

Bridging the Structural Gap: Adapting Autoregressive Generation for Recommendation

Junchao Zeng[†]

Platform and Content Group
Tencent
Shenzhen, China
junchaozeng@tencent.com

Junzhang Zhu[†]

Platform and Content Group
Tencent
Shenzhen, China
johannzhu@tencent.com

Junyang Chen

CSSE
Shenzhen University
Shenzhen, China
junyangchen@szu.edu.cn

Yudong Li*

Platform and Content Group
Tencent
Shenzhen, China
elsonli@tencent.com

Wei Liu*

School of Artificial Intelligence
Sun Yat-sen University
Guangzhou, China
liuw259@mail.sysu.edu.cn

Chengxiang Zhuo

Platform and Content Group
Tencent
Shenzhen, China

Zang Li

Platform and Content Group
Tencent
Shenzhen, China

Abstract—Generative Recommendation (GR) has emerged as a new paradigm for sequential recommendation, in which a representative line of work encodes items into hierarchical semantic IDs via residual quantization and predicts the IDs token by token. However, this generative formulation still exhibits structural gaps with respect to the recommendation task: flattening multi-token IDs into a single sequence destroys item-level structure, and the inconsistency between training and inference over a hierarchical codebook gives rise to *semantic drift*. To bridge these two gaps, we propose BARGE, which employs Item Context-Aware Attention (ICA) to restore item-level structure during encoding, and Hierarchical Path Reranking (HPR) together with Dual-Path Decoding (DPD) to suppress semantic drift from two complementary angles during decoding. Extensive experiments and analytical studies on public benchmarks and a large-scale offline test demonstrate that BARGE achieves superior recommendation performance. An online A/B test on a Tencent platform yields improvements of 0.60% in click-through rate, 1.34% in click unique visitors, and 1.70% in total reading time, confirming the practical value of BARGE in industrial-scale recommendation.

Index Terms—Generative Recommendation, Semantics, Beam Search

I. INTRODUCTION

Sequential recommendation aims to predict the next item that a user will interact with, given the chronologically ordered history of past interactions. The dominant approach has long been discriminative. Models learn to score every candidate item against the encoded user history and select the highest-scoring ones. However, the per-item scoring nature of this paradigm causes its storage and computation costs to scale linearly with the size of the item set, which fundamentally limits its scalability. Generative approaches sidestep this bottleneck by directly producing item identifiers autoregressively.

Generative Recommendation (GR) [1]–[6] offers a fundamentally different paradigm. Instead of scoring all candidates, GR generates the identifier of the target item token by token.

To instantiate this generative idea, the mainstream practice [3], [7]–[12] represents each item as a tuple of L hierarchical semantic ID tokens ($c_1, c_2, \dots, c_L$) obtained via residual quantization methods such as RQ-VAE [13], [14]. This formulation allows semantically similar items to share ID prefixes and reduces the inference complexity from linear in the size of the item set to logarithmic in the depth of the codebook.

Despite this progress, the generative formulation was borrowed from natural language, where each token carries stand-alone lexical semantics. A hierarchical semantic codeword c_l , however, only acquires meaning when combined with its prefix $c_{<l}$. This mismatch manifests as two structural gaps:

(P1) Loss of item-level structure. Recommendation is fundamentally an item-level task, yet existing GR models decompose each item into L semantic ID tokens and flatten all tokens into an undifferentiated sequence. The self-attention mechanism treats every token identically regardless of item membership, which effectively reduces item-level recommendation to token-level sequence modeling. As illustrated on the left of Fig. 1, once items are flattened into a token sequence, the boundaries between items are no longer visible to the encoder, and tokens belonging to different items become indistinguishable from intra-item ones in terms of structural cues. Existing GR methods [3], [7], [8] universally adopt this flat-sequence strategy and leave item boundaries to be recovered implicitly from positional signals alone.

(P2) Semantic drift in hierarchical decoding. Multi-level semantic IDs form a tree-like codebook in which the codeword selected at each layer constrains all subsequent layers, so that an error at any layer redirects the search into a wrong subtree. We refer to this family of failure modes collectively as *semantic drift*. As illustrated on the right of Fig. 1, once decoding deviates from the correct branch, the target leaf is no longer reachable along the chosen path, while standard beam search keeps selecting codewords by locally normalized probabilities and is unaware of such cumulative drift. On

[†]These authors contributed equally to this work.

*Corresponding authors.

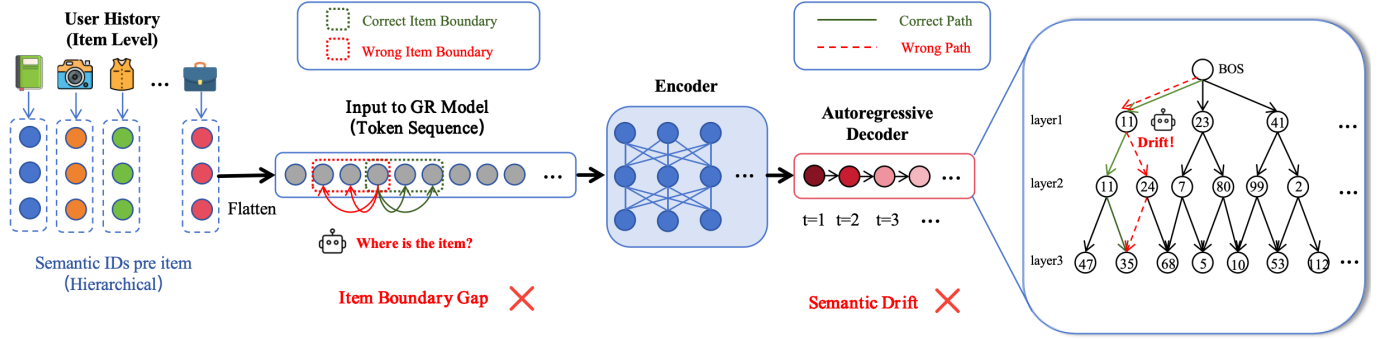


Fig. 1. Two structural gaps in generative recommendation. **Left (Item-Boundary Gap, P1):** after each item is tokenized into L hierarchical semantic IDs and flattened into a single sequence, item boundaries vanish from the input to the GR model, and the encoder can no longer tell which tokens belong to the same item (“Where is the item?”). **Right (Semantic Drift, P2):** on the hierarchical codebook tree, an error at any layer redirects decoding into a wrong subtree, so that the target leaf becomes unreachable along the chosen path no matter how the remaining tokens are scored.

TIGER, the per-layer accuracy of c_3 drops from 77.0% (correct prefix) to 0.6% (erroneous prefix), a $128\times$ gap (Section V-F). This drift can be attacked from two orthogonal angles: intra-path, by reranking candidates within a single channel via global path-level coherence; and cross-channel, by exposing the decoder to a second, structurally orthogonal quantization channel so that an item missed by one channel can still be recovered through the other via OR-style fusion.

To bridge these two gaps, we propose **BARGE** (**B**ridging **A**utoregressive **G**eneration for **r**ecommendation), with three lightweight modules: **ICA** closes the encoder-side gap (P1), while **HPR** and **DPD** jointly close the decoder-side drift gap (P2) from the two complementary angles above. The main contributions of this work are summarized as follows:

- We formalise the long-standing semantic-fidelity gap of GR as two structural gaps along the GR pipeline: an encoder-side item-boundary gap, and a decoder-side semantic-drift gap. We further show that the latter can be mitigated from two complementary, orthogonal angles.
- We propose **BARGE** with three lightweight and mutually orthogonal modules. **ICA** restores item-level structure at the encoder; **HPR** suppresses semantic drift via per-layer dual-tower contrastive reranking within a single decoding channel; and **DPD** further mitigates the same drift from a complementary angle via a Dual-Decoder coupled by an OSQ-VAE tokenizer.
- Extensive experiments demonstrate consistent improvements over strong baselines. BARGE has also been deployed on Tencent’s commercial media platform, bringing consistent improvements across core engagement metrics in real-world industrial scenarios.

II. RELATED WORK

A. Discriminative and Generative Recommendation

Mainstream sequential recommendation methods are discriminative: they encode the interaction history into a user representation and score candidate items, with architectures evolving from RNN-based GRU4Rec [15] and

CNN-based Caser [16] to Transformer-based SASRec [17], BERT4Rec [18], FDSA [19], S³-Rec [20], and more recent variants based on diffusion [21], sparse attention [22], and zero-shot transfer [23]. Despite their architectural diversity, these methods all rely on a candidate-scoring paradigm.

Generative Recommendation (GR) [24]–[27] instead formulates recommendation as autoregressive generation of discrete item identifiers [1], [2], [28]–[31]. TIGER [3] established a representative framework that encodes items into hierarchical semantic IDs via RQ-VAE [14], and subsequent work improves upon it through collaborative signals (LETTER [7]), hybrid representations (COBRA [8]), knowledge integration (MVIGER [32]), and alternative indexing schemes [33], [34]. More recently, HSTU [35], OneRec [36], [37], RPG [38], DACT [39], DELRec [40], and Reg4Rec [41] further advance GR through scalable architectures, end-to-end modeling, efficient generation, continual tokenization, knowledge distillation, and reasoning enhancement, respectively. Unlike Reg4Rec, the proposed DPD explicitly derives two complementary channels from a learnable orthogonal rotation of encoder features, providing a structural guarantee of channel diversity.

B. Structural Awareness and Semantic Drift Suppression

Structural awareness in the encoding stage. When each item is decomposed into L semantic ID tokens, the self-attention in the encoder treats all tokens as an undifferentiated flat sequence and consequently loses item-level grouping. In NLP, ALiBi [42] and Longformer [43] introduce structured attention via relative position bias and local-global patterns, but neither is tailored to semantic IDs. TrieRec [44] explores leveraging the prefix tree structure to introduce structural bias. Orthogonal to these efforts, the proposed ICA enriches every token with item-level semantics prior to encoding, via cross-attention pooling followed by a gated residual injection.

