SHORTCUT OR SEMANTICS? UNDERSTANDING CONDITIONING INTERFACES FOR STEERABLE GENERATIVE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural-language steering enables recommender systems to adapt to explicit user requests that go beyond historical interaction data. In generative retrieval, this requires conditioning a semantic ID (SID) decoder on textual preferences, but how best to interface language representations with the SID decoder remains unresolved. To address this, we propose a decoupled dual-branch interface that injects a frozen language model's final-layer representations via pooled global modulation and token-level cross-attention. We evaluate this design against end-to-end joint conditioning across three e-commerce domains. Joint conditioning reaches the highest retrieval accuracy, but suffers from catastrophic forgetting of general reasoning; meanwhile, the dual-branch interface trails by only 0.34 to 0.79 percentage points while preserving pretrained capabilities. Across meaning-preserving rephrasings, joint conditioning degrades sharply under synonym substitution, revealing that the model memorizes literal word matches instead of learning actual semantic intent. The dual-branch interface, in contrast, remains anchored in the frozen model's semantic space, exhibiting 55–74% lower degradation under lexical substitution across the three domains. These results establish our proposed dual-branch interface as a robust, semantically grounded foundation for steerable generative retrieval.

1 INTRODUCTION

Users' immediate requests can often go beyond interests inferred from interaction histories. Natural-language steering enables recommender systems to adapt to explicit requests directly (Carroll et al., 2025; Zhou et al., 2026), as in Google Discover (Loew, 2026) and Amazon Shopping (Lloyd, 2025). Prevalent approaches retrieve historical candidates before LLM reranking (Sun et al., 2023; Hou et al., 2024; Yue et al., 2023). This adds inference cost and risks candidate starvation: downstream rerankers cannot recommend relevant items if the initial retrieval stage fails to retrieve them (Dong et al., 2026). Conditioning retrieval directly on the request resolves this bottleneck.

Semantic-ID (SID) generative retrieval frames recommendation as autoregressively generating discrete item identifiers from learned codebooks (Rajput et al., 2023). Extending this to natural-language steering (Paischer et al., 2025; He et al., 2025; D'Amico et al., 2026) raises a fundamental question: **how should the language-to-retrieval interface be structured?** A common approach is end-to-end joint conditioning, which serializes text instructions and discrete SIDs into a single causal Transformer (Zheng et al., 2024; De Nadai et al., 2026; Zhou et al., 2025). Nevertheless, updating language representations on retrieval losses risks learning surface keyword shortcuts while catastrophically forgetting general reasoning. Alternatively, to mitigate catastrophic forgetting, we propose a modular dual-branch interface that decouples a frozen language tower from a trainable SID generator. It injects final-layer representations via pooled global modulation and token-level cross-attention (Figures 1 and 2), benchmarked against joint conditioning, the cross-attention baseline widely adopted in generative retrieval (Paischer et al., 2025), and depth controls.

We evaluate retrieval across Amazon Beauty, Sports and Outdoors, and Toys and Games (He & McAuley, 2016), prioritizing preference recommendation and directional sentiment following

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.

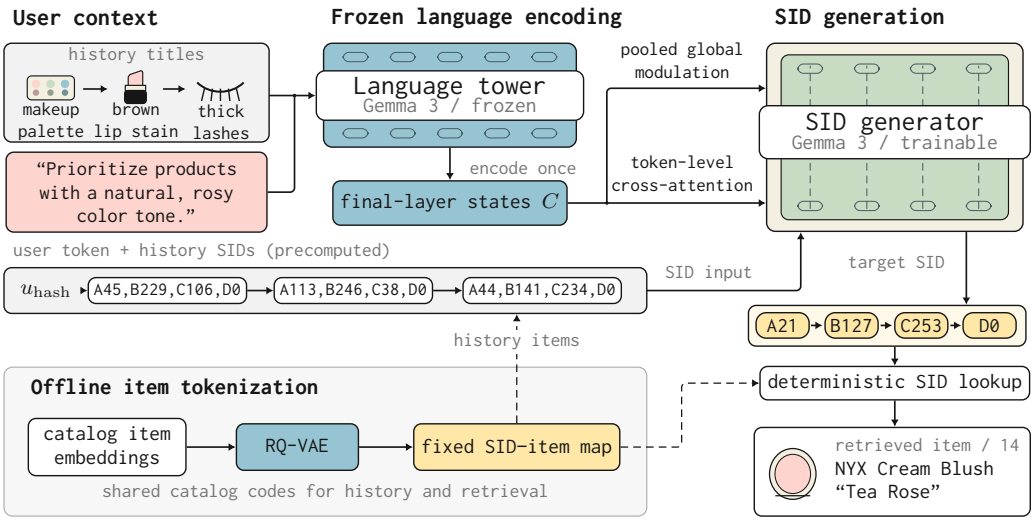


Figure 1: **Steerable generative retrieval with the dual-branch final-state interface.** History titles and a user preference enter the frozen language tower. Its final-layer states condition every block of the trainable SID generator through pooled global modulation and token-level cross-attention. Given the user token and history SIDs, the generator autoregressively produces a target SID for deterministic item lookup. Offline RQ-VAE tokenization supplies the history codes and SID–item map (dashed arrows). The Beauty example uses abbreviated history titles to illustrate the pipeline.

within a five-task benchmark (Paischer et al., 2025). We evaluate rephrasing robustness across all 3 domains, and examine representation depth and language capability retention on Amazon Beauty.

Our main empirical findings and architectural contributions are summarized as follows:

1. **Decoupled dual-branch interface:** We propose a zero-initialized interface combining pooled global modulation (AdaRMS) and token-level cross-attention, improving preference Recall@10 to 8.73% (6.40% AdaRMS, 8.34% cross-attention) and trailing joint conditioning by only 0.34–0.79 percentage points across domains. In the context of SID generation, broadcasting final-layer states significantly outperforms layer-wise routing (6.40% vs. 6.02%), showing that compact high-level summaries avoid intermediate lexical noise.
2. **Surface shortcuts vs. semantic grounding:** On Beauty, joint conditioning loses 34.3% of its Recall@10 under synonym substitution, exposing reliance on surface keyword shortcuts supported by an attention case study. In contrast, our dual-branch interface stays anchored in the frozen semantic space, dropping only 15.3%. After synonym substitution, it achieves 9.49% Recall@10 versus 7.62% for joint conditioning, an advantage of 1.87 percentage points. Smaller drops are also observed for the dual-branch design on Sports and Toys.
3. **General-capability preservation:** Retrieval specialization in joint conditioning causes catastrophic forgetting, wiping out general reasoning (e.g., 0% accuracy on BBH and GSM8K). In contrast, freezing the language tower prevents forgetting, retaining full reasoning capabilities while maintaining competitive retrieval accuracy.

2 Related Work

Semantic-ID retrieval. Generative recommendation decodes discrete Semantic-IDs (SIDs) directly from interaction histories (Rajput et al., 2023), with subsequent advances in collaborative alignment (Zheng et al., 2024; Wang et al., 2024), scaling (He et al., 2025), unified pretraining (Zhou et al., 2025), and reasoning (He et al., 2026). Yet, these models operate in a closed behavioral universe, lacking mechanisms for real-time natural-language steering.

Natural-language steering in recommendation. Language guidance enables user control via instruction tuning (Bao et al., 2023; Zhang et al., 2025), profile editing (Ramos et al., 2024; Lu et al., 2024), or dense compatibility scoring (Carroll et al., 2025). In generative retrieval, incorporating text instructions divides architectures into single-sequence joint autoregression (De Nadai et al., 2026; D'Amico et al., 2026) and decoupled cross-attention (Paischer et al., 2025). While joint conditioning maximizes task adaptation, it triggers catastrophic forgetting; conversely, decoupled cross-attention preserves language representations but lacks global modulation. We evaluate this tension under a unified Gemma 3 backbone (Gemma Team, 2025).

Conditioning across heterogeneous representations. Bridging continuous text with discrete generative decoders parallels Diffusion Transformers, where conditioning factorizes into global adaptive normalization for distribution shifts and cross-attention for granular alignments (Peebles & Xie, 2023; Chen et al., 2024). We introduce this dual-branch design to generative retrieval, pairing pooled AdaRMS modulation with token cross-attention. Furthermore, SID decoding thrives on final-layer semantic abstraction over layer-wise routing, filtering out superficial lexical noise.

Phrasing robustness and capability retention. Steerable recommenders face two interrelated vulnerabilities: sensitivity to surface phrasing variations (Gu et al., 2023; Srikanth et al., 2024; Fu & Barez, 2025) and catastrophic forgetting of general reasoning during retrieval specialization (Hu et al., 2022; Li et al., 2024; De Nadai et al., 2026). We show that shortcut reliance and catastrophic forgetting are twin symptoms of representational distortion, critically mediated by the conditioning interface. Using meaning-preserving rephrasings and reasoning benchmarks, we test whether models learn invariant semantic intent over keyword shortcuts and trace their capability Pareto frontiers.

3 Steering Generative Retrieval

3.1 Retrieval Task

Let $\mathcal{I}$ denote the item corpus. For a given user u, their chronological interaction history consists of an item sequence $(i_1, \ldots, i_n)$, aligned titles $c_u^{\text{text}} = (t_1, \ldots, t_n)$, and SIDs $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$, where each $\mathbf{s}_j = (s_{j,1}, \ldots, s_{j,4})$ is a 4-token SID tuple. Following Rajput et al. (2023), each user is assigned a hashed identifier $u_{\text{hash}} \in \{1, \ldots, 2000\}$ (serialized using Gemma 3's native vocabulary) prepended to the history SID sequence to form the discrete behavioral context $c_u^{\text{SID}} = (u_{\text{hash}}, \mathbf{s}_1, \ldots, \mathbf{s}_n)$. These constitute the persistent user context $c_u = (c_u^{\text{text}}, c_u^{\text{SID}})$. While c_u captures long-term behavioral history, a natural-language utterance r expresses the immediate request. The target item $i \in \mathcal{I}$ is uniquely indexed by its hierarchical 4-token SID $y_i = (s_{i,1}, s_{i,2}, s_{i,3}, s_{i,4})$. The autoregressive generation probability under model parameters θ is factorized as

$$\Pr_\theta(y_i \mid c_u, r) = \prod_{k=1}^{4} \Pr_\theta(s_{i,k} \mid c_u, r, s_{i,<k}). \tag{1}$$

At inference time, unconstrained beam search (with beam width 20) explores this fixed-length decoding tree and ranks candidate token tuples by cumulative log-probabilities. Generated candidate sequences are matched directly against the ground-truth target SID y_i. Any generated SID sequence that does not correspond to a valid catalog item naturally fails to match the ground truth and is recorded as a retrieval miss without post-hoc remapping.

3.2 Steering Benchmark

We construct a steering benchmark across Amazon Beauty, Sports and Outdoors, and Toys and Games (He & McAuley, 2016), following TIGER's evaluation domains (Rajput et al., 2023). Starting from the Preference Discerning artifact (Paischer et al., 2025), we build a separate interaction split, item corpus, and RQ-VAE codebook for each domain (documented in Appendix A).

Table 1 summarizes the five tasks used in our experiments. We adapt four dimensions from Preference Discerning (Paischer et al., 2025): preference recommendation, fine- and coarse-grained steering, and intent following. Our fifth task, sentiment following, pairs find and avoid requests for the same anchor item under identical browsing history. For negative requests (r^-), training uses

standard positive SFT on history-compatible alternative targets $\tilde{i}$ (Appendix A.4), while evaluation strictly checks whether the forbidden anchor is excluded from top ranks. Due to synthetic shortcuts in Fine and Coarse steering (lexical overlap and target concentration; Appendix A.2), we prioritize Preference Recommendation and Sentiment Following as our primary steering benchmarks. Intent following is evaluation-only, appearing in validation and test without training records.

Table 1: **Steering dimensions in the benchmark suite.** Evaluation tasks across Amazon domains; Fine and Coarse steering serve as diagnostic checks due to synthetic shortcuts (Appendix A.2).

