

TAAL: MITIGATING EARLY BEAM PRUNING IN GENERATIVE RECOMMENDATION VIA TEMPORAL AUTOREGRESSIVE ALIGNMENT

Lianjie Li, Zhiying Tu, Dianhui Chu & Hongliang Sun *

Department of Computer Science, Weihai & Qingdao Research Institute
Harbin Institute of Technology
Weihai, China

24s030110@stu.hit.edu.cn, {tzy_hit, chudh, sunhl}@hit.edu.cn

ABSTRACT

Generative recommendation encodes items as hierarchical semantic identifiers (SIDs) and retrieves the next item through autoregressive decoding. Standard next-token prediction, however, does not explicitly cover the multimodal transitions present in interaction sequences, leaving the ground-truth SID vulnerable to irreversible pruning at early beam-search branches. Across three public benchmarks, we find that 91.9%–96.6% of retrieval failures occur within the first two decoding steps. We therefore propose Temporal Autoregressive Alignment (TAAL). During training, TAAL constructs a joint (c_1, c_2) soft target from historical transitions and aligns the early-prefix distribution with a forward KL objective. During inference, it calibrates candidate scores with pointwise mutual information (PMI) to reduce the influence of globally frequent prefixes. On Amazon Beauty, Instruments, and Yelp, TAAL improves NDCG@10 over the standard baseline by 39.5%, 6.7%, and 28.6%, respectively, while increasing full-SID survival by 3.9%–16.6%. Beam-width analysis further shows that the relative survival gain grows as the beam narrows, reaching 39.4% at $B = 5$.

1 INTRODUCTION

Sequential recommendation predicts the next item from a user’s interaction history and has long been dominated by discriminative sequence models (Hidasi et al., 2016; Kang & McAuley, 2018; Sun et al., 2019). Generative recommendation has recently introduced a different paradigm. Items are encoded as hierarchical semantic identifiers (SIDs) through residual vector quantization (van den Oord et al., 2017; Zeghidour et al., 2021), and an encoder–decoder recommender autoregressively generates the identifier of the next item (Rajput et al., 2023). Subsequent improvements to codebooks and training objectives (Wang et al., 2024) have further integrated item representation learning and ranked retrieval within a unified language-modeling framework.

Most advances in this paradigm focus on the tokenizer. LETTER improves codebook quality through contrastive learning (Wang et al., 2024); LC-Rec, SEED, and Co-design jointly quantize language representations and collaborative signals (Zheng et al., 2024; Hou et al., 2024; Sun et al., 2024); ETEGRec aligns the tokenizer and recommender through knowledge distillation (Liu et al., 2025); and DIGER uses a Gumbel–Softmax relaxation to learn differentiable SIDs (Fu et al., 2026). A recent line of work instead optimizes beam-search behavior in the recommender. APAO identifies vulnerable prefixes from model states and adaptively emphasizes them (Yu et al., 2026), while BEAR introduces a beam-aware ranking constraint that pushes positive code tokens into the Top- B set (Yang et al., 2026). Both methods primarily preserve the single positive target path. How corpus-level, history-conditioned transition distributions can provide coverage supervision for multiple plausible prefixes remains underexplored.

Our analysis of decoding paths reveals the severity of early pruning (section 2). With beam width 20, the standard generative baseline misprunes the ground-truth target at the first code token c_1 in

*Hongliang Sun is the corresponding author

53.0%–71.1% of cases (56.4% on Beauty, 53.0% on Instruments, and 71.1% on Yelp), followed by another 21.3%–30.2% at the second token. Among all final misses, **91.9%–96.6% fail within the first two steps**, corresponding to 83.2%–92.4% of the full test set. These errors are irreversible under trie-constrained beam search: once an early prefix leaves the beam, the entire subtree containing the ground-truth item is removed. Although the autoregressive backbone conditions on the full history X , the pointwise cross-entropy objective for next-token prediction (NTP) does not explicitly encourage the predictive distribution to cover the multiple transition modes observed across the corpus. The model may therefore collapse onto a few frequent paths at the earliest and widest branches.

This observation motivates a direct design: because retrieval success is largely determined by (c_1, c_2) , training should preserve plausible joint transition modes at this granularity, while inference should normalize transition evidence by global frequency before reranking model candidates. TAAL implements these complementary operations:

- **Joint-prefix alignment during training.** An exponentially decayed history aggregation forms a joint-prefix soft target. A chain-rule decomposition and an unbiased Monte Carlo estimator with $K_{\text{mc}} = 2$ align the model’s (c_1, c_2) distribution with empirical transitions through forward KL.
- **PMI calibration during inference.** Pointwise mutual information normalizes global prefix frequency and extracts the temporal collaborative lift used to rerank candidates within the beam in $O(B)$ time.

Our contributions are as follows:

- **A hierarchical diagnosis of early pruning.** By tracking ground-truth survival across SID levels, we find a first-token mispruning rate of 53.0%–71.1%, with the first two steps accounting for 91.9%–96.6% of all failures. Joint (c_1, c_2) alignment increases full-SID survival by improving coverage at the vulnerable early branches.
- **Joint training and prior calibration with TAAL.** We introduce a mode-covering joint KL objective based on the chain rule and unbiased Monte Carlo sampling with $K_{\text{mc}} = 2$, improving full-SID survival by 3.9%–16.6%. A lightweight PMI calibration then normalizes prefix frequency and improves candidate ranking.
- **Structure controls and beam-width analysis.** Global-marginal and history-shuffled controls separate generic structured supervision from history-conditioned alignment. Across $B \in \{5, 10, 20, 50\}$, TAAL’s relative survival gain increases as the beam narrows and reaches 39.4% at $B = 5$, consistent with its focus on early search truncation.

2 PRELIMINARIES AND THE EARLY-PRUNING CLIFF

2.1 PROBLEM FORMULATION AND HIERARCHICAL SEMANTIC IDENTIFIERS

Let $\mathcal{U}$ and $\mathcal{V}$ denote the user and item sets, with $N = |\mathcal{V}|$ items. For a user $u \in \mathcal{U}$, the chronological interaction history is $S_u = (v_1, v_2, \dots, v_n)$. Sequential recommendation aims to predict the next item v_{n+1} .