Semantic drift suppression in hierarchical codebook decoding. Hierarchical semantic drift is inherent to any GR method that decodes multi-level semantic IDs, which distinguishes

it from exposure bias in flat-vocabulary generation. Existing methods apply standard beam search without consistency-aware modification. PROMISE [45] explores applying the process reward idea from LLMs to rerank step-by-step decoding. APAO [46] addresses the gap between training and inference by introducing prefix-level pointwise and pairwise ranking losses together with an adaptive worst-prefix weighting strategy that aligns training with beam-search inference. In contrast, the proposed HPR is a lightweight, label-free auxiliary. At each decoder layer, it scores the semantic compatibility between the history-aggregated hidden state and the cumulative path embeddings via a dual-tower contrastive objective trained with symmetric InfoNCE.

III. PRELIMINARY

Let the user set be $\mathcal{U}$ and the item set be $\mathcal{V}$. For a user $u \in \mathcal{U}$, the chronologically ordered historical interaction sequence is denoted as $\mathbf{s}_u = (v_1, v_2, \dots, v_T)$, where $v_t \in \mathcal{V}$. The goal is to predict the next item v_{T+1} .

Semantic IDs via RQ-VAE. In the generative recommendation framework, each item v is assigned a discrete semantic ID of length L via Residual-Quantized VAE (RQ-VAE) [13], [14], which iteratively quantizes the residual of a pretrained item embedding across L codebooks $\{\mathcal{C}_l\}_{l=1}^L$:

$$v \rightarrow \mathbf{s}^{(v)} = (c_1^{(v)}, c_2^{(v)}, \dots, c_L^{(v)}), \quad (1)$$

where $c_l^{(v)} \in \mathcal{C}_l$ is the codeword assigned at layer l . Shallower layers capture coarse-grained semantic categories, whereas deeper layers progressively refine the representation within narrower subtrees. The codebook configuration and training details are discussed in Section V-H.

Generative recommendation. The user history is represented as a flattened token sequence $\mathbf{X} = [\mathbf{s}^{(v_1)}; \dots; \mathbf{s}^{(v_T)}] \in \mathbb{R}^{(T \cdot L) \times d}$, which is fed into an encoder [47] that produces the encoded history $\mathbf{H} \in \mathbb{R}^{(T \cdot L) \times d}$ and autoregressively generates the target semantic ID of the next item:

$$P(v_{T+1} | \mathbf{s}_u) = \prod_{l=1}^L P_l(c_l | c_{<l}, \mathbf{H}), \quad (2)$$

where $\mathbf{H}$ is the encoder output and P_l is the decoder distribution over codebook $\mathcal{C}_l$ at decoding layer l .

IV. METHODOLOGY

A. Model Overview

As illustrated in Fig. 2, BARGE addresses the two structural gaps identified in Section I with three modules: one that targets the encoder-side gap, and two complementary modules that jointly close the decoder-side drift gap. **(1) Item Context-Aware Attention (ICA)** aggregates all token embeddings within each item into a single item-level context $\mathbf{z}^{(i)}$ and fuses it back into every token via a gating network before the encoder block, which closes the item-boundary gap (P1). **(2) Hierarchical Path Reranking (HPR)** augments beam search with per-layer dual-tower contrastive scoring trained via the

symmetric InfoNCE loss ($\mathcal{L}_{\text{HPR}}$) alongside the standard next-token prediction loss ($\mathcal{L}_{\text{NTP}}$), in order to suppress the intra-path facet of decoder-side drift along a single quantization channel. **(3) Dual-Path Decoding (DPD)** replaces the single RQ-VAE tokenizer with an OSQ-VAE tokenizer. The OSQ-VAE decomposes the encoder feature into two complementary subspaces via a learnable orthogonal rotation R , and emits two semantic-ID tuples per item. On top of a shared encoder, DPD runs a Dual-Decoder with channel-private HPR, and fuses their per-channel candidate items in the item-id space at inference time. In this way, the cross-channel facet of decoder-side drift is suppressed through orthogonal quantization channels.

B. Structure-Aware Encoding via Item Context-Aware Attention (ICA)

ICA adopts an aggregate-then-fuse strategy (Fig. 2, bottom-left): for each item, it first aggregates all token embeddings into a single item-level context, and then fuses this context back into every token via a gating network.

Cross-attention pooling. Given the L token embeddings $\{\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_L^{(i)}\}$ of item v_i , ICA uses a learnable query vector $\mathbf{q} \in \mathbb{R}^d$ to compute the item-level context representation via cross-attention, in which $\mathbf{q}$ serves as the query and the tokens of the item serve as the keys and values:

$$\mathbf{z}^{(i)} = \text{LayerNorm} \left(\text{CrossAttn}(\mathbf{q}, \mathbf{X}^{(i)}, \mathbf{X}^{(i)}) \right), \quad (3)$$

where $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}; \dots; \mathbf{x}_L^{(i)}] \in \mathbb{R}^{L \times d}$ is the matrix of L token embeddings for item v_i , and $\mathbf{z}^{(i)} \in \mathbb{R}^d$ is the resulting item-level representation. Cross-attention is used here so that the learnable query $\mathbf{q}$ can adaptively and dynamically weight the contribution of each token when forming the item-level context, since different layers of a semantic ID encode relative information at different granularities.

Context projection. The context vector undergoes multi-layer nonlinear transformation:

$$\hat{\mathbf{z}}^{(i)} = W_2 \cdot \text{GELU}(W_1 \cdot \mathbf{z}^{(i)} + \mathbf{b}_1) + \mathbf{b}_2, \quad (4)$$

where $W_1 \in \mathbb{R}^{d_f \times d}$, $W_2 \in \mathbb{R}^{d \times d_f}$, d_f is the feed-forward hidden dimension, and $\hat{\mathbf{z}}^{(i)} \in \mathbb{R}^d$ is the projected context vector after the nonlinear transformation.

Gated residual fusion. For each token $l \in \{1, \dots, L\}$ of item v_i , a gating network controls how much item-level context is fused back into each token:

$$\mathbf{g}_l^{(i)} = \sigma \left(W_g \cdot [\mathbf{x}_l^{(i)} \| \hat{\mathbf{z}}^{(i)}] + \mathbf{b}_g \right), \quad (5)$$

$$\hat{\mathbf{x}}_l^{(i)} = \mathbf{x}_l^{(i)} + \mathbf{g}_l^{(i)} \odot \hat{\mathbf{z}}^{(i)}, \quad (6)$$

where $W_g \in \mathbb{R}^{d \times 2d}$, $\mathbf{g}_l^{(i)} \in \mathbb{R}^d$ is the gate vector, $\sigma(\cdot)$ is the sigmoid function, $\|$ denotes concatenation, and $\odot$ denotes element-wise multiplication. The gated residual form allows each token to decide how much item-level context to admit, so that the injected context never overwhelms the positional and layer-specific information already encoded in $\mathbf{x}_l^{(i)}$. When the gate approaches zero, ICA reduces to the identity mapping and preserves the original token representation; when the gate

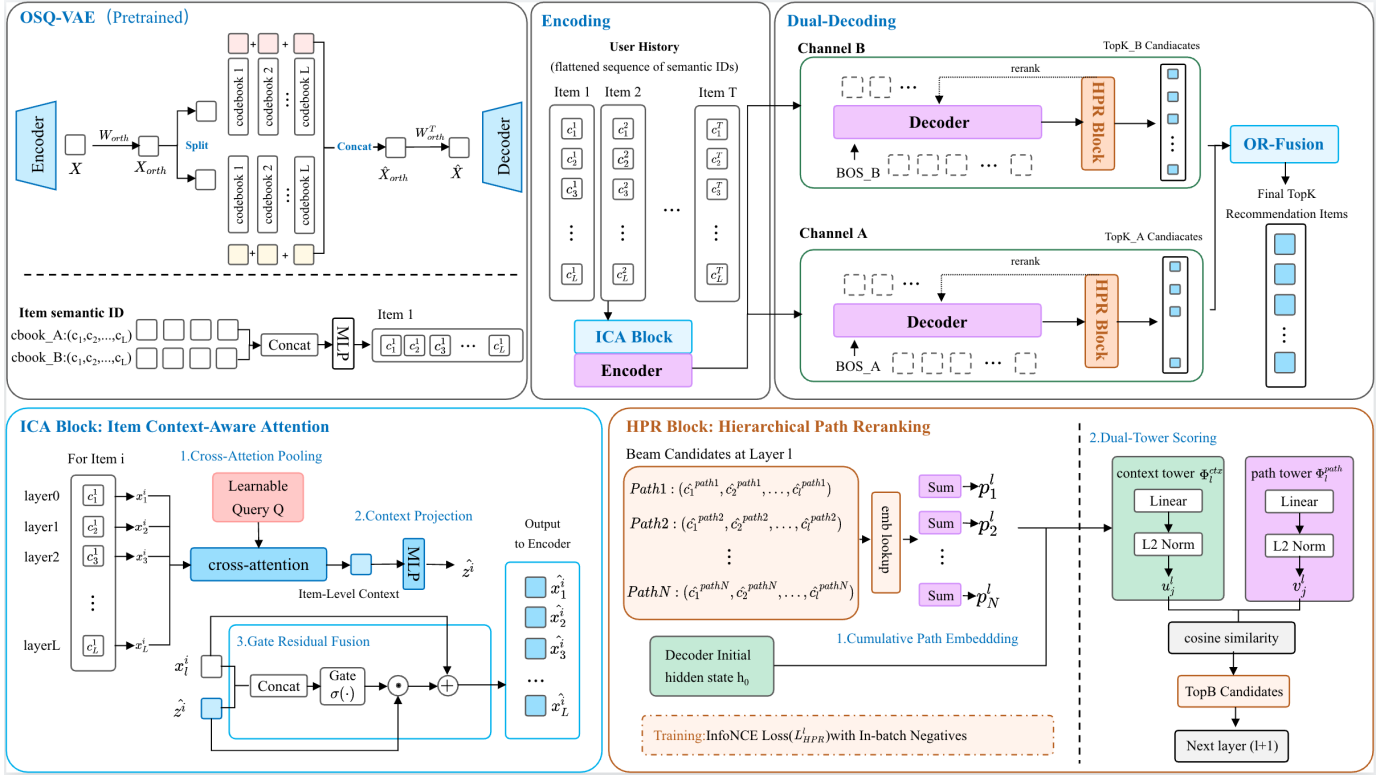


Fig. 2. Overview of BARGE. (Top-left) The pretrained OSQ-VAE tokenizes each item into two channel-specific semantic-ID tuples (cbook_A and cbook_B) via orthogonal rotation and dual codebook stacks; the dual IDs are projected by an MLP into the flattened token sequence. (Top-middle) The user history passes through the ICA before the shared encoder. (Top-right) Two independent decoders (Channel A and Channel B) each followed by an HPR run in parallel; their Top-K candidate lists are merged by OR-Fusion to produce the final recommendations. (Bottom-left) ICA: cross-attention pooling with a learnable query, MLP projection, and gated residual fusion. (Bottom-right) HPR: per-layer dual-tower scoring of cumulative path embeddings against the decoder hidden state h_0 , trained with $\mathcal{L}_{HPR}$.

approaches one, the token is fully merged with the item context. We empirically observe that the learned gate stabilises around 0.35–0.38 across all four layers (Section V-I), which confirms that the network indeed learns a moderate and layer-consistent fusion strength.