Dimension	Core Objective / Description	Target Behavior
Preference recommendation	Predict next item given browsing history and user's preference.	Retrieve target item
Fine-grained steering	Redirect retrieval to a semantically close item via fine-grained preference cues.	Retrieve fine-steered target
Coarse-grained steering	Redirect retrieval to a semantically distant item via category-level preference shift.	Retrieve coarse-steered target
Intent following	Identify the target-relevant intent from 5 candidate preferences to retrieve item.	Filter distractors & retrieve target
Sentiment following	Distinguish directional control signals (*Find* vs. *Avoid*) given identical context.	**Find**: Promote target **Avoid**: Suppress target

3.3 METRICS

Let $\mathcal{D}$ denote the evaluation distribution over triples (c_u, r, i), where c_u is the user context, r is the preference utterance, and $i \in \mathcal{I}$ is the ground-truth target item. Let $C_K(r, c_u)$ denote the ordered set of top-K items retrieved by beam search, and let $\rho_i(r, c_u)$ be the 1-indexed rank of i if $i \in C_K(r, c_u)$, and ∞ otherwise. For preference recommendation, fine- and coarse-grained steering, and intent following, retrieval quality is measured by Recall@K and nDCG@K with $K \in \{5, 10\}$:

$$\begin{aligned} \text{Recall@}\,K &= \mathbb{E}_{(c_u,r,i)\sim\mathcal{D}}\left[\mathbf{1}\left[i \in C_K(r, c_u)\right]\right], \\ \text{nDCG@}\,K &= \mathbb{E}_{(c_u,r,i)\sim\mathcal{D}}\left[\frac{\mathbf{1}\left[\rho_i(r, c_u) \le K\right]}{\log_2(\rho_i(r, c_u) + 1)}\right]. \end{aligned} \tag{2}$$

Because each evaluation instance specifies a single ground-truth target, Recall@K represents the top-K hit rate, while nDCG@K discounts successful retrievals positioned lower in the ranked list.

For sentiment following, evaluation instances are drawn as quadruples $(c_u, r^+, r^-, i) \sim \mathcal{P}$, where r^+ is a positive request to find target i, and r^- is a negative request to avoid i, under identical context c_u. The positive retrieval accuracy is given by $\text{Find Recall@10} = \mathbb{E}_{\mathcal{P}}[\mathbf{1}[i \in C_{10}(r^+, c_u)]]$, while negative suppression failure is measured by $\text{Avoid Hit@10} = \mathbb{E}_{\mathcal{P}}[\mathbf{1}[i \in C_{10}(r^-, c_u)]]$ (where lower values reflect superior avoidance). To quantify the joint success of simultaneously promoting the target under r^+ and suppressing it under r^-, we define Sentiment@10:

$$\text{Sentiment@10} = \mathbb{E}_{(c_u,r^+,r^-,i)\sim\mathcal{P}}\left[\mathbf{1}\left[i \in C_{10}(r^+, c_u)\right] \cdot \mathbf{1}\left[i \notin C_{10}(r^-, c_u)\right]\right]. \tag{3}$$

3.4 SHARED TRAINING OBJECTIVE

Across all interfaces, the model receives context c_u and preference r, and is trained to generate the 4-token target SID sequence $y_i = (s_{i,1}, \ldots, s_{i,4})$ using shifted teacher forcing:

$$\mathcal{L}_{\text{SID}}(\theta) = -\sum_{k=1}^{4} \log \Pr_\theta(s_{i,k} \mid c_u, r, y_{i,<k}). \tag{4}$$

Tokens in the prompt sequence (c_u, r) are masked from the loss calculation. For negative steering requests (r^-), the target sequence y_i in Equation (4) is substituted with a history-compatible alternative target item $\tilde{y}_i$ satisfying the negative constraint (Appendix A), supervising the model to steer toward valid alternatives rather than reproducing the rejected anchor.

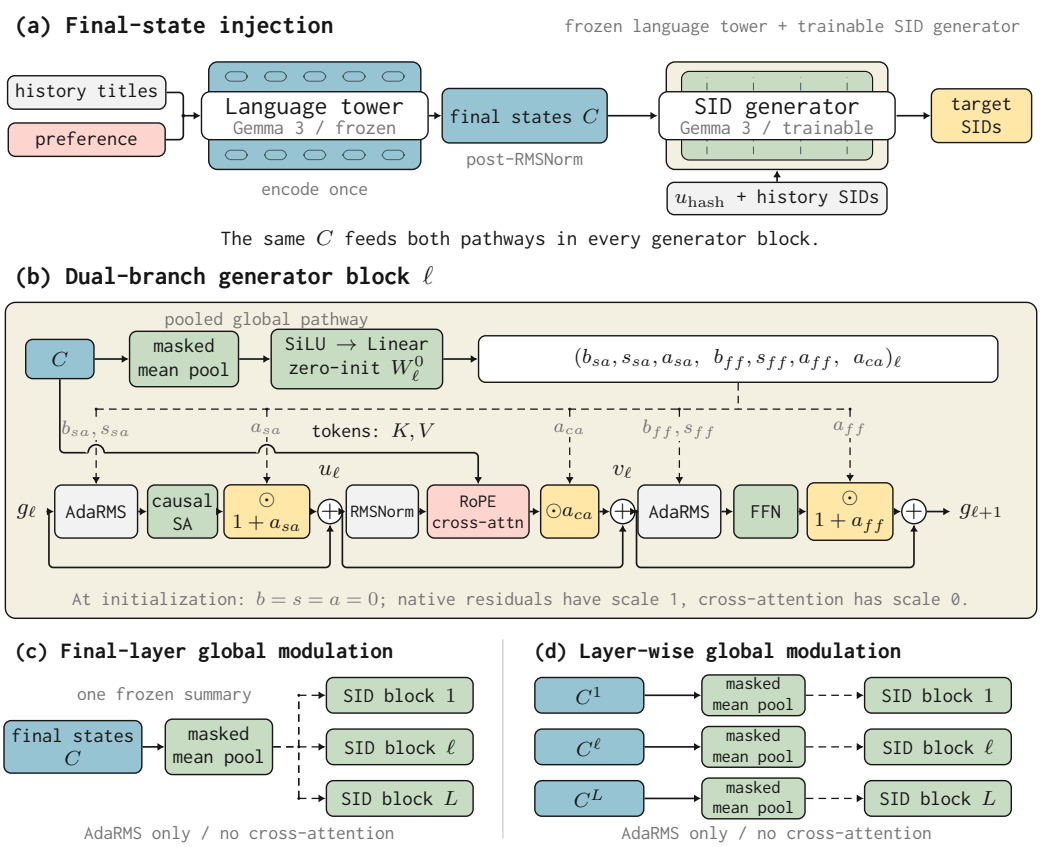


Figure 2: **Language-to-SID conditioning interfaces.** **(a)** Final-state injection broadcasts the same post-RMSNorm language states C to every generator block. **(b)** Each block combines pooled modulation (dashed parameter routes) and token cross-attention (solid route), with residual scales $1+a_{sa}$, a_{ca}, and $1 + a_{ff}$. Zero-initialized modulation preserves the pretrained block at initialization. **(c)–(d)** Depth-routing controls use global modulation only: (c) broadcasts one final-layer summary to all SID blocks; (d) pools each frozen layer's states C^ℓ to modulate the matching SID block, with L denoting final depth (dashed routes: modulation; full equations in Appendix C).

4 Language-to-SID Conditioning Interfaces

We compare our dual-branch final-state interface against joint conditioning and a cross-attention baseline. Throughout, *final-state injection* denotes the dual-branch model unless a component ablation is specified. A global-only control tests final-layer broadcasting against layer-wise routing.

4.1 Joint Conditioning

Joint conditioning concatenates textual history, preference prompt, and discrete SIDs into a single causal sequence, formatted as $[c_u^{\text{text}} \to r \to c_u^{\text{SID}}] \to y_i$. Language understanding and SID decoding share the same trainable Transformer backbone, allowing retrieval gradients to update language representations end to end.

4.2 Final-State Representation Injection

Final-state injection pairs a frozen language tower with a trainable SID generator, both initialized from Gemma 3 1B IT checkpoint. The tower encodes (c_u^{text}, r) once into final-layer, post-RMSNorm states $C \in \mathbb{R}^{L_c \times D}$, with sequence length L_c and hidden width D. The generator embeds c_u^{SID} into its initial residual stream g_0 and generates target SIDs. Every generator block receives the same condition sequence C through two pathways:

Pooled Global Modulation. Mean-pooling the non-padding states of C yields a global context vector $\bar{c}$. Each block projects $\mathrm{SiLU}(\bar{c})$ into channel-wise shifts b, scales s, and residual gates a. Shifts and scales condition self-attention and feed-forward sublayers through Adaptive RMSNorm (Peebles & Xie, 2023):

$$\mathrm{AdaRMS}(x; s, b) = \mathrm{RMSNorm}(x) \odot (1 + s) + b. \tag{5}$$

Token-Level RoPE Cross-Attention. A cross-attention sublayer between self-attention and the feed-forward network lets generator queries access individual condition states in C. Rotary position embeddings (RoPE) place condition and generator tokens in a continuous coordinate system to preserve their relative positions across streams.

Dual-Branch Gated Block. With causal self-attention SA_ℓ, condition cross-attention CA_ℓ, and feed-forward network FF_ℓ, block ℓ updates its residual stream g_ℓ as:

$$\begin{aligned} u_\ell &= g_\ell + (1 + a_{sa,\ell}) \odot \mathrm{SA}_\ell(\mathrm{AdaRMS}(g_\ell; s_{sa,\ell}, b_{sa,\ell})), \\ v_\ell &= u_\ell + a_{ca,\ell} \odot \mathrm{CA}_\ell(\mathrm{RMSNorm}(u_\ell), C), \\ g_{\ell+1} &= v_\ell + (1 + a_{ff,\ell}) \odot \mathrm{FF}_\ell(\mathrm{AdaRMS}(v_\ell; s_{ff,\ell}, b_{ff,\ell})). \end{aligned} \tag{6}$$

Zero-initializing the projection sets all shifts, scales, and gates to zero, preserving pretrained block weights at initialization. Appendix C.1 details masked pooling, the seven-vector projection, and cross-attention; Appendix C.2 specifies RoPE alignment.

4.3 Cross-Attention Baseline and Depth Control

The **cross-attention-only baseline** represents prior generative retrieval models such as Mender (Paischer et al., 2025), which guide SID generation through cross-attention to text representations. To test this design fairly, we implement it directly within our Gemma 3 backbone: it disables pooled modulation, retains cross-attention to frozen final states, and learns a zero-initialized residual gate (Appendix C.1), isolating whether cross-attention alone suffices for steering.

To compare **source representation depth**, we hold the fusion mechanism to global modulation without cross-attention (Figure 2c–d). One variant pools the final-layer states once and broadcasts that summary to every SID block; the other pools each frozen language layer's states separately to modulate the corresponding SID block at matched depth.

5 Empirical Evaluation and Analysis

We evaluate multi-domain steering accuracy, followed by component ablations, language capability retention, and phrasing robustness on all 3 domains (training and selection details in Appendix D).

5.1 Multi-Domain Steering Accuracy

Table 2 summarizes test performance on our two primary tasks across the three domains. We report means and sample standard deviations over three independent runs for validation-selected configurations. Language-conditioned models are selected by a composite criterion covering preference recommendation, sentiment following, and intent following; the SID-history-only baseline is selected by Preference Recall@10. Appendix D.2 gives the selection details. Full validation/test metrics, Fine and Coarse steering, and Intent following are reported in Appendix E.4.

Retrieval accuracy with a frozen language tower. Across all three domains, our decoupled final-state interface achieves competitive retrieval while keeping the language tower frozen, reaching 8.73%, 5.06%, and 7.00% test preference Recall@10 on Beauty, Sports, and Toys (Table 2). It trails fully fine-tuned joint conditioning by merely 0.70, 0.34, and 0.79 pp, while consistently outperforming cross-attention alone by 0.39, 0.15, and 0.66 pp. Recall@5 and nDCG show a similar pattern (Appendix Table 10). As shown later, joint conditioning's marginal edge comes at the cost of catastrophic forgetting (Section 5.3) and superficial keyword shortcuts (Section 5.4).

Table 2: **Core steering results across three domains.** Test results (mean±std over three independent runs, in %) for validation-selected configurations. Bold marks the best mean per domain; lower is better for Avoid Hit@10. Full validation/test results and auxiliary tasks appear in Appendix E.4.

