogeneous inputs into a shared continuous latent space that can be consistently consumed by downstream tokenization and generation. The main problem is that different sources often disagree or have different noise and bias; accordingly, prior work designs source-specific alignment objectives to make the latent space coherent, semantically grounded, and consistent with user preference signals.

Text signals.

Text signals refer to item-side natural language information such as titles, descriptions, and reviews, which provides a semantic prior for representing and indexing items. A key challenge is that raw text is often sparse, noisy, and incomplete, so the resulting embeddings may be insufficient for stable tokenization and preference modeling. To address this, GREAM Hong *et al.* (2025) uses an LLM to fuse multiple textual sources into higher-fidelity item embeddings and trains collaborative–semantic alignment tasks to inject interaction evidence into the induced SID space. In parallel, MINIONEREC Kong *et al.* (2025) treats LLM text embeddings as a world-knowledge prior and aligns this semantic structure to the hierarchical SID token space while modeling users as chronological SID sequences, grounding SID generation in both enriched semantics and behavioral supervision.

 doi:10.20944/preprints202605.0619.v1

Multimodal signals.

Integrating multimodal signals faces the challenge of semantic inconsistency, where naive fusion often leads to conflicting cues and unstable SIDs. To address this mismatch, UTGRec Zheng *et al.* (2025) aligns text and visual representations in a continuous space before discretization to ensure a coherent manifold. Similarly, FORGE Fu *et al.* (2025) enforces cross-modal consistency during quantization to prevent conflict propagation into code assignments. Alternatively, to avoid forcing a single fused representation, MQL4GRec Zhai *et al.* (2025) maintains modality-specific tokenization and coordinates them at the generative stage. Along this line, MACRec Zhang *et al.* (2025a) introduces alignment objectives over discrete codes, explicitly reducing collisions to improve consistency.

Collaborative signals.

These signals directly reflect to the user-item interaction patterns, complementing textual item descriptions. However, using these signals is difficult for two reasons. First, interaction data is often sparse. Second, the data is biased toward popular items. To address this, some researchers use interaction data to build hierarchical structures. For example, SEATERSi *et al.* (2024) organizes items into a tree where items with similar user patterns are grouped together. Other methods, such as LETTER, take a different approach by adding specific rules during training. These rules ensure that the generated IDs stay consistent with known user patterns while reducing the unfair influence of popular items. MMQ-v2Xu *et al.* (2025) use a "mixture" strategy to adaptively balance text and behavior.

5. Tokenization

Tokenization is a semantic compression step. It takes a continuous item representation and converts it into a discrete tokens, which we call a structured identifier (SID). This step is needed because the generative backbone operates on discrete tokens, while item representations are usually learned as dense vectors. A good SID should preserve the main semantic signals so that similar items receive similar token sequences. At the same time, it should be short and use a limited vocabulary, so inference only requires generating a few tokens. We group existing strategies into three paradigms based on how they discretize representations, as shown in the Figure 3.

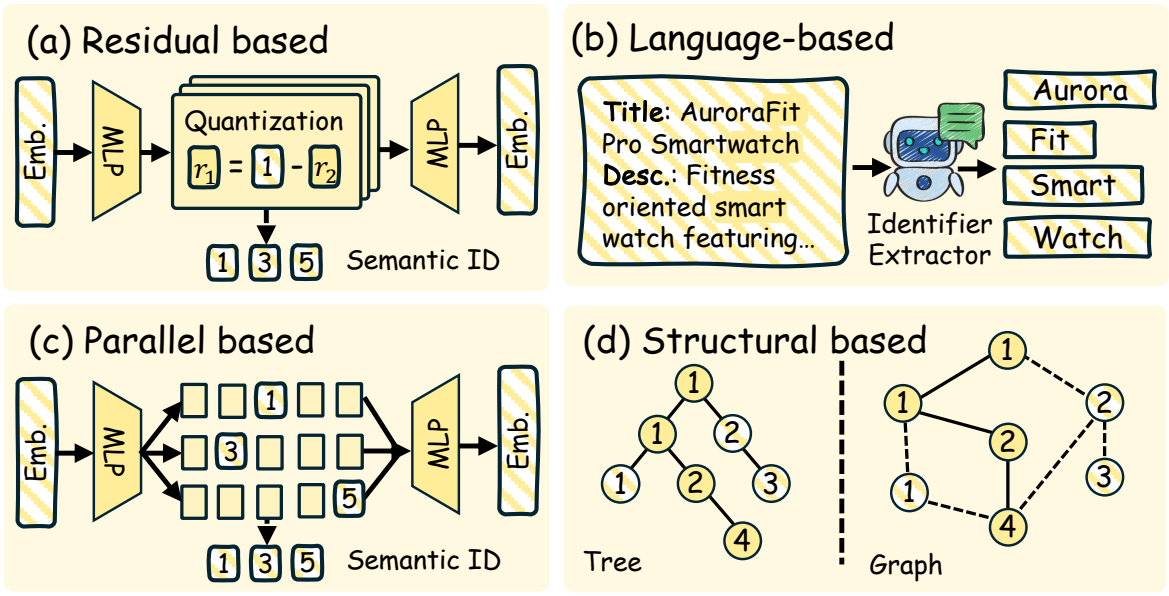

Figure 3. Taxonomy of Tokenization Strategies for Semantic ID Construction

Residual-based Quantization.

