FREQUENCY-GRADED DIRECTIONAL COLLAPSE IN FULL-SOFTMAX SEQUENTIAL RECOMMENDERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Full-softmax sequential recommenders degrade on rare items, but the failure's locus and the mechanism of content-based repair remain unclear. We identify frequency-graded directional concentration after cross-entropy training in the *output*, rather than input, embedding table: rare-item directions form a low-rank cone anti-aligned with typical user states. Untied probes, capacity controls, and a frequency-capping ladder with volume-matched controls support an imbalance-driven optimization explanation. History dropout strengthens tail retrieval, whereas a frozen semantic residual favors popular items, consistent with an anisotropy-associated popularity prior. Whitening the semantic matrix *before* training reduces this anisotropy; tested post-hoc centering, normalization and gating do not reproduce the training-time gains. Across three Amazon corpora, the repaired retriever raises Recall@10 by 20–29% over the matched d=192 pure-ID control. A 17.4K-parameter residual then reranks only the retrieved top-10 using training-set co-occurrence, preserving Recall@10 exactly. Under matched full-catalog evaluation, the final system achieves state-of-the-art performance against ten reproduced baselines, improving all four metrics across three corpora by approximately 17–48% over the strongest five-seed baseline per metric. Repeated tests reject frequency gating; excluding consumed items alone raises NDCG@10 by 28.7%. These results connect output-side geometry to ranking behavior under a frozen evaluation protocol.

1 INTRODUCTION

Sequential recommendation models user preferences from interaction histories to rank the next item. Approaches range from self-attentive encoders (Kang & McAuley, 2018) to context-aware generative tokenization (Hou et al., 2025), yet learning useful representations for rarely observed items remains difficult (Dang et al., 2026). These items receive few positive interactions, although they must compete with popular alternatives when a personalized list is constructed.

Recent methods address sparsity through tail-aware augmentation (Dang et al., 2026), LLM-derived semantic enrichment (Liu et al., 2024; Li et al., 2026), and global semantic alignment (Wang et al., 2026). These approaches enrich the supervision or representations available to the recommender. Complementary studies show that embedding collapse limits naive model scaling (Guo et al., 2024) and contributes to popularity bias in graph collaborative filtering (Zhang et al., 2023). Together, these advances make representation geometry an important object of study. They also motivate a more specific question: which part of a full-softmax sequential recommender carries the rare-item deficit, and which intervention changes that representation?

The distinction matters because an item embedding serves two different roles. On the input side, it represents an observed event from which the sequence encoder constructs a user state. On the output side, it acts as a classifier vector whose compatibility with that state determines the candidate's score. Weight tying shares parameters between these roles, but their functions and training signals differ. A useful history representation therefore need not imply a useful scoring direction, making the output table a distinct target for diagnosis.

Full-catalog softmax competition provides a concrete reason to examine this distinction. The softmax loss contributes attractive gradient terms for the target output vector and repulsive terms for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

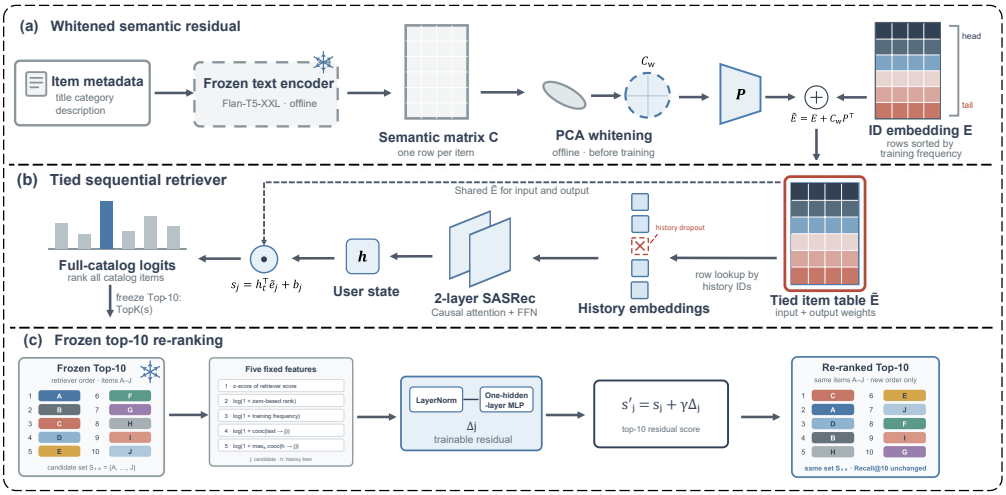


Figure 1: System overview. (a) Frozen item semantics are whitened before training and projected into ID space. (b) Weight tying routes the residual through the output table. (c) A trained count-based residual reranks only the retrieved top-10, preserving Recall@10. Frequency gating is excluded after five-seed tests (§4.4).

competing vectors. Rare items receive the attractive signal infrequently, while repeatedly participating as alternatives to other targets. This asymmetry motivates testing whether imbalance shapes output directions, beyond simply limiting the amount of positive supervision. Untying the tables lets us measure their geometry separately and distinguish an output-side failure from degradation already present in history representations.

Our untied probes reveal *frequency-graded directional collapse*: rare-item output vectors form a low-rank cone anti-aligned with typical user states, while rare-item input vectors remain dispersed. Concentration varies across frequency bands, rather than forming two uniform groups. Large norms do not preclude this failure: anti-alignment can still produce negative dot-product contributions to item scores. We probe its origin through capacity controls, frequency-capping with volume-matched controls, and a layer-peeled analysis building on minority collapse (Fang et al., 2021). These complementary experiments support an imbalance-driven optimization account. Unlike untied probes, a shared table cannot distinguish the geometries learned for the two roles.

This diagnosis motivates a repair that separates history regularization, semantic geometry and within-list ordering (Figure 1). History dropout strengthens tail retrieval, while a frozen semantic residual adds item content to the tied table. Raw semantics introduce their own directional concentration, so we apply PCA whitening *before* training the recommender. Relative to raw semantics, whitening leaves the text encoder, semantic dimension and trainable parameter count unchanged. We then measure the learned scoring table, connecting preprocessing to the representation actually used for ranking. The fixed semantic matrix still allows its projection, ID embeddings and sequence encoder to adapt jointly during training. This tests whether the initial geometry supports a better learned scoring space, rather than treating isotropy of pretrained features as sufficient. Tested post-hoc transformations do not reproduce the training-time gains.

Retrieving a relevant item and placing it near the top of the list are distinct objectives. We therefore refine the retrieved top-10 with a small residual reranker using scores, ranks, item frequency and training-set co-occurrence counts. The count features provide collaborative information beyond the retriever's scalar score. The candidate set remains fixed, so reranking preserves Recall@10 exactly and isolates changes in within-list ordering. All count features are computed from training interactions, excluding validation and test targets.

Our contributions are threefold. First, we provide an **output-side diagnosis** that separates history representations from candidate-scoring geometry. Second, we present **controlled imbalance evidence** linking frequency-graded concentration to optimization under full-softmax competition.

Third, we develop a **repair-channel decomposition** that separates the roles of history regularization, semantic geometry and within-list ordering in a simple, testable pipeline.

We evaluate on Beauty, Toys and Sports under one frozen full-catalog protocol, using five retriever seeds and ten reproduced baselines. The final system achieves state-of-the-art performance in this matched comparison, outperforming the strongest baseline on all twelve dataset–metric combinations, with relative gains of 17–48%. Matched ablations, post-training geometry measurements and frequency-group analyses test the explanation beyond aggregate gains. Frequency-gating controls guide the ungated final design, while group metrics characterize the accuracy–coverage trade-off and protocol audits quantify sensitivity to evaluation choices.

2 Related Work

Neural and minority collapse in classification. Neural collapse characterizes convergence of class means and classifier vectors toward a simplex frame (Papyan et al., 2020). Layer-peeled analysis predicts *minority collapse* under imbalance, where rare-class classifier vectors merge (Fang et al., 2021). We study next-item prediction with skewed frequencies and catalogs larger than the embedding dimension. Our diagnosis localizes frequency-graded concentration to the output table and examines how weight tying obscures it.

Embedding collapse in recommenders. Recent studies examine collapse that limits scaling (Guo et al., 2024), spectral concentration in attention stacks (Cui et al., 2026), and popularity-biased dimensional collapse in graph collaborative filtering (Zhang et al., 2023). Our untied input–output comparison distinguishes history representations from candidate classifiers. It shows that dispersed input embeddings can coexist with concentrated, anti-aligned output directions, motivating measurements at the scoring table alongside hidden-representation diagnostics.

Feature-level sequential models. FDSA separates feature-level and item-level attention (Zhang et al., 2019), while S3Rec uses self-supervised objectives over attributes (Zhou et al., 2020). These methods supply important comparisons to LLM-derived representations. In our ten-baseline panel, FDSA leads baseline NDCG@5 on all three corpora, whereas ActionPiece leads baseline Recall@10. Both feature-aware and generative methods therefore enter the matched comparison.

Long-tail sequential recommendation and LLM enhancement. Long-tail methods include mutual head–tail transfer (MELT (Kim et al., 2023)), tail-aware data augmentation (TADA (Dang et al., 2026)), LLM-derived semantic embeddings with dual-view modeling (LLM-ESR (Liu et al., 2024)), and global semantic alignment (SAGE (Wang et al., 2026)). In parallel, generative recommenders build discrete semantic IDs from frozen text encoders, including TIGER (Rajput et al., 2023), GRID (Ju et al., 2025), LIGER (Yang et al., 2025) and ActionPiece (Hou et al., 2025). Related approaches include masked diffusion (MaskGR, a preprint (Shah et al., 2025)) and metadata-rich late fusion (GRAM (Lee et al., 2025)). Our comparisons use common chronological splits and full-catalog evaluation; Appendix M details the implementations.

3 Methodology

Given a history $H_t = (i_1, \ldots, i_t)$ and catalog $\mathcal{I}$, the task is to rank the next item. Let $E \in \mathbb{R}^{|\mathcal{I}| \times d}$ be the trainable ID embedding table and $C \in \mathbb{R}^{|\mathcal{I}| \times 4096}$ the frozen item-semantic matrix. Figure 1 separates the method into semantic preprocessing, tied sequential retrieval and within-list reranking. The design addresses output-side concentration and refines the ordering of retrieved candidates. Sections 4.2 and 4.5 evaluate the output geometry and frequency-group performance.

3.1 Whitened Semantic Residual

A frozen flan-t5-xxl encoder maps item metadata to rows of C offline. To reduce directional concentration before recommendation training, we center the columns and compute $C_c = USV^\top$. With n catalog items, the final recipe uses PCA whitening:

$$C_w = C_c V \operatorname{diag}\left(\frac{1}{\max(S_{kk}/\sqrt{n}, 10^{-8})}\right), \qquad \widetilde{E} = E + C_w P^\top, \tag{1}$$

where $P \in \mathbb{R}^{d\times 4096}$ is a trainable, bias-free projection. For item j, the enriched vector is $\tilde{e}_j = e_j + Pc_{w,j}$. All 4096 components are retained; whitening changes their scaling and coordinates rather than reducing dimensionality. Each corpus's matrix is computed once and reused across seeds, while E and P are learned with the recommender. The raw-semantic control replaces $c_{w,j}$ with c_j. PCA coordinates are used in all final three-corpus comparisons; historical ZCA diagnostics are identified separately.

3.2 Tied Sequential Retriever

The retriever looks up history embeddings in $\widetilde{E}$ and processes them with a two-layer causal SASRec encoder to obtain user state h_t. During training, history dropout independently removes each history event with probability p, keeping at least one. This produces perturbed history views without increasing the number of tail targets. The same enriched table scores every catalog item:

$$s_j = h_t^\top \tilde{e}_j + b_j, \qquad \mathcal{L}_{\text{rec}} = -\log \frac{\exp(s_y)}{\sum_{j\in\mathcal{I}} \exp(s_j)}, \tag{2}$$

where y is the next-item target and b_j is a learned item-specific bias. Weight tying thus routes the semantic residual through both the history representation and the output classifier. The frozen semantic matrix is not updated; its projection, the ID table, encoder and item biases are optimized under full-softmax cross-entropy. Diagnostic probes untie the two tables to distinguish their geometry (Section 4.2); the deployed retriever remains tied.