Domain	Interface	Preference recommendation		Sentiment following		
		Recall@10 ↑	nDCG@10 ↑	Find R@10 ↑	Avoid Hit@10 ↓	Sentiment@10 ↑
Beauty	SID-history only	7.76±0.10	4.36±0.04	5.87±0.18	5.87±0.18	0.00±0.00
	Joint conditioning	**9.43±0.09**	**5.06±0.04**	**7.49±0.14**	**1.00±0.17**	**7.01±0.16**
	Cross-attention only	8.34±0.07	4.52±0.06	6.05±0.17	2.37±0.26	4.63±0.27
	Final-state injection	8.73±0.18	4.80±0.03	6.50±0.14	1.43±0.21	5.59±0.12
Sports	SID-history only	3.38±0.86	1.84±0.49	2.58±0.65	2.58±0.65	0.00±0.00
	Joint conditioning	**5.40±0.08**	**2.83±0.02**	**7.20±0.25**	1.61±0.13	**6.15±0.27**
	Cross-attention only	4.91±0.02	2.60±0.02	5.23±0.00	**1.29±0.29**	4.45±0.15
	Final-state injection	5.06±0.16	2.66±0.06	6.73±0.13	1.56±0.03	5.65±0.06
Toys	SID-history only	7.26±0.16	4.03±0.11	4.72±0.20	4.72±0.20	0.00±0.00
	Joint conditioning	**7.79±0.37**	**4.11±0.23**	**5.88±0.22**	**0.79±0.19**	**5.33±0.18**
	Cross-attention only	6.34±0.14	3.31±0.12	3.94±0.06	1.40±0.13	3.13±0.09
	Final-state injection	7.00±0.19	3.65±0.09	5.08±0.28	1.54±0.25	3.94±0.10

Directional steering and negation control. In directional steering (sentiment following, Table 2), where models separate positive ("find") from negative ("avoid") requests under fixed history, joint conditioning achieves higher Sentiment@10 (7.01%, 6.15%, 5.33%) via higher Find Recall. However, adding global modulation to cross-attention boosts Sentiment@10 by 0.96, 1.20, and 0.81 pp across domains. While cross-attention alone yields lower Avoid Hit@10 in Sports and Toys, it under-retrieves positive items; our dual-branch design achieves a far superior trade-off between target promotion and anchor suppression.

History baseline and auxiliary tasks. Following TIGER (Rajput et al., 2023), the SID-history-only baseline receives no natural language and decodes targets solely from history SIDs (c_u^{SID}). We include this baseline to calibrate benchmark difficulty and measure how much accuracy comes purely from behavioral history. On preference recommendation, it remains strong: in Toys, it reaches 7.26% Recall@10 (4.03% nDCG@10), showing that sequential co-occurrence alone predicts many targets. For directional steering, however, it scores 0.0% Sentiment@10 across all domains because identical inputs cannot differentiate find and avoid requests. Auxiliary tasks (Fine and Coarse steering, Intent following) serve as diagnostic checks and are detailed in Appendix E.4.

5.2 Conditioning Paths and Source Depth

Complementary conditioning paths and dynamic gating. On Beauty, independently validation-selected global-only, cross-attention-only, and dual-branch models reach 6.40%, 8.34%, and 8.73% preference Recall@10 (Appendix Table 9). Global modulation injects broad user-level attribute shifts via AdaRMS, while cross-attention enables token-level alignment. Crucially, cross-attention strength is regulated by a learnable gate ($a_{ca,\ell}$ in Equation 6), allowing each block to dynamically balance global modulation against token interactions from a zero-initialized identity baseline.

Final-layer broadcasting vs. layer-wise conditioning in SID generation. With global modulation fixed and cross-attention disabled, final-layer broadcasting reaches $6.40 \pm 0.10\%$ test preference Recall@10 on Beauty versus $6.02 \pm 0.12\%$ for layer-wise conditioning (+0.38 pp). Across 120 matched hyper-parameters, last-layer broadcasting wins 102, ties 18, and loses none (Appendix E.2). In SID generation, intermediate layers inject unabstracted lexical noise that disrupts discrete token decoding, whereas the final layer provides a compact semantic summary that guides autoregressive generation far more effectively.

5.3 Retrieval Specialization and General-Capability Retention

Across eight language benchmarks (Hendrycks et al., 2021; Cobbe et al., 2021; Suzgun et al., 2023), the joint-conditioning model selected for steering accuracy on Beauty exhibits severe forgetting: BBH and GSM8K drop to zero, TriviaQA to 0.02%, and MMLU from 40.02% to 25.28% (Table 3).

Table 3: **Retrieval and capability retention on Beauty** (scores in %). Retrieval reports validation Preference Recall@10; language scores evaluate individual models. The steering-selected joint model is compared with an alternative preserving higher language capability at lower retrieval.

Model	Pref. R@10	ARC-C	BBH	BoolQ	GSM8K	HellaSwag	MMLU	MMLU-Pro	TriviaQA
Pretrained Gemma	–	38.57	29.78	75.90	20.07	52.88	40.02	14.86	19.63
Joint (selected)	11.71	32.99	0.00	57.75	0.00	44.21	25.28	10.83	0.02
Joint (higher retention)	6.63	40.28	22.26	74.75	17.73	53.85	38.77	14.51	8.28

Evaluating the configuration cloud reveals a sharp Pareto tension in joint training (Figure 7): optimizing retrieval destroys pretrained reasoning. Preserving language competence (e.g., "Joint (higher retention)" retaining $\geq 96.88\%$ on five tasks) drops validation Recall@10 from 11.71% to 6.63%. In contrast, our decoupled interface completely resolves this dilemma: freezing the language tower preserves full pretrained general reasoning while delivering 8.73% test Recall@10 (11.21% on validation), surpassing the capability-preserving joint baseline (details in Appendix E.5).

5.4 Robustness to Preference Rephrasing

Paired rephrasing evaluation protocol. We compare single-tower joint conditioning and dual-branch final-state injection on paired validation corpora from Beauty, Sports, and Toys (Appendix B), with three seeds per interface and domain. For each instance, user identity, browsing history, target item, and underlying intent remain fixed, while the original prompt r_0 undergoes three semantics-preserving transformations: (1) **Lexical substitution** (r_{lex}), replacing content words with contextual synonyms under identical sentence structure; (2) **Syntactic paraphrase** (r_{syn}), restructuring clauses while preserving exact semantics; and (3) **Conversational rewrite** (r_{conv}), adapting formal prompts into colloquial user queries. Letting M_0 and M_w denote preference Recall@10 under original and rewritten prompts, respectively, relative degradation is $D_w = 1 - M_w/M_0$, computed per seed and summarized by its mean and sample standard deviation (lower is better).

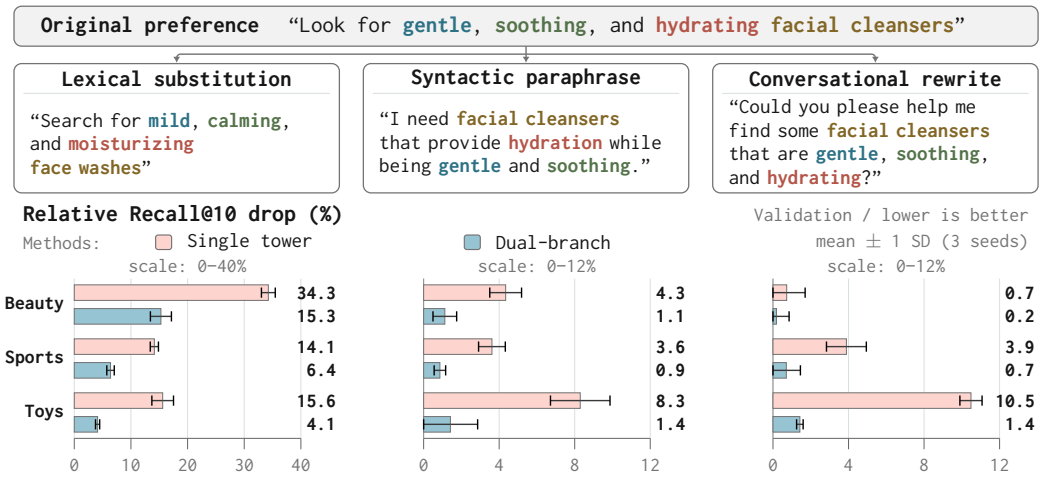


Figure 3: **Robustness to preference rephrasing.** Top: a Beauty example illustrates three parallel transformations, with matching attributes highlighted. Bottom: relative preference Recall@10 drops on Beauty, Sports, and Toys for single tower and dual-branch (lower is better). Bars show means and whiskers ± 1 sample standard deviation across three seeds, normalizing within each seed. Note the different axis scales.

Empirical stability under phrasing shifts. Figure 3 shows smaller mean drops for dual-branch in all nine domain–rewrite combinations. **Lexical substitution** reduces single-tower Recall@10 by 34.3%, 14.1%, and 15.6% on Beauty, Sports, and Toys, versus 15.3%, 6.4%, and 4.1% for dual-branch. The Beauty conversational difference is small relative to seed variability.

In absolute Recall@10, single tower leads on original Beauty preferences (11.60% vs. 11.21%), but dual-branch leads after lexical substitution (9.49% vs. 7.62%): a 0.39 pp deficit becomes a 1.87 pp advantage. Dual-branch also leads on Sports (6.04% vs. 5.61%) and Toys (8.65% vs. 8.02%) after lexical substitution.

Insights on lexical overfitting vs. semantic grounding. This selective breakdown under meaning-preserving lexical substitution demonstrates *lexical shortcut learning* in joint conditioning. On Beauty, syntactic and conversational rewrites preserve core attribute keywords, causing small mean drops (4.3% and 0.7%). Conversely, substituting keywords with synonyms triggers a 34.3% drop, proving that joint conditioning exploits memorized keywords rather than invariant semantic intent.

Target-aligned attention provides qualitative support (Appendix E.1): during target SID generation, the joint model places overwhelming attention on exact prompt keywords shared with the target title. By anchoring in the frozen model's semantic manifold, our dual-branch interface avoids this lexical trap, achieving robust semantic steering.

6 Discussion, Limitations, and Conclusion

Architectural implications. Our findings establish three representational principles for steerable generative retrieval. First, *the specialization dilemma*: adapting causal LLMs directly to discrete tokens triggers catastrophic forgetting of reasoning and incentivizes superficial keyword memorization, imposing an unacceptable capability tax for marginal retrieval gains. Second, *modular decoupling as a semantic anchor*: freezing the language tower forces the SID decoder to ground in the invariant pretrained language topology, preventing representation collapse and reducing degradation under lexical phrasing shifts. Third, *structural synergy across heterogeneous spaces*: broad attribute shifts require global AdaRMS modulation, while granular constraints demand gated cross-attention. Furthermore, discrete SID decoding thrives on final-layer semantic abstraction; intermediate representations inject unabstracted lexical noise that disrupts codebook generation.

Limitations. Our empirical study operates under four primary boundaries. First, the benchmark relies on synthetic preference instructions across three Amazon categories; real-world deployment must address ambiguous requests, implicit feedback, and colloquial conversational noise. Second, dual backbones increase serving memory during prefill, though prefix-caching frozen language states keeps autoregressive SID decoding latency minimal. Third, steering granularity is fundamentally bounded by the semantic resolution of the upstream RQ-VAE codebook. Finally, our evaluation focuses on single-turn steering, leaving multi-turn session critiques to future work.

Future directions. These findings open promising avenues for generative recommendation. One compelling frontier is *continuous latent steering*, projecting preference intents into compact steering vectors that modulate generation without re-encoding full text. Another vital direction is *adaptive codebook co-design*, jointly optimizing discrete quantizers with the language manifold to align codebook hierarchies with steerable semantic axes. Additionally, extending the frozen branch to vision-language backbones would enable multimodal steering, while speculative decoding over SIDs could unlock near-instantaneous inference-time scaling for production serving.

Conclusion. We examined the fundamental tension between end-to-end joint adaptation and modular representation injection in steerable generative retrieval. While joint conditioning achieves slightly higher retrieval accuracy, it suffers catastrophic capability loss and relies on superficial keyword shortcuts. In contrast, our zero-initialized dual-branch interface trails by merely 0.34–0.79 pp in Recall@10, preserves 100% of pretrained reasoning, and shows semantic robustness across user phrasings. Decoupled representation injection thus demonstrates that domain specialization need not sacrifice semantic generalization, offering a principled foundation for steerable recommendation.

AI USE STATEMENT

We use synthetic preferences from the Mender/Preference Discerning release (Paischer et al., 2025), generated by its authors using Llama-3-70B-Instruct. In this work, Gemini 3 Flash Preview was used to select history-compatible alternative targets, generate associated positive intents, and produce preference rephrasings (Appendix B). LLMs were also used for coding assistance and manuscript language polishing. The authors inspected all generated artifacts and take full responsibility for all content.