The most common approach, pioneered by TIGER Rajput *et al.* (2023), adopts Residual Quantization (RQ-VAE). Intuitively, it follows a coarse-to-fine process: it decomposes an item vector into

 doi:10.20944/preprints202605.0619.v1

a sequence of codes, where the first code captures a broad prototype and subsequent codes progressively refine details by correcting the residual errors left by earlier stages. Many works build on this foundation to address specific challenges. First, to align the learned SIDs with the original semantic representations, COST Zhu *et al.* (2024) and MACREC Zhang *et al.* (2025a) introduce contrastive objectives to align semantic spaces, while DAS Ye *et al.* (2025b) and ALIGN3GR Ye *et al.* (2025a) incorporate collaborative signals so that codes better reflect real interaction patterns. Second, to handle complex inputs such as multi-modality and cross-domain data, MQL4GREC Zhai *et al.* (2025) employs dual-stream quantizers for multimodal signals, and GENCDR Hu *et al.* (2025) leverages shared codebooks with lightweight adapters for cross-domain transfer. Finally, to alleviate codebook collapse, RECBASE Zhou *et al.* (2025) adopts curriculum learning, and SIIT Chen *et al.* (2024) uses a self-improvement loop to iteratively refine assignments. Recent hybrid designs such as COBRA Yang *et al.* (2025) further extend this line by cascading sparse IDs with dense vectors.

Table 1. A Unified Summary of Representative Generative Recommendation Methods.

Method	Venue	Representation	Tokenization	Backbone	Training	Inference
I. Foundation & Single-Modality						
TIGER Rajput *et al.* (2023)	NIPS'23	Text	RQ-VAE	Enc-Dec	2-Stage	Beam Search
IDGenRec Tan *et al.* (2024)	SIGIR'24	Text	LLM-Vocab	Enc-Dec	2-Stage	Constr. Beam
SEATER Si *et al.* (2024)	SIGIR'24	CF	Tree-Cluster	Enc-Dec	2-Stage	Constr. Beam
CoST Zhu *et al.* (2024)	RecSys'24	Text	Contrastive RQ	Enc-Dec	2-Stage	Beam Search
ActionPiece Hou *et al.* (2025b)	ICML'25	Text	OPQ	Enc-Dec	2-Stage	Ensemble
ETEGRec Liu *et al.* (2025a)	SIGIR'25	CF	Learnable RQ	Enc-Dec	Alternating	Beam Search
TokenRec Qu *et al.* (2024)	TKDE'25	CF	Masked VQ	Encoder	2-Stage	Scoring
RecBase Zhou *et al.* (2025)	EMNLP'25	Text	Curriculum RQ	Dec-Only	2-Stage	Beam Search
HiD-VAE Fang *et al.* (2025)	Arxiv'25	Text	Hierarchical RQ	Enc-Dec	2-Stage	Constr. Beam
PSI Zhang *et al.* (2025b)	Arxiv'25	Text	Post-hoc RQ	Enc-Dec	2-Stage	Beam Search
SIIT Chen *et al.* (2024)	Arxiv'24	Text	Self-Improved	Enc-Dec	Iterative	Constr. Beam
COBRA Yang *et al.* (2025)	Arxiv'25	Text+SID	Cascaded RQ	Dec-Only	E2E Joint	BeamFusion
II. Multimodal, Cross-Domain & Parallel						
MQL4GRec Zhai *et al.* (2025)	ICLR'25	Text+Img	Dual RQ-VAE	Enc-Dec	2-Stage	Scoring
RPG Hou *et al.* (2025a)	KDD'25	Text+Img	Parallel OPQ	Non-AR	2-Stage	Graph Search
LLaDA-Rec Shi *et al.* (2025)	Arxiv'25	Text	Parallel VQ	Non-AR	Diffusion	Diffusion
GenCDR Hu *et al.* (2025)	AAAI'26	Cross-Dom	RQ+LoRA	Dec-Only	2-Stage	Constr. Beam
MACRec Zhang *et al.* (2025a)	AAAI'26	Text+Img	Cross RQ-VAE	Enc-Dec	2-Stage	Scoring
GMC Jin *et al.* (2025)	Arxiv'25	Cross-Dom	Unified RQ	Enc-Dec	2-Stage	Beam Search
UTGRec Zheng *et al.* (2025)	Arxiv'25	Multi-Dom	Tree	Enc-Dec	2-Stage	Beam Search
UniTok Hou and Shin (2025)	Arxiv'25	Multi-Dom	TokenMoE	-	-	-
MMQ-v2 Xu *et al.* (2025)	Arxiv'25	Text+Img+CF	MoE-VQ	-	-	-
FORGE Fu *et al.* (2025)	Arxiv'25	Text+Img	RQ-VAE	Dec-Only	2-Stage	Dyn. Beam
MME-SID Wang *et al.* (2025)	CIKM'25	Text+Img+CF	MM-RQ-VAE	Dec-Only	2-Stage	Scoring
III. Alignment, Reasoning & Collaborative Fusion						
LC-Rec Zheng *et al.* (2024)	ICDE'24	Text	RQ-VAE	Dec-Only	Multi-Task	Beam Search
EAGER Wang *et al.* (2024b)	KDD'24	Text+CF	Dual H-KMeans	Dual Enc-Dec	Joint	Fusion
ColaRec Wang *et al.* (2024c)	CIKM'24	Text+CF	H-KMeans	Enc-Dec	Joint	Constr. Beam
LETTER Wang *et al.* (2024a)	CIKM'24	Text+CF	Learnable RQ	Enc-Dec	2-Stage	Beam Search
GRAM Lee *et al.* (2025)	ACL'25	Text+CF	H-KMeans	Enc-Dec	E2E	Constr. Beam
DAS Ye *et al.* (2025b)	CIKM'25	Text+CF	Dual RQ-VAE	Enc-Dec	Joint	Beam Search
OneRec Deng *et al.* (2025)	Arxiv'25	Text+CF	R-KMeans	Enc-Dec	NTP+DPO	Session Beam
Align3GR Ye *et al.* (2025a)	AAAI'26	Text+CF	Dual RQ-VAE	Dec-Only	SFT+DPO	Beam Search
GREAM Hong *et al.* (2025)	Arxiv'25	Text+CF	RQ-KMeans	Dec-Only	SFT+RL	Beam Search
MiniOneRec Kong *et al.* (2025)	Arxiv'25	Text	RQ-VAE	Dec-Only	SFT+GRPO	Beam Search
OneRec-Think Liu *et al.* (2025b)	Arxiv'25	Text+CF	H-KMeans	Dec-Only	SFT+RL	CoT
LIGER Yang *et al.* (2024)	Arxiv'24	Text+SID	RQ-VAE	Enc-Dec	Dual-Obj.	Hybrid Rank