In generative recommendation, a quantization model such as RQ-VAE (van den Oord et al., 2017; Zeghidour et al., 2021) maps each item $v \in \mathcal{V}$ to a fixed-length hierarchical code tuple $c^v = (c_1, c_2, \dots, c_L)$. The level- l token c_l is selected from a codebook C_l of size $|C_l| = K$. We use c_1 for the first token and coarse-grained category, $c_{12} = (c_1, c_2)$ for the two-level joint prefix and intermediate subcategory, and $c_{1:l} = (c_1, \dots, c_l)$ for the level- l prefix. Deeper prefixes identify items at progressively finer granularity.

Given the SID sequence of historical items, $X = (c^{v_1}, \dots, c^{v_n})$, a T5-based generative recommender autoregressively produces the target SID:

$$p(c^{v_{n+1}} | X) = \prod_{l=1}^L p_{\theta}(c_l | X, c_{<l}). \quad (1)$$

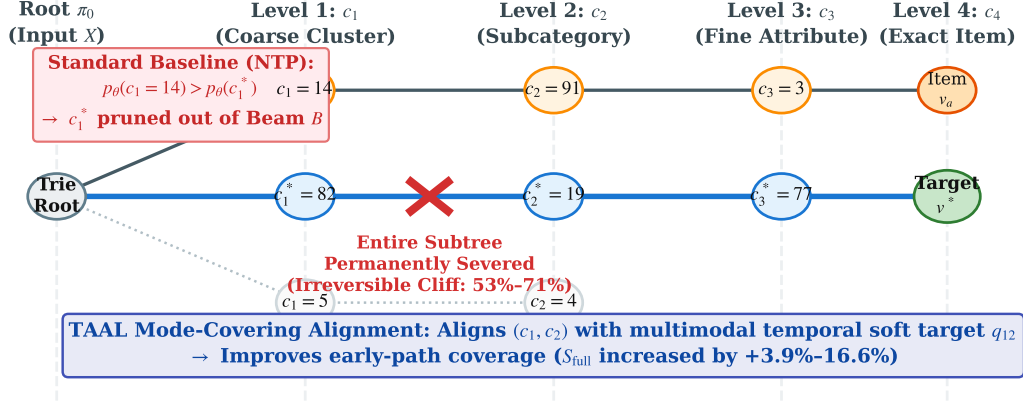


Figure 1: Hierarchical SID trie and the mechanism of early pruning. A deviation by standard NTP at c_1 removes the subtree containing the ground-truth target. TAAL assigns probability mass to historically plausible transitions through joint-prefix alignment, increasing the chance that the ground-truth branch remains within a finite beam.

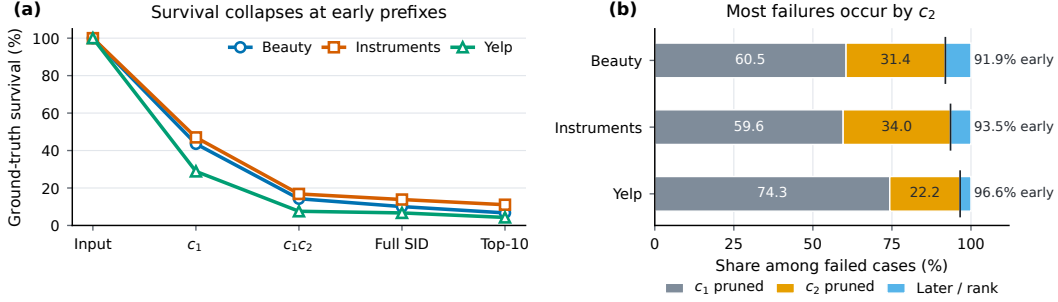


Figure 2: Ground-truth survival across beam-search decoding steps for the standard baseline with $B=20$. The first-token mispruning rate is 53.0%–71.1%, and the first two steps account for 91.9%–96.6% of all failures. Under trie-constrained beam search, these early pruning errors are irreversible.

Standard training minimizes the next-token prediction (NTP) loss:

$$\mathcal{L}_{\text{NTP}} = - \sum_{l=1}^L \log p_{\theta}(c_l^* | X, c_{<l}^*). \quad (2)$$

The hierarchy determines the role of each position: c_1 selects a coarse semantic branch, and c_2 refines the choice within that branch. Because later legal tokens depend on this two-level prefix, these positions carry the largest branching decisions and directly determine whether the ground-truth path survives subsequent search.

2.2 THE EARLY-PRUNING CLIFF

Inference uses trie-constrained beam search to guarantee valid SIDs. Once a prefix leaves the beam, all items below that prefix are permanently removed, making early errors structurally irreversible (fig. 1). Let S_{c_l} be the fraction of examples whose ground-truth prefix remains in the beam after expanding code token l , and let S_{full} be the fraction whose complete ground-truth SID remains in the final candidate set. We define the post-survival rank as the average rank of the ground-truth candidate conditioned on full-SID survival. We track these quantities for the standard LETTER-TIGER baseline at $B=20$ (section 4.1). As shown in fig. 2, first-step survival is 43.6%/47.0%/28.9% on Beauty/Instruments/Yelp, corresponding to mispruning rates of 56.4%/53.0%/71.1%. The second step removes another 21.3%–30.2%. Overall, the first two steps account for 91.9%–96.6% of all retrieval failures (91.9% on Beauty, 93.5% on Instruments, and 96.6% on Yelp), or 83.2%–92.4% of

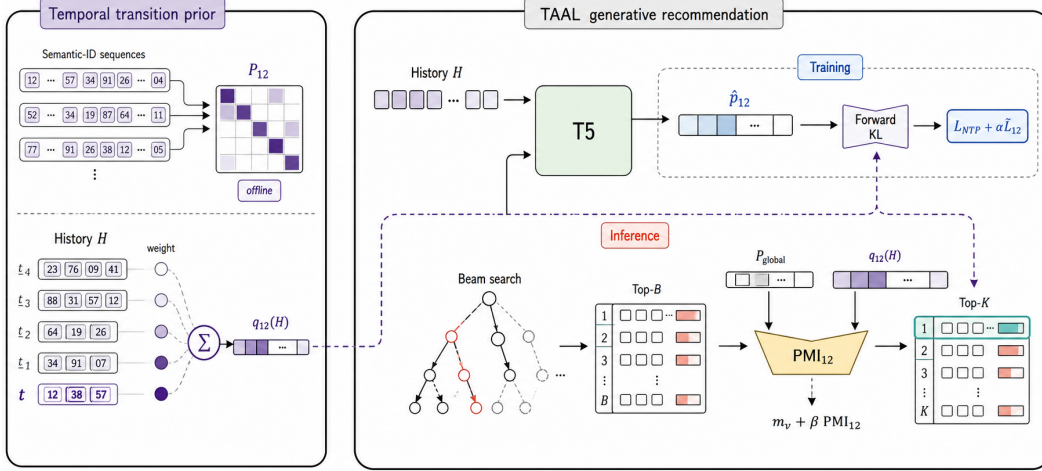


Figure 3: TAAL joint training and prior calibration. During training, exponentially decayed history aggregation constructs a joint (c_1, c_2) soft target, and an unbiased Monte Carlo forward-KL estimator aligns the recommender’s joint distribution to mitigate early pruning. During inference, PMI calibration normalizes global frequency and reranks candidates within the beam in $O(B)$ time.

the complete test set. Later decoding and post-survival ranking together account for the remaining 3.4%–8.1% of failures.