ETHICS STATEMENT

Our benchmark is constructed from publicly available Amazon review data (He & McAuley, 2016) and synthetic preferences (Paischer et al., 2025). All user identifiers are anonymized into contiguous integers, and item metadata contains no personally identifiable information (PII). Synthetic preference instructions may inherit inductive biases or stylistic artifacts from the generating language models. The released benchmark complies with original terms of use and is distributed under the MIT license strictly for academic research.

REPRODUCIBILITY STATEMENT

Appendix A details the source data, three-domain interaction splits, and domain-specific RQ-VAE semantic-ID codebooks. Section 4 and Appendix C specify the block-level execution pipelines, adaptive RMSNorm formulas, continuous RoPE coordinate conventions, and parameter derivations. Section 3.4 gives the training objective, while Appendix D documents hyperparameter search spaces, validation selection, and repeated-seed evaluation, with selected configurations in Table 8.

REFERENCES

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. TALLRec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1007–1014, 2023. doi: 10.1145/3604915.3608857.

Micah Carroll, Adeline Foote, Kevin Feng, Marcus Williams, Anca Dragan, W. Bradley Knox, and Smitha Milli. CTRL-Rec: Controlling recommender systems with natural language. *arXiv preprint arXiv:2510.12742*, 2025. URL https://arxiv.org/abs/2510.12742.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.00426.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Edoardo D'Amico, Marco De Nadai, Praveen Chandar, Divita Vohra, Shawn Lin, Max Lefarov, Paul Gigioli, Gustavo Penha, Ilya Kopysitsky, Ivo Joel Senese, Darren Mei, Francesco Fabbri, Oguz Semerci, Yu Zhao, Vincent Tang, Brian St. Thomas, Alexandra Ranieri, Matthew N. K. Smith, Aaron Bernkopf, Bryan Leung, Ghazal Fazelnia, Mark VanMiddlesworth, Timothy Christopher Heath, Petter Pehrson Skiden, Alice Y. Wang, Doug J. Cole, Andreas Damianou, Maya Hristakeva, Reid Wilbur, Tarun Chillara, Vladan Radosavljevic, Pooja Chitkara, Sainath Adapa, Juan Elenter, Bernd Huber, Jacqueline Wood, Saaketh Vedantam, Jan Stypka, Sandeep Ghael, Martin D. Gould, David Murgatroyd, Yves Raimond, Mounia Lalmas, and Paul N. Bennett. Deploying semantic ID-based generative retrieval for large-scale podcast discovery at Spotify. *arXiv preprint arXiv:2603.17540*, 2026. doi: 10.48550/arXiv.2603.17540. URL https://arxiv.org/abs/2603.17540.

Marco De Nadai, Edoardo D'Amico, Max Lefarov, Alexandre Tamborrino, Divita Vohra, Mark VanMiddlesworth, Shawn Lin, Jacqueline Wood, Jan Stypka, Eliza Klyce, Keshi Dai, Timothy Christopher Heath, Martin D. Gould, Yves Raimond, Sandeep Ghael, Tony Jebara, Andreas Damianou, Vladan Radosavljevic, Paul N. Bennett, Mounia Lalmas, and Praveen Chandar. A unified language model for large scale search, recommendation, and reasoning. *arXiv preprint arXiv:2603.17533*, 2026. doi: 10.48550/arXiv.2603.17533. URL https://arxiv.org/abs/2603.17533.

Zhe Dong, Fang Qin, Manish Shah, and Yicheng Wang. Diagnosing and mitigating retrieval bottlenecks in LLM-based cold-start recommendation. *arXiv preprint arXiv:2606.29947*, 2026. doi: 10.48550/arXiv.2606.29947.

Tingchen Fu and Fazl Barez. Same question, different words: A latent adversarial framework for prompt robustness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31305–31319, 2025. doi: 10.18653/v1/2025.emnlp-main.1595. URL https://aclanthology.org/2025.emnlp-main.1595/.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786. URL https://arxiv.org/abs/2503.19786.

Google. Gemini 3 Flash Preview. Gemini API documentation, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview. Model identifier: gemini-3-flash-preview; accessed 2026-08-10.

Jiasheng Gu, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei, and Wenpeng Yin. Robustness of learning from task instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13935–13948, 2023. doi: 10.18653/v1/2023.findings-acl.875. URL https://aclanthology.org/2023.findings-acl.875/.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 507–517, 2016.

Ruining He, Lukasz Heldt, Lichan Hong, Raghunandan Keshavan, Shifan Mao, Nikhil Mehta, Zhengyang Su, Alicia Tsai, Yueqi Wang, Shao-Chuan Wang, Xinyang Yi, Lexi Baugher, Baykal Cakici, Ed Chi, Cristos Goodrow, Ningren Han, He Ma, Romer Rosales, Abby Van Soest, Devansh Tandon, Su-Lin Wu, Weilong Yang, and Yilin Zheng. PLUM: Adapting pre-trained language models for industrial-scale generative recommendations. *arXiv preprint arXiv:2510.07784*, 2025. doi: 10.48550/arXiv.2510.07784. URL https://arxiv.org/abs/2510.07784.

Yingzhi He, Yan Sun, Junfei Tan, Yuxin Chen, Xiaoyu Kong, Chunxu Shen, Xiang Wang, An Zhang, and Tat-Seng Chua. Reasoning over semantic IDs enhances generative recommendation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3818122. URL https://arxiv.org/abs/2603.23183.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval*, pp. 364–381, 2024. doi: 10.1007/978-3-031-56060-6_24.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of*

the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 8235–8246. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.722/.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308, 2024. doi: 10.18653/v1/2024.findings-emnlp.249.

Daniel Lloyd. Amazon's AI-powered 'Interests' feature automatically finds new products that match your passions and hobbies. Amazon News, 2025. URL https://www.aboutamazon.com/news/retail/artificial-intelligence-amazon-features-interest.

Mrinalini Loew. Personalize the content you see on Search, Discover, and News. Google Blog, August 2026. URL https://blog.google/products-and-platforms/products/search/personalize-search-discover-news/.

Wensheng Lu, Jianxun Lian, Wei Zhang, Guanghua Li, Mingyang Zhou, Hao Liao, and Xing Xie. Aligning large language models for controllable recommendations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8159–8172, 2024. doi: 10.18653/v1/2024.acl-long.443.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, 2022. doi: 10.18653/v1/2022.findings-acl.146.

Fabian Paischer, Liu Yang, Linfeng Liu, Shuai Shao, Kaveh Hassani, Jiacheng Li, Ricky T. Q. Chen, Zhang Gabriel Li, Xiaoli Gao, Wei Shao, Xue Feng, Nima Noorshams, Sem Park, Bo Long, and Hamid Eghbalzadeh. Preference discerning with LLM-enhanced generative retrieval. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=74mrOdhvvT.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10299–10315, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/20dcab0f14046a5c6b02b61da9f13229-Abstract-Conference.html.

Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. Transparent and scrutable recommendations using natural language user profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13971–13984, 2024. doi: 10.18653/v1/2024.acl-long.753. URL https://aclanthology.org/2024.acl-long.753/.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55854–55875. PMLR, 2025. URL https://proceedings.mlr.press/v267/skean25a.html.

Neha Srikanth, Marine Carpuat, and Rachel Rudinger. How often are errors in natural language reasoning due to paraphrastic variability? *Transactions of the Association for Computational Linguistics*, 12:1143–1162, 2024. doi: 10.1162/tacl_a_00692. URL https://aclanthology.org/2024.tacl-1.63/.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, 2023. doi: 10.18653/v1/2023.emnlp-main.923.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2400–2409, 2024. doi: 10.1145/3627673.3679569.

Zhenrui Yue, Sara Rabhi, Gabriel de Souza Pereira Moreira, Dong Wang, and Even Oldridge. LlamaRec: Two-stage recommendation using large language models for ranking. In *First Workshop on Personalized Generative AI at CIKM 2023*, 2023. URL https://arxiv.org/abs/2311.02089.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Recommendation as instruction following: A large language model empowered recommendation approach. *ACM Transactions on Information Systems*, 43(5):1–37, 2025. doi: 10.1145/3708882.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 1435–1448, 2024. doi: 10.1109/ICDE60146.2024.00118. URL https://arxiv.org/abs/2311.09049.

Guorui Zhou, Honghui Bao, Jiaming Huang, Jiaxin Deng, Jinghao Zhang, Junda She, Kuo Cai, Lejian Ren, Lu Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rongzhou Zhang, Ruiming Tang, Shiyao Wang, Wuchao Li, Xiangyu Wu, Xinchen Luo, Xingmei Wang, Yifei Hu, Yunfan Wu, Zhanyu Liu, Zhiyang Zhang, Zixing Zhang, Bo Chen, Bin Wen, Chaoyi Ma, Chengru Song, Chenglong Chu, Defu Lian, Fan Yang, Feng Jiang, Hongtao Cheng, Huanjie Wang, Kun Gai, Pengfei Zheng, Qiang Wang, Rui Huang, Siyang Mao, Tingting Gao, Wei Yuan, Yan Wang, Yang Zhou, Yi Su, Zexuan Cheng, Zhixin Ling, and Ziming Li. OpenOneRec technical report. *arXiv preprint arXiv:2512.24762*, 2025. doi: 10.48550/arXiv.2512.24762. URL https://arxiv.org/abs/2512.24762.

Joyce Zhou, Weijie Zhou, Doug Turnbull, and Thorsten Joachims. SteerEval: A framework for evaluating steerability with natural language profiles for recommendation. *arXiv preprint arXiv:2601.21105*, 2026. URL https://arxiv.org/abs/2601.21105.

APPENDIX

A BENCHMARK CONSTRUCTION

A.1 SOURCE, SPLITS, AND EXPERIMENTAL SUBSET

The public suite reformats the Preference Discerning release (Paischer et al., 2025) over the Beauty, Sports and Outdoors, and Toys and Games categories of the Amazon 5-core reviews and product metadata (He & McAuley, 2016). Upstream preprocessing iteratively enforces the 5-core condition, remaps reviewer and item identifiers to contiguous integers, orders interactions by timestamp, and retains at most the 20 most recent interactions per user. We use chronological leave-two-out splits: all earlier prefixes produce training targets, the penultimate interaction is validation, and the final interaction is test. Consequently, every user occurs in all three splits. The published artifact contains six dimensions; our experiments exclude only the inherited, history-free legacy sentiment split. Table 4 reports overall totals, while Table 5 details example counts per dimension and split.

Cross-split audit confirms that exact-content duplicates under chronological splitting are negligible (fewer than 15 rows across 400,000+ interactions in each domain), naturally mirroring real-world repeat interactions.

Table 4: **Dataset statistics across the three benchmark domains.** Interaction counts reflect our five-axis experimental splits; item corpus, user pools, and RQ-VAE codebooks are domain-specific.

Domain	Public total	Train	Validation	Test	Items	Users
Beauty	429,902	207,228	87,428	86,328	12,101	22,363
Sports and Outdoors	618,645	294,293	132,504	131,898	18,357	35,598
Toys and Games	356,138	171,924	76,446	75,012	11,924	19,412

Table 5: **Example counts per dimension and split across domains.** Interaction counts across evaluation splits; intent following is evaluated zero-shot (Train = 0).

	Beauty			**Sports and Outdoors**			**Toys and Games**		
Dimension	Train	Val	Test	Train	Val	Test	Train	Val	Test
Preference recommendation	114,104	22,363	22,363	175,425	35,598	35,598	95,582	19,412	19,412
Intent following	0	22,363	22,363	0	35,598	35,598	0	19,412	19,412
Sentiment following	44,896	9,298	9,690	45,764	9,852	10,326	28,882	5,686	6,288
Fine steering	24,114	16,702	15,956	36,552	25,728	25,188	23,730	15,968	14,950
Coarse steering	24,114	16,702	15,956	36,552	25,728	25,188	23,730	15,968	14,950
Legacy sentiment (excl.)	34,210	7,310	7,398	42,820	8,546	8,584	23,054	4,684	5,018

A.2 Preference and Axis Construction

Preference provenance. The inherited preference text was generated with Llama-3-70B-Instruct from up to 5 recent reviewed items, requesting short search instructions describing likes, dislikes, and attributes to avoid (excluding brands, named products, delivery, and price) using temperature 0.6, nucleus probability 0.9, and at most 512 new tokens.

Preference recommendation and intent following. For each chronological next-item target, preference recommendation retrieves instructions attached to the immediately preceding item, selecting the candidate with the highest Sentence-T5 dot product with the target title (Ni et al., 2022). Intent following concatenates up to the first 5 candidate preferences associated with the previous item, requiring the model to identify the target-relevant intent among distractors. It is evaluated exclusively on validation and test splits without contributing training records.