Legend: *Rep (Representation):* **CF** (Collaborative Filtering), **Img** (Image), **Cross/Multi-Dom** (Cross/Multi-Domain). *Tok (Tokenization):* **RQ** (Residual Quantization), **VQ** (Vector Quantization), **OPQ** (Optimized Product Quantization), **H/R-KMeans** (Hierarchical/Residual K-Means), **MoE** (Mixture-of-Experts). *Backbone:* **Enc-Dec** (Encoder-Decoder), **Non-AR** (Non-Autoregressive). *Train:* **E2E** (End-to-End), **SFT** (Supervised Fine-Tuning), **NTP** (Next Token Prediction), **DPO** (Direct Preference Optimization), **RL** (Reinforcement Learning), **GRPO** (Group Relative Policy Optimization). *Inf (Inference):* **Constr** (Constrained), **Dyn** (Dynamic), **CoT** (Chain-of-Thought).

 doi:10.20944/preprints202605.0619.v1

Structural & Parallel Discretization.

To address the speed and collision issues associated with RQ-VAE, alternative approaches pursue structure-based or parallel discretization. **Structure-based methods**, such as SEATER Si *et al.* (2024), OneRec Deng *et al.* (2025), and EAGER Wang *et al.* (2024b), replace gradient-based quantization with hierarchical clustering (e.g., Residual K-Means). They impose an explicit tree structure where each item maps to a unique leaf, ensuring that no two items share the same ID and thereby avoiding collisions. In contrast, parallel discretization is designed for high-speed inference: it factorizes the code into independent subspaces so that multiple SID tokens can be predicted in parallel, reducing generation latency compared to autoregressive, token-by-token decoding. Models such as RPG Hou *et al.* (2025a) and LLaDA-Rec Shi *et al.* (2025) partition the vector into independent subspaces via Product Quantization (e.g., OPQ) or multi-head VQ. This independence enables all tokens to be predicted simultaneously rather than sequentially. In addition, TokenRec Qu *et al.* (2024) introduces Masked VQ, and MMQ-v2 Xu *et al.* (2025) proposes Mixture-of-Quantization to adapt code capacity dynamically according to item frequency.

Natural Language-based Identification.

A distinct paradigm avoids learning new codebooks and instead reuses the LLM's native vocabulary as the identifier space. Approaches such as IDGenRec Tan *et al.* (2024) and GRAM Lee *et al.* (2025) generate human-readable textual IDs or keywords by leveraging a *lexical interpreter/extractor* that maps item semantics into **meaning-bearing vocabulary tokens** (e.g., attribute keywords or concept phrases) rather than arbitrary indices. Since these tokens are already grounded in pretrained LLM embeddings, the resulting identifiers inherit lexical priors and tend to transfer better to unseen items in a zero-shot setting.