C. Hierarchical Path Reranking (HPR)

HPR augments beam search with a per-layer reranking mechanism that evaluates the semantic compatibility between the initial hidden state of the decoder h_0 and the cumulative path embedding at each decoding layer (Fig. 2, bottom-right). The key insight is that h_0 , produced by the cross-attention of the decoder over the full encoder output before any token is generated, serves as a holistic representation of the historical preferences of the user, which makes it a natural and suitable anchor for assessing whether a candidate path aligns with the intent of the user.

Cumulative path embedding. Unlike token-level scoring that evaluates individual codewords in isolation, HPR operates explicitly at the path level to better capture dependencies between different layers. For each candidate at layer l , the

cumulative path embedding is computed as

$$\mathbf{p}^{(l)} = \sum_{j=1}^l \mathbf{e}_{c_j}, \quad (7)$$

where $\mathbf{e}_{c_j} \in \mathbb{R}^{d_{\text{emb}}}$ is the learnable embedding of codeword c_j . This cumulative representation captures the semantic trajectory of the partial path and enables the reranker to assess global coherence rather than evaluating each token independently.

Per-layer dual-tower scoring. As illustrated in Fig. 2 (bottom-right), HPR maintains a separate dual-tower scorer for each layer l . Each scorer projects the initial hidden state of the decoder and the cumulative path embedding into a shared low-dimensional space, and then computes their cosine similarity scaled by a learnable temperature:

$$r_l(\mathbf{h}_0, \mathbf{p}^{(l)}) = \cos(\phi_l^{\text{ctx}}(\mathbf{h}_0), \phi_l^{\text{path}}(\mathbf{p}^{(l)})) \cdot e^{\tau_l}, \quad (8)$$

where $\phi_l^{\text{ctx}}: \mathbb{R}^{d_{\text{attn}}} \rightarrow \mathbb{R}^{d_{\text{proj}}}$ and $\phi_l^{\text{path}}: \mathbb{R}^{d_{\text{emb}}} \rightarrow \mathbb{R}^{d_{\text{proj}}}$ are layer-specific linear projections followed by L_2 normalisation, and τ_l is the learnable log-temperature. The dual-tower architecture allows efficient scoring of multiple candidate paths, as the context projection is computed only once per sample.

Training via symmetric InfoNCE. HPR is jointly trained with the main model using the InfoNCE loss. For each layer l ,

within a batch of B samples, the positive pair consists of the initial hidden state of the decoder $\mathbf{h}_0^{(i)}$ and the ground-truth cumulative path embedding $\mathbf{p}^{(l,i)}$ for sample i . In addition to the in-batch negatives, we incorporate prefix-aware negatives that explicitly expose the model to plausible but erroneous prefixes. Concretely, at each decoding step, we draw high-probability candidates that are not ground truth from the NTP distribution, and further introduce business-level negatives such as impressed but unclicked items. These prefix-conditioned negatives emulate the drift patterns encountered during inference and strengthen the model’s ability to discriminate correct semantic prefixes from misleading or user-irrelevant ones. The per-layer loss is defined as

$$\mathcal{L}_{\text{HPR}}^{(l)} = \frac{1}{2} \left[\mathcal{L}_{\text{c2p}}^{(l)} + \mathcal{L}_{\text{p2c}}^{(l)} \right], \quad (9)$$

where $\mathcal{L}_{\text{c2p}}^{(l)}$ (context-to-path) and $\mathcal{L}_{\text{p2c}}^{(l)}$ (path-to-context) are standard cross-entropy losses over the similarity matrix. Taking $\mathcal{L}_{\text{c2p}}^{(l)}$ as an example,

$$\mathcal{L}_{\text{c2p}}^{(l)} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(r_l^{(i,i)})}{\sum_{j=1}^B \exp(r_l^{(i,j)})}, \quad (10)$$

where $r_l^{(i,j)} = r_l(\mathbf{h}_0^{(i)}, \mathbf{p}^{(l,j)})$ denotes the reranker score between the i -th context and the j -th path. The symmetric formulation ensures that both projection heads are well-calibrated: the context tower learns to produce representations that are discriminative across different paths, while the path tower learns representations that are discriminative across different user contexts. The total reranker loss is averaged across all L layers as $\mathcal{L}_{\text{HPR}} = \frac{1}{L} \sum_{l=1}^L \mathcal{L}_{\text{HPR}}^{(l)}$.

Joint scoring during inference. Standard beam search keeps a set of B partial paths and, at layer l , expands them into $B \times |\mathcal{C}_l|$ candidates scored solely by their cumulative log-probability; the top- B are kept for the next layer. Such greedy local selection ignores global semantic coherence, and simply enlarging B is costly, since the KV cache and the expanded score tensor all scale linearly with B . We therefore follow prior generative works (e.g., $B = 20$ in TIGER [3]) and instead correct erroneous paths without enlarging the beam.

HPR intervenes after beam expansion at each layer. From the $B \times |\mathcal{C}_l|$ expanded candidates, we first keep the top- N (with $B < N \ll B \times |\mathcal{C}_l|$) ranked by generation log-probability to form a scoring pool, which retains semantically reasonable but locally lower-probability paths that vanilla beam search would prematurely discard. Subsequently, each candidate in the pool is rescored by fusing its original generation log-probability with the score from the HPR reranker.

$$\begin{aligned} \text{score}(c, l) &= \log p(c \mid c_{<l}, \mathbf{s}_u) \\ &\quad + \lambda \cdot \log \text{softmax}\left(r_l(\mathbf{h}_0, \mathbf{p}^{(l)})\right), \end{aligned} \quad (11)$$

where c is a candidate codeword at layer l , $\mathbf{p}^{(l)}$ is the cumulative path embedding obtained by appending c to $c_{<l}$, $p(c \mid c_{<l}, \mathbf{s}_u)$ is the softmax generation probability and $\lambda \geq 0$ controls the reranking strength ($\lambda = 0$ recovers vanilla beam

search). The top- B paths under this fused score advance to layer $l+1$, so the outgoing beam width is unchanged and only a lightweight dual-tower scoring is added per layer, while the enlarged pool N allows semantically inconsistent paths to be corrected before their errors propagate deeper.

D. Dual-Path Decoding (DPD)

ICA and HPR both operate within a single quantization channel, and neither addresses the cross-channel facet. A single RQ-VAE structure commits the decoder to one factorization of an item, projecting its rich semantics onto a single quantization axis; any semantic facet of the item not captured by that axis is permanently locked outside the corresponding semantic subtree. DPD tries to close this gap with three components: an OSQ-VAE tokenizer, a Dual-Decoder, and an OR-fusion inference procedure.

OSQ-VAE (Orthogonal Split-and-Quantize VAE) tokenizer. Let $\mathbf{z} \in \mathbb{R}^D$ be the pretrained item embedding. The OSQ-VAE tokenizer first applies a learnable orthogonal rotation

$$\tilde{\mathbf{z}} = R\mathbf{z}, \quad R \in \mathbb{R}^{D \times D}, \quad R^\top R = I_D, \quad (12)$$

in which R is parameterized as a product of Householder reflections, so that $R^\top R = I_D$ holds by construction throughout training without any auxiliary loss term. The rotated feature is then split into two equal halves

$$\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}^{(A)} \parallel \tilde{\mathbf{z}}^{(B)}], \quad \tilde{\mathbf{z}}^{(A)}, \tilde{\mathbf{z}}^{(B)} \in \mathbb{R}^{D/2}, \quad (13)$$

and each half is quantized by an independent L -layer residual codebook stack, which yields per-channel semantic-ID tuples $\mathbf{s}^{(c)} = (c_1^{(c)}, \dots, c_L^{(c)})$ for $c \in \{A, B\}$. Because $R^\top R = I_D$ and the split is coordinate-aligned in the rotated frame, the supporting subspaces of the two channels satisfy $S_A \perp S_B$ and $S_A \oplus S_B = \mathbb{R}^D$, which is a hard architectural invariant. The OSQ-VAE is trained end-to-end with the standard reconstruction and commitment losses applied independently to each channel:

$$\begin{aligned} \mathcal{L}_{\text{OSQ}} &= \underbrace{\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2}_{\mathcal{L}_{\text{recon}}} + \sum_{c \in \{A, B\}} \underbrace{\|\text{sg}[\tilde{\mathbf{z}}^{(c)}] - \hat{\tilde{\mathbf{z}}}^{(c)}\|_2^2}_{\text{codebook loss}} \\ &\quad + \beta \sum_{c \in \{A, B\}} \underbrace{\|\tilde{\mathbf{z}}^{(c)} - \text{sg}[\hat{\tilde{\mathbf{z}}}^{(c)}]\|_2^2}_{\text{commitment loss}}, \end{aligned} \quad (14)$$

where $\hat{\mathbf{z}} = R^\top [\hat{\tilde{\mathbf{z}}}^{(A)} \parallel \hat{\tilde{\mathbf{z}}}^{(B)}]$ is the rotated-back reconstruction, $\text{sg}[\cdot]$ denotes the stop-gradient operator, and β is the commitment weight. The first term $\mathcal{L}_{\text{recon}}$ is the global reconstruction loss between the encoder output $\mathbf{z}$ and the dequantized reconstruction $\hat{\mathbf{z}}$. Inside the summation, the two per-channel terms differ only in the direction of stop-gradient: the codebook loss freezes the encoder output $\tilde{\mathbf{z}}^{(c)}$ and pulls the codebook embedding $\hat{\tilde{\mathbf{z}}}^{(c)}$ towards it, while the commitment loss freezes the codebook embedding and forces the encoder output to commit to it, weighted by β .