Causal sentiment following vs. legacy construction. We replace the unconditioned sentiment task from Preference Discerning (Paischer et al., 2025) with a strictly causal, history-conditioned formulation. In the legacy construction, instructions were evaluated without user browsing history and could be paired with preferences drawn from future interactions. Our construction anchors on each chronological negative-review interaction, materializing its strict prefix as user history. It selects the candidate instruction most aligned with the anchor item via Sentence-T5 from anchor or strictly prior positions, and obtains its positive counterpart by lexical inversion. Both find and avoid rows share identical user history and anchor targets, making Sentiment@10 a controlled intervention where request polarity varies under fixed behavioral context.

Fine steering shortcut and lexical control. Fine steering pairs source preferences with target items via Sentence-T5 nearest-neighbor matching in the same embedding space used to train the discrete RQ-VAE codebooks. As detailed in Table 6, this synthetic alignment induces substantial lexical and semantic overlap: prompt–target title overlap reaches 60.7–84.7% for fine steering compared to 10.0–24.1% for preference recommendation. To quantify the extent to which keyword overlap alone exposes the exact target without neural retrieval, we evaluate a model-free baseline using sublinear TF-IDF vectors $\widehat{\phi}(x)$ with cosine similarity:

$$\text{Recall@10}_{\text{TFIDF}} = \mathbb{E}_{(r_u, i_u^*)\sim\mathcal{D}}\left[\mathbf{1}\left[i_u^* \in \text{Top}_{10}(r_u)\right]\right], \tag{7}$$

Table 6: **Audit of preference recommendation vs. fine steering (values in %).** *Last-history R@10* uses TF-IDF on the most recent history title; *Train-prior R@10* uses the most frequent training targets; and *Context–target Prefix-1* is the fraction sharing the first RQ-VAE code with the context item (preceding item for Pref. rec., paired anchor for Fine).

Domain	Task	Prompt–title overlap	Preference-only R@10	Last-history R@10	Train-prior R@10	Context–target Prefix-1
Beauty	Pref. rec.	24.08	1.09	6.45	1.40	10.25
	Fine	84.67	21.72	3.16	0.33	69.70
Sports	Pref. rec.	9.97	0.90	3.51	0.93	7.67
	Fine	80.71	30.56	2.19	0.23	73.92
Toys	Pref. rec.	10.13	0.63	8.53	0.90	16.03
	Fine	60.68	16.69	4.73	0.54	67.17

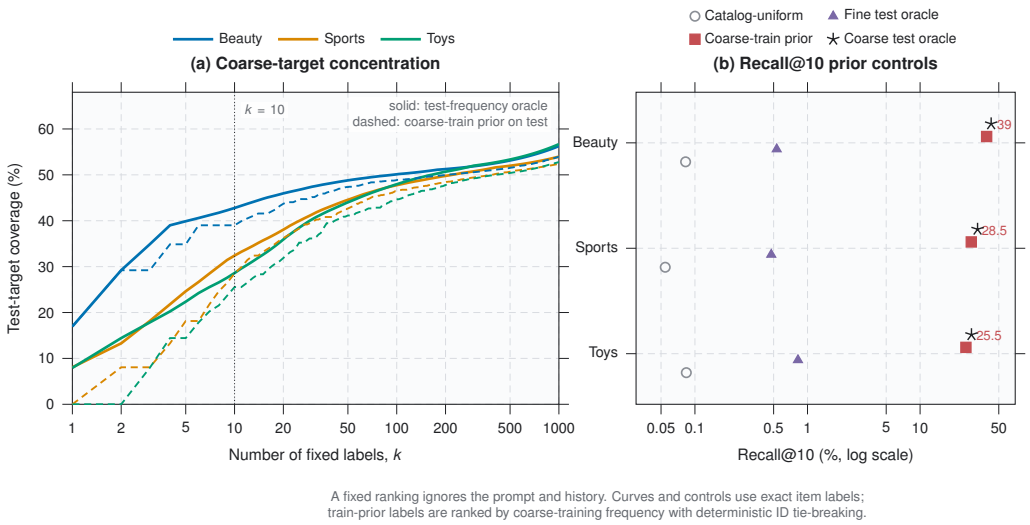


Figure 4: **Target-prior audit for coarse steering on the complete test splits.** **(a)** Cumulative target coverage under test-frequency oracle rankings (solid) and coarse-training frequency rankings (dashed). At $k = 10$, oracle coverage is 42.8%, 32.4%, and 28.6% on Beauty, Sports, and Toys. **(b)** Context-free training priors achieve 39.0%, 28.5%, and 25.5% Recall@10, respectively; catalog-uniform chance remains below 0.1%.

where $\mathrm{Top}_{10}(r_u) = \arg\mathrm{top}^{10}_{i\in\mathcal{I}} \cos(\widehat{\phi}(r_u), \widehat{\phi}(t_i))$ denotes the 10 catalog items whose title TF-IDF vectors $\widehat{\phi}(t_i)$ have highest cosine similarity with $\widehat{\phi}(r_u)$. Table 6 also reports *Last-history R@10* (retrieval via the last interaction's title) and *Train-prior R@10* (a static list of top training targets). This preference-only TF-IDF baseline achieves 16.7–30.6% Recall@10 on Fine versus only 0.6–1.1% on preference recommendation, confirming that elevated Fine Recall reflects a surface lexical recovery shortcut rather than unconstrained discovery.

Coarse steering target-prior concentration. Coarse steering pairs source preferences with least-similar items via Sentence-T5 farthest-neighbor matching, swapping trajectories to redirect retrieval. Each test split contains equally sized anchor and transformed halves. The 10 most frequent targets cover 57.1–85.4% of the transformed half across the three domains; on the complete test splits, top-10 target coverage is 28.6–42.8%. A context-free top-10 ranking learned from coarse-training target frequencies achieves 25.5–39.0% Recall@10 on these complete test splits (Figure 4), showing that substantial coarse retrieval performance can be obtained from target frequencies alone.

A.3 SEMANTIC IDS AND USER TOKENS

Each domain trains a separate RQ-VAE over standardized 768-dimensional Sentence-T5-XXL embeddings of concatenated title and brand metadata (Ni et al., 2022; Rajput et al., 2023). Its MLP encoder has widths $768 \to 768 \to 512 \to 256 \to 128$ with a symmetric decoder, LayerNorm, ReLU, residual projections, and dropout 0.1. The 3 residual codebooks contain 256 vectors each and use sequential k-means initialization. With reconstruction, embedding, and commitment mean-squared errors, the objective is

$$\mathcal{L}_{\mathrm{RQ}} = \mathcal{L}_{\mathrm{recon}} + \mathcal{L}_{\mathrm{embed}} + 0.25\,\mathcal{L}_{\mathrm{commit}}. \tag{8}$$

The recorded optimization uses AdamW with learning rate 10^{-3}, weight decay 0.1, batch size 2,048, 8,000 epochs, and a 5% RQ-VAE validation split with seed 42. The 4th code is a deterministic collision index, assigned in numerical item order, that disambiguates items sharing the same 3-code prefix. All 42,382 items have 4-token SIDs; the maximum collision indices are 9, 13, and 24 for Beauty, Sports, and Toys, respectively.

The joint-conditioning and modular loaders follow TIGER's user-token construction, mapping each user ID to one of 2,000 hash buckets $u_{\mathrm{hash}} \in \{1, \ldots, 2000\}$ via $1 + (\text{SHA-1}(\text{user_id}) \bmod 2000)$, serialized as subword text tokens using Gemma 3's native vocabulary (Rajput et al., 2023). The public chronological split contains the same users in training, validation, and test, and therefore does not evaluate user cold start.

A.4 ALTERNATIVE-TARGET SUPERVISION FOR NEGATIVE PREFERENCES

Our sentiment-following task pairs positive ("find") and negative ("avoid") requests for the same anchor item i under identical browsing context. Unlike naive loss-reversal heuristics that maximize cross-entropy on rejected items without providing replacement candidates (Paischer et al., 2025), we convert negative preferences into positive supervision over history-compatible alternatives.

For each negative request, candidate replacement items are scored by Gemini 3 Flash Preview for compliance with the negative constraint and alignment with the user's browsing pattern. Retaining eligible alternatives scoring ≥ 7, we augment the base training set with a fixed sample of $0.3N$ alternative-target pairs ($1.3N$ total records). Each sampled item $\tilde{i}_j$ provides standard positive-likelihood SFT supervision without reversing the training loss. Candidate selection strictly respects split boundaries, querying neither validation/test interactions nor future user behavior. Alternative targets affect training only; validation and test evaluation retains the ground-truth anchor i to measure exact retrieval and exclusion.

A.5 FIXED PROMPT LAYOUT

The data loader preserves three logical inputs across all evaluation records: chronological history titles, the user preference utterance, and the discrete history sequence S. The hashed user token u_{hash} forms the first token of the discrete history. In the serialized prompt, the natural-language preference includes a fixed directive instructing the model to generate the semantic ID of the next item based on the interaction history and preference constraint.

Both conditioning paradigms share the exact same natural-language context, differing only in architectural serialization: (1) **Joint conditioning (single causal stream):** all three components are serialized sequentially as [history titles] $\to$ [preference] $\to$ [u_{hash} + history SIDs], followed immediately by autoregressive target SID generation; (2) **Modular interfaces (separate language and SID towers):** the textual context [history titles] $\to$ [preference] enters the frozen language tower as the condition sequence, while the discrete history [u_{hash} + history SIDs] enters the trainable SID generator. Complete prompts and rephrasings appear in Appendix B.

B PROMPT TEMPLATES

B.1 JOINT-CONDITIONING RETRIEVAL PROMPT

The Beauty validation instance (sample 3252) below is the source of Figure 5, formatted with native Gemma 3 turn tokens (<bos>, <start_of_turn>, <end_of_turn>). Ellipses reflect title truncation by the loader.

Joint-conditioning input prompt and supervised target — Beauty validation sample 3252

<bos><start_of_turn>user
The user 918 has interacted with the following items in chronological order:
1. Toppik Hair Building Fibers (.42 oz.) – Medium Brown
2. EyeLash Extensions Disposable Micro Brushes 100 pack
3. 10 Pairs Handmade Natural Soft False Eyelashes Fake Eye Lash
4. Beauty False Fake Eyelash Glue Applicator & Clip
User preference: Search for false eyelashes that are soft, natural, and easy to apply. Based on the ordered interaction history and the user preference, generate the semantic ID of the next item to recommend.
918 :<A180><B39><C8><D0><A27><B78><C182><D0><A44><B180><C183><D0>
<A44><B172><C183><D0><end_of_turn>
<start_of_turn>model
<A44><B141><C183><D0><end_of_turn>

The target item title is "20 Pairs of Natural & Regular Long False Eyelashes." Although not appearing verbatim, *false eyelashes* and *natural* occur in both the preference and target title, alongside related history titles. This lexical overlap makes Figure 5 useful for diagnosing potential keyword-matching shortcuts rather than asserting semantic understanding.

B.2 HISTORY-AWARE ALTERNATIVE-TARGET CONSTRUCTION

To synthesize history-compatible replacement candidates for negative constraints (Section A.4), the pipeline calls the Gemini 3 Flash Preview API (Google, 2026). The template below faithfully restates instructions with batch payloads omitted; JSON schemas are line-wrapped for readability.

Alternative-target selection and positive-intent synthesis — Gemini 3 Flash Preview API

You are an expert Amazon e-commerce recommendation system. You are given a batch of users. For each user, I will provide their browsing history, a negative constraint, the rejected item, and a pool of candidate items. For each user, select up to 10 of the **BEST** items from **THEIR** candidate pool that you think they will actually take, avoiding their specific constraint.

Crucially, calculate a distance_score from 0 to 10 for each selection:
• 0 means the candidate violates the constraint.
• 10 means the candidate is completely free of the negative constraint, while still making logical sense.

Additionally, for each selection, synthesize a positive_intent string (e.g., 'find a soothing ceramic styling tool') based on the user's history and constraint.

CRITICAL NO DATA LEAKAGE RULE: The synthesized positive_intent **MUST NOT** contain the exact title or brand name of the selected candidate item. Describe only its general product type, feature, or category intent.

Output format **MUST** be strictly JSON with no surrounding text or markdown formatting:

```
{"users_selections": [{"user_index": integer, "selections": [{
"selected_index": integer, "distance_score": integer,
"reasoning": "brief explanation",
"positive_intent": "find a soothing ceramic tool"}]}]}
```

This prompt constructs positive-likelihood SFT targets for avoid requests; its distance score serves as a construction-time quality filter rather than an evaluation metric. Candidates scoring below 7 are discarded during dataset construction to ensure that negative constraints are strictly satisfied without introducing noisy pseudo-targets.