Tokenization determines the system's effective resolution and often becomes the primary bottleneck in GenRec. The chosen strategy governs how distinctly and meaningfully items are represented. Residual-based methods provide strong compression but can lead to uneven code usage and degraded capacity in parts of the codebook. Structural methods enforce an explicit hierarchy to guarantee unique identifiers, while parallel strategies maximize throughput by sacrificing token dependencies. Ultimately, if tokenization cannot produce a discrete token space that is both collision-resistant and preference-relevant, the downstream generative backbone will struggle to learn robustly, no matter how far it is scaled.

6. Generative Backbones

The **Generative Backbone** functions as the core probabilistic engine of the SID-based framework, responsible for modeling the sequential dependencies within user history and estimating the conditional likelihood of target item tokens. Existing backbones are mainly grouped by architecture into two streams, encoder–decoder and decoder-only.

Encoder-Decoder Architectures.

The foundational wave, represented by TIGER Rajput *et al.* (2023), uses a T5-style encoder-decoder backbone to cast recommendation as seq2seq generation: the encoder encodes the user history into a global context, and the decoder autoregressively generates the target item's SID tokens conditioned on that context. This backbone cleanly separates encoding and generation, so later works can add extra training signals on encoder states. ETEGRec Liu *et al.* (2025a) makes the encoder learn a sequence state that matches the target item semantics, and makes the decoder learn a preference state that is close to the target item representation. GRAM Lee *et al.* (2025) encodes a short user ID sequence for overall intent and encodes each history item with rich text for details, then lets the decoder fuse these two views through cross-attention instead of concatenating them into one long input. MACRec Zhang *et al.* (2025a) adds multimodal latent alignment in the encoder to keep text and image signals consistent during generation.

 doi:10.20944/preprints202605.0619.v1

Decoder-Only Architectures.

More recent work shifts to decoder-only backbones such as LLaMA and Qwen to better leverage foundation models' language understanding and general knowledge. A representative line is LC-Rec Zheng *et al.* (2024), which does not simply "prompt" an LLM with item tokens; instead, it designs a set of SID–LLM alignment mechanisms so that discrete SIDs become compatible with the model's lexical space and generation behavior, largely establishing the practical LLM-as-SID-generator paradigm. Building on this direction, GenCDR Hu *et al.* (2025) further uses LLM backbones with lightweight adaptation such as LoRA and MoE-style routing to support cross-domain, multi-scenario recommendation under a unified generator. Besides, OneRec-Think Liu *et al.* (2025b) and GREAM Hong *et al.* (2025) explicitly introduce reasoning tokens, using the LLM's inference capability to produce rationales that help interpret and better align SID generation with user preference.

7. Learning Paradigms

Once the backbone is established, the training paradigm defines how the parameters are optimized. Research has progressed along four distinct trajectories, reflecting different exploration goals such as stability, tighter coupling, better alignment with user utility, and faster generation.