3.3 Fixed-Top-10 Residual Reranking

The second stage holds the retriever and its candidate set fixed. For $S_{10} = \text{Top10}(s)$, a small residual head computes

$$s'_j = s_j + \gamma\Delta_j, \qquad j \in S_{10}. \tag{3}$$

Its five features are the within-list standardized retriever score, log rank, log training frequency, the log transition count from the last history item, and the log maximum history–target co-occurrence count. Rank is zero-based and each logarithmic feature uses $\log(1+x)$. LayerNorm and an MLP map these features to Δ_j; the final map is zero-initialized and γ starts at one. The head has 17.4K parameters. The final ranking permutes only the first ten positions and appends the unchanged suffix, rather than globally sorting modified scores. Consequently, Recall@10 is invariant, although within-list metrics can change.

Candidate lists for head training are generated out of fold with user-disjoint retrievers. Frequency and co-occurrence features use only training prefix–target pairs, excluding validation and test targets. Candidate generation is cross-fitted; count features are pooled across the training split. The head uses listwise cross-entropy on examples whose target appears in S_{10} and is selected by validation NDCG@5. Appendix J gives the fold construction and optimization settings; Appendix L tests the feature contributions. Appendix N separates candidate-generation and head-training costs.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation. Main comparisons use Beauty, Toys and Sports with chronological leave-two-out splits after 5-core filtering: the last event is the test target and the penultimate event is validation. We retain the last L earlier events before expanding training prefix–target pairs. We rank the full catalog without removing consumed items and report Recall and NDCG at 5 and 10 on the complete test set. All models use the exported splits and a common evaluator. The Amazon history window is $L = 20$; ML-1M is used only for a window-length protocol control. Dataset statistics and detailed evaluation settings are in Table 9 and Appendix I.

Models and statistical reporting. We compare ten baselines using the model-specific implementations and training settings in Appendix M. Our $d = 64$ pure-ID trunk supports diagnosis; final retrieval uses $d = 192$, history dropout and PCA-whitened flan-t5-xxl semantics, without a gate.

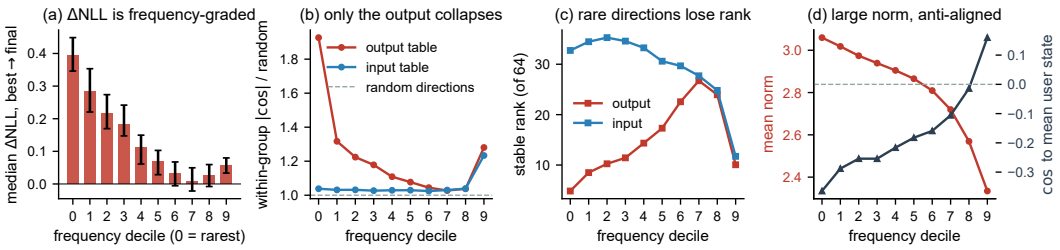


Figure 2: Output-side collapse on Beauty (d=64). (a) Target NLL degradation in a tied model, with 95% paired-bootstrap intervals. (b–d) Untied probes. (b) Pairwise $|\cos|$ relative to random directions: only the output table forms a rare-item cone. (c) Stable rank. (d) Rare-item output vectors have larger norms and are anti-aligned with the mean user state. Probe settings and uncertainty definitions are in Appendix E.

The final system adds the fixed-top-10 residual. Main tables report mean and sample SD over retriever seeds $\{2020, \ldots, 2024\}$; one trained reranker per corpus is shared across these seeds. SD thus reflects retriever-seed variation with a fixed reranker. Retriever training uses at most 300 epochs, patience 30 and a 30-epoch minimum. Selection criteria and model-specific settings are specified in Appendices G and M; statistical procedures are in Appendix J.

4.2 Output Embedding Collapse

Frequency-dependent loss degradation. We first localize the degradation in likelihood space. Comparing each example's NLL between the best-validation checkpoint and the final checkpoint of a converged SASRec-class model on Beauty, the paired degradation is sharply frequency-graded. The median ΔNLL is $+0.394$ for targets in the rarest decile against $+0.057$ for the most popular decile, an approximately $6.9\times$ gap. The absolute tail NLL of 13.4 exceeds the uniform-distribution reference of $\log|\mathcal{I}| = 9.40$, showing poor likelihood for rare targets. The uniform reference is frequency-agnostic; we next examine output-table geometry to characterize how the learned representations accompany this loss.

Input–output geometry. Figure 2 compares the two tables: the input and output tables diverge by frequency decile, and only the output side collapses. For each frequency group, let D contain the unit-normalized item vectors as rows. We measure stable rank as $\mathrm{srank}(D) = \|D\|_F^2/\|D\|_2^2$, separating directional concentration from embedding norms. Pairwise absolute cosine excludes self-pairs and is divided by the approximate isotropic reference $\sqrt{2/(\pi d)}$; values above one indicate excess concentration. For the standard tied model, within-group pairwise cosine of the rarest decile reaches $1.39\times$ the random-direction reference ($\sqrt{2/\pi d} \approx 0.0997$ at d=64) with stable rank $7/64$. Untying the input and output tables separates the two candidate loci cleanly. The *input* table is near the random reference in the first nine frequency groups (1.02–$1.04\times$), rising to $1.23\times$ in the hottest group. The *output* table instead collapses in a frequency-graded manner: tail cone $1.93\times$ random, stable rank $4.8/64$, mean cosine to the average test-time user state -0.36, and a norm–frequency ordinal-rank correlation of -0.62. Tail items therefore acquire *large-norm, anti-aligned* directions, consistent with negative dot-product contributions. Complete logits also include learned item biases. The diagnostics describe the checkpoint's joint learned state, with direction, norm and bias coupled in the observed target ranks. The width control shows a metric trade-off rather than uniform recovery. In Table 6, widening the pure-ID trunk from d=64 to 128 raises NDCG@5 but lowers Recall@10 on Beauty. Increasing width alone therefore does not recover retrieval in this comparison. The whitened model benefits more consistently from widening (§4.4). Untying reveals different geometries for history lookup and candidate scoring; a tied table alone cannot separate these roles.

Controlled imbalance experiments. Figure 4 summarizes the imbalance controls. To examine imbalance we build a *frequency-capping ladder*, where rung C caps every item's training occurrences at C. The ladder ratio is $R = C/n_{\mathrm{tail}}$, where n_{tail} is the fixed tail group's median count in the natural corpus. Each rung is paired with a volume-matched control that deletes the same

Table 1: Beauty results (mean±SD, five seeds). Bold/underline denote best/runner-up values. Ours, final uses one shared fixed reranker across five retriever seeds; Recall@10 is unchanged.

Model	R@5	R@10	N@5	N@10
GRU4Rec (Hidasi et al., 2016)	0.0401 ±.0017	0.0583 ±.0025	0.0279 ±.0011	0.0338 ±.0013
SASRec (Kang & McAuley, 2018)	0.0475 ±.0009	0.0643 ±.0013	0.0327 ±.0004	0.0381 ±.0005
BERT4Rec (Sun et al., 2019)	0.0043 ±.0004	0.0076 ±.0005	0.0027 ±.0002	0.0038 ±.0002
FDSA (Zhang et al., 2019)	0.0488 ±.0014	0.0683 ±.0020	0.0355 ±.0012	0.0418 ±.0014
FMLP-Rec (Zhou et al., 2022)	0.0268 ±.0010	0.0440 ±.0011	0.0136 ±.0005	0.0192 ±.0005
MELT (Kim et al., 2023)	0.0218 ±.0022	0.0398 ±.0026	0.0129 ±.0010	0.0186 ±.0010
LLM-ESR (Liu et al., 2024)	0.0193 ±.0007	0.0370 ±.0020	0.0114 ±.0006	0.0170 ±.0009
ActionPiece (Hou et al., 2025)	0.0506 ±.0008	<u>0.0776 ±.0016</u>	0.0342 ±.0007	0.0429 ±.0008
GenAIR (Li et al., 2026)	0.0219 ±.0011	0.0429 ±.0018	0.0124 ±.0005	0.0191 ±.0006
TADA (Dang et al., 2026)	0.0355 ±.0017	0.0619 ±.0025	0.0190 ±.0013	0.0275 ±.0015
Ours, retrieval	<u>0.0609 ±.0012</u>	**0.0912 ±.0011**	<u>0.0407 ±.0012</u>	<u>0.0505 ±.0010</u>
Ours, final	**0.0642 ±.0005**		**0.0454 ±.0002**	**0.0541 ±.0003**
Improv. over best baseline	+26.9%	+17.4%	+27.7%	+25.9%

number of raw events uniformly at random, before survivor filtering and history truncation, without item-frequency capping. The fixed tail group reaches cosine 0.33 under the cap-4 random-deletion control. Under capping, its cosine rises from 0.03 at R=1 to 0.095 at R=30. A layer-peeled (unconstrained-features) model also exhibits a sharp transition: at classifier regularization λ_w=0.01 the within-tail mean cosine similarity jumps from 0.19 to 0.93 inside $R \in [2, 5]$; the weaker-regularization sweep does not show this transition. The ladder fixes the tail item group, not the evaluated users or sequence contexts, which change after deletion. Appendix E reports the population counts and diagnostic procedures.

4.3 Recommendation Performance

Main comparison. Tables 1 and 2 report the same ten baselines on all three corpora. On Beauty, ActionPiece leads the baselines on Recall and NDCG@10, while FDSA leads on NDCG@5, motivating a panel that includes both ID-only and feature-aware methods. The retrieval stage improves Recall@10; the fixed top-10 residual improves within-list ordering without changing that recall. Relative improvements use the strongest of the ten baselines for each metric.

Cross-dataset. At the final width d=192, matched controls show Recall@10 gains over pure ID of 22.9%, 19.9% and 29.3% on Beauty, Toys and Sports, respectively (Table 3). With common optimizer, head settings and PCA whitening across corpora (Appendix J), the final system exceeds the strongest reproduced baseline on all twelve dataset–metric combinations. The five-seed mean improvements are 17.4–27.7% on Beauty, 31.6–40.5% on Toys and 37.7–47.8% on Sports.

On fixed seed-2020 retrievers and shared out-of-fold caches, five head initializations isolate the reranker's feature contributions (Appendix L). Score/rank calibration changes validation NDCG@5 by at most 0.4%; adding frequency yields at most 0.8% over the retriever. Full features yield gains of 15.7%, 10.7% and 20.4% on Beauty, Toys and Sports, respectively. Co-occurrence counts supply complementary ranking information beyond retriever scores. The final recipe uses no frequency gate; its negative controls and the protocol audit are reported below.

4.4 Ablation and Sensitivity Analysis

Component contributions. Table 3 compares pure ID, raw semantics and PCA-whitened semantics at the final width d=192. Whitening improved all four five-seed mean metrics over raw semantics on every corpus. Raw semantics increased Recall@10 but lowered NDCG@5 on Beauty and Toys relative to pure ID. The pure-ID component-removal control also removes the projection parameters. Raw and whitened semantics share the same encoder, projection size and trainable parameter count, isolating the preprocessing choice. Appendix G reports paired tests and historical dropout and gate controls (Tables 5 and 7).

Table 2: Toys and Sports results with the same complete baseline panel and conventions as Table 1 (mean±SD, five seeds). Final rows share one fixed reranker across retriever seeds. ActionPiece's Sports results retain the reproducible degenerate seed (Appendix M.1).