B.3 PREFERENCE-REPHRASING GENERATION

To evaluate phrasing robustness under semantics-preserving linguistic variations (Section 5.4), candidate preference rephrasings are generated using the Gemini 3 Flash Preview API (Google, 2026). The three fields specify distinct semantic transformations with descriptive aliases.

Meaning-preserving preference rephrasings — Gemini 3 Flash Preview API

You are an expert NLP assistant for e-commerce recommendation systems. I will provide a batch of user preference instructions from a recommendation dataset. For each preference instruction, generate 3 paraphrased versions:

- "`lexical_substitution`": Replace key nouns/adjectives/verbs with synonyms while keeping the **EXACT** sentence structure.
- "`syntactic_paraphrase`": Rephrase the syntax structure (e.g., change "Consider products with..." to "I am looking for items that...") while keeping the exact same meaning.
- "`conversational_rewrite`": Reformulate as a natural conversational or polite user request while keeping the exact same meaning.

CRITICAL SEMANTIC CONSTRAINT:

- **DO NOT** change the semantic intent or target product attributes!
- Keep the meaning 100% equivalent so that the recommended product should remain identical.

Output format **MUST** be strictly JSON with no surrounding text or markdown formatting:

```
{"results": [{"index": integer, "lexical_substitution": "...",
"syntactic_paraphrase": "...", "conversational_rewrite": "..."}]}
```

Manual spot-checks over the generated rephrasings confirm high semantic fidelity, with negation scope and core product attributes accurately preserved across transformations. Evaluating models across these variants probes sensitivity to surface phrasing; performance alone does not establish a particular semantic mechanism.

C ARCHITECTURE DETAILS

Joint conditioning uses Gemma 3's native local sliding-window and global attention patterns without additional fusion parameters. The modular interfaces retain a frozen Gemma language tower and add conditioning to a separate trainable generator as specified below.

C.1 FINAL-STATE BLOCK

Both towers are initialized from the Gemma 3 1B IT checkpoint. The language tower remains frozen with its native 262,144-token vocabulary, which also represents all text and hashed user numbers u_{hash}. The trainable SID generator expands this vocabulary with 1,024 newly initialized SID tokens, corresponding to 4 hierarchical codebook positions of cardinality 256 (initialized from $\mathcal{N}(0, 0.02^2)$ with seed 42).

The final-state condition sequence $C = (c_1, \ldots, c_{L_c}) \in \mathbb{R}^{L_c \times D}$ is extracted after the frozen tower's final RMSNorm layer. Sequences are right-padded to the batch maximum length L_c, with binary mask m_j marking non-padding tokens. For the global modulation branch, non-padding condition representations are pooled via masked mean:

$$\bar{c} = \frac{\sum_{j=1}^{L_c} m_j c_j}{\max\left(\sum_{j=1}^{L_c} m_j,\ 10^{-5}\right)}, \qquad m_j = \mathbf{1}[\text{token}_j \neq 0]. \tag{9}$$

A block-specific linear projection W_ℓ^0, initialized with zero weights and zero bias, transforms $\text{SiLU}(\bar{c})$ into seven D-dimensional modulation and gating vectors:

$$(b_{sa}, s_{sa}, a_{sa}, b_{ff}, s_{ff}, a_{ff}, a_{ca})_\ell = W_\ell^0 \,\text{SiLU}(\bar{c}). \tag{10}$$

Here b_{sa}, s_{sa} and b_{ff}, s_{ff} shift and scale the self-attention and feed-forward inputs, while a_{sa}, a_{ff}, a_{ca} control their residual branches and the cross-attention residual. The modulated sublayers apply Adaptive RMSNorm (Equation 5), and the gated residual connections (Equation 6) ensure that all blocks initially reproduce the unmodified pretrained Gemma layer.

Table 7 summarizes the routing configurations across the three evaluated final-state ablations. The global-only variant disables cross-attention and leaves the 7th projection chunk $a_{ca,\ell}$ inactive. The cross-attention-only variant disables global pooling and parameterizes $a_{ca,\ell}$ as an independent zero-initialized trainable vector in $\mathbb{R}^D$. The combined variant activates both pathways, extracting $a_{ca,\ell}$ directly from the 7-way projection.

Table 7: **Final-state conditioning modes.** Routing pathways and residual gate parameterizations across the three ablation variants.

Mode	Pooled global path	Token cross-attention	Cross residual gate
Global only	Yes	No	Unused 7th chunk
Cross-attention only	No	Yes	Separate zero-init vector
Both	Yes	Yes	7th projected chunk

For completeness, let $\tilde{u}_{\ell,i} = \mathrm{RMSNorm}(u_{\ell,i})$ be a generator state at right-stream segment position π_i, and let m_j be the condition mask defined above. Standalone cross-attention computes

$$
\begin{aligned}
q_i &= \tilde{u}_{\ell,i} W_Q^\ell, & k_j &= c_j W_K^\ell, & v_j &= c_j W_V^\ell, \\
p_i^Q &= n_c + \pi_i, & p_j^K &= j, & n_c &= \sum_j m_j, \\
\widetilde{q}_i &= \mathrm{RoPE}_\ell(q_i, p_i^Q), & \widetilde{k}_j &= \mathrm{RoPE}_\ell(k_j, p_j^K), &&
\end{aligned}
$$

$$
\alpha_{ij} = \mathrm{softmax}_{j:m_j=1}\left(\frac{\widetilde{q}_i \widetilde{k}_j^\top}{\sqrt{d_h}}\right), \qquad \mathrm{CA}_\ell(\tilde{u}_{\ell,i}, C) = \left(\sum_j \alpha_{ij} v_j\right) W_O^\ell. \tag{11}
$$

The implementation uses independent W_Q, W_K, W_V, W_O projections with Gemma's grouped-query head layout, applying no extra query/key normalization and no cross-attention dropout. Unrotated values and invalid right-padded condition logits are masked before softmax. For local sliding-window blocks, RoPE_ℓ adopts the configured local base frequency and scale factor; global blocks use their global configuration. Cross-stream rotary positions are thus continuous while preserving layer-dependent geometry.

The added trainable parameters can be stated independently of a particular Gemma size. With embedding width D, H query heads, H_{kv} key/value heads, and head width d_h, and including projection biases, the 7-way projection contributes $7D^2 + 7D$ parameters per block. Standalone cross-attention plus its pre-cross-attention RMSNorm contributes $2D(H + H_{kv})d_h + Hd_h + 2H_{kv}d_h + 2D$ parameters per block, with the cross-attention-only variant retaining an additional D-dimensional gate parameter.

Layer-wise adaptive RMSNorm control. This control takes the hidden-state sequence produced at each frozen language layer ℓ, mean-pools non-padding tokens, and modulates generator block ℓ via adaptive RMSNorm without token cross-attention. The matched last-layer baseline likewise uses global modulation only, but pools post-final-RMSNorm states once and broadcasts that summary to all blocks. This pair compares final-state broadcasting with depth-aligned routing under the same pooled modulation mechanism; it does not rank all individual source layers.

C.2 Token and Position Alignment

The dual-branch and cross-attention-only interfaces enforce a continuous rotary coordinate system across the language and generator streams. Prompts are right-padded with token ID 0, with non-padding tokens determining the condition length $n_c = \sum_j \mathbf{1}[\mathrm{token}_j \neq 0]$. Language tokens occupy coordinates $[0, n_c)$. In the generator stream, the discrete token sequence (comprising the user hash token, history SIDs, and target SIDs, with total length $n_s = |c_u^{\mathrm{SID}}| + |y_i|$) occupies contiguous coordinates $[n_c, n_c{+}n_s)$, where each generator token at segment-local index $\pi_i \in [0, n_s)$ is assigned the global rotary coordinate $n_c + \pi_i$. Each block preserves Gemma 3's layer-dependent rotary geometry, adopting the configured local base frequency and scaling factor for sliding-window layers and global settings for full-attention layers.

Condition keys retain contiguous prompt indices $j \in [0, n_c)$ with padding keys masked during cross-attention, while generator queries evaluate at $n_c + \pi_i$.

Autoregressive causal decoding. Generator self-attention is causal. Every decoding step can attend to all non-padding condition states through cross-attention, while generated SID tokens attend only to antecedent generator positions in the four-step trajectory $(s_1, \ldots, s_4)$.

D Training and Evaluation

D.1 Sweep Grid

To establish fair operating points, we sweep peak learning rate, batch size, and training budget across all three architectures: (1) **Joint conditioning:** learning rates $\{5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}, 2\times10^{-2}\}$, batch sizes $\{8, 12, 16, 32, 48\}$, and budgets of $\{30\text{k}, 50\text{k}, 70\text{k}\}$ steps (75 configurations per domain); (2) **Final-state conditioning:** learning rates $\{10^{-4}, 2\times10^{-4}, 5\times10^{-4}, 10^{-3}, 2\times10^{-3}, 5\times10^{-3}, 10^{-2}, 2\times10^{-2}\}$, batch sizes $\{8, 12, 16, 32, 48\}$, and budgets of $\{30\text{k}, 50\text{k}, 70\text{k}\}$ steps (120 configurations per domain); and (3) **Cross-attention only:** the same 8 learning rates, 5 batch sizes, and 3 step budgets as the dual-branch model (120 configurations per domain).

In total, these primary-interface sweeps comprise 945 language-conditioned training runs across the three benchmark domains (225 joint-conditioning, 360 dual-branch, and 360 cross-attention-only runs). In addition, the text-free SID-history baseline evaluates 75 runs per domain (225 runs total). Headline evaluations and rephrasing robustness span all three domains; layer-depth ablations, general language capabilities, and attention analyses are conducted on Beauty.

D.2 Selection and Uncertainty

Hyperparameters for each interface and domain are chosen strictly on validation splits by maximizing a composite score: $0.45 \times \text{Recall@10}_{\text{pref}} + 0.45 \times \text{Sentiment@10} + 0.10 \times \text{Recall@10}_{\text{intent}}$ (excluding fine and coarse steering due to synthetic shortcuts and label priors). Test splits are held out entirely and never accessed during tuning. For the text-free SID-history baseline, selection maximizes validation preference-recommendation Recall@10.

To assess variance, each selected configuration is retrained across three independent random seeds (42, 43, and 44). Table 10 reports the resulting means and sample standard deviations across all five steering dimensions, and Table 8 lists the chosen hyperparameters for each interface and domain.

Table 8: **Hyperparameter configurations of selected operating points.** Optimal training settings and selected hyperparameters for the multi-domain benchmark results in Table 10.

Domain	Interface	Peak LR	Batch	Steps
Beauty	SID-history only	1×10^{-2}	16	30k
	Joint conditioning	5×10^{-3}	16	30k
	Cross-attention only	5×10^{-3}	12	50k
	Final-state injection	5×10^{-3}	32	30k
Sports	SID-history only	2×10^{-2}	32	30k
	Joint conditioning	1×10^{-2}	12	50k
	Cross-attention only	5×10^{-3}	32	30k
	Final-state injection	5×10^{-3}	32	30k
Toys	SID-history only	5×10^{-3}	12	30k
	Joint conditioning	2×10^{-2}	12	30k
	Cross-attention only	5×10^{-3}	32	30k
	Final-state injection	5×10^{-3}	16	50k

E Additional Analyses and Ablations

E.1 Target-Aligned Attention Analysis and Extraction

To complement the rephrasing results in Section 5.4, we inspect attention routing in a representative Beauty validation instance under joint conditioning (Figure 5). The complete prompt is provided in Appendix B.1. The attention visualization uses an evaluation-only, teacher-forced forward pass on the complete validation sequence. Let $A_{qk}^{(\ell,h)}$ be the causal self-attention weight from target-SID query position q to context position $k \leq q$ in layer ℓ and head h. To capture maximum routing intensity across depth $\mathcal{C}$ (all 26 layers), we extract the envelope S_{qk} and normalize with per-query reference scale $S_q^{\text{ref}} = \max_{k>0} S_{qk}$ (excluding the $k = 0$ attention sink, with $\epsilon = 10^{-8}$):

$$S_{qk} = \max_{\ell \in \mathcal{C},\, h} A_{qk}^{(\ell,h)} \quad (k \leq q), \qquad z_{qk} = \text{clip}\left(\frac{\log(S_{qk} + \epsilon) - \log S_q^{\text{ref}} + 6}{6}, 0, 1 \right). \tag{12}$$

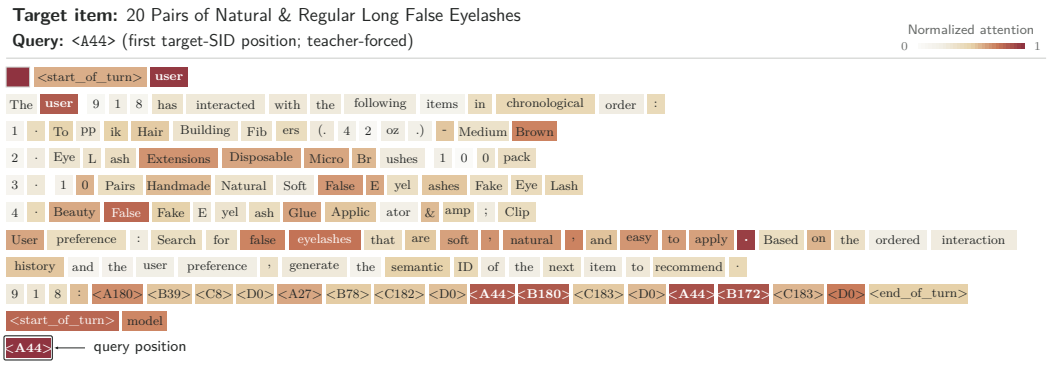


Figure 5: **Target-aligned attention in a joint-conditioning example.** Attention from the first target-SID token position (<A44>) in a teacher-forced sequence, maximized across heads and layers and normalized as in Equation 12. The outlined token marks the query. Preference keywords shared with the target title (*false*, *eyelashes*, *natural*) receive high normalized attention.

