LEARNING TO NAVIGATE SEMANTIC IDS WITH JOINT-PREFIX BEHAVIORAL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic identifiers (SIDs) turn generative recommendation into an irreversible path-search problem: once an early prefix is pruned, the target item becomes unreachable. We identify a training–search mismatch in SID retrieval: behavioral L2 supervision is normalized separately within each L1 prefix (parent branch), whereas beam search compares cumulative scores across L1L2 prefixes rooted at different L1 codes. We propose Joint-Prefix Behavioral-Neighbor Distillation (JP-BND), which projects training-partition behavioral evidence onto joint L1L2 prefixes and distills it over the joint prefixes that compete under the current student, without changing the SID mapping or inference procedure. A joint-KL decomposition separates within-parent conditional alignment from cross-parent marginal alignment. On Delayed development, Full-SID Hit@500 increases from 3.82% for the Hard-example BND (H-BND) baseline to 8.60% with a matched multi-parent conditional control and 10.66% with core JP-BND. Fixed-candidate KL diagnostics show that conditional training primarily reduces within-parent mismatch, whereas joint normalization additionally reduces cross-parent marginal mismatch. A 5,000-instance constrained beam-search audit further reveals selective cross-parent reordering rather than a uniform lift in target-prefix margins. The gains remain stable across three training seeds and a separate Delayed held-out evaluation, persist at every tested inference beam width from 20 to 500, and transfer at the SID-path level to Amazon Books and Home public datasets. Together, these results show that aligning behavioral supervision with the prefixes that actually compete during beam search improves generative retrieval over fixed SID spaces.

1 INTRODUCTION

Generative recommenders (Rajput et al., 2023; Deng et al., 2025) retrieve items by autoregressively generating semantic identifiers (SIDs). Each SID defines a path through a hierarchical search space, where beam-search pruning irreversibly removes parts of the catalog (Wang et al., 2026b). A model may therefore assign high probability to the remaining tokens of a target SID once the correct prefix is given, yet fail to keep that prefix alive during constrained decoding. We study how to improve this navigation while keeping the item-to-SID mapping, legal trie, model architecture, and inference procedure fixed. The training–search mismatch is structural and becomes especially consequential when the model must be evaluated on behavior emerging after the training window while the SID hierarchy remains unchanged.

The central difficulty is a mismatch between how prefix decisions are supervised during training and how they compete during beam decoding. At L2, second-level code distributions are supervised separately within each first-level prefix (parent branch), whereas beam search compares cumulative path scores across prefixes rooted at different first-level codes. A second-level code can be locally preferred within its parent and still be pruned because the resulting joint prefix has insufficient cumulative probability. A surviving prefix may also lose all matching descendants before complete paths are returned. Thus, within-parent prediction quality does not by itself characterize multi-parent competition during decoding. We distinguish *immediate prefix survival* from *final-path coverage*: the former identifies whether the target prefix survives a given pruning step, whereas the latter measures whether it remains represented among returned complete paths.

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.

Behavioral-Neighbor Distillation (BND) projects a training-partition history-to-target affinity graph onto the SID hierarchy to supervise L1 and within-parent L2 distributions. The graph provides behavioral alternatives across different parents, but BND still applies its L2 supervision separately within each parent. We introduce **Joint-Prefix BND (JP-BND)**, which projects the same behavioral evidence onto joint L1L2 prefixes and distills it over the prefixes competing under the current student. Behavior-supported prefixes are therefore trained against cross-parent competitors rather than only siblings within one parent. We focus joint distillation on L1L2 because this is the first depth at which children from multiple surviving parents compete through cumulative path scores, while the candidate set induced by the current student remains tractable for joint normalization during training. Extending the same joint normalization to deeper SID levels is left for future work.

A joint-KL decomposition separates within-parent conditional alignment from cross-parent marginal alignment. Joint-score gradients can change aggregate parent mass, whereas a within-parent KL has zero-sum score gradients within its normalization domain. This motivates a controlled question: *Does joint normalization help beyond exposing the same behavioral supervision across multiple parents?* We answer it with matched conditional and joint objectives, then separate distributional from search effects by having all models rescore the same frozen set of legal L1L2 prefixes and by tracing prefix survival under constrained beam search.