Dataset	Model	R@5	R@10	N@5	N@10
Toys	GRU4Rec	0.0329 ±.0003	0.0464 ±.0006	0.0238 ±.0003	0.0281 ±.0004
	SASRec	0.0527 ±.0013	0.0680 ±.0014	0.0366 ±.0015	0.0415 ±.0016
	BERT4Rec	0.0012 ±.0003	0.0022 ±.0008	0.0006 ±.0002	0.0009 ±.0003
	FDSA	0.0495 ±.0011	0.0664 ±.0016	0.0374 ±.0007	0.0429 ±.0008
	FMLP-Rec	0.0310 ±.0008	0.0447 ±.0006	0.0161 ±.0005	0.0205 ±.0002
	MELT	0.0199 ±.0011	0.0353 ±.0005	0.0120 ±.0007	0.0169 ±.0005
	LLM-ESR	0.0209 ±.0008	0.0381 ±.0018	0.0120 ±.0007	0.0175 ±.0006
	ActionPiece	0.0498 ±.0013	0.0751 ±.0020	0.0329 ±.0002	0.0410 ±.0005
	GenAIR	0.0259 ±.0009	0.0479 ±.0016	0.0140 ±.0003	0.0211 ±.0005
	TADA	0.0473 ±.0015	0.0740 ±.0007	0.0242 ±.0007	0.0328 ±.0003
	Ours, retrieval	0.0687 ±.0018	**0.0989 ±.0016**	0.0473 ±.0013	0.0570 ±.0011
	Ours, final	**0.0717 ±.0018**		**0.0515 ±.0011**	**0.0603 ±.0010**
	Improv. over best baseline	+36.0%	+31.6%	+37.5%	+40.5%
Sports	GRU4Rec	0.0183 ±.0008	0.0296 ±.0009	0.0120 ±.0006	0.0156 ±.0006
	SASRec	0.0251 ±.0005	0.0352 ±.0007	0.0164 ±.0004	0.0197 ±.0004
	BERT4Rec	0.0039 ±.0004	0.0064 ±.0007	0.0023 ±.0002	0.0031 ±.0003
	FDSA	0.0263 ±.0006	0.0386 ±.0004	0.0183 ±.0008	0.0223 ±.0005
	FMLP-Rec	0.0110 ±.0004	0.0182 ±.0004	0.0055 ±.0002	0.0079 ±.0002
	MELT	0.0180 ±.0004	0.0300 ±.0004	0.0114 ±.0005	0.0153 ±.0004
	LLM-ESR	0.0141 ±.0002	0.0251 ±.0004	0.0085 ±.0002	0.0120 ±.0002
	ActionPiece	0.0261 ±.0115	0.0405 ±.0167	0.0173 ±.0074	0.0219 ±.0091
	GenAIR	0.0157 ±.0008	0.0278 ±.0007	0.0093 ±.0005	0.0132 ±.0005
	TADA	0.0202 ±.0008	0.0358 ±.0008	0.0105 ±.0005	0.0156 ±.0006
	Ours, retrieval	0.0367 ±.0010	**0.0559 ±.0018**	0.0234 ±.0004	0.0296 ±.0006
	Ours, final	**0.0388 ±.0007**		**0.0268 ±.0005**	**0.0323 ±.0009**
	Improv. over best baseline	+47.8%	+37.7%	+46.4%	+45.2%

Table 3: Matched d=192 retrieval controls, mean±sample SD across five seeds. All rows use history dropout 0.15 and validation NDCG@10 checkpoint selection and early stopping. No row includes the top-10 reranker.

Dataset	Variant	R@5	R@10	N@5	N@10
Beauty	Pure ID	.05193±.00083	.07418±.00093	.03640±.00087	.04357±.00082
	Raw	.05457±.00162	.08188±.00177	.03614±.00049	.04494±.00046
	PCA-whitened	.06087±.00120	.09120±.00114	.04072±.00117	.05046±.00097
Toys	Pure ID	.06015±.00067	.08243±.00123	.04325±.00051	.05044±.00036
	Raw	.06031±.00102	.08734±.00136	.04074±.00088	.04945±.00075
	PCA-whitened	.06871±.00184	.09888±.00156	.04726±.00131	.05698±.00109
Sports	Pure ID	.02811±.00086	.04321±.00143	.01796±.00057	.02281±.00059
	Raw	.03252±.00114	.05251±.00083	.01836±.00060	.02480±.00059
	PCA-whitened	.03668±.00102	.05585±.00175	.02344±.00039	.02959±.00057

Embedding width. Table 6 reports PCA retrieval controls at widths 128 and 192. Widening to 192 increases retrieval NDCG@5 but reduces Recall@10 on all three corpora. The same frozen reranker changes this trade-off: width 128 has higher test means on all four metrics on Toys and Sports (Table 13). We use the width-192 recipe selected on Beauty validation across all three corpora.

Encoder ablation. Earlier Beauty d=64 controls compare frozen encoders under the same recommender settings, with encoder-dependent input dimensions and projection sizes. Five-seed Recall@10 is 0.0831 for bge-m3 and 0.0857 for raw flan-t5-xxl. The whitened flan-t5-xxl config-

uration, also switching selection from Recall@10 to NDCG@10, reaches 0.0934 ± 0.0011. The final d=192 controls isolate whitening under matched selection; Appendix F identifies the historical configurations.

Frequency gating. In the earlier diagnostic configurations (PCA on Beauty, ZCA on Toys/Sports), attenuating the semantic residual by $g(n) = n/(n + \lambda)$ during training reduces all four metrics on all three Amazon corpora. The paired intervals exclude zero, and the loss extends to rare items: on Beauty tail Recall@10 drops from 0.0370 to 0.0269, below even the pure-ID trunk. In these controls, attenuating the whitened residual by rarity removes useful signal, and neither a corpus-relative scale nor an inverted gate rescues it (Appendix C).

Additional checkpoint-selection and fusion controls are reported in Appendix G. The ungated configuration outperforms both tested gate directions.

4.5 Geometry and Head–Tail Trade-offs

Scoring-table geometry. Naively, the semantic residual appears to fix the collapse, since the raw table's tail cone drops sharply. But the *scoring* table is $E + CP^\top$, and frozen sentence embeddings are themselves strongly anisotropic. Measured on the encoder the method actually uses, the projected semantic channel Pc forms a cone far tighter than the collapse it is meant to repair. Its tail pairwise $|\cos|$ is $7.6\times$ the random reference at stable rank 1.3, and the scoring table inherits that at $4.9\times$ and stable rank 2.0. The learned scoring table thus inherits the semantic channel's anisotropy. Separating the semantic direction from its orthogonal complement clarifies this effect. The component *perpendicular* to the semantic direction is less concentrated ($|\cos| = 1.5$–$1.8\times$ random, with tail stable rank 9.6–11 above the head's 5.6).

In these diagnostic checkpoints, whitening the same matrix reduces the semantic cone from $7.6\times$ to $1.3\times$ random and lifts its stable rank from 1.3 to 7.6; the scoring table follows, from $4.9\times$ at rank 2.0 to $1.1\times$ at rank 16.7. Whitening therefore reduces anisotropy in the learned scoring table. The final-width matched analysis in Appendix G tests this geometric effect across three corpora and five seeds.

Final-width geometry and long-tail performance. The final d=192 controls show aggregate gains alongside reduced scoring-table concentration (Appendix G). In the matched PCA controls, whitening improved tail Recall@10 relative to raw semantics on all corpora, but Sports remained below pure ID (0.01430 versus 0.01667). Tail coverage was lower than pure ID on every corpus. Here, tail denotes the bottom half of items by training-window frequency, not the rarest decile used for the geometric diagnosis. Tail coverage measures the fraction of tail items recommended at least once, whereas tail Recall@10 measures recovery of held-out tail targets. Improved target recovery can therefore coexist with a narrower set of recommended items.

Across the 15 matched raw/whitened checkpoint pairs, the rarest decile's scoring-table stable rank rises from 2.05 to 49.69 on Beauty, 2.01 to 43.95 on Toys and 1.85 to 31.37 on Sports (five-seed means). The corresponding absolute-cosine ratios fall from 8.40, 8.61 and 9.36 to 1.17, 1.18 and 1.22. These post-training measurements show that whitening improves the learned scoring-table geometry across corpora.

Training-time versus post-hoc interventions. The sweeps in Appendix B compare training-time and post-hoc interventions. Post-hoc centering, row normalization and gating do not reproduce the gains from whitening before training. The contrast supports learning dynamics as a contributor to the repair.

The final system also improves mean NDCG@5 over its retriever in every user-activity tertile (Appendix H).

4.6 Case Study

We compared exact target-item ranks for the same Beauty test examples under three earlier configurations: pure ID, whitened semantics and its frequency-gated counterpart. These seed-2020 diagnostics are distinct from the final d=192 five-seed system and exclude its top-10 reranker.

Case A: recovery of rare targets. A replacement brush-head pack with two training observations moved from rank $1{,}838$ under pure ID to rank 2 with whitened semantics. A detangling brush with six observations moved from $1{,}053$ to rank 1. Both targets entered the top-10, improving retrieval rather than only within-list ordering.

Case B: rare-target displacement and gating response. A Kiehl's serum with six training observations instead fell from rank 33 to 148 with whitened semantics; frequency gating recovered it to rank 4. A Macadamia curl cream followed the same pattern, $10 \rightarrow 47 \rightarrow 5$. These examples show that semantics can also displace rare targets and that gating can reverse individual losses.

Five-seed aggregate metrics and tail Recall@10 favor the ungated configuration in these diagnostic comparisons (§4.4), motivating the final design. Appendix K.3 gives tail definitions and cross-seed checks.

4.7 Efficiency and Protocol Sensitivity

Efficiency accounting. Appendix N reports per-epoch training and inference costs alongside model size and offline encoding measurements. Retriever optimization, offline semantic encoding, out-of-fold candidate generation and residual-head training are distinct costs. Candidate generation requires additional retriever fits, whereas head training uses cached lists. Hardware and measurement scope accompany each timing table.

Protocol audit. Removing consumed items raises the same Beauty checkpoint's NDCG@10 from 0.0432 to 0.0556 ($+28.7\%$). Evaluating ML-1M with $L{=}20$ yields 58% of its Recall@10 at $L{=}200$. History exclusion changes candidate competition; window truncation changes history and user state even with fixed weights. Appendix A audits hard and soft history exclusion, window length and candidate sampling.

5 Conclusion

We identify frequency-graded directional collapse in the output embeddings of full-softmax sequential recommenders. Untied probes localize the failure, and imbalance experiments support an optimization explanation. Untying distinguishes the input and output roles. The resulting repair separates tail retrieval, semantic geometry and within-list precision. History dropout strengthens tail retrieval; whitening before training reduces concentration in the learned scoring table. The repaired retriever improves Recall@10 by 20–29% over matched $d{=}192$ pure-ID controls. A 17.4K-parameter residual reranker further improves ordering while preserving Recall@10 exactly. Across three corpora, the final system outperforms all ten reproduced baselines on all four metrics, with relative gains of approximately 17–48% under matched full-catalog evaluation. This work establishes output-table geometry as a practical diagnostic target and connects its repair to substantial gains in recommendation accuracy.

Limitations. Our mechanistic evidence centers on full-softmax SASRec, with final-system evaluation on three Amazon corpora. Learned directions, norms and item biases remain coupled, so their individual contributions to ranking are not fully isolated. Whitening improves aggregate accuracy but can reduce tail coverage, while fixed-top-10 reranking cannot recover items missed during retrieval.

Future work. Extending the analysis to sampled objectives, other backbones and larger catalogs would test the generality of the observed collapse. Targeted norm and bias interventions could further isolate the optimization mechanism. Coverage-aware training and broader candidate retrieval offer directions for improving the accuracy–coverage trade-off beyond within-list reranking.

AI USE STATEMENT

Generative AI tools assisted with implementation, hypothesis refinement, experiment design and interpretation, data-processing and analysis code, literature search, manuscript drafting, translation and editing, figure-generation code, and reference formatting. Benchmark interactions and reported metric values were not synthesized by generative AI. Reported numbers were computed by executing the experimental and analysis code; statistical procedures are described in Appendix J. Verification included implementation tests and comparisons with source records. The authors take responsibility for the submitted text, claims, references, code and experimental results.