Dual-Decoder. Given a user history, BARGE encodes it once through the shared ICA-enriched encoder. Two decoder towers, $\text{Dec}^{(A)}$ and $\text{Dec}^{(B)}$, then run in parallel on top of the shared encoder output. Each tower has its own input

projection, per-layer output heads tied to its channel-specific codebooks, and its own HPR scorer with channel-private projection heads $(\phi_l^{\text{ctx},(c)}, \phi_l^{\text{path},(c)})$. DPD follows a two-stage training procedure. In the first stage, the OSQ-VAE is pre-trained offline via $\mathcal{L}_{\text{OSQ}}$ (Eq. (14)) to produce two sets of channel-specific semantic IDs $\{s^{(A)}\}$ and $\{s^{(B)}\}$ for all items, after which the weights of the OSQ-VAE are frozen. In the second stage, the Dual-Decoder is trained on top of the fixed semantic IDs. Decoder $\text{Dec}^{(c)}$ is trained to autoregressively predict the channel- c semantic ID:

$$P(v_{T+1}^{(c)} | \mathbf{s}_u) = \prod_{l=1}^L P_l^{(c)}(c_l^{(c)} | c_{<l}^{(c)}, \mathbf{H}). \quad (15)$$

The overall training loss sums the next-token prediction loss and the HPR contrastive loss over both channels:

$$\mathcal{L}_{\text{total}} = \sum_{c \in \{A, B\}} (\mathcal{L}_{\text{NTP}}^{(c)} + \mathcal{L}_{\text{HPR}}^{(c)}). \quad (16)$$

Because the two towers do not share the decoder parameters or the HPR scorers, each tower specializes in the codebook statistics of its own channel, while the shared encoder benefits from gradients flowing in from both sides.

Inference: OR-fusion in the item-id space. At inference time, each decoder tower runs an independent beam search of width B and produces a ranked list of channel-specific semantic IDs. These IDs are then mapped back to items via the OSQ-VAE codebook of the corresponding channel. Because each item carries two semantic IDs (one per channel), the two ranked lists live in the same item-id space and are merged by an OR-fusion operator:

$$s(v) = f(s^{(A)}(v), s^{(B)}(v)), \quad (17)$$

where $s^{(c)}(v)$ denotes the score of item v under channel c . An item is recovered as long as at least one channel ranks it highly, while items absent from both beams are uniformly rejected. This OR semantics directly attacks the cross-channel facet of semantic drift, and its effectiveness relies on the two channels proposing complementary top- K candidates rather than on any postulated independence between them. We emphasize that OR-fusion does not enlarge the candidate budget: each tower still runs a beam of width B identical to that of the single-tower baselines, and the merged ranked item list from which $\text{Recall}@K$ and $\text{NDCG}@K$ are computed is truncated at the same K , so all reported metrics are evaluated under exactly the same protocol. We empirically validate this complementarity together with the choice of fusion function f in Section V-G.

E. Design Rationale and Verifiable Conditions

We complement the design of BARGE with a module-wise analysis. For each module, we derive a condition that characterizes when it improves recommendation accuracy, and link that condition to the empirical section in which it is directly measured.

ICA: identity-preserving property. Since $\hat{\mathbf{x}}_l^{(i)} = \mathbf{x}_l^{(i)} + \mathbf{g}_l^{(i)} \odot \hat{\mathbf{z}}^{(i)}$ with $\mathbf{g}_l^{(i)} \in [0, 1]^d$, we have $\|\hat{\mathbf{x}}_l^{(i)} - \mathbf{x}_l^{(i)}\|_2 \leq \|\hat{\mathbf{z}}^{(i)}\|_2$, and

the gate can shrink to $\mathbf{0}$ to recover the vanilla encoder. ICA therefore augments rather than overwrites the original signal.

HPR: decomposing per-layer errors. At layer l , let g_l be the autoregressive likelihood, r_l the HPR reranker score, and $\pi_l \propto g_l \cdot r_l^\lambda$ the fused score used by HPR. Denote the per-layer miss events of vanilla beam search and of HPR by $\varepsilon_l^{\text{van}} = \Pr[c_l^* \notin \text{Top}_B(g_l)]$ and $\varepsilon_l^{\text{HPR}} = \Pr[c_l^* \notin \text{Top}_B(\pi_l)]$, respectively. Reranking with π_l moves the boundary of the surviving top- B in two directions: it can pull a previously discarded ground-truth codeword into the top- B (*rescue*), but it can also push a previously surviving ground-truth codeword out of it (*damage*). Formally, define

$$\begin{aligned} \text{Rescue}_l &\triangleq \{c_l^* \in \text{Top}_B(\pi_l) \setminus \text{Top}_B(g_l)\}, \\ \text{Damage}_l &\triangleq \{c_l^* \in \text{Top}_B(g_l) \setminus \text{Top}_B(\pi_l)\}. \end{aligned}$$

Then the following identity holds without any assumption on r_l :

$$\varepsilon_l^{\text{van}} - \varepsilon_l^{\text{HPR}} = \Pr[\text{Rescue}_l] - \Pr[\text{Damage}_l]. \quad (18)$$

Eq. (18) converts the question “when does HPR help?” into a directly measurable one: HPR is net beneficial at layer l iff $\Pr[\text{Rescue}_l] > \Pr[\text{Damage}_l]$. We do not assume this inequality a priori. Two pieces of evidence already in the paper are consistent with it. First, the InfoNCE training of r_l maximises a lower bound on the mutual information between the user context and the cumulative path [48], which biases Rescue_l above Damage_l for ground-truth-aligned paths. Second, the inverted-U behavior of λ in Section V-G (Fig. 3) is the predicted shape under Eq. (18): at small λ the reranker barely perturbs the boundary, so $\Pr[\text{Damage}_l] \approx 0$ and any non-zero rescue translates into a gain; at overlarge λ the reranker overrides the likelihood and inflates $\Pr[\text{Damage}_l]$, eventually erasing the gain.

DPD: an OR-fusion gain identity. Let $E^{(A)}$ and $E^{(B)}$ denote the events that channel A and channel B respectively miss the ground-truth item in their per-channel top- K list, and let $\kappa \triangleq \Pr[E^{(B)} | E^{(A)}]$ denote the conditional miss rate of channel B given that channel A already misses. OR-fusion misses the ground truth iff both channels miss, so the gain of OR-fusion over channel A is $\Pr[E^{(A)}] - \Pr[E^{(A)} \cap E^{(B)}]$. Applying the chain rule of probability gives the *identity*

$$\underbrace{\Pr[E^{(A)}] - \Pr[E^{(A)} \cap E^{(B)}]}_{\text{OR-fusion gain over channel A}} = (1 - \kappa) \cdot \Pr[E^{(A)}]. \quad (19)$$

Eq. (19) makes no independence assumption and turns the design question of DPD into a measurable one: how small is κ in practice? The orthogonal rotation R in OSQ-VAE ($R^\top R = I$, $S_A \perp S_B$) is precisely what drives κ down. We verify $\kappa < 1$ directly in Section V-G: the Jaccard overlap between the two top- K pools is only 0.18 on Beauty and 0.17 on Sports, and 15–24% of all OR-fusion top- K hits come from a single channel, both implying a non-trivial $1 - \kappa$ and explaining the gains in Table II.

Stacking along orthogonal failure dimensions. Eq. (18) and Eq. (19) act on disjoint failure modes: HPR rescues a ground-truth codeword that is still reachable within a single channel,

whereas DPD recovers an item that is no longer reachable in one channel by exposing it through the other. Because the two mechanisms reduce different terms of the overall miss probability, we expect their gains to be largely additive, which is consistent with the component-wise ablation in Section V-D.

V. EXPERIMENT

A. Experimental Setup

Datasets. We evaluate the proposed method on three datasets: (1) Amazon Beauty and (2) Amazon Sports and Outdoors, two widely used sequential recommendation benchmarks with 5-core filtering [49]; and (3) a large-scale offline test from Tencent’s commercial media platform that comprises millions of users and hundreds of millions of interactions over 11 days (the first 10 days are used for training and the last day for evaluation). Statistics of the Amazon datasets are summarized in Table I. For the Amazon datasets, we follow the preprocessing of P5 [2] and adopt the leave-one-out evaluation protocol.

TABLE I
STATISTICS OF THE AMAZON DATASETS.

Dataset	#Users	#Items	#Interactions	Sparsity
Amazon Beauty	22,363	12,101	198,502	99.93%
Amazon Sports	25,598	18,357	296,337	99.95%

Metrics. All datasets are evaluated with Recall@K and NDCG@K, with $K \in \{5, 10\}$. All metrics are computed on the full item set.

Baselines. We compare the proposed method against two groups of competitors.

Non-generative sequential recommenders:

- **P5** [2]: reformulates recommendation tasks as text-to-text generation under a shared language model.
- **Caser** [16]: treats the recent interaction history as an image and applies horizontal and vertical convolutions to capture sequential patterns.
- **HGN** [50]: models long- and short-term user interests through hierarchical gating networks.
- **GRU4Rec** [15]: adopts a GRU-based recurrent network to encode the interaction sequence.
- **BERT4Rec** [18]: pretrains a bidirectional Transformer with a masked-item prediction objective.
- **FDSA** [19]: introduces a feature-level self-attention block that complements item-level attention.
- **SASRec** [17]: uses a unidirectional self-attention network for next-item prediction.
- **S³-Rec** [20]: pretrains a sequential encoder with mutual-information-based self-supervision over items, attributes, and segments.