The resulting mismatch is clear. Retrieval is determined largely by the first two steps, yet the point-wise NTP objective does not explicitly use the corpus-level multimodal transition distribution as coverage supervision. At these high-branching and high-cost decisions, the predictive distribution can concentrate on a few frequent paths. Existing prefix optimization identifies the pruning phenomenon, as in APAO, but does not inject a data-driven, multimodal temporal transition distribution into both recommender training and decoding. This motivates our central question: can the same statistical prior correct branching bias at training and inference without changing the end-to-end generative paradigm?

3 TAAL: JOINT TRAINING AND PMI PRIOR CALIBRATION

The analysis in section 2 shows that retrieval failures concentrate at the first two SID positions. Token c_1 selects a coarse semantic branch, while c_2 refines it and constrains all subsequent legal paths. This creates two related challenges: training must cover multiple plausible transitions under the current history, and inference must prevent empirical priors from being dominated by globally frequent prefixes. TAAL addresses them through joint (c_1, c_2) prefix alignment during training and PMI calibration during inference (fig. 3). Forward KL preserves probability mass across transition modes, while PMI subtracts the global marginal frequency before reranking candidates. The tokenizer, T5 backbone, and beam-search procedure remain unchanged.

3.1 JOINT-PREFIX TRANSITION MODELING WITH EXPONENTIAL HISTORY DECAY

Let $\pi_{12}(v) = (c_1^v, c_2^v)$ denote the two-level joint prefix of item v , and let $\mathcal{P}_{12}$ be the set of legal joint prefixes observed in the training set. We count joint-prefix transitions between adjacent items in each training interaction sequence. The final two validation and test items are excluded to prevent leakage. This yields $\text{Count}_{12} \in \mathbb{R}^{|\mathcal{P}_{12}| \times |\mathcal{P}_{12}|}$ and the Laplace-smoothed empirical transition distribution

$$P(\pi_{12}^{\text{next}} = b \mid \pi_{12}^{\text{prev}} = a) = \frac{\text{Count}_{12}[a][b] + \varepsilon}{\sum_{b' \in \mathcal{P}_{12}} \text{Count}_{12}[a][b'] + \varepsilon |\mathcal{P}_{12}|}, \quad (3)$$

where $\varepsilon > 0$ is a smoothing constant. The use of (c_1, c_2) is motivated directly by the measurements in section 2.2: the first two steps already account for 83.2%–92.4% of the complete test set.

For a training history $(h_1, \dots, h_n)$, we aggregate transition distributions from the most recent K_{hist} items with exponentially decaying weights. Recent behavior receives greater weight while longer-range transitions retain nonzero probability mass:

$$q_{12}(b \mid \mathcal{H}) = \frac{\sum_{k=1}^{K_{\text{hist}}} \gamma^{k-1} P(\pi_{12}^{\text{next}} = b \mid \pi_{12}^{h_n-k+1})}{\sum_{k=1}^{K_{\text{hist}}} \gamma^{k-1}}. \quad (4)$$

When the history is empty, q_{12} falls back to the global joint-prefix marginal in the training set, $P_{\text{global}}(b) = \sum_a P(b \mid a)P(a)$.

3.2 TRAINING: MODE-COVERING FORWARD-KL ALIGNMENT AND MONTE CARLO SAMPLING

During the T5 forward pass, the first position directly produces $p_{\theta,1}(b_1 \mid X)$. Enumerating and decoding the second token for every legal c_1 candidate would be expensive. By the probability chain rule, the joint autoregressive forward KL decomposes exactly into a first-token marginal and a conditional expectation:

$$\mathcal{L}_{12} = \text{KL}(q_{12} \parallel \hat{p}_{\theta}) = \text{KL}(q_1 \parallel p_{\theta,1}(\cdot \mid X)) + \mathbb{E}_{b_1 \sim q_1} [\text{KL}(q(\cdot \mid b_1) \parallel p_{\theta,2}(\cdot \mid X, b_1))], \quad (5)$$

where $q_1(b_1) = \sum_{b_2} q_{12}(b_1, b_2)$ is the first-token marginal target and $q(b_2 \mid b_1) = q_{12}(b_1, b_2)/q_1(b_1)$ is the conditional target. The first term in equation 5 is computed exactly over the complete C_1 vocabulary. To estimate the second term without enumerating all candidates, we sample $K_{\text{mc}} = 2$ branches from q_1 and decode the corresponding conditional c_2 probabilities:

$$\tilde{\mathcal{L}}_{12} = \text{KL}(q_1 \parallel p_{\theta,1}(\cdot \mid X)) + \frac{1}{K_{\text{mc}}} \sum_{k=1}^{K_{\text{mc}}} \text{KL}(q(\cdot \mid b_1^{(k)}) \parallel p_{\theta,2}(\cdot \mid X, b_1^{(k)})), \quad b_1^{(k)} \sim q_1. \quad (6)$$

This estimator satisfies $\mathbb{E}[\tilde{\mathcal{L}}_{12}] = \mathcal{L}_{12}$ and provides an unbiased stochastic-gradient estimate of the conditional KL.

We use forward KL, $\text{KL}(q \parallel \hat{p})$, because its mode-covering behavior penalizes valid transitions for which $q_{12}(b) > 0$ but $\hat{p}_{\theta}(b)$ is close to zero. This encourages the model to preserve probability mass on transition paths supported by the empirical distribution. The joint objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NTP}} + \alpha \cdot \tilde{\mathcal{L}}_{12}, \quad (7)$$

where α controls the auxiliary objective. This term applies only to the first two decoding positions.