ETHICS STATEMENT

This work uses public benchmark corpora (Amazon Reviews 2014 and MovieLens-1M), recruits no participants and collects no additional personal data. The method can move rare items up the ranking, which touches popularity bias; we therefore report per-frequency and catalogue-coverage effects alongside aggregate accuracy, and we state explicitly where the repair concentrates the head. The authors are aware of no conflict of interest.

REPRODUCIBILITY STATEMENT

Section 4.1 specifies the frozen evaluation protocol; Appendix I details its implementation; and Appendix J lists the configurations and statistical procedures behind each table. All datasets are public. The accompanying anonymous archive contains the core implementation, preprocessing and evaluation scripts, recorded configurations, efficiency measurements, per-seed metrics, statistical comparisons and core mechanism-diagnostic outputs. Appendix E documents the mechanism experiments. Third-party datasets, pretrained language models and trained checkpoints are not bundled in this code archive.

REFERENCES

Yu Cui, Yi Xu, Jiahao Wang, Hao Zhang, Yu Zhang, Xiaoyi Zeng, Can Wang, Jinxin Hu, and Jiawei Chen. SpecFormer: Mitigating embedding and attention collapse via spectral-aware transformer for recommendation. *arXiv preprint arXiv:2607.24025*, 2026.

Yizhou Dang, Zhifu Wei, Minhan Huang, Lianbo Ma, Jianzhe Zhao, Guibing Guo, and Xingwei Wang. Tail-aware data augmentation for long-tail sequential recommendation. In *Proceedings of the ACM Web Conference 2026 (WWW)*, pp. 5886–5897, 2026. doi: 10.1145/3774904.3792171.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021. doi: 10.1073/pnas.2103091118.

Xingzhuo Guo, Junwei Pan, Ximei Wang, Baixu Chen, Jie Jiang, and Mingsheng Long. On the embedding collapse when scaling up recommendation models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16891–16909, 2024.

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, Jianmo Ni, Zhankui He, Noveen Sachdeva, Wang-Cheng Kang, Ed H. Chi, Julian McAuley, and Derek Zhiyuan Cheng. ActionPiece: Contextually tokenizing action sequences for generative recommendation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24004–24024, 2025.

Clark Mingxuan Ju, Liam Collins, Leonardo Neves, Bhuvesh Kumar, Louis Yufeng Wang, Tong Zhao, and Neil Shah. Generative recommendation with semantic IDs: A practitioner's handbook. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 6420–6425, 2025. doi: 10.1145/3746252.3761612.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018.

Kibum Kim, Dongmin Hyun, Sukwon Yun, and Chanyoung Park. MELT: Mutual enhancement of long-tailed user and item for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 68–77, 2023. doi: 10.1145/3539618.3591725.

Sunkyung Lee, Minjin Choi, Eunseong Choi, Hye-young Kim, and Jongwuk Lee. GRAM: Generative recommendation via semantic-aware multi-granular late fusion. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33294–33312, 2025. doi: 10.18653/v1/2025.acl-long.1596. URL https://aclanthology.org/2025.acl-long.1596/.

Yifan Li, Jiahong Liu, Xinni Zhang, Hao Chen, Yankai Chen, Wenhao Yu, Jianting Chen, and Irwin King. Generative archetype-grounded item representations for sequential recommendation. In *Proceedings of the ACM Web Conference 2026*, pp. 6642–6653, 2026. doi: 10.1145/3774904.3792587.

Qidong Liu, Xian Wu, Yejing Wang, Zijian Zhang, Feng Tian, Yefeng Zheng, and Xiangyu Zhao. LLM-ESR: Large language models enhancement for long-tailed sequential recommendation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10299–10315, 2023.

Kulin Shah, Bhuvesh Kumar, Neil Shah, and Liam Collins. Masked diffusion for generative recommendation. *arXiv preprint arXiv:2511.23021*, 2025.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1441–1450, 2019. doi: 10.1145/3357384.3357895.

Maolin Wang, Tongshu Bian, Ziyan Wang, Xiaotong Jiang, Binhao Wang, Derong Xu, Wanyu Wang, Ruocheng Guo, and Xiangyu Zhao. SAGE: Global semantic alignment with LLMs for long-tail sequential recommendation. In *Proceedings of the ACM Web Conference 2026 (WWW)*, pp. 6433–6444, 2026. doi: 10.1145/3774904.3792456.

Liu Yang, Fabian Paischer, Kaveh Hassani, Jiacheng Li, Shuai Shao, Zhang Gabriel Li, Yun He, Xue Feng, Nima Noorshams, Sem Park, Bo Long, Robert D. Nowak, Xiaoli Gao, and Hamid Eghbalzadeh. Unifying generative and dense retrieval for sequential recommendation. *Transactions on Machine Learning Research*, 2025.

Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S. Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, and Xiaofang Zhou. Feature-level deeper self-attention network for sequential recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4320–4326, 2019. doi: 10.24963/ijcai.2019/600.

Yifei Zhang, Hao Zhu, Yankai Chen, Zixing Song, Piotr Koniusz, and Irwin King. Mitigating the popularity bias of graph collaborative filtering: A dimensional collapse perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1893–1902, 2020. doi: 10.1145/3340531.3411954.

Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. Filter-enhanced MLP is all you need for sequential recommendation. In *Proceedings of the ACM Web Conference (WWW)*, pp. 2388–2399, 2022.

APPENDIX

A PROTOCOL AUDIT

Evaluation choices that touch the candidate set or frequency mix can move reported numbers more than typical method deltas. We examine four such relaxations on frozen checkpoints; candidate sampling is treated separately in Appendix D.

History exclusion, removing consumed items from the candidate set, inflates the same Beauty checkpoint from 0.0432 to 0.0556 NDCG@10, a $+28.7\%$ shift, comparable to the final system's NDCG@10 improvement over the strongest reproduced baseline in Table 1. **Window choice** also matters: on ML-1M, the Amazon-style $L{=}20$ window yields Recall@10 0.1449, only 58% of the $L{=}200$ value on the same checkpoints.

History-aware score calibration differs from hard candidate exclusion: it changes scores while retaining previously consumed items. Chronological leave-two-out alone does not guarantee that a target is absent from its history when repeated interactions are retained. The effect of such calibration therefore depends on repeat behavior as well as the scoring rule. Candidate filtering, score calibration, context window, split chronology and selection metric should be specified separately when comparing results.

B WHERE THE POPULARITY PRIOR LIVES

B.1 POST-HOC DIRECTION AND MAGNITUDE CONTROLS

The cone hypothesis admits two candidate carriers of the implicit popularity prior: the shared *direction* (the mean semantic vector) and the frequency-graded *magnitude* (the norm gradient $25 \to 31$). We test centering and row normalization of the frozen semantic matrix. Centering does not produce a detectable aggregate change across five seeds (paired NDCG@10 $+0.0001$, exploratory 95% CI $[-0.0004, +0.0006]$), and does not restore tail coverage. Row normalization likewise yields no clear improvement in these controls. Both controls act on the frozen semantic matrix; direction and magnitude remain coupled in the learned combined table.

B.2 TRAINING-TIME AND POST-HOC INTERVENTIONS

The following are earlier single-seed diagnostic probes, not the final $d{=}192$ matched comparison in Appendix G. First, **whitening** the semantic matrix before training improves Recall@10 by $+2.9\%$ over the strongest un-whitened variant on a single seed. The per-decile decomposition shows a trade-off: tail Recall@10 falls by 16% and catalog coverage drops from 0.91 to 0.82 despite the overall gain. Second, a **frequency gate** $g(n) = n/(n + \lambda)$ attenuates the semantic residual on rare items. The idea was that the tail would keep its dropout-densified ID geometry while the head kept its content. On single-seed Beauty probes the gate did appear to recover the NDCG columns that whitening alone sacrifices ($N@5$ $0.0337 \to 0.0370$) together with most of the coverage loss. Five matched seeds on three corpora instead show consistent gating losses (§4.4), motivating its exclusion from the final model. Applying the *same* gate post hoc to an already-trained model strictly degrades every metric, at every tested λ and insertion point. The inverse gate, which attenuates the head instead, also reduces accuracy in the tested settings.

The training-time and post-hoc interventions behave differently in these probes, supporting a role for learning dynamics in the observed effects. Appendix G separates final PCA results from historical long-tail controls.

C FREQUENCY GATING IN FULL

Frequency gating tests whether attenuating rare-item semantics can improve the balance between head and tail accuracy. The gate scales the semantic residual by $g(n) = n/(n + \lambda)$ during training, so rare items receive an attenuated content signal. Figure 3 collects the evidence against it.

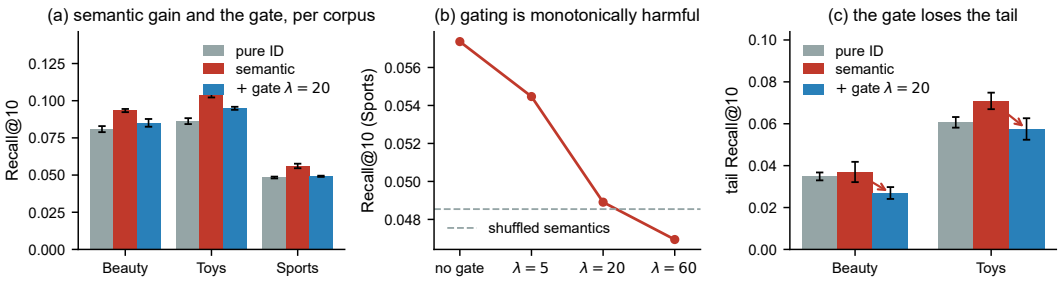


Figure 3: Frequency gating fails in the historical controls (PCA on Beauty, ZCA on Toys/Sports). Five-seed means in (a,c) use ± 1 s.d.; (b) is a seed-2020 sweep. (a) Gating lowers Recall@10 on all corpora. (b) The Sports sweep is monotonically harmful and approaches a shuffled-text control. (c) It also lowers tail Recall@10 below both the ungated recipe and pure-ID trunk on Beauty and Toys.

For the λ=20 controls, at five matched seeds per corpus with everything except the gate held fixed, gating loses on all four metrics on all three Amazon corpora. A paired bootstrap over users with seeds matched puts every interval clear of zero. On Beauty the ungated recipe leads by $+0.0052$ Recall@5 (95% CI $[+0.0039, +0.0066]$), $+0.0083$ Recall@10 ($[+0.0066, +0.0099]$), $+0.0017$ NDCG@5 ($[+0.0009, +0.0025]$) and $+0.0027$ NDCG@10 ($[+0.0019, +0.0035]$); on Toys the corresponding Recall@10 and NDCG@5 gaps are $+0.0089$ ($[+0.0070, +0.0108]$) and $+0.0012$ ($[+0.0002, +0.0022]$). Sports agrees.

Per-decile results show that gating lowers both head accuracy and tail coverage. On Beauty, moving from the ungated recipe to λ=20 drops tail Recall@10 from 0.0370 to 0.0269, head Recall@10 from 0.1854 to 0.1741, catalog coverage from 0.810 to 0.723, and tail slot share from 0.113 to 0.092. Toys shows the same direction of change. The ungated whitened residual improves both measured head and tail groups: relative to the pure trunk it raises tail Recall@10 by 6.0% on Beauty and 16.8% on Toys while also raising head Recall@10 by 15.9% and 20.9%. Gating attenuates that signal in proportion to rarity, which removes most of it where it is scarcest and some of it everywhere else. Appendix Table 10 shows the same conclusion from the other direction: on Sports the gated and shuffled-matrix controls have similar Recall@10 (0.0489 and 0.0485) in this single-seed comparison.

We also tested corpus-relative scaling and the inverse gate. The median positive training count is 6 on all three corpora and λ=20 suppresses 87–89% of items on each. Setting λ to that median instead of the tuned 20 gives $0.0562/0.0865/0.0357/0.0454$ on Beauty over five seeds, which a paired bootstrap cannot separate from λ=20 on three of four metrics, and both remain below the ungated recipe. Inverting the gate to attenuate the head also lowers all four metrics (§G).