Generative recommenders:

- **TIGER** [3]: represents each item by an RQ-VAE-based hierarchical semantic ID and generates the target ID autoregressively.

- **HSTU** [35]: reformulates recommendation as a generative modeling task over user actions and proposes the Hierarchical Sequential Transduction Unit, a linear-complexity architecture that scales generative recommendation to industrial settings.
- **COBRA** [8]: couples sparse semantic IDs with dense embeddings and retrieves candidates via approximate nearest neighbour search at inference time.
- **APAO-pointwise** [46]: introduces a prefix-aware optimization objective to mitigate the training–inference mismatch of beam search. We adopt the pointwise variant rather than the pairwise one because the pairwise mode incurs a prohibitive inference cost, and the original paper of APAO also reports that its production deployment uses the pointwise mode.
- **ActionPiece** [5]: tokenizes user actions with a context-aware sub-action vocabulary that adapts the granularity of generation.

The baseline results on the Amazon datasets are taken from [5], [8] under the same evaluation protocol, and the results of HSTU and ActionPiece are directly sourced from [5].

Implementation details. All experiments are conducted on $2 \times$ NVIDIA H20 GPUs. We adopt a semantic ID depth of $L = 4$ with a layer-wise decreasing codebook configuration $(|\mathcal{C}_1|, |\mathcal{C}_2|, |\mathcal{C}_3|, |\mathcal{C}_4|) = (512, 256, 128, 64)$ inspired by PLUM [51]. All four layers are learned semantic codebooks. The codebook configuration analysis is presented in Section V-H. The encoder and each of the two DPD decoder towers share the same Transformer configuration with 2 layers and 4 attention heads. The embedding dimension is 128, the attention dimension is 512, and the FFN hidden size is 1024. The model is trained for 200 epochs at a batch size of 256, using the Adam optimizer with warmup. Early stopping is applied based on the validation performance, and all results are averaged over 3 random seeds. In addition to the full BARGE, we also report **BARGE-base**, a variant that keeps all model components but replaces our 4-layer learned codebook with TIGER’s 3-layer codebook plus a random collision-resolving ID, so as to isolate the gain of the three structural modules from that of our codebook design.

B. Overall Performance Comparison

Table II reports the overall performance of BARGE against discriminative and generative baselines on the two datasets. We summarize the key observations below:

- **Overall superiority of BARGE.** BARGE achieves the best score on every metric across both datasets, with representative gains of +19.6% R@10 on Beauty and +8.8%/+16.7% R@10/N@10 on Sports over the strongest baseline. The gains are uniform over K and over both recall and ranking metrics, indicating a structural improvement rather than a mere shift of operating point.

- **Advantage over discriminative recommenders.** Discriminative baselines trail the generative family by a clear margin, and BARGE further enlarges this gap. Representing each item

TABLE II
OVERALL PERFORMANCE COMPARISON ON TWO AMAZON DATASETS. BEST RESULT IN BOLD; THE STRONGEST BASELINE IS UNDERLINED.

Methods		Beauty				Sports and Outdoors			
		R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
Non-Generative	P5	0.0163	0.0107	0.0254	0.0136	0.0061	0.0041	0.0095	0.0052
	Caser	0.0205	0.0131	0.0347	0.0176	0.0116	0.0072	0.0194	0.0097
	HGN	0.0325	0.0206	0.0512	0.0266	0.0189	0.0120	0.0313	0.0159
	GRU4Rec	0.0164	0.0099	0.0283	0.0137	0.0129	0.0086	0.0204	0.0110
	BERT4Rec	0.0203	0.0124	0.0347	0.0170	0.0115	0.0075	0.0191	0.0099
	FDSA	0.0267	0.0163	0.0407	0.0208	0.0182	0.0122	0.0288	0.0156
	SASRec	0.0337	0.0225	0.0536	0.0289	0.0177	0.0111	0.0309	0.0153
	S ³ -Rec	0.0359	0.0218	0.0613	0.0299	0.0251	0.0161	0.0385	0.0204
Generative	TIGER	0.0454	0.0321	0.0648	0.0384	0.0264	0.0181	0.0400	0.0225
	COBRA	0.0537	0.0395	0.0725	0.0456	0.0305	0.0215	0.0434	0.0257
	HSTU	0.0469	0.0314	0.0704	0.0389	0.0258	0.0165	0.0414	0.0215
	ActionPiece	0.0511	0.0340	0.0775	0.0424	0.0316	0.0205	0.0500	0.0264
	APAO-pointwise	0.0530	0.0368	0.0795	0.0453	0.0283	0.0186	0.0444	0.0237
	BARGE-base	0.0598	0.0420	0.0896	0.0515	0.0337	0.0229	0.0513	0.0285
	BARGE	0.0654	0.0460	0.0927	0.0547	0.0369	0.0252	0.0544	0.0308

with a single ID embedding forces items with related but non-identical semantics to be aligned through interaction signals alone, which is fragile under data sparsity. BARGE instead encodes each item as a hierarchical semantic ID and predicts it layer by layer, so coarse-grained aspects such as category and brand are shared across related items and only the fine-grained distinctions need to be learned from behavior. This yields stronger few-shot generalization on cold and tail items.

- **Advantage over generative recommenders.** Compared with prior generative recommenders, BARGE consistently ranks first, and we attribute the improvement to a more structured treatment of the hierarchical semantic ID on both the encoder and the decoder sides. Existing methods are trained purely under a per-token likelihood objective, so shallow-layer errors are not penalised beyond their own step and tend to cascade into deeper layers. In addition, they commit to a single deterministic decoding path, which makes off-path items effectively unreachable no matter how sharp the token-level distribution is. Remedies such as attaching an extra ANN retrieval stage or refining the training loss only address part of the picture, and often reintroduce the full-catalogue cost and other overheads that a purely generative paradigm was meant to remove. BARGE tackles these two weaknesses jointly through three complementary modules.

- **Disentangling the codebook from the model design.** BARGE-base already surpasses every prior generative baseline on both datasets, isolating the contribution of ICA, HPR, and DPD from that of our codebook design. The remaining gap to the full BARGE shows that the 4-layer codebook in Section V-H is complementary to the structural modules rather than a substitute for them.

On the Tencent commercial media platform offline test (Table III), BARGE consistently outperforms all baselines. The two deployed baselines Graph and NANN [52] both operate on atomic item embeddings and cannot express the shared coarse-grained structure among semantically related items, which is especially costly on a catalogue with hundreds of thousands of

TABLE III
PERFORMANCE COMPARISON ON THE TENCENT COMMERCIAL MEDIA PLATFORM OFFLINE TEST. * DEPLOYED INDUSTRIAL BASELINE.

	Hit@5	Hit@10	Hit@20	Hit@50
Graph*	0.2932	0.3743	0.4650	0.5951
NANN*	0.4416	0.4946	0.5636	0.6760
OneRec	<u>0.5459</u>	<u>0.6132</u>	<u>0.6729</u>	<u>0.7348</u>
BARGE	0.6015	0.6510	0.6967	0.7520

long-tail items. OneRec narrows this gap through a generative formulation over semantic IDs, but its single-path token-level objective still leaves shallow-layer errors and off-path items as the dominant failure modes. BARGE keeps the generative backbone and repairs these two weaknesses, with the margin most pronounced at the top of the ranked list. This advantage also carries over to a catalogue orders of magnitude larger than the academic benchmarks. Such scalability suggests that path-level supervision and dual-path decoding grow in importance as the codebook has to cover a more heterogeneous item population in real-world recommendation scenarios.

TABLE IV
EFFICIENCY COMPARISON BETWEEN TIGER AND BARGE.

	Params	Train (s/epoch)	Infer (s/epoch)
TIGER	22.71 M	22	17
BARGE	19.91 M	24	18

C. Efficiency Analysis

To assess whether ICA, HPR, and DPD introduce non-trivial overhead, we compare BARGE against TIGER on Amazon Beauty under the same training and inference configuration. As shown in Table IV, although BARGE adds three structural modules on top of the backbone, its total parameter count is comparable to TIGER. This is because we use a 2-layer encoder rather than the 4-layer encoder in TIGER: the saving

on the encoder side absorbs the extra parameters of ICA, the HPR scorers, and the DPD dual-decoder, so that the introduced structural changes do not translate into a noticeable increase in model size. The added per-epoch cost is also modest, since the two DPD towers share the same encoder and run in parallel. Overall, BARGE delivers its accuracy gains without inflating model size or wall-clock cost.

D. Ablation Study

We conduct ablation studies along three complementary axes, all reported in Table V. The first axis (*Component*) progressively adds one of the three modules, in order to isolate the individual contribution of ICA, HPR, and DPD. The second axis (*OR-fusion function*) fixes the full BARGE backbone and varies the fusion operator $f(\cdot, \cdot)$ in Eq. (17), in order to verify that the OR semantics, rather than a particular implementation of it, is what drives the gain of DPD. We compare four representative instantiations: (1) **LSE**, $f = \log(\exp s^{(A)} + \exp s^{(B)})$, a soft-OR that smoothly favors the channel with the higher score; (2) **Max**, $f = \max(s^{(A)}, s^{(B)})$, which most directly implements the OR semantics by trusting whichever channel ranks the item higher; (3) **Mean**, $f = \frac{1}{2}(s^{(A)} + s^{(B)})$, an AND-style average that requires both channels to agree; and (4) **RRF**, the reciprocal-rank fusion $f = \frac{1}{k + \text{rank}^{(A)}(v)} + \frac{1}{k + \text{rank}^{(B)}(v)}$, which is rank-based. The third axis (*Rotation*) replaces the learned orthogonal rotation R in DPD with a fixed random orthogonal matrix, in order to verify that the gain of DPD comes from *learning* a task-specific orthogonal decomposition rather than merely from any orthogonal split.