3.3 INFERENCE: PMI-BASED TEMPORAL PRIOR CALIBRATION

At inference, trie-constrained beam search returns the Top- B candidate set $\mathcal{B} = \{v_1, \dots, v_B\}$ and model log probabilities $m_v = \log p_{\theta}(c^v \mid X)$. The autoregressive model naturally contains a popularity bias toward globally frequent items. Directly adding the empirical temporal prior $\log q_{12}(c_{12}^v \mid \mathcal{H})$ can therefore overemphasize globally frequent prefixes in dense domains.

TAAL instead uses **pointwise mutual information (PMI) calibration**. Dividing by the global marginal frequency measures the collaborative lift of a local temporal association over its overall popularity:

$$\text{PMI}_{12}(c_{12}^v; \mathcal{H}) = \log \left(\frac{\tilde{P}(c_{12}^v \mid \mathcal{H})}{P_{\text{global}}(c_{12}^v)} \right), \quad (8)$$

where $\tilde{P}(c_{12}^v \mid \mathcal{H})$ uses empirical-Bayes shrinkage for low-support contexts:

$$\tilde{P}(c_{12}^v \mid \mathcal{H}) = \frac{N_{\mathcal{H}}}{N_{\mathcal{H}} + \tau} q_{12}(c_{12}^v \mid \mathcal{H}) + \frac{\tau}{N_{\mathcal{H}} + \tau} P_{\text{global}}(c_{12}^v), \quad (9)$$

and $N_{\mathcal{H}}$ is the observation count of the user’s recent context in the transition table. We set the shrinkage strength to $\tau = 30$. Each candidate $v \in \mathcal{B}$ receives the calibrated score

$$\text{Score}(v) = m_v + \beta \cdot \text{PMI}_{12}(c_{12}^v; \mathcal{H}), \quad (10)$$

where β controls prior strength and is set to 0.02. Calibration reranks only the completed Top- B beam in $O(B)$ time and requires no change to the autoregressive decoder.

3.4 DIVISION OF LABOR AND COMPLEXITY

Training and inference address complementary error sources:

- **Candidate generation during training.** Mode-covering joint alignment encourages the model to cover multiple empirically supported paths at the (c_1, c_2) branches, increasing the probability that the ground-truth target enters the final candidate set.
- **Candidate ranking during inference.** PMI calibration removes global marginal frequency and adjusts candidate ranks according to temporal associations with the current history.

4 EXPERIMENTS

We evaluate TAAL from three perspectives: overall effectiveness, component contributions, and consistency with the proposed pruning mechanism. Main experiments compare recommendation performance across three benchmarks. Ablations isolate joint-prefix training and PMI calibration. Survival diagnostics, structure controls, and beam-width sweeps test whether the observed gains follow the early-pruning hypothesis, while sensitivity curves characterize the range of effective training and inference weights.

4.1 EXPERIMENTAL SETUP

Datasets and protocol. We use Amazon Beauty, Amazon Instruments, and Yelp. The two Amazon datasets contain category-specific product interaction sequences and have relatively concentrated item categories. Yelp contains interactions with local businesses, a larger catalog, more heterogeneous behavior, and lower interaction density. Following DIGER (Fu et al., 2026), we use leave-one-out splitting and full-catalog ranking with $\text{Hit}@K$ and $\text{NDCG}@K$. Under leave-one-out evaluation, $\text{Hit}@K$ is equivalent to $\text{Recall}@K$. Dataset sizes and densities are reported in Appendix table 5; these statistics are computed on the complete original interaction files before the validation/test holdout.

Backbone and baselines. We use LETTER-TIGER (Wang et al., 2024; Rajput et al., 2023), comprising RQ-VAE SIDs and T5, with beam width $B=20$ and codebook size $K=256$. LETTER ($\alpha=0, \beta=0$) and APAO are rerun under the same protocol. Results for the remaining baselines are those compiled by DIGER (Fu et al., 2026) under the same data splits and full-catalog ranking protocol. We report the jointly trained $c_1 c_2$ model as TAAL, and retain a c_1 -only variant as a diagnostic of first-prefix coverage. Within each dataset, the no-reranking and PMI-calibrated results use the same training checkpoint. Models are trained independently across datasets.

TAAL hyperparameters and protocol. Training uses $K_{\text{hist}}=5$, $\gamma=0.7$, smoothing constant $\varepsilon=10^{-4}$, and $K_{\text{mc}}=2$. We select α separately for each dataset from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ on the validation set, obtaining 0.2/0.2/0.3 for Beauty/Instruments/Yelp. Inference uses $\beta=0.02$, shrinkage strength $\tau=30$, and the five most recent history items for PMI calibration on all three datasets. This inference window affects only post-checkpoint candidate calibration and does not change the trained model. All reported results use a fixed seed of 42.

4.2 MAIN RESULTS

Table 1 reports the main results. Without inference reranking, joint $c_1 c_2$ training reaches $\text{NDCG}@10$ values of 0.0477/0.0874/0.0286 on Beauty/Instruments/Yelp. PMI calibration further improves them to 0.0484/0.0881/0.0293, corresponding to gains of 39.5%, 6.7%, and 28.6% over the standard baseline. Under the same splits and full-catalog ranking protocol, TAAL with PMI improves over DIGER by 30.2%, 4.4%, and 28.9% in $\text{NDCG}@10$. APAO (Yu et al., 2026) optimizes vulnerable positive paths based on the model’s prefix states and achieves 0.0337/0.0811/0.0216. TAAL instead uses corpus-level, history-conditioned joint transitions as coverage supervision and obtains higher values in our reruns. The methods differ primarily in the source of supervision and the paths covered.

Table 1: Recommendation performance on three benchmarks under leave-one-out full-catalog ranking, where H@10 is equivalent to Recall@10. LETTER ($\alpha=0, \beta=0$) and APAO (Yu et al., 2026) are retrained and evaluated under the same protocol. Within each dataset, the final two rows use the same jointly trained c_1c_2 checkpoint, and PMI changes only candidate ranking. Bold indicates the best result in each column.