The paired results distinguish the consistent gating loss from configuration-level variability. These results support excluding the gate from the final model while retaining it as a mechanistic control.

D Candidate Sampling as a Protocol Relaxation

Ranking against 100 sampled negatives removes most of the competition that a rare target faces in the full catalog. This changes the evaluation problem because our diagnosis concerns output directions relative to the entire catalog. We therefore use full-catalog evaluation throughout. Appendix A quantifies the effects of history exclusion and context-window changes under fixed checkpoints.

E Protocols for the Mechanistic Diagnostics

The diagnostics in Figures 2 and 4 are historical Beauty experiments, separate from the final five-seed, width-192 retrieval comparison. We describe their measurement units and training settings below. Run-level provenance and reproduction details are provided in `REPRODUCIBILITY.md` in the anonymous code supplement.

E.1 LIKELIHOOD AND EMBEDDING GEOMETRY

Checkpoints and training. Figure 2(a) compares two states from one tied, pure-ID SASRec run with seed 2020. The selected state is epoch 54, chosen by validation Recall@10, and the final state is epoch 79. Panels (b–d) instead use the selected epoch-46 state of one untied run with seed 2020, which stopped at epoch 71. Both use width 64, two causal Transformer layers, one attention head, dropout 0.5 and history length 20. The stored settings specify learning rate 0.005, weight decay 0.001, batch size 1024, gradient clipping at 5, 500 warmup steps and a learning-rate floor of 0.1 of the peak. They allow at most 300 epochs, with patience 25 and a minimum of 30 epochs. The implementation uses AdamW. These settings differ from the patience-30 final retrieval recipe. Both diagnostics evaluate all 22,363 Beauty test histories against the 12,101-item catalog, with consumed items retained.

Frequency groups and paired likelihood changes. Training-window frequency counts occurrences in each eligible user's last 20 events before the two held-out targets; it is not the number of expanded prefix–target training rows. For panel (a), the test examples are binned by quantiles of their target's training-window count. Duplicate quantile boundaries are removed, and boundary ties enter the lower bin. These bins therefore need not contain equal numbers of examples. For each example we calculate $\Delta\mathrm{NLL} = \mathrm{NLL}_{79} - \mathrm{NLL}_{54}$ before aggregation. Each displayed interval is a marginal 95% percentile interval for the bin's median paired difference, obtained from 1,000 resamples of its examples with replacement and a fixed bootstrap generator seed of 0. The intervals condition on the two fitted states and do not quantify training-seed uncertainty. In the retained record, the rarest bin contains 3,504 targets with counts 0–3, and the most frequent bin contains 2,233 targets with counts 64–361. Their median differences are 0.39434 and 0.05732. The rarest bin's median NLL at the selected state is 13.36946, compared with the uniform-catalog reference $\log(12{,}101) = 9.40104$.

Geometry and reference distribution. For panels (b–d), catalog items are sorted by training-window count with stable tie handling and divided into ten nearly equal-sized item groups. This item-based grouping differs from the target-example quantiles in panel (a). For a group with normalized row directions D, we compute $\mathrm{srank}(D) = \|D\|_F^2/\sigma_{\max}(D)^2$ on the entire group, without centering the vectors. Pairwise cosine statistics omit the diagonal and use all pairs among at most 800 sampled items per group; the measurement generator starts at seed 0. Absolute cosine is divided by $\sqrt{2/(\pi d)}$, the large-d approximation to the expected absolute inner product of independent uniform unit directions. At $d = 64$ this reference is 0.0997356; it is an analytic reference, not another trained model. Norms are measured before row normalization.

The context direction is the normalized mean of the encoder states over the complete test set. Panel (d) reports the mean inner product between that direction and the normalized output vectors in each item group. The rarest output group's value is -0.36462. The reported norm–frequency association uses the implementation's ordinal ranks, obtained by sorting each variable twice, and their Pearson correlation; frequency ties are not assigned average ranks.

Interpretation of the geometric measurements. These measurements describe the learned parameterization. With user states frozen, adding the same vector a to every output vector adds $h^\top a$ to every candidate logit. Rankings and softmax probabilities are unchanged, although uncentered angular statistics and stable rank can change. Consequently, negative alignment and its change after whitening are parameter-state diagnostics, not an isolated causal decomposition of relative ranking suppression. Directions, norms and learned item biases remain coupled in the complete scores.

E.2 FREQUENCY-CAPPING LADDER

Construction and retained populations. For each user, the builder separates the original sequence into the training region (all but its last two events) and the two held-out targets. For each cap $C \in \{4, 8, 20, 40, 120\}$, it chooses excess occurrences of each item uniformly without replacement and deletes them from the training region. The retained events keep their original order, and the unchanged validation and test targets are appended. The control arm selects the same total number of training-region events uniformly without replacement across all user-event positions. This matches the raw retained event total before survivor filtering and history truncation, not necessarily the final number of training examples.

The split routine subsequently excludes rebuilt sequences with fewer than five events and retains at most 20 training-region events per surviving user before expanding prefixes. Therefore different rungs and controls have different training and evaluation populations (Table 4). The plotted records use this original construction, not the optional fixed-user mode provided by the builder. Context, survivor composition and the effective training volume remain coupled to the intervention.

Table 4: Recorded Beauty ladder populations. The event total is matched between arms before survivor filtering and the 20-event training-window limit. Each arm has one training seed (2020). Epochs list rung/control training lengths; users are the actual complete-split evaluation counts.

C	R	Raw retained events	Rung users	Control users	Epochs
4	1	46,162	6,010	5,680	55/62
8	2	73,877	11,134	10,170	63/51
20	5	106,997	16,657	15,431	74/43
40	10	128,846	19,497	18,853	46/34
120	30	149,983	21,801	21,839	32/34

Fixed item groups and training. The tail group fixes 4,566 item IDs whose natural-corpus training-window counts are 3–5, with median 4. Thus $R = C/4$ uses a fixed natural reference; it does not assert that each tail item's realized count remains four after deletion. A separate head group contains the 1,000 most frequent natural items. Figure 4(a) plots the signed within-tail cosine of the untied output table, rather than the absolute-cosine ratio in Figure 2(b).

Each of the ten arm–cap runs uses training seed 2020, width 64, two layers, one head, dropout 0.5, learning rate 0.005, weight decay 0.001 and batch size 1024. Checkpoint selection uses validation Recall@10, with a 120-epoch cap, patience 15 and a 30-epoch minimum. The plotted geometry comes from the selected state of each run. These are single-training-run diagnostics without between-seed error bars. The builder uses random generator seeds $1000s + C$ and $1000s + C + 1$ for the rung and control, respectively. Construction metadata and cache provenance are documented in the anonymous code supplement.

E.3 Layer-peeled model

Objective and solver. The accompanying unconstrained-features solver represents each selected class by one free feature vector $\mu_k \in \mathbb{R}^d$ and one classifier row $w_k \in \mathbb{R}^d$, with no sequence encoder or item bias. Given selected training-window counts n_k and $\pi_k = n_k / \sum_{\ell=1}^{K} n_\ell$, its objective is

$$\mathcal{L}_{\mathrm{UFM}} = \sum_{k=1}^{K} \pi_k \left[-\mu_k^\top w_k + \log \sum_{\ell=1}^{K} \exp(\mu_k^\top w_\ell) \right] + \lambda_w \|W\|_F^2 + \lambda_h \sum_{k=1}^{K} \pi_k \|\mu_k\|_2^2. \tag{4}$$

The supplied solver initializes both matrices with independent draws from $\mathcal{N}(0, 0.1^2)$ using the run's random generator, computes in float64, and optimizes both matrices jointly with full-batch Adam at learning rate 0.05 for a fixed number of updates. It applies the penalties explicitly in Equation 4, without a separate neural backbone. The feature penalty is parameterized as $\lambda_h = \rho \lambda_w$.

Class sampling and recorded settings. For each rung, the implementation obtains counts from its rebuilt corpus and samples a class subset. Up to half of the class budget is drawn uniformly from a configured tail-count band. Remaining classes are frequency-sorted, divided into strata, and sampled once per stratum. The same solver seed controls both class sampling and initialization. This tail definition is applied to each rung's counts, not to the 4,566 fixed item IDs used in the transformer ladder. The saved aggregate record confirms $K = 2400$, $d = 64$, 4,000 updates and $\lambda_w \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. The five ladder ratios are $1, 2, 5, 10, 30$. The natural-corpus point is $R = 50.693$, defined by the 99.9th percentile of positive natural counts divided by the median count in the configured tail band; it is not the cap-based ratio of a sixth rung.

Reported summaries. For each solve, the summary is the mean off-diagonal cosine among normalized tail classifier rows. Figure 4(b) displays the saved mean and one standard deviation across

solver replicates, computed with divisor n (`numpy.std`, `ddof=0`). This variability jointly includes class sampling and initialization. These bars describe solver-run dispersion, not confidence intervals or five-seed final-model uncertainty. The saved values at $\lambda_w = 0.01$ are 0.19453 at $R = 2$ and 0.93041 at $R = 5$. The archived solves provide a descriptive mechanism diagnostic rather than an estimate of a universal transition threshold. Their run-level metadata are detailed in the supplementary `REPRODUCIBILITY.md`.

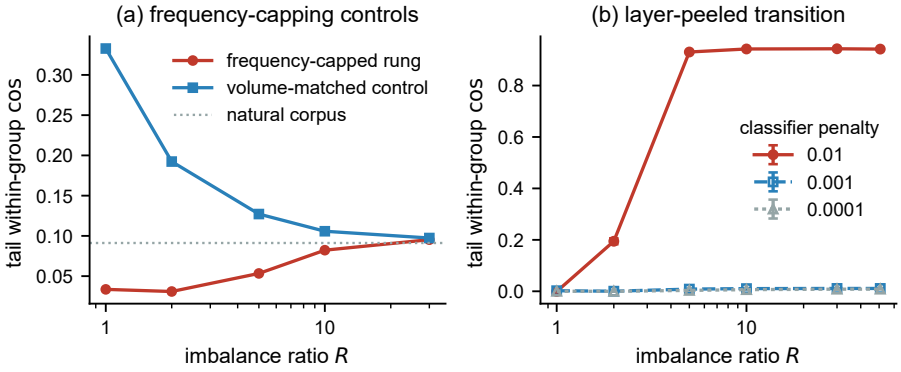


Figure 4: Imbalance controls. (a) Tail cosine under frequency capping and volume-matched random deletion. The tail item group is fixed; surviving users and contexts vary (Table 4). (b) In the layer-peeled model, mean tail cosine rises from 0.19 to 0.93 within $R \in [2, 5]$ at $\lambda_w{=}10^{-2}$; the weaker-regularization sweep does not show this transition. Summary statistics are defined in Appendix E.

F Encoder Ablation in Full

The historical Beauty $d{=}64$ encoder comparison holds history dropout at 0.15 and uses five matched seeds with validation Recall@10 selection. It varies frozen representations, including their input dimensions and projection sizes. Recall@10 reaches 0.0831 with bge-m3 and 0.0857 with flan-t5-xxl (sample SDs 0.0015 and 0.0010, respectively). A separate seed-2020 bge-base probe reaches 0.08295; it is not a five-seed result. A legacy llama matrix with unrecorded encoder provenance reaches 0.0863 and is listed separately as a historical reference. The whitened flan-t5-xxl configuration additionally switches checkpoint selection to validation NDCG@10, reaching Recall@10 0.0934 ± 0.0011, 9.0% above the raw configuration. This historical comparison therefore combines whitening and selection; Table 3 isolates whitening at the final width with selection held fixed. Relative to the historical pure-ID trunk, the whitened configuration raises tail Recall@10 by 6.0%, while catalog coverage decreases from 0.930 to 0.810. It retrieves more held-out tail targets while recommending fewer distinct catalog items.