• **Component-level results.** All three single-module variants outperform TIGER, confirming that each module delivers gains independently. Among them, BARGE w/ DPD attains the strongest single-module performance, as decoupling an item into two orthogonal channels directly enlarges the recoverable semantic coverage. BARGE w/ HPR ranks second, showing that path-level reranking effectively suppresses intra-path drift even within a single channel, while BARGE w/ ICA confirms the value of restoring item-level structure at the encoder. The full BARGE further surpasses all single-module variants, demonstrating the complementarity of the three modules: ICA, HPR, and DPD address non-overlapping failure sources at the encoder, the path level, and across channels, respectively.

• **OR-fusion function results.** LSE attains the best results across both datasets and all four metrics, validating the soft-OR formulation as the default choice for f . The three OR-style operators (LSE, Max, RRF) all clearly outperform the AND-style Mean, corroborating the design principle of DPD: an item should be recovered as long as at least one channel ranks it highly. LSE consistently dominates Max because Max relies on a hard $\arg \max$ that is sensitive to score-scale differences between channels, whereas LSE smoothly aggregates both sides and remains robust when the two scores are close. RRF matches LSE on Beauty but slightly trails on Sports, suggesting that score-based soft-OR carries finer-

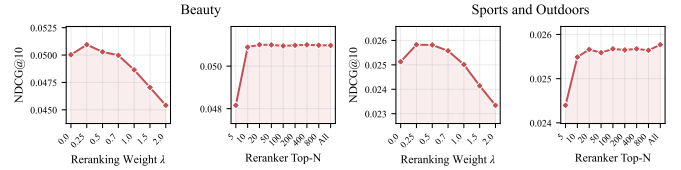


Fig. 3. Sensitivity of HPR hyperparameters (NDCG@10).

grained signal than rank-based fusion when channel scores are well-calibrated by the per-channel HPR.

• **Rotation results.** Replacing the learned R with a frozen random orthogonal matrix consistently degrades performance on both datasets, since a random R splits the embedding space along directions unrelated to the recommendation objective. This confirms that the gain of DPD stems from *learning* a task-aware orthogonal decomposition, not from any arbitrary orthogonal split.

E. Hyperparameter Sensitivity Analysis

We study the sensitivity of two key HPR hyperparameters: the reranking weight λ and the reranker Top-N.

Reranking weight λ . As shown in Fig. 3 (left two panels), we vary $\lambda \in \{0.0, 0.25, 0.5, 0.7, 1.0, 1.5, 2.0\}$. The performance exhibits an inverted-U pattern: a moderate λ balances the generation likelihood (which captures local token-level fluency) and the reranking signal (which captures global path-level semantic coherence), while an excessively large λ causes the reranker to dominate and override valid generation probabilities. We select $\lambda = 0.25$.

Reranker Top-N. As shown in Fig. 3 (right two panels), the performance improves rapidly as Top-N increases from small values, and then plateaus beyond approximately 400. This is expected: once the scoring pool is large enough to include the correct path with high probability, further expansion introduces only low-scoring candidates that do not affect the final ranking. We select Top-N = 400.

F. Analysis of Hierarchical Semantic Drift

To understand how prefix errors propagate across decoding layers, we compare the prediction quality on the TIGER baseline under two modes: Teacher Forcing (TF), which feeds the ground-truth prefix to the decoder, and Autoregressive (AR), which uses the model-selected prefix as in standard inference. The experiments are conducted on the Beauty test set with greedy decoding. The TF-AR gap directly measures the damage caused by prefix errors.

As shown in Table VI, the deepest layer c_3 achieves a target probability of 0.787 under TF but collapses to 0.015 under AR, a $52\times$ drop driven entirely by the propagation of prefix errors. The bottom block further isolates the cascade mechanism: at c_3 , the per-layer accuracy is 77.0% when the AR prefix is entirely correct but drops to 0.6% when prefix errors exist, yielding a $128\times$ gap that is dominated by prefix correctness rather than by the capacity of the layer itself. While beam search ($B > 1$) partially mitigates the issue

TABLE V
ABLATION STUDY ON THE TWO AMAZON DATASETS.

Variant		Beauty				Sports and Outdoors			
		R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
Component	BARGE (full)	0.0654	0.0460	0.0927	0.0547	0.0369	0.0252	0.0544	0.0308
	BARGE w/ ICA	0.0577	0.0393	0.0859	0.0483	0.0276	0.0181	0.0446	0.0235
	BARGE w/ HPR	0.0592	0.0413	0.0864	0.0500	0.0304	0.0202	0.0481	0.0258
	BARGE w/ DPD	0.0629	0.0437	0.0913	0.0529	0.0350	0.0237	0.0527	0.0294
OR-fusion function	LSE	0.0654	0.0460	0.0927	0.0547	0.0369	0.0252	0.0544	0.0308
	Max	0.0634	0.0451	0.0913	0.0540	0.0351	0.0239	0.0528	0.0296
	Mean	0.0621	0.0436	0.0888	0.0522	0.0341	0.0229	0.0526	0.0288
	RRF	0.0654	0.0451	0.0931	0.0540	0.0360	0.0245	0.0538	0.0303
Rotation	Random R (frozen)	0.0591	0.0420	0.0845	0.0502	0.0309	0.0213	0.0472	0.0265

by maintaining multiple candidate paths, it does not address the root cause: without any mechanism to assess cross-layer semantic coherence, erroneous paths continue to propagate, and enlarging B is impractical in production because inference latency scales linearly with B while GPU memory grows as $B \times L$. These observations motivate HPR as a principled reranking mechanism that corrects semantically inconsistent paths without relying on brute-force beam expansion.

TABLE VI
PER-LAYER PREDICTION QUALITY UNDER TEACHER FORCING (TF) AND AUTOREGRESSIVE (AR) MODES ON AMAZON BEAUTY. THE LAST TWO ROWS SPLIT THE AR RESULTS AT c_3 BY WHETHER THE AUTOREGRESSIVE PREFIX IS ENTIRELY CORRECT.

Layer / Mode	Mismatch $\downarrow$	Rank $\downarrow$	Prob $\uparrow$
c_1 (TF / AR)	0.915	66.9	0.046
c_2 (TF)	0.864	25.3	0.102
c_2 (AR)	0.981	108.2	0.015
c_3 (TF)	0.195	7.4	0.787
c_3 (AR)	0.984	119.5	0.015
c_3 (AR, prefix correct)	0.230	3.7	0.738
c_3 (AR, prefix error)	0.994	121.1	0.006

G. DPD Analysis

We verify two properties of DPD that justify its design without relying on idealised assumptions: (i) the learnable rotation R is used in a non-trivial way, and (ii) the two channels propose empirically complementary top- K candidate sets, so that a non-trivial fraction of OR-fusion hits is contributed exclusively by a single channel.

Diagnostic of the learned rotation R . We probe the trained R on Beauty and Sports along two axes: a structural test (how exactly orthogonality holds and how far R deviates from I_D) and a functional test (replacing R with I_D at inference time), as summarized in Table VII.

The Householder parameterisation enforces $R^\top R = I_D$ to numerical precision ($\sim 10^{-6}$), while $\|R - I\|_F / \sqrt{D} \approx 1.2\text{--}1.4$ confirms that R moves substantially away from the identity. The trained R consistently reduces the reconstruction loss by 0.04–0.05 relative to the $R := I_D$ baseline on both datasets,

TABLE VII
DIAGNOSTIC OF THE LEARNED ROTATION R ($D=32$).

	Beauty	Sports
$\ R - I\ _F / \sqrt{D}$	1.361	1.243
$\ R^\top R - I\ _F$	9.6e-6	9.0e-6
Recon loss with $R := I$	0.2067	0.2500
Recon loss with R	0.1621	0.2012
Δ Recon loss	−0.0445	−0.0488

showing that R encodes information actively used by the OSQ-VAE quantizer rather than collapsing to a trivial solution.

Empirical complementarity of the two channels. We next measure on the Amazon test sets whether the two channel-private decoders make complementary mistakes, which is the empirical property that OR-fusion relies on. For every test sample we record (i) whether each channel hits the ground-truth item at $K=10$ and (ii) each channel’s top- K candidate pool $V_K^{(c)}$. From these we report in Table VIII the marginal hit rates of the two channels and of OR-fusion, the Jaccard similarity $|V_K^{(A)} \cap V_K^{(B)}| / |V_K^{(A)} \cup V_K^{(B)}|$, and a decomposition of the OR top- K hits into items shared by both channels and items exclusively contributed by a single channel.

TABLE VIII
EMPIRICAL COMPLEMENTARITY OF THE TWO DPD CHANNELS ON THE AMAZON TEST SETS AT $K=10$.

	Beauty	Sports
Hit rate, view A	0.0879	0.0499
Hit rate, view B	0.0873	0.0496
Hit rate, OR-fusion	0.0928	0.0544
Jaccard($V^{(A)}, V^{(B)}$)	0.183	0.172
OR top- K hits, Shared	1759	1470
OR top- K hits, Excl. A	174	229
OR top- K hits, Excl. B	151	244

Three consistent findings emerge across both datasets. First, OR-fusion strictly improves the hit rate over either channel alone, and these gains directly translate into the improvements reported in Table II. Second, the two top- K pools are highly complementary, with an average Jaccard of only 0.18 on

Beauty and 0.17 on Sports. Third, a non-trivial fraction of OR top- K hits is contributed exclusively by a single channel ($\approx 15\%$ on Beauty and $\approx 24\%$ on Sports), confirming that the two channels rescue genuinely different ground-truth items rather than redundantly agreeing on easy ones. Together, these measurements support the OR-fusion design empirically: as long as the channels do not collapse to identical failure patterns, OR-fusion yields a measurable gain.

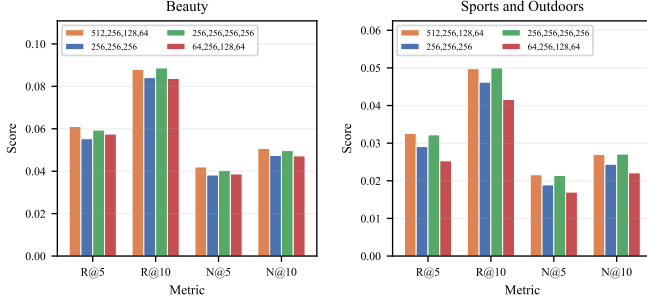


Fig. 4. Codebook configuration analysis on Beauty and Sports. Four fully-learned configurations (no random collision ID) are compared across R@5, N@5, R@10, and N@10.