Empirically, on a separate Delayed held-out evaluation, core JP-BND raises mean Full-SID Hit@500 from 4.57% for the H-BND baseline to 11.74% across three training seeds. On Delayed development, the matched multi-parent conditional objective reaches 8.60%, while core JP-BND reaches 10.66%. Distributional and beam-search diagnostics associate this additional gain with reduced cross-parent marginal mismatch and selective cross-parent reordering rather than a uniform lift in target-prefix margins. The gain persists across inference beam widths from 20 to 500 and transfers at the SID-path level to Amazon Books and Home (Hou et al., 2024) public datasets.

Our contributions are:

- We identify a structural training–search mismatch in SID generative recommendation between within-parent supervision and multi-parent competition during constrained decoding.
- We propose JP-BND, which distills graph-derived behavioral evidence over competing joint prefixes, and derive a joint-KL decomposition that separates cross-parent marginal alignment from within-parent conditional alignment.
- We show stable gains across three seeds and Delayed held-out evaluation, isolate joint normalization with matched controls, connect it to selective cross-parent reordering through beam-search audits, and demonstrate SID-path transfer to Amazon Books and Home.

2 Related Work

Semantic identifiers and behavioral supervision. Semantic identifiers represent items with hierarchical discrete codes and enable autoregressive generative retrieval (Lee et al., 2022; Rajput et al., 2023; Deng et al., 2025). Subsequent work incorporates collaborative information either into identifier construction or into training supervision. LC-Rec and LETTER (Zheng et al., 2024; Wang et al., 2024a), for example, learn collaborative item representations or tokenizers that alter the resulting SID geometry, while TCA4Rec (Lin et al., 2026) projects item-level collaborative scores into token-level soft supervision. Our setting is complementary: we keep the item-to-SID mapping and legal hierarchy fixed and instead use behavioral evidence to supervise competition among joint L1L2 prefixes across different parents.

Search-aware training and reachability. Beam-search optimization (Wiseman & Rush, 2016) and prefix-oriented ranking (Zeng et al., 2023) directly optimize prefix-level search behavior. RIPOR's POR component ranks positive and negative prefixes across multiple depths using teacher margins, while its full pipeline additionally includes its own identifier construction and staged negative mining. SmartGR (Zhang et al., 2026) combines hierarchical distillation with beam-aware prefix ranking using teacher-generated candidates and student scores. BEAR (Yang et al., 2026) instead regularizes teacher-forced target paths using the current student's local top-B legal-token threshold, providing a beam-survival signal without changing inference. DASO (Yu et al., 2026) addresses difficult prefixes in a post-SFT on-policy reinforcement-learning stage through guided rollout allocation, while ISD (Wang et al., 2026a) introduces external item evidence during decoding.

JP-BND differs in both its supervision source and probability domain. It derives a multi-target teacher from behavioral relations and distills that distribution over the current student's competing joint prefixes, including competitors with zero teacher mass. The method preserves the fixed SID mapping, legal trie, model architecture, and constrained decoder, and requires no external support at inference.

3 METHODOLOGY

3.1 PROBLEM SETUP AND TRAINING–SEARCH MISMATCH

Let i index an instance with history H_i and target item $v_i^\star \in \mathcal{I}$. The fixed, possibly non-injective mapping $g : \mathcal{I} \to \mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{C}_3$ assigns each item a three-level SID:

$$g(v) = (a_v, b_v, c_v), \qquad a_v \in \mathcal{C}_1,\ b_v \in \mathcal{C}_2,\ c_v \in \mathcal{C}_3. \tag{1}$$

The target SID is $(a_i^\star, b_i^\star, c_i^\star) = g(v_i^\star)$. The model generates it from H_i under a legal trie and is trained with the standard autoregressive negative log-likelihood $\mathcal{L}_{\mathrm{AR}}$ (Appendix A.2).