Category	Method	Beauty		Instruments		Yelp	
		H@10	N@10	H@10	N@10	H@10	N@10
Sequential/ collaborative	MF (Koren et al., 2009)	.0474	.0191	.0735	.0412	.0381	.0190
	Caser (Tang & Wang, 2018)	.0347	.0176	.0710	.0409	.0263	.0134
	HGN (Ma et al., 2019)	.0512	.0266	.1048	.0774	.0326	.0159
	LightGCN (He et al., 2020)	.0511	.0260	.1000	.0728	.0407	.0207
	SASRec (Kang & McAuley, 2018)	.0588	.0313	.0947	.0690	.0296	.0152
	BERT4Rec (Sun et al., 2019)	.0347	.0170	.0822	.0608	.0291	.0159
Generative baselines	BIGRec (Bao et al., 2025)	.0299	.0198	.0576	.0491	.0169	.0142
	P5-SID (Geng et al., 2022)	.0584	.0335	.0964	.0730	.0324	.0170
	P5-CID (Geng et al., 2022)	.0597	.0347	.0987	.0751	.0347	.0181
	TIGER (Rajput et al., 2023)	.0610	.0331	.1058	.0797	.0407	.0213
Tokenizer/decoding improvements	STE (Huh et al., 2023)	.0134	.0067	.0554	.0360	.0147	.0077
	ETEGRec (Liu et al., 2025)	.0615	.0335	.1106	.0810	.0415	.0214
	LETTER ($\alpha=0, \beta=0$) (Wang et al., 2024)	.0671	.0347	.1107	.0825	.0429	.0228
	DIGER (Fu et al., 2026)	.0683	.0372	.1138	.0844	.0432	.0227
	APAO (Yu et al., 2026)	.0641	.0337	.1092	.0811	.0409	.0216
Ours	TAAL (c_1c_2 , no reranking)	.0844	.0477	.1167	.0874	.0515	.0286
Ours	TAAL + PMI calibration	.0856	.0484	.1176	.0881	.0525	.0293

Table 2: Ablation of the training objective and inference calibration. The no-reranking NTP baseline is compared with TAAL under no reranking, the direct conditional transition prior $\log q_{12}$, and PMI calibration $\log(P/P_{\text{global}})$. Bold indicates the best result in each column.

Training model	Inference scheme	Beauty			Instruments			Yelp		
		H@1	H@10	N@10	H@1	H@10	N@10	H@1	H@10	N@10
$c_1c_2, \alpha = 0$	No reranking	.0118	.0671	.0347	.0623	.1107	.0825	.0081	.0429	.0228
$c_1c_2, \alpha > 0$	No reranking	.0204	.0844	.0477	.0658	.1167	.0874	.0117	.0515	.0286
$c_1c_2, \alpha > 0$	+ conditional prior ($\log q_{12}$)	.0206	.0850	.0482	.0664	.1163	.0875	.0123	.0524	.0292
$c_1c_2, \alpha > 0$	+ PMI calibration	.0208	.0856	.0484	.0664	.1176	.0881	.0118	.0525	.0293

4.3 ABLATION: JOINT TRAINING, CONDITIONAL PRIOR, AND PMI CALIBRATION

To isolate the training and inference contributions, table 2 compares the no-reranking baseline ($\alpha = 0$) with TAAL ($\alpha > 0$) under no reranking, a direct conditional prior ($\log q_{12}$), and PMI calibration ($\log(P/P_{\text{global}})$). We make three observations:

1. **Joint training provides most of the gain.** Comparing the no-reranking rows for $\alpha > 0$ and $\alpha = 0$, NDCG@10 improves by 37.5%, 5.9%, and 25.7% on Beauty, Instruments, and Yelp. Joint-prefix alignment therefore improves candidate generation independently of inference calibration.
2. **PMI provides a small but consistent NDCG@10 improvement in the matched runs.** Relative to the direct conditional prior, PMI is slightly higher on Beauty (.0484 vs. .0482), Instruments (.0881 vs. .0875), and Yelp (.0293 vs. .0292). The difference is smallest on Yelp.
3. **The magnitude of calibration gains depends on the dataset.** PMI improves the no-reranking TAAL model by 1.5%, 0.8%, and 2.3% in NDCG@10 on Beauty, Instruments, and Yelp, respectively. These modest changes are consistent with PMI acting as a lightweight post-generation ranking correction rather than changing the candidate set.

4.4 BEAM SURVIVAL AND EARLY-PRUNING MITIGATION

To test whether joint KL mitigates early pruning, fig. 4 compares the failure decomposition and full-SID survival of the standard baseline and joint c_1c_2 alignment. The c_1 -only variant isolates first-

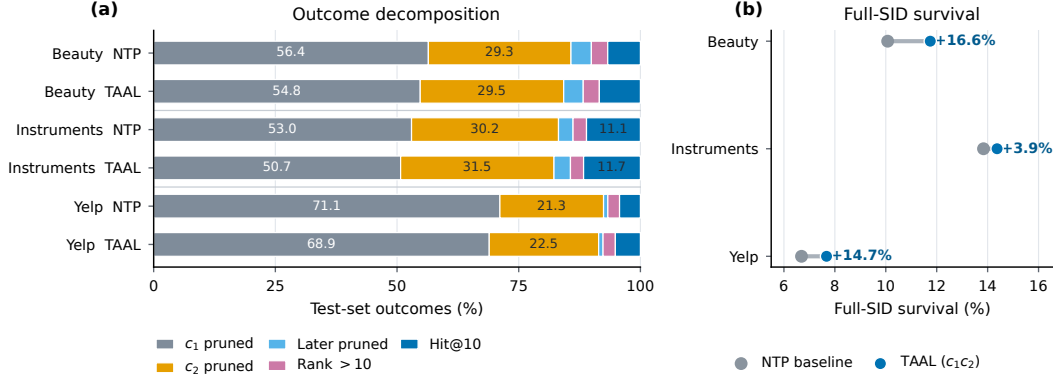


Figure 4: Failure decomposition and full-SID survival for TAAL. (a) Five-stage outcome decomposition for standard NTP and the jointly trained TAAL model. (b) Ground-truth full-SID survival S_{full} , with relative gains of 16.6%, 3.9%, and 14.7%. Complete values are reported in Appendix table 6.

prefix coverage: it asks whether improving the first prefix is sufficient for later c_2 and complete-SID survival, whereas joint c_1c_2 alignment covers the coupled early path. In the survival table, c_1 -only raises S_{c_1} on all three datasets but leaves S_{c_2} and S_{full} close to the standard baseline, whereas joint alignment improves the subsequent stages as well. Appendix table 6 reports the complete values and post-survival ranking under PMI. The main observations are:

1. **Joint c_1c_2 alignment improves the complete path.** Joint alignment increases S_{c_2} to 15.76%/17.75%/8.58% and improves S_{full} over the baseline by 3.9%–16.6%. The direction of these changes agrees with the NDCG gains in the main results.
2. **PMI changes only post-survival ranking.** PMI runs after beam search and therefore leaves the candidate set and all three survival rates unchanged. Its effect appears in the average rank of surviving targets, which improves from 7.21 to 7.11 on Beauty and from 5.13 to 5.02 on Instruments.