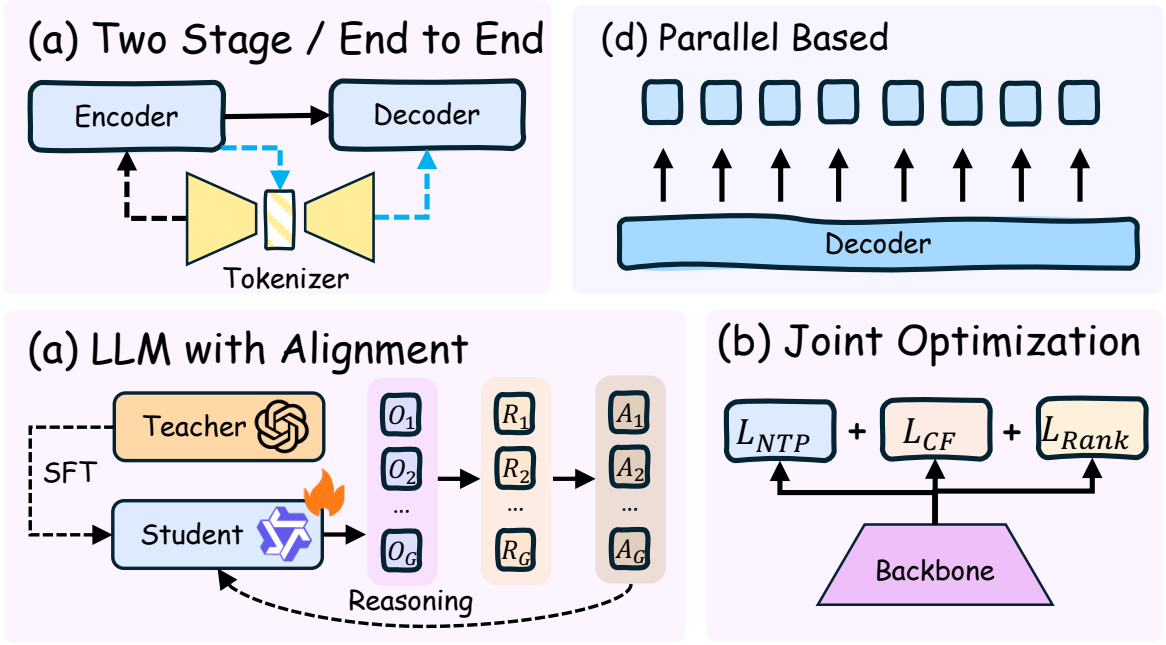

Figure 4. Taxonomy of Learning Paradigms for SID-based Generative Recommendation

Two-Stage Paradigms.

Two-stage training is a widely used paradigm in SID-based GenRec that decouples SID construction from generative modeling: the tokenizer is trained first to map each item to a fixed-length SID sequence, and this item-to-SID assignment is then frozen so the backbone can be trained on SIDs as its discrete vocabulary, typically with autoregressive next-token prediction. TIGER Rajput *et al.* (2023) is a representative example, where an RQ-VAE tokenizer produces SIDs and an encoder–decoder model learns to generate the target SID conditioned on the SID-serialized user history. This design is stable and modular, but its performance is largely bounded by the quality and preference faithfulness of the fixed SIDs.

End-to-End and Joint Paradigms.

To reduce the mismatch between SID construction and downstream generation, recent work moves beyond the disjoint pipeline and couples tokenization with the generative recommender during training. ETEGRec Liu *et al.* (2025a) is a representative end-to-end design, where the tokenizer

 doi:10.20944/preprints202605.0619.v1

and backbone are updated in a coordinated manner so the backbone's learned collaborative/context signals can be injected back into SID learning and progressively refine the codes. In parallel, joint optimization via multi-task learning couples generation with auxiliary objectives under a shared backbone: COLAREC Wang *et al.* (2024c) jointly trains user-item generation with an item-side indexing task (often with auxiliary ranking/contrastive regularization) to align content understanding with collaborative IDs, while EAGER Wang *et al.* (2024b) jointly trains dual generation streams (behavior vs semantic) with cross-stream transfer objectives so two complementary token spaces co-evolve rather than being learned in isolation.

Alignment-Driven Paradigms.

Beyond next-token accuracy, recent works incorporate alignment objectives to better reflect recommendation utility. Preference optimization (e.g., DPO) is adopted to separate "good" and "bad" recommendations (e.g., ONEREC Deng *et al.* (2025), ALIGN3GR Ye *et al.* (2025a)). Some works further augment generation with explicit rationales (CoT) and apply RL-style updates (e.g. GRPO) to align intermediate reasoning with final outcomes (e.g., ONEREC-THINK Liu *et al.* (2025b), GREAM Hong *et al.* (2025), MINIONEREC Kong *et al.* (2025)).

Non-Autoregressive and Parallel Paradigms.

Addressing the efficiency bottleneck, alternative paradigms explore parallel mechanisms. RPG Hou *et al.* (2025a) proposes a fully parallel Transformer that utilizes multi-head projections to predict all ID tokens simultaneously, treating generation as a multi-label classification task. Taking a different approach, LLADA-REC Shi *et al.* (2025) introduces Discrete Diffusion, modeling generation as an iterative denoising process via a bidirectional Transformer. These approaches discard the rigid left-to-right order, offering a flexible trade-off between inference speed and generation quality.

Overall, learning paradigms in GenRec reveal a set of recurring tensions. Two-stage pipelines prioritize stability and controllability, but often cap performance because the generated SIDs are not fully adapted to what the backbone ultimately needs. Joint and end-to-end optimization pushes the field toward tighter coupling between tokenization and generation, aiming for task-shaped SIDs and stronger downstream gains, at the cost of more delicate training dynamics. Preference and RL-style alignment signals a shift from "predicting likely tokens" to "generating useful recommendations," reframing optimization around utility rather than likelihood. Efficiency-driven directions emphasize scalability under serving constraints, motivating more parallel or low-latency generation mechanisms to make SID generation practical at industrial scale.

8. Inference and Decoding Strategies

The inference module serves as the operational bridge mapping generative probability distributions to actionable recommendation lists. As GenRec evolves, decoding strategies have transcended simple likelihood search, shifting towards sophisticated decision processes that navigate the trilemma of **validity, efficiency, and alignment**.

Constrained Autoregressive Decoding.