H. Codebook Configuration Analysis

BARGE adopts a fully learned 4-layer codebook ($L=4$) with layer-wise decreasing sizes ($|C_1|, \dots, |C_4|$) = (512, 256, 128, 64), which differs from the TIGER convention [3] of a 3-layer uniform codebook plus a collision-resolving ID. Two design choices warrant explanation.

Why no random collision-resolving ID at the last layer. Item-level structural modelling requires *every* semantic-ID layer, including the last, to carry genuine semantic information; appending a random ID at the deepest layer would inject a non-semantic bit into the very position where ICA and HPR operate, breaking the layer-wise structural signal that these modules rely on. To assess whether this concession is even necessary, we run a controlled comparison on TIGER between its original 3-layer codebook with a random collision-resolving ID and a variant in which the random bit is replaced by an additional fully-learned semantic layer of the same size: the two settings differ by less than 0.1% on all recommendation metrics, indicating that whatever benefit the random ID brings can be matched by simply turning that slot into a semantic layer.

Why layer-wise decreasing sizes. Following the progressive layer-wise modelling principle of PLUM [51], we let the first layer use a larger codebook to cover diverse coarse-grained categories and shrink subsequent layers as they refine the representation within increasingly constrained semantic subtrees. Beyond the conceptual fit, this configuration is also cheaper: compared with a 4-layer uniform (256 $\times$ 4) codebook, our (512, 256, 128, 64) schedule reduces the total number of codeword embeddings from 1024 to 960, and the per-layer output-head parameters scale accordingly.

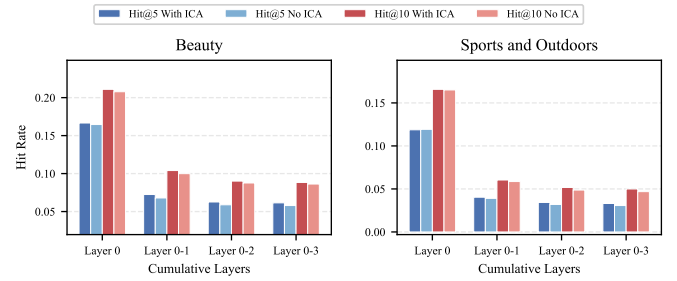


Fig. 5. Cumulative hit rate comparison with and without ICA across semantic ID layers. The advantage of ICA grows as layers accumulate, demonstrating stronger resistance to semantic drift at deeper layers.

To support these two choices empirically, we compare four configurations: (1) (512, 256, 128, 64), our final setting; (2) (256, 256, 256) plus a random collision-resolving ID, which is the original TIGER configuration; (3) (256, 256, 256, 256), a 4-layer uniform configuration with a fully-learned last layer; and (4) (64, 256, 128, 64), a variant with a deliberately small first-layer codebook. Fig. 4 yields two observations consistent with the two design choices above:

(1) *Depth matters, and layer-wise decreasing is the most effective allocation.* The 4-layer uniform consistently outperforms the TIGER-style 3-layer codebook with a collision-resolving ID on both datasets, confirming that an additional semantic layer is the right way to expand the ID space. On top of this, our (512, 256, 128, 64) schedule further outperforms the 4-layer uniform baseline while using fewer parameters, validating the PLUM-style decreasing allocation.

(2) *Capacity should concentrate at coarse-grained layers.* Comparing (512, 256, 128, 64) with (64, 256, 128, 64), shrinking only the first layer from 512 to 64 causes consistent degradation across all metrics on both datasets. This confirms that the first layer, which defines the coarse-grained semantic partitioning, requires sufficient capacity to establish well-separated category boundaries, exactly the regime in which a larger first-layer codebook pays off most.

I. Qualitative Analysis

ICA: Per-layer cumulative hit rate. Fig. 5 compares the cumulative hit rate with and without ICA across semantic ID layers. The advantage of ICA becomes more pronounced as layers accumulate: at c_1 the gap is relatively small, but by c_4 the ICA-equipped model maintains a clearly higher hit rate. This widening gap demonstrates that ICA effectively mitigates semantic drift at deeper layers by injecting item-level context that provides the discriminative signal needed to keep the decoding path on track.

ICA: Gate activation distribution. Fig. 6 visualises the distribution of the gate activations across all four semantic ID layers. The gate values are concentrated around 0.35–0.38 on both datasets, which indicates that the gating network selectively injects item-level context rather than overwriting token representations. The distributions remain consistent across

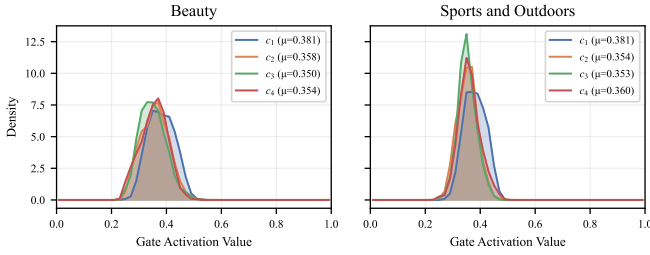


Fig. 6. Gate activation value distribution across semantic ID layers. Gate values are stably concentrated around 0.35–0.38, indicating selective and consistent injection of item-level context.

layers (c_1 through c_4), which suggests that item-level context is stably beneficial at all levels of the hierarchy.

HPR: Drift-recovery case study. To examine when HPR is most beneficial, we sweep the entire Sports test set and identify samples for which vanilla beam search drops the ground-truth (GT) path at some intermediate layer (i.e., its NTP-only beam rank exceeds the beam width $K=20$), while HPR rescues it back into the surviving beam by fusing the per-layer dual-tower path score. Two representative cases are presented below, which illustrate two drift patterns.

Case A: Single-layer drift on a hunting/tactical sequence.

Input: User Interaction History

A user’s last five interactions form a coherent hunting and tactical chain:

- 1) NcStar – Paintball & Airsoft / Paintball Sights
- 2) Plano – Hunting & Fishing / Tackle Boxes
- 3) Mtech USA – Hunting & Tactical Knives / Hunting Knives
- 4) (Unbranded) – Hunting Accessories / Cleaning & Maintenance Products
- 5) Predator – Hunting Accessories / Targets & Accessories

Ground Truth: Next Item

Sports & Outdoors (generic, no fine-grained sub-category)

Semantic ID: (243, 182, 45, 41)

Drift Recovery: Vanilla NTP vs. HPR (beam $K=20$)

Layer	NTP rank	HPR rank	Δ
c_2	33 (dropped)	14	+19
c_3	22 (dropped)	13	+9

After a long hunting-themed history, vanilla NTP overcommits to fine-grained hunting sub-categories at c_2 and pushes the more generic GT outside the top-20. The path-level score of HPR rewards prefixes that remain compatible with the broader category, which recovers the GT to rank 14.

Case B: Multi-layer drift on a mixed-interest sequence.

Input: User Interaction History

The history mixes cycling, swimming and tactical interests:

- 1) Shimano – Cycling / Drivetrain Components / Chains
- 2) (Unbranded) – Boating & Water Sports / Swimming Training Paddles
- 3) Profile Design – Cycling / Handlebars
- 4) (Unbranded) – Swimming / Men’s Swimwear / Jammers
- 5) MTECH USA XTREME – Hunting & Fishing / Tactical Knives

Ground Truth: Next Item

Aquatic Fitness Equipment (Exercise & Fitness / Accessories)

Semantic ID: (118, 158, 30, 15)

Drift Recovery: Vanilla NTP vs. HPR (beam $K=20$)

Layer	NTP rank	HPR rank	Δ
c_2	32 (dropped)	19	+13
c_3	23 (dropped)	17	+6
c_4	23 (dropped)	20	+3

NTP-only decoding consistently drops the ground-truth path at every one of the three deep layers, whereas HPR successfully rescues it at all three layers simultaneously. This confirms that the per-layer reranker repeatedly intervenes whenever the locally most likely prefix misaligns with the global semantic plausibility implied by the full history of the user, effectively correcting beam search errors before they propagate further.

Summary. Across the full test set, we observe such drift-recovery events on a non-trivial fraction of samples, with rank improvements ranging from a few positions to more than 30. This qualitative picture is consistent with the aggregate gains reported in Table II and the cascade-drift statistics in Table VI.

J. Online A/B Test

We further conducted an online A/B test on a Tencent commercial media platform, allocating 6% of the live traffic to BARGE and comparing it against the incumbent multi-stage system. BARGE achieved statistically significant improvements on core engagement metrics, with click-through rate +0.60%, click unique visitors +1.34%, and total reading time +1.70%, confirming that its offline gains translate into tangible business value in industrial-scale recommendation.

VI. CONCLUSION

In this work, we revisit generative recommendation from a structural standpoint and ask where, along the encode-decode pipeline, fidelity to the recommendation task is actually lost. Our answer locates two specific failure points: token flattening dissolves item boundaries inside the encoder, and hierarchical autoregressive decoding accumulates drift across codebook layers. Around this diagnosis we build BARGE as three mutually orthogonal modules. ICA reinstates item-level granularity through cross-attention pooling with gated fusion, HPR rescores beams via per-layer dual-tower contrastive evaluation, and DPD decomposes item embeddings into orthogonal subspaces via OSQ-VAE and decodes them through paired branches over a shared encoder, each with channel-private HPR. Empirically, the three modules act on disjoint error sources and their gains add up rather than overlap, as evidenced on public benchmarks and the large-scale offline test on Tencent’s commercial media platform. Beyond BARGE itself, we plan to further explore the use of hierarchical semantics in designing more powerful structures for generative recommendation.