4.5 STRUCTURE-SPECIFIC MECHANISM DIAGNOSTICS AND CONTROLS

An auxiliary soft target may provide both structured smoothing and history-conditioned supervision, so a performance gain alone does not identify which source is responsible. We separate them on Beauty with seed = 42 and $B = 20$. All rows use the same data split, training recipe, seed, and decoding protocol. The global-marginal control, history-shuffled control, and TAAL additionally share the same forward KL and $K_{\text{mc}} = 2$ sampling procedure; only the joint-prefix soft target differs (table 3):

1. **Global marginal control** ($P_{\text{global}}(c_{12})$). This target retains the nonuniform frequency and popularity structure of joint prefixes but removes dependence on H_i . It tests whether gains arise from generic supervision toward popular prefixes rather than personalized historical transitions. We sweep $\alpha \in \{0.1, 0.2, 0.4\}$ independently.
2. **History-shuffled control** ($q(c_{12} | H_{\pi(i)})$). We first compute the history-conditioned distribution for each sample in a batch, then cyclically shift the soft targets by one sample to break their correspondence with the current histories. This control preserves entropy and multimodality while testing the role of correct history binding. We use the same α sweep.
3. **Temporal TAAL** ($q(c_{12} | H_i)$). TAAL uses the context-aligned temporal distribution and is evaluated with $\alpha \in \{0.2, 0.4\}$.

Table 3 supports two conclusions:

1. **Structured but misaligned supervision provides limited and weight-sensitive gains.** At $\alpha = 0.1$, the global-marginal and history-shuffled controls reach .0382 and .0386, slightly

Table 3: Structure-specific controls and hyperparameter sweeps on Beauty with $B = 20$ and seed 42. All changes are computed against the matched NTP baseline. The global-marginal target retains popularity while removing history dependence; the shuffled target retains the conditional distribution shape while breaking sample correspondence. Both controls can improve the baseline at low weights but degrade as the weight increases. History-aligned TAAL achieves higher results across its tested range.

Control	Soft target	Setting	$S_{c_1}(\%)$	$S_{c_2}(\%)$	$S_{full}(\%)$	Hit@10	NDCG@10	Change vs. NTP
A. NTP baseline	No auxiliary loss	$\alpha = 0$	43.57	14.26	10.07	.0671	.0347	Reference (0.0%)
B. Global marginal ($P_{global}(c_{12})$)	Static catalog prior without context	$\alpha = 0.4$ (high)	37.28	11.75	7.63	.0567	.0296	-14.7%
		$\alpha = 0.2$ (medium)	40.95	13.56	9.44	.0667	.0368	+6.1%
		$\alpha = 0.1$ (best)	42.80	14.30	10.22	.0715	.0382	+10.2%
C. History shuffled ($q(c_{12} H_{\pi(i)})$)	Preserved multimodality; broken history binding	$\alpha = 0.4$ (high)	35.87	11.06	7.13	.0512	.0269	-22.5%
		$\alpha = 0.2$ (medium)	40.88	13.37	9.39	.0669	.0360	+3.8%
		$\alpha = 0.1$ (best)	42.62	14.40	10.20	.0724	.0386	+11.3%
D. TAAL (ours)	Context-aligned temporal transitions + PMI calibration	$\alpha = 0.4$	45.58	15.33	11.54	.0806	.0455	+31.2%
		$\alpha = 0.2$ (main)	45.24	15.76	11.75	.0844	.0477	+37.5%
		$\alpha = 0.2, \beta = 0.02$	45.24	15.76	11.75	.0856	.0484	+39.5%

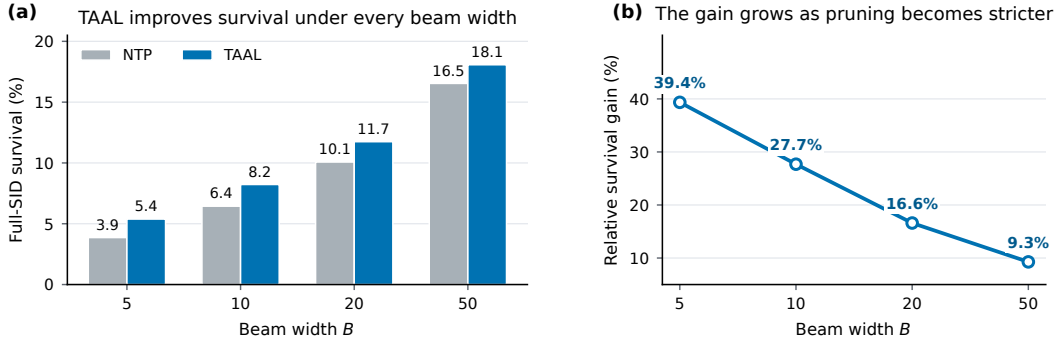


Figure 5: Beam-width sensitivity on Beauty. (a) Ground-truth full-SID survival at $B \in \{5, 10, 20, 50\}$. (b) Relative survival gain of TAAL over NTP. The gain increases as the beam narrows and reaches 39.4% at $B = 5$.

above the matched NTP baseline of .0347. As α increases to 0.4, they fall to .0296 and .0269, respectively. Thus, weak structured supervision can help, but increasing a target that lacks the correct history correspondence can hurt the current sample’s decoding objective.

2. **History alignment is more effective and tolerates larger weights.** TAAL reaches NDCG@10 of .0477 and S_{full} of 11.75% at the main setting $\alpha = 0.2$. At $\alpha = 0.4$, it remains at .0455, whereas both controls are below the matched baseline. This pattern is consistent with correct history-conditioned transitions providing the main advantage beyond generic prefix structure.

4.6 BEAM-WIDTH SENSITIVITY AND PRUNING PRESSURE

To examine behavior under different pruning pressures, we fix the NTP and TAAL model parameters on Beauty and vary only the decoding beam width over $B \in \{5, 10, 20, 50\}$.