Score z_{qk} maps to a light-to-red palette indicating intra-query localization. The target SID is <A44><B141><C183><D0>; its first token, <A44>, already occupies the query position in this teacher-forced pass. Within the preference span, keywords shared with the target title receive high normalized attention: *eyelashes* (0.782), *false* (0.729), and *natural* (0.678). This single-case visualization shows keyword-focused routing consistent with the lexical shortcut behavior observed in Section 5.4.

E.2 Source Representation Depth: Final-Layer versus Layer-Wise Conditioning

Controlled depth-routing comparison. Holding the conditioning mechanism fixed to global-only AdaRMS modulation and disabling cross-attention, we compare two depth-routing strategies on Beauty: injecting the final-layer summary into all generator blocks versus layer-wise one-to-one injection from matching language depths. We evaluate both a 120-cell matched-grid diagnostic across identical hyperparameters and independently selected operating points across seeds 42–44. Across three seeds, validation Recall@10 is $8.95 \pm 0.24\%$ for last-layer versus $8.14 \pm 0.26\%$ for layer-wise; on held-out test, they reach $6.40 \pm 0.10\%$ versus $6.02 \pm 0.12\%$ (advantage of 0.38 ± 0.22 pp).

Interpretation. We hypothesize that final-layer states consolidate task-relevant context into a more useful shared signal. Context-knowledge probes in LLaMA 2 find that higher layers make contextual information accessible across more token positions (Ju et al., 2024), although intermediate layers can perform better on text-embedding tasks (Skean et al., 2025). Our retrieval comparison

favors final-layer broadcasting over depth-aligned summaries; it does not directly measure representation noise or rank every fixed source layer.

E.3 Final-State Conditioning Paths

We evaluate individual components of the final-state interface by isolating global modulation, token cross-attention, and their dual-branch combination using the validation selection protocol (Appendix D.2), evaluated across seeds 42–44 on held-out test. The cross-attention-only variant is also a main baseline in Table 2, evaluating the encoder–decoder conditioning pattern used by Mender (Paischer et al., 2025) under our shared Gemma 3 backbone, SID codebooks, and training objective. Each mode is selected independently: all three selected configurations use a peak learning rate of 5×10^{-3}, with batch size/training steps of $16/70{,}000$ for global-only, $12/50{,}000$ for cross-attention-only, and $32/30{,}000$ for the combined model. The comparison therefore measures validation-selected performance rather than performance at a fixed training budget.

Table 9: **Ablation of final-state conditioning paths on Beauty test.** Mean±std over three seeds (42–44) in %; bold indicates best. Lower is better for Avoid Hit@10 (↓).

Conditioning mode	Pref. Recall@10	Intent Recall@10	Find Recall@10	Avoid Hit@10	Sentiment@10
Global only	6.40 ± 0.10	5.63 ± 0.06	4.20 ± 0.22	3.20 ± 0.27	1.53 ± 0.29
Cross-attention only	8.34 ± 0.07	$\mathbf{5.89 \pm 0.52}$	6.05 ± 0.17	2.37 ± 0.26	4.63 ± 0.27
Global + cross-attention	$\mathbf{8.73 \pm 0.18}$	5.55 ± 0.68	$\mathbf{6.50 \pm 0.14}$	$\mathbf{1.43 \pm 0.21}$	$\mathbf{5.59 \pm 0.12}$

As reported in Table 9, the combined dual-branch path achieves the strongest preference recommendation (8.73%), Find Recall@10 (6.50%), and Sentiment@10 (5.59%), alongside the lowest Avoid hit rate (1.43%). Cross-attention-only conditioning has the highest mean intent-following Recall@10 (5.89%, compared with 5.63% for global-only and 5.55% for the combined model). Relative to cross-attention alone, the combined model gains 0.39 percentage points in preference Recall@10 and 0.96 percentage points in Sentiment@10, while reducing Avoid Hit@10 by 0.94 percentage points. These results support complementary benefits on retrieval and directional control, but not uniform superiority across tasks.

E.4 Full Multi-Domain Steering Results

Table 10 reports validation and test results for all five steering tasks, including Recall and nDCG at both ranks 5 and 10. The configurations and three seeds are identical to the core test results in Table 2. Figure 6 visualizes the five-task test profiles.

Intent filtering under distractor noise. In intent following, the model is presented with five candidate preferences simultaneously and must isolate the target-relevant intent while ignoring four distractors. The difference between joint conditioning and final-state injection varies across domains: joint conditioning maintains an edge in Beauty (6.38% vs. 5.55% Recall@10), while both interfaces achieve comparable test Recall@10 in Sports (2.84% vs. 2.72%) and Toys (4.60% vs. 4.84%, within sample variance). The small and domain-dependent differences do not establish a consistent advantage for either language-conditioned interface on this task.

The SID-history-only baseline reaches Intent Recall@10 of 7.73% in Beauty, 3.40% in Sports, and 7.27% in Toys, exceeding the language-conditioned interfaces despite ignoring the candidate preferences. This illustrates the strength of behavioral-history priors in this evaluation and limits its interpretation as an isolated test of intent filtering.

Fine and Coarse steering diagnostics. Language-conditioned interfaces outperform the SID-history-only baseline on these transformed tasks, but their high absolute scores are amplified by synthetic artifacts. Fine steering has substantial prompt–target lexical overlap, while Coarse steering concentrates targets on a small set of items. Appendix A reports the lexical-retrieval and target-prior audits. These dimensions provide supplementary diagnostics and are excluded from the validation selection score; Preference Recommendation and Sentiment Following remain the primary evaluation tasks.

Table 10: **Multi-domain steering performance across conditioning interfaces.** Results are reported as Val/Test (mean±std over seeds 42–44) across Amazon Beauty, Sports, and Toys. Bold indicates the best performance per split. A dash marks unavailable final-step validation aggregates. *SID-history only* denotes the preference-blind baseline.

(a) Retrieve-target dimensions *each cell: Val/Test (%)*

Domain	Interface	Metric	Pref. rec.	Fine	Coarse	Intent
Beauty	SID-history only	Recall@5	6.91±0.15/5.19±0.09	2.56±0.01/1.77±0.13	0.05±0.02/0.02±0.01	**6.89±0.14/5.17±0.09**
		Recall@10	9.92±0.18/7.76±0.10	4.07±0.07/3.14±0.11	0.10±0.02/0.05±0.03	**9.92±0.18/7.73±0.08**
		nDCG@5	4.65±0.10/3.53±0.07	1.61±0.01/1.11±0.06	0.03±0.01/0.02±0.01	**4.63±0.10/3.51±0.07**
		nDCG@10	5.62±0.12/4.36±0.04	2.09±0.02/1.55±0.05	0.04±0.01/0.02±0.01	**5.61±0.11/4.34±0.05**
	Joint conditioning	Recall@5	**7.83±0.03/6.15±0.07**	**25.26±0.34/25.10±0.14**	**38.64±1.48/35.81±2.57**	5.19±0.23/3.94±0.13
		Recall@10	**11.60±0.10/9.43±0.09**	**35.33±0.39/35.07±0.41**	**45.38±1.24/41.73±2.46**	8.14±0.27/6.38±0.15
		nDCG@5	**5.11±0.03/4.01±0.03**	**17.04±0.17/17.06±0.15**	**31.96±1.94/29.96±2.55**	3.30±0.12/2.49±0.09
		nDCG@10	**6.33±0.06/5.06±0.04**	**20.29±0.12/20.26±0.21**	**34.14±1.84/31.87±2.34**	4.25±0.12/3.28±0.07
	Cross-attention only	Recall@5	7.29±0.16/5.46±0.17	10.22±0.21/9.50±0.10	–/23.47±2.19	5.27±0.36/3.80±0.43
		Recall@10	10.90±0.23/8.34±0.07	16.02±0.32/15.21±0.32	–/26.82±1.82	7.78±0.60/5.89±0.52
		nDCG@5	4.81±0.11/3.60±0.08	6.53±0.08/6.04±0.05	–/20.46±2.43	3.47±0.20/2.49±0.25
		nDCG@10	5.97±0.13/4.52±0.06	8.39±0.11/7.87±0.11	–/21.53±2.29	4.28±0.29/3.16±0.27
	Final-state injection	Recall@5	7.69±0.11/5.75±0.12	18.03±0.42/17.54±0.70	29.90±2.08/28.11±2.40	4.91±0.49/3.51±0.37
		Recall@10	11.22±0.13/8.73±0.18	26.88±0.35/26.26±0.70	34.86±1.98/32.76±2.52	7.26±0.71/5.55±0.68
		nDCG@5	5.10±0.11/3.85±0.05	11.88±0.31/11.48±0.43	25.98±2.33/24.55±2.51	3.26±0.36/2.31±0.26
		nDCG@10	6.23±0.11/4.80±0.03	14.73±0.29/14.29±0.44	27.58±2.30/26.05±2.55	4.02±0.42/2.96±0.36
Sports	SID-history only	Recall@5	2.99±0.59/2.22±0.59	1.01±0.24/0.82±0.28	0.01±0.01/0.01±0.01	**2.99±0.59/2.23±0.58**
		Recall@10	4.43±1.03/3.38±0.86	1.65±0.43/1.36±0.42	0.02±0.00/0.02±0.01	**4.44±1.03/3.40±0.85**
		nDCG@5	2.03±0.39/1.47±0.40	0.65±0.15/0.51±0.16	0.01±0.00/0.00±0.00	**2.03±0.39/1.47±0.39**
		nDCG@10	2.49±0.53/1.84±0.49	0.85±0.21/0.68±0.20	0.01±0.00/0.01±0.00	**2.49±0.53/1.85±0.48**
	Joint conditioning	Recall@5	**4.25±0.10/3.41±0.06**	**33.84±0.34/33.42±0.32**	**33.89±1.22/33.01±2.58**	2.34±0.33/1.82±0.25
		Recall@10	**6.52±0.07/5.40±0.08**	**44.52±0.27/44.08±0.34**	**42.71±1.00/41.75±1.43**	3.65±0.53/2.84±0.37
		nDCG@5	**2.78±0.05/2.19±0.05**	**23.59±0.18/23.27±0.13**	**23.83±0.97/22.85±1.56**	1.48±0.21/1.15±0.16
		nDCG@10	**3.51±0.03/2.83±0.02**	**27.04±0.15/26.73±0.17**	**26.67±0.69/25.67±1.12**	1.91±0.27/1.47±0.20
	Cross-attention only	Recall@5	4.12±0.08/3.13±0.06	11.16±0.14/11.05±0.07	22.56±0.43/19.92±1.91	2.81±0.43/2.06±0.33
		Recall@10	6.23±0.14/4.91±0.02	17.73±0.21/17.59±0.08	28.70±1.54/26.19±3.82	4.16±0.73/3.17±0.61
		nDCG@5	2.72±0.05/2.04±0.03	7.09±0.12/6.97±0.06	16.74±0.42/14.09±1.22	1.84±0.27/1.33±0.21
		nDCG@10	3.40±0.05/2.60±0.02	9.20±0.13/9.08±0.05	18.72±0.60/16.12±1.82	2.28±0.37/1.69±0.30
	Final-state injection	Recall@5	4.18±0.12/3.17±0.06	25.27±0.93/25.08±0.69	32.84±0.19/30.91±0.80	2.39±0.13/1.75±0.05
		Recall@10	6.42±0.14/5.06±0.16	35.99±0.83/35.80±0.69	40.84±0.39/39.20±1.16	3.58±0.17/2.72±0.06
		nDCG@5	2.76±0.05/2.05±0.03	16.90±0.64/16.78±0.51	23.72±0.42/21.78±0.24	1.54±0.09/1.13±0.02
		nDCG@10	3.48±0.06/2.66±0.06	20.36±0.60/20.24±0.52	26.31±0.43/24.46±0.31	1.93±0.10/1.45±0.03
Toys	SID-history only	Recall@5	6.01±0.30/4.77±0.28	3.10±0.21/2.77±0.13	0.04±0.01/0.04±0.01	**5.99±0.27/4.78±0.28**
		Recall@10	8.78±0.29/7.26±0.16	5.04±0.28/4.54±0.20	0.08±0.02/0.09±0.03	**8.75±0.28/7.27±0.14**
		nDCG@5	**4.03±0.25/3.23±0.15**	1.96±0.20/1.68±0.05	0.03±0.01/0.03±0.01	**4.01±0.24/3.24±0.14**
		nDCG@10	4.92±0.25/4.03±0.11	2.59±0.21/2.25±0.07	0.04±0.01/0.04±0.01	**4.89±0.25/4.04±0.10**
	Joint conditioning	Recall@5	**6.17±0.31/4.95±0.36**	**22.87±0.44/23.03±0.89**	21.85±2.38/**27.00±2.57**	3.83±0.79/2.87±0.84
		Recall@10	**9.51±0.38/7.79±0.37**	**33.12±0.45/33.16±0.86**	28.93±1.23/**35.06±1.49**	5.89±1.35/4.60±1.19
		nDCG@5	4.01±0.15/3.20±0.23	**15.32±0.26/15.35±0.61**	14.38±2.09/17.87±1.99	2.39±0.51/1.84±0.58
		nDCG@10	**5.08±0.17/4.11±0.23**	**18.63±0.26/18.62±0.60**	16.70±1.56/20.50±1.31	3.05±0.69/2.39±0.70
	Cross-attention only	Recall@5	5.78±0.08/3.92±0.13	9.11±0.54/8.19±0.22	14.59±2.89/17.41±1.77	4.92±0.31/3.14±0.19
		Recall@10	8.75±0.12/6.34±0.14	14.28±0.78/13.28±0.65	17.59±3.61/20.40±2.11	7.33±0.45/5.06±0.38
		nDCG@5	3.77±0.07/2.53±0.12	5.80±0.27/5.20±0.20	11.23±2.93/13.64±2.02	3.21±0.16/2.05±0.11
		nDCG@10	4.73±0.07/3.31±0.12	7.46±0.34/6.84±0.34	12.19±3.17/14.59±2.11	3.98±0.21/2.67±0.17
	Final-state injection	Recall@5	6.00±0.12/4.41±0.06	20.26±0.42/19.95±0.16	**23.67±1.32**/26.96±1.27	4.38±0.15/3.02±0.25
		Recall@10	9.02±0.09/7.00±0.19	29.98±0.42/29.63±0.21	**30.11±1.21**/33.48±0.93	6.73±0.25/4.84±0.25
		nDCG@5	3.90±0.08/2.81±0.05	13.33±0.33/13.03±0.02	**15.94±0.81/18.67±0.69**	2.85±0.13/1.96±0.15
		nDCG@10	4.88±0.03/3.65±0.09	16.46±0.31/16.15±0.02	**18.03±0.73/20.78±0.55**	3.61±0.16/2.55±0.15