REFERENCES

- [1] N. De Cao, G. Izacard, S. Riedel, and F. Petroni, “Autoregressive entity retrieval,” *arXiv preprint arXiv:2010.00904*, 2020.
- [2] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, “Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5),” in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 299–315.
- [3] S. Rajput, N. Mehta, A. Singh, R. Hulikal Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Tran, J. Samost *et al.*, “Recommender systems with generative retrieval,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 10299–10315, 2023.
- [4] Y. Pang, Z. Liu, Y. Li, S. Zhu, Z. Luo, C. Yu, S. Wu, S. Shen, C. Xu, B. Wang *et al.*, “Higr: Efficient generative slate recommendation via hierarchical planning and multi-objective preference alignment,” *arXiv preprint arXiv:2512.24787*, 2025.
- [5] Y. Hou, J. Ni, Z. He, N. Sachdeva, W.-C. Kang, E. H. Chi, J. McAuley, and D. Z. Cheng, “Actionpiece: Contextually tokenizing action sequences for generative recommendation,” *arXiv preprint arXiv:2502.13581*, 2025.
- [6] X. Guo, B. Chen, S. Wang, Y. Yang, M. Cheng, C. Lei, Y. Ding, and H. Li, “Onesug: The unified end-to-end generative framework for e-commerce query suggestion,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 17, 2026, pp. 14774–14782.
- [7] W. Wang, H. Bao, X. Lin, J. Zhang, Y. Li, F. Feng, S.-K. Ng, and T.-S. Chua, “Learnable item tokenization for generative recommendation,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 2400–2409.
- [8] Y. Yang, Z. Ji, Z. Li, Y. Li, Z. Mo, Y. Ding, K. Chen, Z. Zhang, J. Li, S. Li *et al.*, “Sparse meets dense: Unified generative recommendations with cascaded sparse-dense representations,” *arXiv preprint arXiv:2503.02453*, 2025.
- [9] X. Lin, H. Shi, W. Wang, F. Feng, Q. Wang, S.-K. Ng, and T.-S. Chua, “Order-agnostic identifier for large language model-based generative recommendation,” in *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval*, 2025, pp. 1923–1933.
- [10] E. Liu, B. Zheng, C. Ling, L. Hu, H. Li, and W. X. 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*, 2025, pp. 729–739.
- [11] B. Zheng, H. Lu, Y. Chen, W. X. Zhao, and J.-R. Wen, “Universal item tokenization for transferable generative recommendation,” *arXiv preprint arXiv:2504.04405*, 2025.
- [12] S. Du, J. Guo, B. Li, S. Cui, Z. Xu, Y. Luo, Y. Wei, K. Gai, X. Wang, K. Wu *et al.*, “Vqrae: Representation quantization autoencoders for multimodal understanding, generation and reconstruction,” *arXiv preprint arXiv:2511.23386*, 2025.
- [13] A. Van Den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [14] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, “Autoregressive image generation using residual quantization,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11523–11532.
- [15] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, “Session-based recommendations with recurrent neural networks,” *arXiv preprint arXiv:1511.06939*, 2015.
- [16] J. Tang and K. Wang, “Personalized top-n sequential recommendation via convolutional sequence embedding,” in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 565–573.
- [17] W.-C. Kang and J. McAuley, “Self-attentive sequential recommendation,” in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 197–206.
- [18] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, “Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer,” in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.
- [19] T. Zhang, P. Zhao, Y. Liu, V. S. Sheng, J. Xu, D. Wang, G. Liu, X. Zhou *et al.*, “Feature-level deeper self-attention network for sequential recommendation,” in *IJCAI*, 2019, pp. 4320–4326.
- [20] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, “S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization,” in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 1893–1902.
- [21] Z. Li, A. Sun, and C. Li, “Diffurec: A diffusion model for sequential recommendation,” *ACM Transactions on Information Systems*, vol. 42, no. 3, pp. 1–28, 2023.
- [22] C. Li, Y. Wang, Q. Liu, X. Zhao, W. Wang, Y. Wang, L. Zou, W. Fan, and Q. Li, “Strec: Sparse transformer for sequential recommendations,” in *Proceedings of the 17th ACM conference on recommender systems*, 2023, pp. 101–111.
- [23] J. Chen, H. Wang, Y. Wu, Q. Lin, Y. Diao, and J. Ji, “Zero-shot recommendation: Towards class semantic relation learning for inferring labels of unseen micro-videos,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 24, 2026, pp. 20163–20171.
- [24] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, “Wide & deep learning for recommender systems,” in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016, pp. 7–10.
- [25] P. Covington, J. Adams, and E. Sargin, “Deep neural networks for youtube recommendations,” in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 191–198.
- [26] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, “Deepfm: a factorization-machine based neural network for ctr prediction,” *arXiv preprint arXiv:1703.04247*, 2017.
- [27] J. Qiu, Z. Wang, F. Zhang, Z. Zheng, J. Zhu, J. Fan, T. Zhang, H. Wang, and X. Wang, “One model to rank them all: Unifying online advertising with end-to-end learning,” *arXiv e-prints*, pp. arXiv–2505, 2025.
- [28] Z. Cui, J. Ma, C. Zhou, J. Zhou, and H. Yang, “M6-rec: Generative pretrained language models are open-ended recommender systems,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.
- [29] Y. Li, X. Lin, W. Wang, F. Feng, L. Pang, W. Li, L. Nie, X. He, and T.-S. Chua, “A survey of generative search and recommendation in the era of large language models,” *arXiv preprint arXiv:2404.16924*, 2024.
- [30] K. Bao, J. Zhang, W. Wang, Y. Zhang, Z. Yang, Y. Luo, C. Chen, F. Feng, and Q. Tian, “A bi-step grounding paradigm for large language models in recommendation systems,” *ACM Transactions on Recommender Systems*, vol. 3, no. 4, pp. 1–27, 2025.
- [31] Y. Zhang, W. Xu, X. Zhao, W. Wang, F. Feng, X. He, and T.-S. Chua, “Reinforced latent reasoning for llm-based recommendation,” *arXiv preprint arXiv:2505.19092*, 2025.
- [32] T. Kim, S. Yoon, S. Kang, J. Yeo, and D. Lee, “Mviger: Multi-view variational integration of complementary knowledge for generative recommender,” *arXiv preprint arXiv:2408.08686*, 2024.
- [33] W. Hua, S. Xu, Y. Ge, and Y. Zhang, “How to index item ids for recommendation foundation models,” in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 2023, pp. 195–204.
- [34] W. Sun, L. Yan, Z. Chen, S. Wang, H. Zhu, P. Ren, Z. Chen, D. Yin, M. Rijke, and Z. Ren, “Learning to tokenize for generative retrieval,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 46345–46361, 2023.
- [35] J. Zhai, L. Liao, X. Liu, Y. Wang, R. Li, X. Cao, L. Gao, Z. Gong, F. Gu, M. He *et al.*, “Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations,” *arXiv preprint arXiv:2402.17152*, 2024.
- [36] J. Deng, S. Wang, K. Cai, L. Ren, Q. Hu, W. Ding, Q. Luo, and G. Zhou, “Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment,” *arXiv preprint arXiv:2502.18965*, 2025.
- [37] Z. Liu, S. Wang, X. Wang, R. Zhang, J. Deng, H. Bao, J. Zhang, W. Li, P. Zheng, X. Wu *et al.*, “Onerec-think: In-text reasoning for generative recommendation,” *arXiv preprint arXiv:2510.11639*, 2025.
- [38] Y. Hou, J. Li, A. Shin, J. Jeon, A. Santhanam, W. Shao, K. Hassani, N. Yao, and J. 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*, 2025, pp. 956–966.
- [39] Y. Feng, J. Liu, M. Han, D. Li, H. Gu, P. Zhang, T. Lu, and N. Gu, “Drift-aware continual tokenization for generative recommendation,” *arXiv preprint arXiv:2603.29705*, 2026.
- [40] H. Zhang, G. Sun, J. Lu, G. Liu, and X. S. Fang, “Delrec: Distilling sequential pattern to enhance llms-based sequential recommendation,” in

2025 IEEE 41st International Conference on Data Engineering (ICDE). IEEE, 2025, pp. 1–14.

- [41] H. Xing, H. Deng, Y. Mao, L. Mu, J. Hu, Y. Xu, H. Zhang, J. Wang, S. Wang, Y. Zhang *et al.*, “Reg4rec: Reasoning-enhanced generative model for large-scale recommendation systems,” *arXiv preprint arXiv:2508.15308*, 2025.
- [42] O. Press, N. A. Smith, and M. Lewis, “Train short, test long: Attention with linear biases enables input length extrapolation,” *arXiv preprint arXiv:2108.12409*, 2021.
- [43] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint arXiv:2004.05150*, 2020.
- [44] Z. Xu, J. Chen, S. Chen, Y. He, J. Yang, C. Yuan, K. Ding, and C. Wang, “Trie-aware transformers for generative recommendation,” *arXiv preprint arXiv:2602.21677*, 2026.
- [45] C. Guo, K. Cai, Y. Zhou, Q. Luo, R. Tang, H. Li, K. Gai, and G. Zhou, “Promise: Process reward models unlock test-time scaling laws in generative recommendations,” *arXiv preprint arXiv:2601.04674*, 2026.
- [46] Y. Yu, Y. Wang, W. Ma, Z. Guo, and M. Zhang, “Apao: Adaptive prefix-aware optimization for generative recommendation,” *arXiv preprint arXiv:2603.02730*, 2026.
- [47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [48] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [49] R. He and J. McAuley, “Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering,” in *proceedings of the 25th international conference on world wide web*, 2016, pp. 507–517.
- [50] C. Ma, P. Kang, and X. Liu, “Hierarchical gating networks for sequential recommendation,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 825–833.
- [51] R. He, L. Heldt, L. Hong, R. Keshavan, S. Mao, N. Mehta, Z. Su, A. Tsai, Y. Wang, S.-C. Wang *et al.*, “Plum: Adapting pre-trained language models for industrial-scale generative recommendations,” *arXiv preprint arXiv:2510.07784*, 2025.
- [52] R. Chen, B. Liu, H. Zhu, Y. Wang, Q. Li, B. Ma, Q. Hua, J. Jiang, Y. Xu, H. Deng *et al.*, “Approximate nearest neighbor search under neural similarity metric for large-scale recommendation,” in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3013–3022.