G Ablation Studies in Full

G.1 Matched controls at the final width

Table 3 compares pure ID, raw semantics and whitened semantics at $d{=}192$, before reranking, with the same five training seeds. PCA whitening improves all four five-seed mean metrics over raw semantics on all three corpora. Raw semantics alone is less consistent: on Toys, mean NDCG@5 and NDCG@10 are below pure ID despite higher Recall@10. The pure-ID control removes both the semantics and their projection parameters, as described in Section 4.4.

The matched-user analysis rejects 32 of the 36 component null hypotheses after Holm correction. All 12 whitened-versus-raw comparisons are positive with adjusted $p \leq 0.0072$. NDCG@5 gains are 0.00458, 0.00652 and 0.00509 on Beauty, Toys and Sports, with marginal 95% confidence intervals $[0.00355, 0.00560]$, $[0.00533, 0.00774]$ and $[0.00436, 0.00580]$. These user-level intervals condition on the fitted models (Appendix J).

Geometry after training. Across the 15 matched raw/whitened checkpoint pairs, the rarest frequency decile's final scoring table, not merely its frozen semantic matrix, has higher stable rank after whitening: $2.05 \rightarrow 49.69$ on Beauty, $2.01 \rightarrow 43.95$ on Toys and $1.85 \rightarrow 31.37$ on Sports (five-seed means). Mean absolute cosine relative to a random-direction reference decreases from 8.40 to 1.17, 8.61 to 1.18, and 9.36 to 1.22, respectively. These paired diagnostics show reduced directional concentration in the learned scoring table, using stable rank as defined in Section 4.2.

Matched long-tail controls. All three corpora use the final PCA checkpoints. Tail items are the bottom 50% of catalog items by training-window frequency, distinct from the rarest-decile geometry group. Tail Recall@10 for pure ID/raw/whitened is $0.03306/0.03440/0.03957$ on Beauty, $0.06274/0.06166/0.07285$ on Toys and $0.01667/0.01152/0.01430$ on Sports. Whitening improves over raw semantics in all three cases but does not recover the Sports pure-ID level. Tail coverage also remains below pure ID: $0.9546 \rightarrow 0.8716$ on Beauty, $0.9327 \rightarrow 0.9091$ on Toys and $0.9450 \rightarrow 0.7093$ on Sports.

G.2 Earlier diagnostic probes

Table 5: Historical Beauty configurations at d=64 (mean±SD, five seeds). Rows add components alongside a history-dropout change from 0.20 (pure ID) to 0.15 (semantic variants). The whitening and gated rows also use validation NDCG@10 selection instead of the earlier rows' Recall@10 selection. The final row adds a frequency gate (λ=6). Bold/underline denote best/runner-up values. Matched final-width controls are in Table 3.

Variant (d=64)	R@5	R@10	N@5	N@10
SASRec, pure ID, no history dropout	0.0473 ±.0013	0.0719 ±.0013	0.0295 ±.0012	0.0375 ±.0011
+ history dropout	0.0527 ±.0021	0.0809 ±.0020	0.0332 ±.0019	0.0423 ±.0018
+ raw flan-t5-xxl residual	0.0531 ±.0010	0.0857 ±.0010	0.0316 ±.0007	0.0421 ±.0007
+ whitening	**0.0606 ±.0020**	**0.0934 ±.0011**	**0.0372 ±.0018**	**0.0477 ±.0016**
+ frequency gate	<u>0.0562 ±.0016</u>	<u>0.0865 ±.0018</u>	<u>0.0357 ±.0012</u>	<u>0.0454 ±.0013</u>

Table 5 reports historical Beauty configurations, complemented by the targeted probes in Table 11. These historical probes are exploratory rather than final-width matched controls. Gating fails in both directions. Attenuating the tail loses on every metric (§4.4); inverting it to attenuate the head loses monotonically in λ. The ungated residual outperforms both directions in these controls. The input channels also interact with preprocessing. Joint whitening of concatenated text and collaborative blocks underperforms the displayed whitened-text reference on all four metrics. Separate whitening improves over joint whitening, but exceeds that reference only on NDCG@5 and NDCG@10. A collaborative-only SVD residual retrieves useful signal without text, but its Recall@10 (0.0816) remains below the reference (0.0895).

H Per-Group Metrics in Full

Users are grouped by training-history length at its one-third and two-thirds quantiles, with boundary ties assigned to the lower group. These five-seed means use the final PCA checkpoints and the same fixed reranker as the main tables. Mean NDCG@5 increases after reranking in all nine corpus–activity groups; Recall@10 is unchanged within each group. Item-frequency groups use training-window counts: the bottom 50%, next 40% and top 10% of catalog items. Full per-seed group metrics and coverage accompany the analysis outputs.

Table 6: Retrieval capacity controls, five-seed means, without reranking. All semantic rows use PCA whitening. The width-128 and width-192 rows share the final recipe apart from width; pure-ID and width-64 rows are earlier diagnostic references. Width-128 checkpoints are also used in Table 13.

Dataset	Configuration	R@5	R@10	N@5	N@10
Beauty	pure ID, d=64	0.0527	0.0809	0.0332	0.0423
	pure ID, d=128	0.0529	0.0776	0.0367	0.0446
	recipe, d=64	0.0606	0.0934	0.0372	0.0477
	recipe, d=128	0.0613	0.0948	0.0391	0.0499
	recipe, d=192	0.0609	0.0912	0.0407	0.0505
Toys	pure ID, d=64	0.0594	0.0863	0.0364	0.0450
	recipe, d=128	0.0693	0.1033	0.0438	0.0547
	recipe, d=192	0.0687	0.0989	0.0473	0.0570
Sports	pure ID, d=64	0.0308	0.0484	0.0190	0.0247
	recipe, d=128	0.0386	0.0598	0.0225	0.0293
	recipe, d=192	0.0367	0.0559	0.0234	0.0296

Table 7: Historical component ablation on Toys and Sports at d=64 with ZCA whitening, five seeds, mean±SD, built the same way as Table 5. Except for the no-history-dropout controls, rows use the trunk's own history-dropout rate (0.15 on Toys, 0.20 on Sports) and validation Recall@10 selection, so the whitened row is a like-for-like ablation step rather than the validation-selected recipe of Table 6. Best and runner-up within each block are **bold** and underlined.

Dataset	Variant	R@5	R@10	N@5	N@10
Toys	pure ID, no history dropout	0.0493 ±.0013	0.0724 ±.0020	0.0305 ±.0014	0.0379 ±.0016
	+ history dropout	0.0594 ±.0018	0.0863 ±.0020	0.0364 ±.0018	0.0450 ±.0019
	+ raw flan-t5-xxl residual	0.0622 ±.0009	0.0944 ±.0009	0.0353 ±.0007	0.0457 ±.0006
	+ whitening	**0.0669 ±.0023**	**0.1022 ±.0014**	**0.0392 ±.0019**	**0.0507 ±.0016**
Sports	pure ID, no history dropout	0.0251 ±.0008	0.0398 ±.0007	0.0156 ±.0007	0.0204 ±.0006
	+ history dropout	0.0308 ±.0006	0.0484 ±.0006	0.0190 ±.0004	0.0247 ±.0003
	+ raw flan-t5-xxl residual	0.0326 ±.0008	0.0526 ±.0015	0.0187 ±.0005	0.0251 ±.0007
	+ whitening	**0.0372 ±.0012**	**0.0582 ±.0012**	**0.0220 ±.0007**	**0.0287 ±.0007**

Table 8: Final PCA system by user activity, five-seed mean NDCG@5.

Corpus	Stage	Low	Middle	High
Beauty	Retrieval	.03733	.03617	.04980
	Final	.04182	.04098	.05461
Toys	Retrieval	.05142	.04903	.03944
	Final	.05633	.05339	.04255
Sports	Retrieval	.02474	.02316	.02126
	Final	.02829	.02735	.02376

I Detailed Evaluation Protocol

Appendix A shows that evaluation choices move reported numbers by more than typical method gaps. We therefore fix the protocol for the reported comparisons and state its settings here. Every number in this paper obeys it unless the text marks the run as a deliberate protocol contrast.

Split and candidates. We use chronological leave-two-out per user: the last interaction is the test target and the second-to-last is validation. After 5-core filtering, we retain the last L events of each user's earlier training region, then expand that truncated sequence into prefix–target pairs. Thus, early training events outside this window do not become targets. Validation and test histories use the preceding original events, truncated to L; the test history includes the validation event. Training-

window item frequencies count occurrences in the truncated training sequences, whereas reranker frequency features count target occurrences in the resulting prefix–target pairs. Candidates are the *full* catalog, and consumed items are *not* removed from the candidate set. Our trunk's exact full-catalog scores support uncensored ranks and MRR. External top-K dumps support Recall/NDCG only within the exported depth; truncated reciprocal ranks are not treated as full-catalog MRR. We evaluate the complete test set, and any subsample would be a seeded uniform draw recorded in the run's protocol block.

Context window. The history window follows each corpus's own convention: $L{=}20$ on the Amazon corpora and $L{=}200$ on ML-1M, fixed for all compared models. Appendix A shows that scoring ML-1M with the Amazon window yields only 58% of the Recall@10 obtained on the same checkpoints at $L{=}200$, so a mismatched window silently changes the benchmark rather than the model.

Training budget and seeds. Our retriever trains for at most 300 epochs, with patience 30 and a minimum of 30 epochs; external baselines use the model-specific settings in Appendix M. Main comparisons report mean and sample standard deviation (SD) over seeds $\{2020, \ldots, 2024\}$. Checkpoint selection uses Recall@10 or NDCG@10 as specified for each comparison (§G). Table 9 in the appendix lists the four public corpora after 5-core filtering.

I.1 Development and Configuration Selection

Beauty's test results were inspected repeatedly during development of the diagnosis, repair channels and encoder ablation. Design decisions were checked on validation, including width selection on Beauty. One Toys test run (seed 2020) was inspected after that width decision but before the complete pipeline was frozen; Sports had not been inspected at that stage. Subsequent runs used common settings across corpora. A later seed-2020 validation comparison of PCA and ZCA on all three corpora motivated the common PCA convention. Toys and Sports were rerun with five seeds, while Beauty retained its existing PCA runs.

I.2 Common Evaluation and Implementation

External baselines consume the exported splits, and their score or top-K dumps are evaluated with the same metric implementation. The history-filtering audit compares candidate policies using a fixed checkpoint (Appendix A). Appendix M describes the baseline implementations.

The reference corpus statistics and protocol block of each external comparison target are pinned in code, and corpus counts are matched within a 0.1% tolerance. The exported benchmark files give the external trainer our $116{,}259$ Beauty training samples, our $22{,}363$ validation targets and our $22{,}363$ test targets, so it never re-derives splits from raw sequences. Our trainer tracks the best checkpoint under both Recall@10 and NDCG@10 within a single run, allowing checkpoint selection within the same recorded trajectory without additional optimization. Separate runs with different early-stopping rules may terminate at different epochs. Corpus counts after 5-core filtering appear in Table 9, are reproduced from our own preprocessing, and are checked against the pinned reference statistics at run time.

J Reproducibility

Statistical analysis. Main comparisons use seeds $\{2020, \ldots, 2024\}$ and report mean and sample SD ($n-1$ denominator), with each seed contributing once. The final-system rows apply one fixed trained reranker to five retrievers, so their SD reflects retriever-seed variation conditional on that head. For the final PCA runs, we average matched-seed differences within users, then use 2,000 percentile-bootstrap draws and 5,000 two-sided sign-flip draws. Holm correction covers a 96-test family comparing the final system with GRU4Rec, SASRec, FDSA, FMLP-Rec, LLM-ESR, GenAIR, TADA and its retriever, and a separate 36-test component family. A separate 12-test ActionPiece family uses Bonferroni correction; LIGER is excluded. Resampling seeds are 20260916–20260918 for Beauty, Toys and Sports. Confidence intervals are marginal, not simultaneous; inference assumes independent users and sign exchangeability. The baseline/retriever family rejects 93 of 96 nulls at adjusted $p < 0.05$; the three exceptions are the exactly unchanged Recall@10 after

reranking. User-level resampling holds the fitted models, shared OOF cache and configuration selection fixed, quantifying variation across evaluated users. Exact two-sided five-seed sign-flip tests have minimum attainable $p = 0.0625$; the significance results above use the user-level tests. Older diagnostic-probe intervals remain exploratory.