As shown in fig. 5 and table 4, the relative improvement in full-SID survival follows a monotonic trend:

1. The relative gain in S_{full} is

$$\begin{aligned}
 B = 5 (+39.4\%) &\longrightarrow B = 10 (+27.7\%) \\
 &\longrightarrow B = 20 (+16.6\%) \longrightarrow B = 50 (+9.3\%).
 \end{aligned}
 \tag{11}$$

Under the stronger truncation at $B = 5$, the ground-truth path is more likely to be pruned and joint-prefix alignment provides its largest relative gain. As the beam expands to $B = 50$, the baseline explores more paths and the gap narrows.

Table 4: Ground-truth survival at different beam widths on Beauty, with $\alpha = 0.2$ for TAAL. The relative full-SID survival gain over NTP increases as the beam narrows. Values at $B = 20$ agree with Appendix table 6.

Beam B	Model	$S_{c_1}(\%)$	$S_{c_2}(\%)$	$S_{full}(\%)$	Relative gain
$B = 5$	NTP baseline	27.38	7.12	3.86	
	TAAL	28.95	8.49	5.38	+39.4%
$B = 10$	NTP baseline	35.02	10.19	6.44	
	TAAL	36.48	11.76	8.23	+27.7%
$B = 20$	NTP baseline	43.57	14.26	10.07	
	TAAL	45.24	15.76	11.75	+16.6%
$B = 50$	NTP baseline	55.84	21.40	16.54	
	TAAL	58.69	22.84	18.07	+9.3%

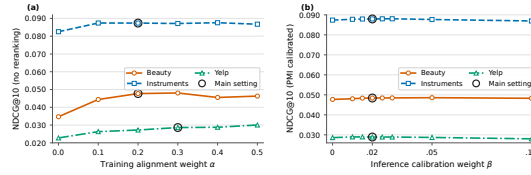


Figure 6: Hyperparameter sensitivity on three datasets. (a) No-reranking NDCG@10 versus the training alignment weight α . (b) PMI-calibrated NDCG@10 versus the inference weight β . Out-lined markers indicate the settings used in the main results ($\alpha = 0.2/0.2/0.3$ and $\beta = 0.02$), rather than the maxima of the test curves.

2. TAAL improves prefix survival at all four beam widths, and the improvement increases with pruning pressure. This pattern is consistent with the method’s focus on early beam-search truncation.

4.7 HYPERPARAMETER SENSITIVITY

Figure 6 reports no-reranking NDCG@10 across the training weight $\alpha \in [0, 0.5]$ and PMI-calibrated NDCG@10 across the inference weight $\beta \in [0, 0.10]$. Over the tested α grid, the highest observed values occur at 0.3, 0.4, and 0.5 for Beauty, Instruments, and Yelp, respectively. Yelp is still increasing at the upper boundary, so this sweep characterizes the tested effective range rather than establishing an interior optimum. The observed β maxima occur at 0.05, 0.03, and 0.015; the common setting $\beta = 0.02$ is within 0.00017 NDCG@10 of all three maxima, indicating a broad effective region.

5 RELATED WORK

Generative sequential recommendation. Early sequential recommenders were primarily discriminative, including GRU4Rec (Hidasi et al., 2016), SASRec (Kang & McAuley, 2018), BERT4Rec (Sun et al., 2019), and graph-based collaborative filtering with LightGCN (He et al., 2020). P5 (Geng et al., 2022) and M6-Rec (Cui et al., 2022) formulate recommendation as a language task. TIGER (Rajput et al., 2023) established the current generative retrieval paradigm by combining RQ-VAE SIDs with autoregressive generation. TAAL operates on the training objective and decoding of the recommender within this paradigm.

Tokenizer optimization and collaborative quantization. LETTER (Wang et al., 2024) improves codebook learning through multitask contrastive objectives. LC-Rec (Zheng et al., 2024) and SEED (Hou et al., 2024) integrate language and collaborative signals into item quantization, while Co-design (Sun et al., 2024) jointly optimizes collaborative and semantic quantization. ETE-GRec (Liu et al., 2025) aligns the tokenizer and recommender through distillation, and DIGER (Fu et al., 2026) introduces differentiable semantic identifiers with Gumbel relaxation. These methods mainly improve offline SID quantization or tokenizer learning. TAAL instead operates after SIDs

are fixed, modifying recommender training and candidate calibration. We use LETTER-TIGER as the backbone, so the two directions are complementary.

Constrained decoding and the training–inference gap. Generative retrieval commonly uses trie-constrained beam search (Tay et al., 2022; Bevilacqua et al., 2022; Cao et al., 2021), and STATIC (Su et al., 2026) accelerates this process on hardware. APAO (Yu et al., 2026) detects vulnerable prefixes from the model’s own prediction states and applies adaptive optimization. BEAR (Yang et al., 2026) uses a beam-aware ranking regularizer so that positive code tokens satisfy the condition for entering the Top- B candidates. Search-error analysis in machine translation provides a related perspective (Stahlberg & Byrne, 2019). APAO and BEAR primarily supervise the positive target path and its model-internal ranking state. TAAL estimates a history-conditioned (c_1, c_2) transition distribution from the training corpus, covers multiple plausible prefixes during training, and removes global marginal frequency with PMI at inference. Its focus is the source and dual use of external collaborative transition information rather than a redefinition of within-beam constraints on the positive path.

Collaborative signal injection. Transition and co-occurrence structure are classical recommendation signals, from item-based collaborative filtering (Sarwar et al., 2001) to factorized personalized Markov chains (Rendle et al., 2010). Recent work also injects collaborative cues into SID generation through personalized natural language at inference (Han et al., 2026). TAAL introduces no additional language model; it uses a statistical transition table to construct both the training soft target and the inference calibration score.

6 CONCLUSION AND LIMITATIONS

We quantify early beam pruning of ground-truth targets in generative recommendation and find that 91.9%–96.6% of retrieval failures across three benchmarks occur within the first two decoding steps. TAAL uses joint-prefix KL during training to cover multimodal, history-conditioned transitions and applies PMI at inference to remove global marginal frequency before reranking candidates. Full-SID survival improves by 3.9%–16.6%, and NDCG@10 reaches 0.04842/0.08810/0.02926 on Beauty/Instruments/Yelp. Global-marginal and history-shuffled controls show that structured soft supervision can provide limited gains, while transitions aligned with the current history are more effective and tolerate larger auxiliary weights. The monotonic beam-width trend also agrees with the proposed early-pruning mechanism.