Let $p_\theta(\cdot \mid \cdot)$ denote the model's next-token distribution. For candidate tokens $a \in \mathcal{C}_1$ and $b \in \mathcal{C}_2$, define the cumulative log-probability score of prefix (a, b) as

$$S_\theta(a, b \mid H_i) = \log p_\theta(a \mid H_i) + \log p_\theta(b \mid H_i, a). \tag{2}$$

At decoding depth d, let $\mathcal{A}_{i,d}^\theta$ denote the legal expansions of the surviving depth-$(d-1)$ frontier. The decoder retains up to B prefixes with highest cumulative scores under its deterministic tie-breaking rule. Let $\tau_{i,d,B}$ denote the score of the last retained prefix, and let $F_{i,d}$ indicate whether the target depth-d prefix belongs to this retained set.

Local supervision versus multi-parent competition. Local L2 supervision optimizes the conditional distribution $p_\theta(b \mid H_i, a)$ within a fixed parent a, whereas beam search compares cumulative joint-prefix scores $S_\theta(a, b \mid H_i)$ across parents. A locally preferred child can still fall below the global beam cutoff. For example, the illustrative paths in Figure 1 have probabilities $0.20 \times 0.80 = 0.16$ and $0.50 \times 0.40 = 0.20$. Although $0.80 > 0.40$, the corresponding joint-prefix score is lower. Local L2 preference alone therefore does not determine survival under multi-parent beam competition.

Immediate survival versus final-path coverage. Let $C_{i,d}$ indicate target-prefix coverage among final decoded paths. L1, L1L2, and Full-SID Hit@500 measure $C_{i,d}$ at depths 1, 2, and 3. Since $C_{i,d} \le F_{i,d}$, final-path coverage alone cannot identify at which depth the target prefix is first pruned. We therefore use constrained beam-search traces for localization.

3.2 BEHAVIORAL JOINT-PREFIX DISTILLATION

As illustrated in Figure 1, we first construct behavior-derived prefix targets and then restrict them to the current student's candidate set.

3.2.1 FROM WITHIN-PARENT TO JOINT BEHAVIORAL TEACHERS

We construct a directed history-to-target affinity graph from training-partition interactions. An edge $(x, v) \in \mathcal{I} \times \mathcal{I}$ links a historical source item $x \in \mathcal{I}$ to a later target item $v \in \mathcal{I}$. We retain edges supported across two disjoint user partitions and apply degree correction to reduce the influence of frequent sources and destinations. Let $\alpha(x, v)$ denote the resulting edge weight. Appendix A.1 gives the graph construction and dataset statistics.

For instance i, we first aggregate graph evidence from its history H_i to each candidate item $v \in \mathcal{I}$:

$$w_i(v) = \sum_{x \in H_i} \alpha(x, v). \tag{3}$$

Thus, $w_i(v)$ measures the behavioral affinity between the observed history and candidate item v.

The behavioral graph provides item-level affinity, whereas the model predicts SID tokens. We therefore convert item-level affinity into token-level supervision by aggregating $w_i(v)$ according to the fixed SID mapping $g(v)$. This changes the behavioral training targets without changing the