(b) Sentiment following *each cell: Val/Test @10 (%)*

Domain	Interface	Find Recall ↑	Avoid Hit@10 ↓	Sentiment ↑
Beauty	SID-history only	7.79±0.19/5.87±0.18	7.79±0.19/5.87±0.18	0.00±0.00/0.00±0.00
	Joint conditioning	**9.60±0.32/7.49±0.14**	**1.14±0.07/1.00±0.17**	**9.04±0.35/7.01±0.16**
	Cross-attention only	8.51±0.39/6.05±0.17	3.01±0.58/2.37±0.26	6.49±0.32/4.63±0.27
	Final-state injection	9.24±0.32/6.50±0.14	1.91±0.28/1.43±0.21	7.87±0.15/5.59±0.12
Sports	SID-history only	3.59±0.69/2.58±0.65	3.59±0.69/2.58±0.65	0.00±0.00/0.00±0.00
	Joint conditioning	**9.51±0.10/7.20±0.25**	1.92±0.16/1.61±0.13	**8.20±0.13/6.15±0.27**
	Cross-attention only	7.06±0.01/5.23±0.00	**1.72±0.29/1.29±0.29**	5.99±0.28/4.45±0.15
	Final-state injection	8.59±0.17/6.73±0.13	1.84±0.34/1.56±0.03	7.46±0.23/5.65±0.06
Toys	SID-history only	5.81±0.37/4.72±0.20	5.81±0.37/4.72±0.20	0.00±0.00/0.00±0.00
	Joint conditioning	**7.50±0.65/5.88±0.22**	**1.15±0.09/0.79±0.19**	**6.76±0.63/5.33±0.18**
	Cross-attention only	6.48±0.49/3.94±0.06	2.03±0.24/1.40±0.13	5.04±0.58/3.13±0.09
	Final-state injection	7.01±0.16/5.08±0.28	2.43±0.33/1.54±0.25	5.31±0.32/3.94±0.10

E.5 Capability Retention, Checkpoint Correspondence, and Serialization Controls

Evaluation protocol. We evaluate 75 joint-conditioning configurations on Beauty across ARC Challenge (ARC-C), BIG-Bench Hard (BBH), BoolQ, GSM8K, HellaSwag, MMLU, MMLU-Pro,

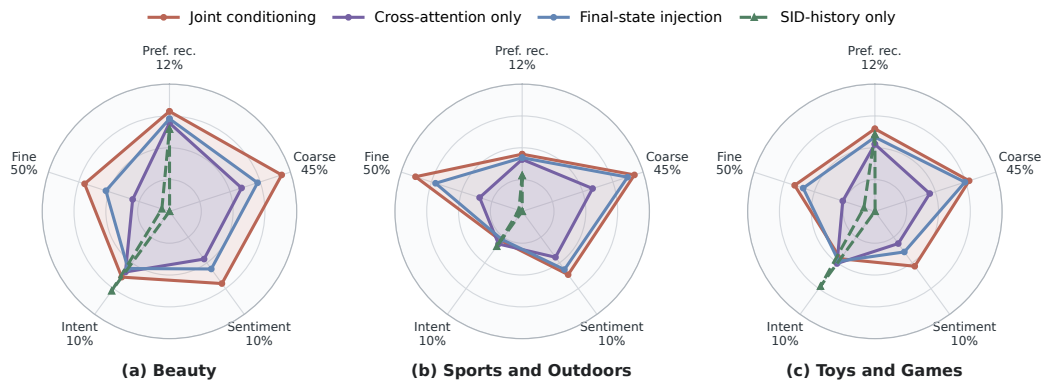


Figure 6: **Multi-domain steering profiles.** Radar charts of test-set performance across five dimensions on Beauty, Sports, and Toys (mean of 3 seeds; Table 10, higher is better; spokes share fixed scale caps). Fine and Coarse scores are diagnostic due to lexical overlap and target concentration, rather than stronger semantic steering.

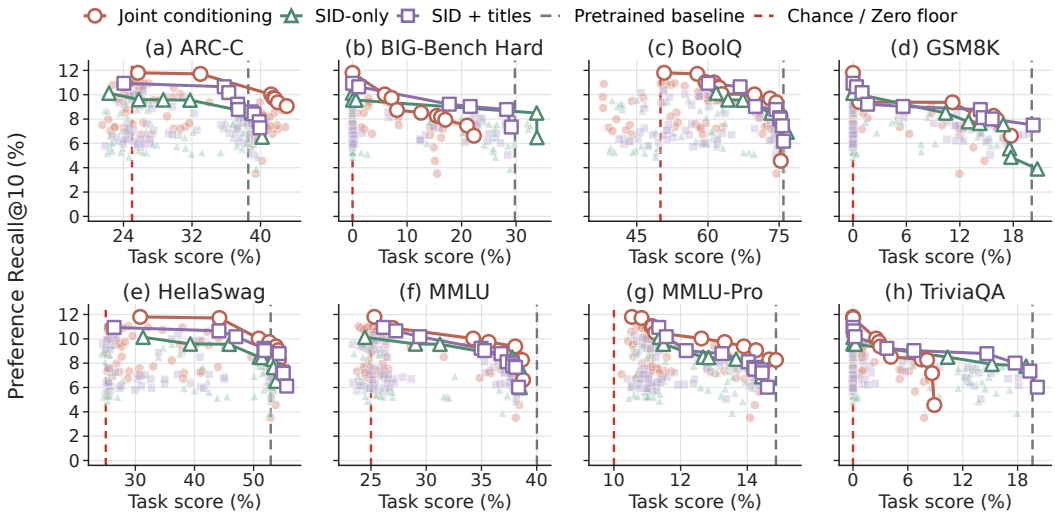


Figure 7: **Retrieval–capability Pareto frontiers.** Trade-off pairing validation Preference Recall@10 on Beauty (y-axis) with accuracy across eight language tasks (x-axis). Frontiers are computed separately for each task and regime; connecting segments do not represent training trajectories; dashed lines mark pretrained baseline (gray) and task floor (red). Top-right is optimal.

and TriviaQA. Each language score is paired with the preference Recall@10 on the validation split. We also evaluate two preference-blind sweeps of 75 configurations each: SID-history only and SID history augmented with chronological item titles. The three regimes share the Gemma 3 1B IT backbone, but their comparison is descriptive rather than a matched-compute superiority test.

Correspondence with the main retrieval results. The "Joint (selected)" model in Table 3 uses a peak learning rate of 0.005, batch size 16, and 30,000 training steps, chosen by the validation rule in Appendix D.2, with 11.71% validation preference Recall@10 and the language scores in Table 3. Table 2 reports three retrainings across seeds 42–44, yielding $11.60 \pm 0.10\%$ validation and $9.43 \pm 0.09\%$ test Recall@10. Language evaluations cover only the original sweep: the correspondence is at the configuration level, without a three-seed estimate of capability loss. Figure 7 likewise uses sweep validation scores, not retrained test results.

Pareto construction and task-dependent retention. For each language task and training regime separately, a configuration is nondominated if no other observed configuration matches or exceeds both its language score and retrieval Recall@10, with a strict improvement in at least one. Connecting segments order these finite observed points; they are neither training trajectories nor claims about attainable interpolations. Retention is the post-training task score divided by the corresponding pretrained Gemma 3 1B IT score, and is not clipped at 100%.

The joint-conditioning sweep reaches its highest validation preference Recall@10 (11.80%) at peak learning rate 0.01, batch size 16, and 30,000 steps, where BBH, GSM8K, and TriviaQA drop to zero and several multiple-choice scores approach their task floors. The validation-selected joint model (learning rate 0.005, batch size 16, 30,000 steps) likewise exhibits severe open-ended capability loss. The "Joint (higher retention)" model in Table 3 corresponds to the configuration trained with learning rate 0.0005, batch size 48, and 30,000 steps, which reaches 6.63% Recall@10 and retains 96.88–104.43% of pretrained scores on ARC-C, BoolQ, HellaSwag, MMLU, and MMLU-Pro (rounded). Its BBH, GSM8K, and TriviaQA retention is 74.75%, 88.34%, and 42.18%, respectively. This illustrates partial retention rather than preservation of all capabilities, and no single operating point should be inferred by combining the best values across different task frontiers.

Item serialization and preference-blind controls. The SID-history-only sweep reaches up to 10.13% validation preference Recall@10, while augmenting history with item titles reaches 10.94%. Both controls exhibit retrieval–capability trade-offs despite receiving no explicit preference, showing that capability loss can accompany SID specialization without preference conditioning. Joint conditioning attains the highest observed retrieval score (11.80%) among these sweeps. These comparisons do not establish that SID prediction or preference input is the sole cause of capability loss, nor that the frozen language tower's retained capabilities transfer unchanged to its trainable SID generator.