Limitations. First, the transition model uses first-order contextual transitions; higher-order and multi-hop temporal behavior remain open directions. Second, the experiments cover three medium-scale public datasets, and generalization to very large industrial catalogs requires further study. Third, the main results use a fixed random seed of 42.

REFERENCES

- Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Chong Chen, Fuli Feng, and Qi Tian. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems*, 3(4):1–27, 2025.
- Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen tau Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as identifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 31668–31683, 2022.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *International Conference on Learning Representations (ICLR)*, 2021.
- Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. M6-Rec: Generative pre-trained language models are open-ended recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.
- Junchen Fu, Xuri Ge, Alexandros Karatzoglou, Ioannis Arapakis, Suzan Verberne, Joemon M. Jose, and Zhaochun Ren. Differentiable semantic ID for generative recommendation. *arXiv preprint arXiv:2601.19711*, 2026.

- Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys)*, pp. 299–315, 2022.
- Changjiang Han, Qingyang Li, Yaqiang Zang, Jikun Kang, and Pinghua Gong. Restoring collaborative signals in semantic-ID generative recommendation via personalized natural language. *arXiv preprint arXiv:2607.27682*, 2026.
- Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 639–648, 2020.
- Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.
- Yupeng Hou, Jiacheng Li, Zhongyi Liu, Wayne Xin Zhao, Lingpeng Kong, and Ji-Rong Wen. Learning vector-quantized item representations for LLM-based recommender systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1112–1122, 2024.
- Minyoung Huh, Brian Cheung, Pulkit Agrawal, and Phillip Isola. Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks. In *International Conference on Machine Learning (ICML)*, pp. 14096–14113, 2023.
- Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018.
- Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
- 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 (SIGIR)*, 2025.
- Chen Ma, Peng Kang, and Xue Liu. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 825–833, 2019.
- Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trong Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Trung Vinh Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.
- Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 811–820, 2010.
- Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW)*, pp. 285–295, 2001.
- Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3356–3362, 2019.
- Zhengyang Su, Isay Katsman, Yueqi Wang, Ruining He, Lukasz Heldt, Raghunandan Keshavan, Shao-Chuan Wang, Xinyang Yi, Mingyan Gao, Onkar Dalal, Lichan Hong, Ed Chi, and Ningren Han. Vectorizing the trie: Efficient constrained decoding for LLM-based generative retrieval on accelerators. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026.

- Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformers. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1441–1450, 2019.
- Peiyan Sun, Le Zhang, Chao Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Hui Liu, Huaiwen Liu, and Xiaoxue Zhao. Co-design of collaborative and semantic quantization for generative recommendation. In *Proceedings of the ACM Web Conference (WWW)*, pp. 1245–1254, 2024.
- Jiaxi Tang and Ke Wang. Personalized top-N sequential recommendation via convolutional sequence embedding. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 565–573, 2018.
- Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Donald Metzler, Vinton G. Carpenter, Tal Schuster, Huaixiu Steven Zheng, Kashyap Gopalakrishnan, Adams Wei Yu, and Luke Zettlemoyer. Transformer memory as a differentiable search index. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 21831–21843, 2022.
- Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.
- Wenjia 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 (CIKM)*, pp. 2400–2409, 2024.
- Weiqin Yang, Bohao Wang, Zhenxiang Xu, Jiawei Chen, Shengjia Zhang, Jingbang Chen, Canghong Jin, and Can Wang. BEAR: Towards beam-search-aware optimization for recommendation with large language models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2026. arXiv:2601.22925.
- Yuanqing Yu, Yifan Wang, Weizhi Ma, Zhiqiang Guo, and Min Zhang. APAO: Bridging the training-inference gap in generative recommendation via adaptive prefix-aware optimization. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026.
- Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.
- Bowen Zheng, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Jun Xu, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 946–954, 2024.

A DATASET STATISTICS AND ADDITIONAL DIAGNOSTICS

A.1 DATASET STATISTICS

Table 5: Statistics of the three benchmark datasets before leave-one-out holdout. Density is the number of interactions divided by the product of the number of users and items.

Dataset	Users	Items	Interactions	Avg. sequence length	Density
Beauty	22,363	12,101	198,502	8.88	0.0734%
Instruments	24,772	9,922	206,153	8.32	0.0839%
Yelp	30,431	20,033	316,354	10.40	0.0519%

A.2 COMPLETE BEAM-SURVIVAL RESULTS

Table 6: Ground-truth survival and ranking diagnostics across decoding stages at beam width 20. PMI calibration runs after candidate generation, so it leaves all three survival rates unchanged and affects only post-survival rank and final ranking metrics.

Dataset	Training and inference	$S_{c_1}(\%)$	$S_{c_2}(\%)$	$S_{full}(\%)$	Post-survival rank	Hit@10	NDCG@10
Beauty	Standard baseline (NTP, $\alpha = 0$)	43.57	14.26	10.07	8.13	.0671	.0347
	c_1 -only ($\alpha = 0.4$)	46.39	14.13	10.10	7.81	.0690	.0377
	+ joint $c_1 c_2$ KL ($\alpha = 0.2$)	45.24	15.76	11.75	7.21	.0844	.0477
	+ PMI calibration ($\beta = 0.02$)	45.24	15.76	11.75	7.11	.0856	.0484
Instruments	Standard baseline (NTP, $\alpha = 0$)	47.02	16.83	13.84	5.31	.1107	.0825
	c_1 -only ($\alpha = 0.2$)	50.77	16.79	13.80	5.40	.1101	.0818
	+ joint $c_1 c_2$ KL ($\alpha = 0.2$)	49.27	17.75	14.37	5.13	.1167	.0874
	+ PMI calibration ($\beta = 0.02$)	49.27	17.75	14.37	5.02	.1176	.0881
Yelp	Standard baseline (NTP, $\alpha = 0$)	28.87	7.57	6.69	8.29	.0429	.0228
	c_1 -only ($\alpha = 0.3$)	32.60	7.55	6.69	8.23	.0435	.0234
	+ joint $c_1 c_2$ KL ($\alpha = 0.3$)	31.06	8.58	7.67	7.82	.0515	.0286
	+ PMI calibration ($\beta = 0.02$)	31.06	8.58	7.67	7.61	.0525	.0293