Our own optimiser and components. The trunk trains with AdamW at learning rate 5×10^{-3}, weight decay 10^{-3}, batch size 1024, gradient clipping at 5.0, 500 warmup steps and a floor of 0.1 of the peak rate, for at most 300 epochs with patience 30 and a 30-epoch minimum. The architecture is two layers, one head, dropout 0.5, history window 20, history dropout 0.15, output embedding tied to the input table. Matched width controls change only d; historical probes also vary dropout or checkpoint selection as specified in their tables. Final retrieval uses d=192. Item embeddings are initialised $\mathcal{N}(0, d^{-1})$.

For whitening, centre the frozen matrix X and compute the thin SVD $X_c = USV^\top$ in float64. Let $D = \mathrm{diag}(1/\max(S/\sqrt{n}, 10^{-8}))$, where n is the item count. All final comparisons use PCA coordinates, $C_w = X_c VD$, retaining all 4096 components; outputs are stored in float32. Each corpus's matrix is held fixed across five training seeds. Earlier Toys/Sports diagnostics used ZCA coordinates, $X_c VDV^\top$, and are explicitly labeled historical. The two conventions differ by an orthogonal rotation and can produce different optimization trajectories under the same initialized projection. A seed-2020 validation comparison favored PCA on NDCG@10 by 0.54%, 0.83% and 0.42% on Beauty, Toys and Sports, respectively. ZCA nevertheless had higher Recall@10 on Toys and Sports. We adopted PCA as the common preprocessing convention across corpora. The development history is in Appendix I.1. The semantic residual is a bias-free linear map from the encoder width (4096 for flan-t5-xxl) to d, initialised $\mathcal{N}(0, 4096^{-1})$, added to the item table before scoring.

The top-10 residual uses five features. For candidate j at zero-based rank r_j, these are its within-list standardized score (SD floored at 10^{-6}), $\log(1+r_j)$, $\log(1+n_j)$, $\log(1+N_{\text{last}}(i_t, j))$, and $\log(1+\max_{h \in H} N_{\text{any}}(h, j))$. Counts use only training prefix–target pairs. Each pair increments the target frequency n_j, its last-item transition, and one history–target count for each distinct history item. All head-training examples share one count table pooled from the complete training split. Candidate generation uses user-disjoint folds; the pooled count table is reused across folds. Validation and test targets are excluded from these counts.

Features pass through LayerNorm and affine maps $5 \to 128 \to 128 \to 1$, with GELU after the first map. The final map starts at zero and γ at one. The final two affine maps have no intervening nonlinearity. The head uses AdamW, learning rate 3×10^{-5}, weight decay 10^{-2}, batch size 512, at most 60 epochs and patience 15, selected by validation NDCG@5. Out-of-fold candidates use five user-disjoint folds. Unique user IDs are shuffled with seed 2020 and assigned round-robin; each fold retriever omits all training prefixes from its held-out users. Only rows whose target is in the first ten candidates enter the listwise cross-entropy objective. One trained head per corpus is fixed across the five final retriever seeds.

Semantic embeddings. Each frozen semantic matrix retains the encoder identity, pooling rule, metadata fields and item order. Embedding rows follow the same item mapping as the interaction data. Encoder comparisons are reported in Section 4.4; Appendix F identifies the legacy reference matrix separately.

Cost accounting. Instrumented runs record parameter counts by module and, where available, training time, peak GPU memory and device information. Historical logs do not uniformly provide all fields. We separate trainable recommender parameters, offline encoder execution and out-of-fold candidate generation. Appendix N specifies the measurement conditions and cost coverage.

External baselines. Baselines use the exported chronological splits and the common evaluator, with the full catalog as the candidate set and consumed items retained. Table 14 lists the model-specific reproduction settings.

BERT4Rec reproduction settings. Our benchmark export writes one row per *(prefix, target)* pair, which is what makes an external trainer consume the same prefix targets as our trainer. That granularity may disadvantage masked objectives. In an initial Beauty audit the median exported prefix

holds three items and 51.5% hold three or fewer, so a cloze objective that masks a fifth of the sequence has little remaining context. BERT4Rec reaches Recall@10 0.0075 under our scorer in that audit. RecBole's own evaluator reports 0.0082 on the same checkpoints, and SASRec trained from the identical files reaches 0.0643. Both evaluators thus report low Recall@10 for this prefix-based masked-model recipe. Whole training sequences and mask-aware augmentation offer alternative training-sample organizations within the same chronological validation and test boundaries.

K Dataset and Sensitivity Details

Table 9: Corpora after 5-core filtering. Density is $|\mathcal{D}|/(|\mathcal{U}| \cdot |\mathcal{I}|)$. The window column gives the protocol's history length L for that corpus.

Corpus	Users	Items	Interactions	Density	L
Amazon Beauty	22,363	12,101	198,502	0.073%	20
Amazon Sports	35,598	18,357	296,337	0.045%	20
Amazon Toys	19,412	11,924	167,597	0.072%	20
ML-1M	6,040	3,416	999,610	4.845%	200

Table 10: Historical Sports ZCA gate controls (n=1, seed 2020), with only the gate changed. "Shuffled" preserves matrix statistics but breaks item–text alignment. The λ=20 gate approaches this uninformative control.

Sports variant	R@5	R@10	N@5	N@10	valid N@10
no gate	**0.0370**	**0.0574**	**0.0224**	**0.0290**	**0.0355**
gate λ=5	0.0341	0.0545	0.0207	0.0273	0.0344
gate λ=20	0.0315	0.0489	0.0191	0.0247	0.0323
gate λ=20, shuffled matrix	0.0309	0.0485	0.0188	0.0244	0.0312
gate λ=60	0.0302	0.0469	0.0182	0.0236	0.0301

Table 11: Beauty diagnostic probes (d=64, n=1, seed 2020). These historical probes vary preprocessing, gating, fusion or checkpoint selection as named in each row; they are not a single-factor ablation chain.

Probe	R@5	R@10	N@5	N@10
Whitened configuration (reference)	0.0558	0.0895	0.0337	0.0446
Gate on raw matrix, no whitening	0.0547	0.0847	0.0348	0.0444
Recall@10 early stopping	0.0563	0.0859	0.0364	0.0459
Semantic matrix centered	0.0526	0.0834	0.0314	0.0414
Semantic matrix row-normalized	0.0535	0.0836	0.0321	0.0418
Gate inverted, λ=4 (tail-boost)	0.0537	0.0824	0.0335	0.0427
Gate inverted, λ=8	0.0504	0.0806	0.0299	0.0397
+ CF channel, joint whitening	0.0479	0.0762	0.0287	0.0378
+ CF channel, separate whitening	0.0551	0.0824	0.0360	0.0448
CF only (co-occurrence SVD, no text)	0.0541	0.0816	0.0331	0.0420

K.1 Hyperparameter sensitivity

We examine four design choices. **History dropout rate**: on the pure trunk, five seeds per value, Recall@10 rises gently and monotonically with p (0.0781, 0.0799, 0.0809, 0.0818 at $p = 0.10, 0.15, 0.20, 0.25$), a total spread of 0.0037 against seed standard deviations of 0.0007–0.0021. The mean range within $[0.15, 0.25]$ is 0.0019; these SDs describe run variation, not a test of equality between dropout settings. **Gate strength**: five-seed comparisons favor the ungated model

on every corpus, the inverse gate degrades monotonically on Beauty, and the Sports sweep worsens as λ grows (Table 10). Corpus-relative scaling does not rescue it (§4.4). **Selection metric**: switching early stopping from Recall@10 to NDCG@10 can move the operating point along the recall–precision frontier. The dual-checkpoint trainer records both within a fixed trajectory without another training pass (Appendix I). **Trunk dimension**: over $d \in \{64, 128, 192\}$ the repaired model improves monotonically on NDCG@5 on Beauty ($0.0372 \rightarrow 0.0391 \rightarrow 0.0407$) while Recall@10 peaks at d=128 ($0.0934 \rightarrow 0.0948 \rightarrow 0.0912$), so width trades recall for precision once the geometry is repaired rather than lifting both; Table 6 gives all three corpora. Section L describes width selection on validation.

K.2 Efficiency and model size

On Beauty, the d=64 semantic variant has 1,150,277 trainable parameters, including the 262,144-parameter projection. The final d=192 retriever has approximately 4.02M trainable parameters and the feature reranker adds 17,420, including feature normalization and the residual scale. Embedding counts depend on catalog size, while the offline text encoder is a separate pipeline cost (Appendix N).

Training the small reranking head on cached candidates does not require retriever forward passes, but generating those out-of-fold candidates requires additional retriever training. Appendix N separates these stages and reports the hardware and timing scope.

K.3 Additional case-study diagnostics

Section 4.6 presents the two exploratory case types from the earlier Beauty configurations. These comparisons use per-target exact ranks under the same evaluation order, not the final d=192 system. Tail here means a target whose training count is at or below the catalog median, which on Beauty is 6 and covers 7,716 of the 22,363 test users.

The gate improves the selected items in Case B but loses on every aggregate metric across five seeds (§4.4). Across seeds 2020–2022, the number of items matching the crowding signature under identical rules ranges from 7 to 128, showing substantial variation in this case pattern. The median-rank ordering is more stable: median tail rank is 4,108–4,288 under pure ID, 2,948–3,116 under whitened semantics and 3,043–3,156 under the gated variant, with 1,181–1,230 head controls whose ranks move by at most five positions across all three. Whitened semantics improves the tail's absolute standing the most and also achieves the highest tail Recall@10.

L Residual Transition Information in the Retrieved List

The retriever's top-10 contains candidates whose ordering can benefit from additional transition information. We freeze this candidate set, reorder it with a zero-initialised residual, and leave every other item untouched:

$$s'_j = s_j + \gamma\, \Delta_j, \qquad \Delta_j\big|_{\text{init}} = 0.$$

The final ranking permutes only the first ten positions and appends the unchanged suffix; it is not a new global sort of the modified scores. Because the candidate set never changes, Recall@10 is invariant by construction, and the measured paired difference across five seeds is exactly zero.

What carries the gain. Table 12 compares five head-initialization seeds per feature set, conditional on one frozen PCA retriever and one shared OOF cache per corpus. Score/rank calibration changes validation NDCG@5 by at most 0.4%; adding frequency yields at most 0.8% relative to the retriever. Full features improve it by 15.7%, 10.7% and 20.4% on Beauty, Toys and Sports. Co-occurrence therefore supplies most of the measured gain. Comparing the two frequency-free variants, removing `cooc-last` reduces the relative gain by approximately seven to eight percentage points.

Test evaluations retain the same validation-selected heads and confirm that co-occurrence features provide larger gains than score/rank calibration (Table 12). Removing frequency yields higher test NDCG@5 on Toys. Removing the last-item count also slightly improves the frequency-free Sports variant. All corpora use the complete feature set selected on validation.

Table 12: Top-10 feature ablation, validation/test NDCG@5 means over five head initializations. Each corpus shares one seed-2020 PCA retriever and OOF cache, unlike the five-retriever-seed main tables. Test sample SD is at most 0.00046, 0.00026 and 0.00032 on Beauty, Toys and Sports. Feature selection uses validation NDCG@5.

	Beauty		Toys		Sports	
Residual features	Valid	Test	Valid	Test	Valid	Test
score, rank, freq, cooc-last, cooc-any	.0584	.04551	.0616	.05134	.0348	.02749
score, rank, cooc-last, cooc-any	.0566	.04529	.0613	.05211	.0332	.02677
score, rank, cooc-any	.0531	.04229	.0576	.04974	.0309	.02705
score, rank, freq	.0505	.04057	.0557	.04775	.0292	.02396
score, rank	.0505	.04065	.0558	.04770	.0290	.02397
rank only	.0505	.04063	.0557	.04778	.0289	.02409

The 17.4K residual improves the 4.02M-parameter trunk's ordering using additional count features, demonstrating their complementary ranking value. Counting successors of the last history item on the training split identifies the validation target for 2.84% of users outright, and places it among that item's recorded successors for 10.94% , against the trunk's own Recall@10 of 11.30%. The two signals are comparable in magnitude and different in content. Count features inform within-list ordering; Recall@10 remains unchanged because the candidate set is fixed, independently of which features the reranker uses.