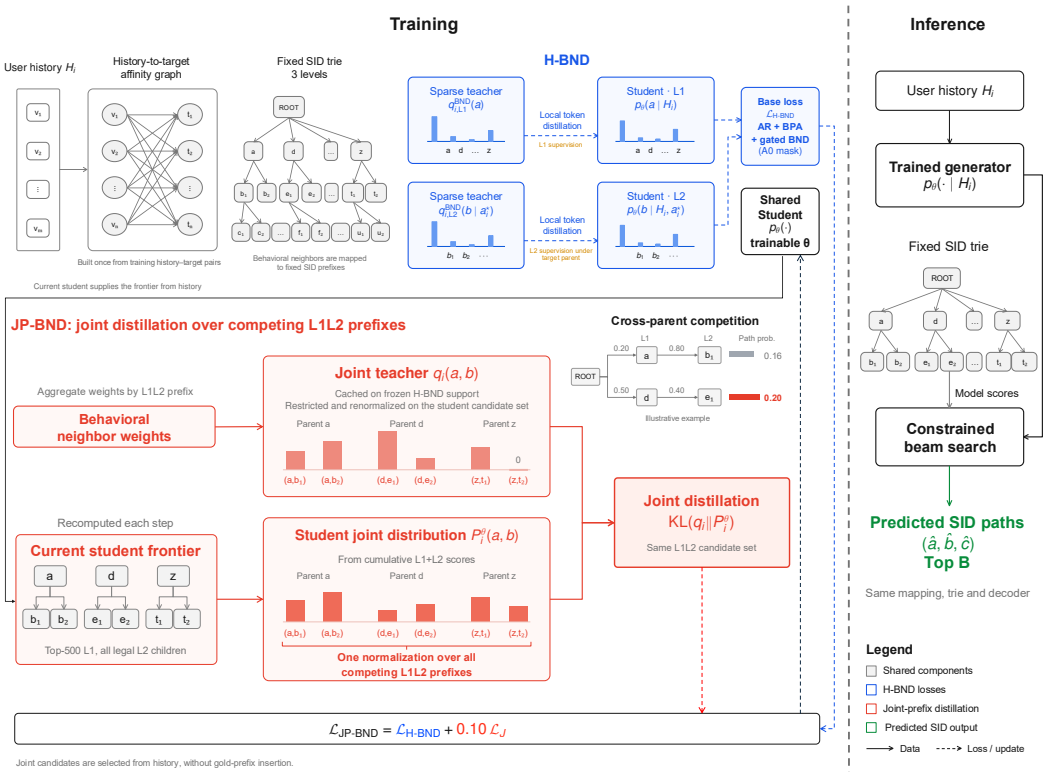

Figure 1: **From within-parent to joint-prefix behavioral supervision.** H-BND supervises L1 and target-parent L2. JP-BND distills behavior-derived supervision over the student's L1L2 candidates, using cached support from frozen H-BND. Teachers and references are discarded at inference.

identifiers themselves. Specifically, Behavioral-Neighbor Distillation (BND) forms an L1 teacher over parents and an L2 teacher over children within the observed target parent $a_i^\star$:

$$q^{\mathrm{BND}}_{i,\mathrm{L1}}(a) = \frac{\sum_{v:a_v=a} w_i(v)}{\sum_v w_i(v)}, \qquad q^{\mathrm{BND}}_{i,\mathrm{L2}}(b \mid a_i^\star) = \frac{\sum_{v:(a_v,b_v)=(a_i^\star,b)} w_i(v)}{\sum_{v:a_v=a_i^\star} w_i(v)}. \tag{4}$$

A teacher is active only when its denominator is positive. L1 mass spans parents, but L2 supervision remains within the target parent.

Let $\mathcal{B}$ be a training batch, z_i a binary BND exposure mask, $s_{i,d}$ indicate a nonempty teacher, and $p_{\theta,i,d}$ denote the student distribution at L1 or target-parent L2. BND loss minimizes the KL divergence between the behavioral teachers in Equation 4 and their corresponding student distributions:

$$\mathcal{L}_{\mathrm{BND}}(z) = \mathrm{Avg}_{i\in\mathcal{B},\, d:\, z_i s_{i,d}=1} \mathrm{KL}\big(q^{\mathrm{BND}}_{i,d} \| p_{\theta,i,d}\big), \qquad d \in \{\mathrm{L1}, \mathrm{L2}\}. \tag{5}$$

The average is over active instance–depth pairs and is zero when no pair is active. Behavioral Prefix Alignment (BPA) is the graph-free prefix-alignment loss. Hard-example BND (H-BND) sets $z_i^H = 1$ only when target L1 falls outside the legal top-B L1 frontier of a frozen AR-only reference, giving

$$\mathcal{L}_{\text{H-BND}} = \mathcal{L}_{\mathrm{AR}} + \lambda_{\mathrm{BPA}}\mathcal{L}_{\mathrm{BPA}} + \lambda_{\mathrm{BND}}\mathcal{L}_{\mathrm{BND}}(z^H). \tag{6}$$

Details on BPA, loss reductions, and hard-example mask construction are provided in Appendix A.2.