As the mainstream paradigm, Autoregressive (AR) decoding generates tokens sequentially. Early works like TIGER Rajput *et al.* (2023) established the standard Beam Search workflow. However, to mitigate the risk of generating invalid hallucinations in sparse semantic spaces, Trie-based Constrained Decoding has become the de facto standard. Methods such as GENCDR Hu *et al.* (2025), COLAREC Wang *et al.* (2024c), and HID-VAE Fang *et al.* (2025) construct a Valid Prefix Tree at each step to force the model to sample only from legal subtrees, ensuring validity. Advanced variants further optimize this process: SEATER Si *et al.* (2024) employs balanced trees for uniform retrieval paths, FORGE Fu *et al.* (2025) introduces Dynamic Beam Search to adaptively prune hypotheses, and IDGENREC Tan *et al.* (2024) utilizes Diverse Beam Search to prevent redundancy in candidate generation.

 doi:10.20944/preprints202605.0619.v1

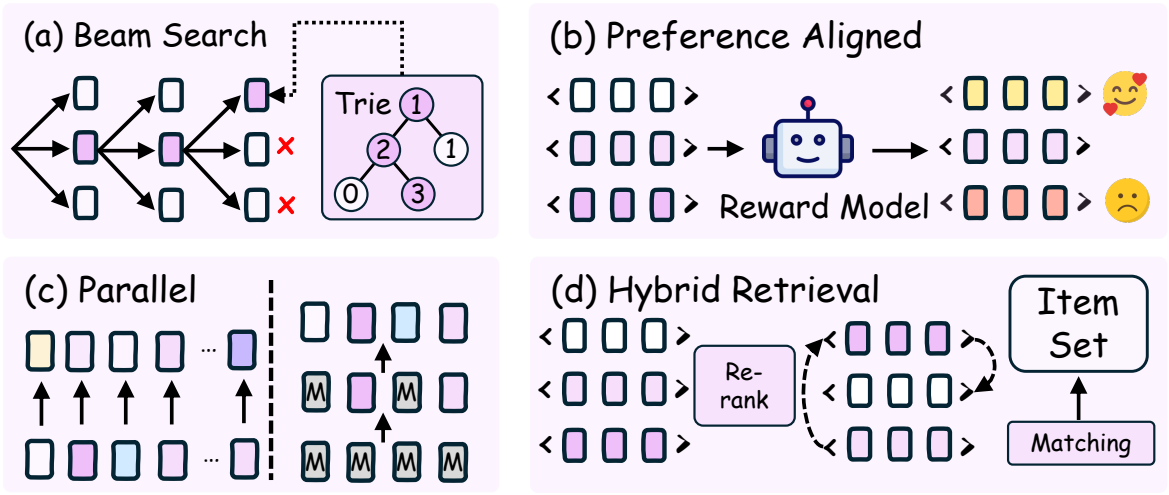

Figure 5. Taxonomy of Inference Strategies for Generative Recommendation

Preference-Aligned Decoding.

A key inference issue is that likelihood mainly reflects what the model thinks is most probable, but real platforms care about what users truly prefer under business constraints, and these two signals can differ. To optimize preference more directly, ONEREC Deng *et al.* (2025) applies DPO on preference pairs built from its generated item-token candidates, so decoding is pushed toward user-preferred results rather than purely high-probability ones. ALIGN3GR Ye *et al.* (2025a) further improves stability by training DPO progressively and introducing real feedback pairs to better match online signals.

Parallel and Diffusion Decoding.

To reduce the sequential latency of AR generation, alternative paradigms predict tokens in parallel or through iterative refinement. RPG Hou *et al.* (2025a) predicts multiple SID positions simultaneously and may incorporate structural constraints (e.g., graph-based constraints) to improve consistency. LLADA-REC Shi *et al.* (2025) formulates SID generation as discrete diffusion, where inference performs iterative denoising and can be combined with search to control the quality–efficiency trade-off.

Hybrid and Retrieval-Enhanced Decoding.

To meet industrial constraints, some pipelines combine generation with retrieval signals or multi-stage execution. For reasoning-intensive systems, ONEREC-THINK Liu *et al.* (2025b) and GREAM Hong *et al.* (2025) decouple offline reasoning from online decoding to reduce serving overhead. Other works integrate dense retrieval signals into decoding (e.g., LIGER Yang *et al.* (2024), COBRA Yang *et al.* (2025)) or reuse generator representations for efficient dot-product scoring, treating the generative model as a high-capacity retriever (e.g., MME-SID Wang *et al.* (2025), TOKENREC Qu *et al.* (2024)).

Overall, decoding in GenRec has moved from simply searching for the highest-likelihood sequence toward a decision procedure that must satisfy three practical requirements in deployment: validity, scale, and preference. Trie constraints and related variants make validity a hard guarantee by restricting generation to legal paths in the item index, rather than relying on post-hoc filtering. As catalogs grow and latency budgets tighten, inference is also moving beyond strict left-to-right generation, with parallel and diffusion-style decoding exploring faster generation that trades some dependency modeling for higher throughput. At the same time, the goal of decoding is changing: instead of returning what the model thinks is most probable, recent work increasingly combines decoding with preference alignment so that the produced candidates better match what users actually choose under platform constraints. The frontier is no longer only how to find the most probable SID, but how to decode efficiently at scale while staying aligned with real user utility.

 doi:10.20944/preprints202605.0619.v1

9. Emerging Directions

Despite rapid progress, SID-based GenRec still faces open questions. We highlight four emerging directions suggested by recent progress across the pipeline.