Interaction and capacity controls. Letting each candidate cross-attend to the history sequence, so that candidates read per-item history rather than sharing one compressed user vector, reaches +13.4% with 202K parameters; adding candidate-to-candidate attention gives +13.1% with 268K (five seeds each, at the best of three learning rates; at the feature residual's own rate of 3×10^{-5} both attention variants fail to train, +0.3%). Both trail the hand-featured residual. Only 12.8% of training rows have their target inside the top-10, which is 14,932 listwise examples. Limited supervision may contribute to the larger variants' weaker results. The feature residual is stable over the tested hyperparameters, at five seeds per setting: learning rate over two decades (10^{-5} to 10^{-3}), weight decay from 0 to 0.1 and hidden width from 64 to 256 all land within 0.0579–0.0586 validation NDCG@5 (the selected setting gives 0.0584 ± 0.0004), a spread of about one seed standard deviation. Reranking the top-20 instead gives up the Recall@10 invariance for +1.3% Recall@10 and a smaller NDCG@5 gain (+13.5% against +15.7%).

Width–reranking interaction. Table 13 applies each corpus's fixed width-192-trained head to both PCA retriever widths. Width 128 has higher means on all four metrics on Toys and Sports; Beauty favors width 192 only on NDCG@5. The head uses width-independent scalar features, so no retraining is needed. Reranking preserves each retriever's Recall@10 exactly. All three corpora use width 192, selected on Beauty validation.

Table 13: PCA width comparison with the same frozen residual, five-seed means. Best and runner-up values within each dataset are **bold** and underlined. The reported d=192 width was selected on Beauty validation; Appendix I.1 records the development history.

Dataset	Width	R@5	R@10	N@5	N@10
Beauty	d=192	0.0642	0.0912	**0.0454**	0.0541
	d=128	**0.0643**	**0.0948**	0.0448	**0.0546**
Toys	d=192	0.0717	0.0989	0.0515	0.0603
	d=128	**0.0746**	**0.1033**	**0.0530**	**0.0623**
Sports	d=192	0.0388	0.0559	0.0268	0.0323
	d=128	**0.0419**	**0.0598**	**0.0292**	**0.0349**

Full-catalog interaction variants. We tested a low-rank bilinear term $h^\top(I + AB^\top)e$ at ranks 32 and 64, multiple interest queries with $s = \max_m h_m^\top e$ at $M \in \{3, 5\}$, and a learnable logit temperature. None improved validation NDCG@5 by more than the reported 3.4% relative seed SD

of the reference, a descriptive comparison rather than a significance test. All five reduced Recall@10 by 2–11%. Restricting reranking to the retrieved top-10 preserves Recall@10 by construction.

M BASELINE IMPLEMENTATIONS

We use the model-specific settings in Table 14 with the common chronological split, history window and full-catalog candidate policy. LLM-ESR's native HR@10 and NDCG@10 were also cross-checked. Ten baselines define the cross-corpus reference panel, with five seeds per corpus. BERT4Rec's masked-objective reproduction uses the prefix-based training organization described in Appendix J. The MELT runs use five second-stage seeds with a shared backbone pretraining checkpoint. The common-score audit covers its exported predictions; native-evaluator verification and complete per-seed checkpoint provenance remain unverified. Improvement calculations use all ten baselines. Padded or unmapped prediction entries remain misses at their original positions and are excluded from catalog-coverage counts; lists are not compacted.

Text encoder and training settings. GenAIR uses the repository's input- and generation-side hidden-state means, PCA-384 for each, then concatenation. We use Llama-3.2-3B because the released recipe leaves the LLM unspecified, with a shared 300-epoch cap and patience 30.

Table 14: Baseline settings from RecBole or released repositories, in chronological order. All rows have five training seeds per corpus; MELT's five second-stage runs share one pretraining checkpoint. "None" denotes no text encoder.

Baseline	Venue	Optimiser / settings used	Semantic encoder
GRU4Rec	ICLR 2016	Adam, 10^{-3}, batch 256	None
SASRec	ICDM 2018	Adam, 10^{-3}, batch 128	None
BERT4Rec	CIKM 2019	Adam, 10^{-4}, batch 256	None
FDSA	IJCAI 2019	Adam, 10^{-3}, batch 256	attribute fields
FMLP-Rec	WWW 2022	repo default	None
MELT	SIGIR 2023	repo default, λ_u=0.1, λ_i=0.4	None
LLM-ESR	NeurIPS 2024	repo default	LLM embeddings
ActionPiece	ICML 2025	repo config, fixed evaluation seed 2020	metadata tokens
GenAIR	WWW 2026	repo default, τ=4, α=10^{-4}	Llama-3.2-3B (recipe LLM)
TADA	WWW 2026	repo default	None

M.1 ADDITIONAL GENERATIVE-BASELINE EVIDENCE

ActionPiece has five completed and verified seeds on each corpus, shown in Tables 1 and 2. The Sports seed-2020 checkpoint produces the same ordered top-10 for all 35,598 users. Full-set checkpoint reevaluation reproduces the original predictions exactly; a native-generation probe also yields a constant list despite different histories and nonidentical encoder states. The recorded hyperparameters and prepared caches match the other seeds. The constant output is reproducible in both native generation and export; its origin within training or decoding remains unresolved. Main results include all five seeds. In a separate sensitivity analysis, the other four Sports seeds average 0.03124/0.04802/0.02064/0.02603 on R@5/R@10/N@5/N@10. This four-seed sensitivity summary quantifies the effect of the constant-output run.

The five-seed mean comparisons favor the PCA final system on all four metrics in each corpus. The corresponding 12 conditional user-level tests have Bonferroni-adjusted $p \leq 0.0024$. This family is separate from the 96-test baseline/retriever audit (Appendix J). The LIGER row is a historical seed-42, epoch-209, K=100 report; its source and candidate-policy provenance are less complete than those of the main panel. LIGER is excluded from headline improvement calculations and significance tests.

N COMPUTATIONAL COST

Table 16 separates the trainable trunk from the offline encoder.

Table 15: Generative-baseline audit summary. n counts training seeds. ActionPiece uses all five seeds and enters the main comparison; the single-seed LIGER row is historical context only.

Method	Dataset	n	R@5	R@10	N@5	N@10
ActionPiece	Toys	5	.04982±.00131	.07514±.00196	.03289±.00024	.04104±.00047
ActionPiece	Sports	5	.02610±.01150	.04055±.01670	.01731±.00744	.02194±.00913
LIGER	Beauty	1	.05442	.08662	.03459	.04490

Table 16: Beauty validation cost (n=1, seed 2020). The offline encoder runs once before training and is absent from recommender training and inference. Encoder sizes denote full pretrained model families; encoder-only execution counts are reported separately below. Best and runner-up NDCG@5 are **bold** and underlined.

Configuration	Offline encoder	Trainable	valid N@5
d=128	flan-t5-xxl, 11 B	2.49 M	0.0477
d=192	flan-t5-xxl, 11 B	4.02 M	<u>0.0505</u>
d=192	t5-small, 60 M	3.33 M	0.0470
d=192 + rerank	flan-t5-xxl, 11 B	4.02 M + 17.4 K	**0.0587**

The item table accounts for 58% of the Beauty trunk; the reranker adds 0.4% and uses training-split counts, not a separate text encoder. The flan-t5-xxl matrix has $12{,}101 \times 4096$ entries. Switching to t5-small (512-dimensional input) reduces trainable parameters to 3.33M and validation NDCG@5 by 7.0%. The encoder comparison jointly varies pretraining and representation size; it is not a parameter-matched control.

Measured inference and encoding costs. Table 17 measures seed-2020 checkpoints on an exclusive RTX 4080 (PyTorch 2.12.1, CUDA 13.0): medians of 20 repetitions after three warm-ups. Per-batch latency includes history encoding, full-catalog scoring, top-10 selection, CPU features and reranking with a cached output table; it excludes model loading, offline encoding, count construction and OOF fitting. Batch-1 histories have lengths 4, 4 and 3 for Beauty, Toys and Sports; batch-256 padded width is 20.

Table 17: Implementation cost for the final PCA seed-2020 checkpoint on each corpus. Peak allocated memory is measured at batch 256 and excludes the offline encoder.

Corpus	Retriever params.	Batch 1 (ms)	Batch 256 (ms)	Peak (MiB)
Beauty	4,016,453	7.12	30.08	371.5
Toys	3,982,292	7.23	33.76	367.1
Sports	5,223,861	7.31	39.13	498.7

The frozen encoder executes $4{,}762{,}310{,}656$ parameters, versus 11B for the full pretrained family. Encoding 256 evenly spaced catalog items takes median 12.84/12.98/13.34 seconds on Beauty/Toys/Sports (three repetitions, fp16, batch 16, at most 256 tokens, masked-mean pooling), with peak allocation 11.92 GiB. This includes tokenization, transfer, encoding, pooling and CPU output, excluding metadata parsing and the 6.27-second model load.

Per-epoch training and baseline inference. Table 18 summarizes 33 method–corpus measurements, with one benchmark run per cell, separately from the final-system measurements above. Training time averages three complete epochs after one warm-up epoch, including data loading, transfers, forward and backward passes, and optimizer updates. It excludes validation, test evaluation and checkpoint saving. Inference latency is the median of 20 repetitions after three warm-ups, using saved weights and fixed validation histories at batch 256. The accompanying records also report batch-1 latency and individual repetitions.

SASRec uses training batch 128; BERT4Rec uses learning rate 10^{-4} and training batch 256, matching their reported configurations. FDSA Beauty uses its saved seed-2022 checkpoint; all benchmarks fix measurement randomness to seed 2020.

Table 18: Per-epoch training time (seconds) and inference latency (milliseconds per batch of 256). Each cell is one benchmark run. Our inference includes reranking. Each method uses its recorded hardware, software environment and native training batch size, as detailed below.

	Beauty		Toys		Sports	
Method	s/epoch	ms/batch	s/epoch	ms/batch	s/epoch	ms/batch
GRU4Rec	1.79	0.56	1.55	0.54	2.82	0.92
SASRec	17.34	3.14	14.38	3.19	25.85	3.14
FDSA	17.18	6.15	14.54	2.29	26.74	2.24
FMLP-Rec	2.78	5.93	1.12	4.46	4.60	6.55
TADA	3.20	2.10	2.79	2.26	5.18	5.40
LLM-ESR	66.69	2089.58	24.22	2086.17	45.52	3237.83
GenAIR	38.26	130.32	41.01	136.50	60.93	196.43
ActionPiece	69.44	4527.38	56.18	4759.82	115.15	4998.12
BERT4Rec	10.36	7.60	8.64	10.51	15.97	10.61
MELT	11.51	13.83	9.40	14.95	17.95	17.90
Ours	2.72	28.74	2.12	29.41	3.83	28.16

Measurements use RTX 4080 and RTX 4080 SUPER GPUs with per-method environments and native training batch sizes, under exclusive GPU use. Our training time covers the retriever; TADA and MELT cover their final recommendation stage. Offline encoding, tokenizer construction, auxiliary pretraining and OOF fitting are excluded from the per-epoch column. Native chunked prediction in LLM-ESR and GenAIR can repeat sequence encoding; ActionPiece includes native beam search and ensemble inference. ActionPiece Sports timing uses the retained seed-2020 checkpoint discussed above.

Each of the five GenAIR Toys checkpoints contains 9,604,224 parameters: 445,056 trainable and 9,159,168 in the frozen item table, excluding the offline Llama model and preprocessing resources.