We next construct the joint-prefix teacher. A separately trained, frozen H-BND model provides the L1 reference frontier $\mathcal{F}_i^H$ for the joint teacher. It is distinct from the AR-only reference used for the mask. Joint behavioral mass aggregates items sharing an L1L2 prefix:

$$W_i(a,b) = \sum_{v:(a_v,b_v)=(a,b)} w_i(v). \tag{7}$$

We then retain behavior-supported legal prefixes whose parents lie in $\mathcal{F}_i^H$:

$$\mathcal{S}_i^H = \big\{(a,b) : a \in \mathcal{F}_i^H,\ (a,b) \text{ is legal},\ W_i(a,b) > 0\big\}. \tag{8}$$

For nonempty $\mathcal{S}_i^H$, the cached joint teacher on this set is

$$q_i^{\mathrm{JP}}(a,b) = \frac{W_i(a,b)}{\sum_{(a',b')\in\mathcal{S}_i^H} W_i(a',b')}. \tag{9}$$

We set $q_i^{\mathrm{JP}}(a,b) = 0$ outside $\mathcal{S}_i^H$, including when $\mathcal{S}_i^H$ is empty. Notably, $\mathcal{S}_i^H$ depends only on the history, training-partition behavior graph, frozen H-BND frontier, and fixed SID mapping, without inserting the current target prefix.

3.2.2 Distillation over Competing Prefixes

The teacher is constructed from the frozen reference frontier $\mathcal{F}_i^H$, whereas the current student recomputes its own legal top-B L1 frontier $\mathcal{F}_i^\theta$ at each training step. The student itself then expands all legal L2 children under these parents:

$$\Omega_i^\theta = \left\{(a,b) : a \in \mathcal{F}_i^\theta,\ b \in \mathcal{C}_2,\ (a,b) \text{ is legal}\right\}. \tag{10}$$

Training fixes $B = 500$, matching the primary decoding protocol. Beam-width evaluations change only inference B, without retraining. The cumulative scores in Equation 2 are normalized over this candidate set to form the student's joint distribution over competing L1L2 prefixes:

$$P_i^\theta(a,b) = \frac{\exp S_\theta(a,b \mid H_i)}{\sum_{(a',b')\in\Omega_i^\theta} \exp S_\theta(a',b' \mid H_i)}. \tag{11}$$

We restrict the cached teacher to the current candidate set and define its retained mass as

$$Z_i = \sum_{(a,b)\in\Omega_i^\theta} q_i^{\mathrm{JP}}(a,b). \tag{12}$$

For instances with $Z_i > 0$, the restricted teacher is

$$q_i(a,b) = \frac{q_i^{\mathrm{JP}}(a,b)}{Z_i}, \qquad (a,b) \in \Omega_i^\theta. \tag{13}$$

We do not augment Ω_i^θ with the target L1 or additional parents from the frozen reference frontier. All legal prefixes in Ω_i^θ remain in the normalization of P_i^θ, including those with zero teacher mass. Thus, behavior-supported prefixes are trained against their actual competitors in Ω_i^θ. Support statistics are provided in Appendix A.3, and the training procedure and access constraints in Appendix A.4.

With the teacher and student distributions defined on the same candidate set, Joint-Prefix BND (JP-BND) minimizes the joint-prefix distillation objective

$$\mathcal{L}_J = \frac{1}{|\mathcal{B}_J|} \sum_{i\in\mathcal{B}_J} \mathrm{KL}\big(q_i \| P_i^\theta\big), \qquad \mathcal{B}_J = \{i \in \mathcal{B} : Z_i > 0\}, \tag{14}$$

where $\mathcal{B}_J$ contains instances with positive retained teacher mass, and both q_i and P_i^θ are distributions over Ω_i^θ. We set $\mathcal{L}_J = 0$ when $\mathcal{B}_J$ is empty.