High-Capacity and Flexible Tokenization. Current residual quantization methods often bottleneck information density and rely on rigid hierarchical structures that may not optimally capture item semantics. Future research should explore high-capacity identifiers, such as parallel token streams or extended sequences, to encode richer details. This requires developing alternative quantization paradigms beyond standard residual frameworks, allowing the system to represent complex multimodal signals with higher fidelity and without precision loss.

Recommendation Architectures. The autoregressive generation paradigm inherited from natural language processing imposes latency penalties incompatible with high-throughput recommendation. To address this, the field must evolve towards architectures explicitly designed for set-oriented generation. Promising directions include non-autoregressive decoding, discrete diffusion models, and sparse mixture-of-experts, which enable models to decouple inference latency from parameter scaling to meet industrial requirements.

Preference Alignment. The goal of preference alignment is to prioritize user satisfaction over data imitation. However, standard likelihood maximization inherently conflates *occurrence frequency* with *user preference*, leading to models that over-generate globally popular SIDs at the expense of personalized utility. Future research must bridge this gap by shifting towards objectives that explicitly decouple popularity from preference, ensuring the model ranks a preferred item above a popular but irrelevant alternative rather than simply predicting the most frequent token.

10. Conclusion

SID-based GenRec reformulates recommendation as a unified sequence generation task, uniquely unlocking the potential for Neural Scaling Laws. Our analysis confirms that this paradigm enables predictable performance gains through expanded semantic capacity. To realize "Foundational Recommender Models," future research must transcend current bottlenecks of autoregressive latency and static tokenization, converging toward efficient, dynamic architectures for scalable industrial deployment. We hope this survey serves as a definitive roadmap, accelerating the community's transition from fragmented exploration to the systematic development of next-generation universal recommender systems.

References

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

Runjin Chen, Mingxuan Ju, Ngoc Bui, Dimosthenis Antypas, Stanley Cai, Xiaopeng Wu, Leonardo Neves, Zhangyang Wang, Neil Shah, and Tong Zhao. Enhancing item tokenization for generative recommendation through self-improvement. *arXiv preprint arXiv:2412.17171*, 2024.

Aminu Da'u and Naomie Salim. Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review*, 53(4):2709–2748, 2020.

Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 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, 2024.

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

Dengzhao Fang, Jingtong Gao, Chengcheng Zhu, Yu Li, Xiangyu Zhao, and Yi Chang. Hid-vae: Interpretable generative recommendation via hierarchical and disentangled semantic ids. *arXiv preprint arXiv:2508.04618*, 2025.

 doi:10.20944/preprints202605.0619.v1

Kairui Fu, Tao Zhang, Shuwen Xiao, Ziyang Wang, Xinming Zhang, Chenchi Zhang, Yuliang Yan, Junjun Zheng, Yu Li, Zhihong Chen, et al. Forge: Forming semantic identifiers for generative retrieval in industrial datasets. *arXiv preprint arXiv:2509.20904*, 2025.

Minjie Hong, Zetong Zhou, Zirun Guo, Ziang Zhang, Ruofan Hu, Weinan Gan, Jieming Zhu, and Zhou Zhao. Generative reasoning recommendation via llms. *arXiv preprint arXiv:2510.20815*, 2025.

Yu Hou and Won-Yong Shin. Tokenize once, recommend anywhere: Unified item tokenization for multi-domain llm-based recommendation. *arXiv preprint arXiv:2511.12922*, 2025.

Yupeng Hou, Jiacheng Li, Ashley Shin, Jinsung Jeon, Abhishek Santhanam, Wei Shao, Kaveh Hassani, Ning Yao, and Julian McAuley. Generating long semantic ids in parallel for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 956–966, 2025.

Yupeng Hou, Jianmo Ni, Zhankui He, Noveen Sachdeva, Wang-Cheng Kang, Ed H Chi, Julian McAuley, and Derek Zhiyuan Cheng. Actionpiece: Contextually tokenizing action sequences for generative recommendation. *arXiv preprint arXiv:2502.13581*, 2025.

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

Jinqiu Jin, Yang Zhang, Fuli Feng, and Xiangnan He. Generative multi-target cross-domain recommendation. *arXiv preprint arXiv:2507.12871*, 2025.

Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

Xiaoyu Kong, Leheng Sheng, Junfei Tan, Yuxin Chen, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Minionerec: An open-source framework for scaling generative recommendation. *arXiv preprint arXiv:2510.24431*, 2025.

Sunkyung Lee, Minjin Choi, Eunseong Choi, Hye-young Kim, and Jongwuk Lee. Gram: Generative recommendation via semantic-aware multi-granular late fusion. *arXiv preprint arXiv:2506.01673*, 2025.

Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. Large language models for generative recommendation: A survey and visionary discussions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10146–10159, 2024.

Xiaopeng Li, Bo Chen, Junda She, Shiteng Cao, You Wang, Qinlin Jia, Haiying He, Zheli Zhou, Zhao Liu, Ji Liu, et al. A survey of generative recommendation from a tri-decoupled perspective: Tokenization, architecture, and optimization. 2025.

Siqi Liang and Yudi Zhang. Generative recommendation: A survey of models, systems, and industrial advances. 2025.

Enze Liu, Bowen Zheng, Cheng Ling, Lantao Hu, Han Li, and Wayne Xin Zhao. Generative recommender with end-to-end learnable item tokenization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 729–739, 2025.

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

Haohao Qu, Wenqi Fan, Zihuai Zhao, and Qing Li. Tokenrec: Learning to tokenize id for llm-based generative recommendation. corr abs/2406.10450 (2024), 2024.

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

Teng Shi, Chenglei Shen, Weijie Yu, Shen Nie, Chongxuan Li, Xiao Zhang, Ming He, Yan Han, and Jun Xu. Llada-rec: Discrete diffusion for parallel semantic id generation in generative recommendation. *arXiv preprint arXiv:2511.06254*, 2025.

Zihua Si, Zhongxiang Sun, Jiale Chen, Guozhang Chen, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, Jun Xu, and Kun Gai. Generative retrieval with semantic tree-structured identifiers and contrastive learning. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 154–163, 2024.

Juntao Tan, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Zelong Li, and Yongfeng Zhang. 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, 2024.

 doi:10.20944/preprints202605.0619.v1

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

Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, et al. 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, 2024.

Yidan Wang, Zhaochun Ren, Weiwei Sun, Jiyuan Yang, Zhixiang Liang, Xin Chen, Ruobing Xie, Su Yan, Xu Zhang, Pengjie Ren, et al. Content-based collaborative generation for recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2420–2430, 2024.

Yuhao Wang, Junwei Pan, Xinhang Li, Maolin Wang, Yuan Wang, Yue Liu, Dapeng Liu, Jie Jiang, and Xiangyu Zhao. Empowering large language model for sequential recommendation via multimodal embeddings and semantic ids. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3209–3219, 2025.

Yi Xu, Moyu Zhang, Chaofan Fan, Jinxin Hu, Xiaochen Li, Yu Zhang, Xiaoyi Zeng, and Jing Zhang. Mmq-v2: Align, denoise, and amplify: Adaptive behavior mining for semantic ids learning in recommendation. *arXiv preprint arXiv:2510.25622*, 2025.

Liu Yang, Fabian Paischer, Kaveh Hassani, Jiacheng Li, Shuai Shao, Zhang Gabriel Li, Yun He, Xue Feng, Nima Noorshams, Sem Park, et al. Unifying generative and dense retrieval for sequential recommendation. *arXiv preprint arXiv:2411.18814*, 2024.

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

Wencai Ye, Mingjie Sun, Shuhang Chen, Wenjin Wu, and Peng Jiang. Align3gr: Unified multi-level alignment for llm-based generative recommendation. *arXiv preprint arXiv:2511.11255*, 2025.

Wencai Ye, Mingjie Sun, Shaoyun Shi, Peng Wang, Wenjin Wu, and Peng Jiang. Das: Dual-aligned semantic ids empowered industrial recommender system. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6217–6224, 2025.

Jianyang Zhai, Zi-Feng Mai, Chang-Dong Wang, Feidiao Yang, Xiawu Zheng, Hui Li, and Yonghong Tian. Multimodal quantitative language for generative recommendation. *arXiv preprint arXiv:2504.05314*, 2025.

Fuwei Zhang, Xiaoyu Liu, Dongbo Xi, Jishen Yin, Huan Chen, Peng Yan, Fuzhen Zhuang, and Zhao Zhang. Multi-aspect cross-modal quantization for generative recommendation. *arXiv preprint arXiv:2511.15122*, 2025.

Ruohan Zhang, Jiacheng Li, Julian McAuley, and Yupeng Hou. Purely semantic indexing for llm-based generative recommendation and retrieval. *arXiv preprint arXiv:2509.16446*, 2025.

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

Bowen Zheng, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. Universal item tokenization for transferable generative recommendation. *arXiv preprint arXiv:2504.04405*, 2025.

Sashuai Zhou, Weinan Gan, Qijiong Liu, Ke Lei, Jieming Zhu, Hai Huang, Yan Xia, Ruiming Tang, Zhenhua Dong, and Zhou Zhao. Recbase: Generative foundation model pretraining for zero-shot recommendation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15598–15610, 2025.

Jieming Zhu, Mengqun Jin, Qijiong Liu, Zexuan Qiu, Zhenhua Dong, and Xiu Li. Cost: Contrastive quantization based semantic tokenization for generative recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 969–974, 2024.