3.2.3 Decomposition of Joint-Prefix Supervision

To characterize the effect of joint normalization, we introduce a matched multi-parent conditional control. H-BND supervises L2 only within the target parent. The conditional control supervises all teacher-supported parents, normalizing L2 separately within each parent and weighting by the teacher marginal. It therefore aligns within-parent conditionals, whereas JP-BND additionally aligns probability mass across parents.

Let $q_{i,\mathrm{L1}}$ and $P_{i,\mathrm{L1}}^\theta$ be the parent marginals,

$$q_{i,\mathrm{L1}}(a) = \sum_b q_i(a,b), \qquad P_{i,\mathrm{L1}}^\theta(a) = \sum_b P_i^\theta(a,b), \tag{15}$$

where sums range over children in Ω_i^θ. On positive teacher-mass parents, $q_i(b \mid a) = q_i(a,b)/q_{i,\mathrm{L1}}(a)$ and $P_i^\theta(b \mid a) = P_i^\theta(a,b)/P_{i,\mathrm{L1}}^\theta(a)$. The matched multi-parent conditional objective is

$$\mathcal{L}_{\mathrm{cond}} = \frac{1}{|\mathcal{B}_J|} \sum_{i\in\mathcal{B}_J} \sum_{a:q_{i,\mathrm{L1}}(a)>0} q_{i,\mathrm{L1}}(a)\, \mathrm{KL}\big(q_i(\cdot \mid a) \| P_i^\theta(\cdot \mid a)\big). \tag{16}$$

Proposition 1 (joint-prefix decomposition). For any instance $i \in \mathcal{B}_J$ with fixed candidate set Ω_i^θ,

$$\mathrm{KL}(q_i \| P_i^\theta) = \mathrm{KL}(q_{i,\mathrm{L1}} \| P_{i,\mathrm{L1}}^\theta) + \sum_{a: q_{i,\mathrm{L1}}(a) > 0} q_{i,\mathrm{L1}}(a)\, \mathrm{KL}\big(q_i(\cdot \mid a) \| P_i^\theta(\cdot \mid a)\big)\,. \tag{17}$$

Therefore, under the same active-instance set and reduction,

$$\boxed{\mathcal{L}_J = \mathcal{L}_{\mathrm{cond}} + \mathcal{L}_{\mathrm{marg}}} \qquad \text{with} \qquad \mathcal{L}_{\mathrm{marg}} = \frac{1}{|\mathcal{B}_J|} \sum_{i \in \mathcal{B}_J} \mathrm{KL}(q_{i,\mathrm{L1}} \| P_{i,\mathrm{L1}}^\theta). \tag{18}$$

Thus $\mathcal{L}_{\mathrm{cond}}$ aligns child distributions within each surviving parent, while $\mathcal{L}_J$ additionally aligns the parent-marginal distribution across competing L1 parents. This decomposition isolates cross-parent marginal alignment as the additional component in $\mathcal{L}_J$. We therefore construct a matched intervention that replaces $\mathcal{L}_J$ with $\mathcal{L}_{\mathrm{cond}}$ while preserving the offline behavioral support, candidate-generation rule, initialization, and training budget. The independently trained models may induce different student frontiers as their parameters diverge. Both $\mathcal{L}_{\mathrm{cond}}$, and $\mathcal{L}_{\mathrm{marg}}$ are set to zero when $\mathcal{B}_J$ is empty. Appendix B.1 gives the complete derivation and conditions.

At the score-gradient level, the distinction is explicit. For fixed Ω_i^θ, the common root factor $p_\theta(a \mid H_i)$ cancels within each conditional normalization, so $\mathcal{L}_{\mathrm{cond}}$ has no direct derivative with respect to the root score. By contrast, $\mathcal{L}_J$ contributes the aggregate parent residual $P_{i,\mathrm{L1}}^\theta(a) - q_{i,\mathrm{L1}}(a)$ (Appendix B.2). This does not mean that conditional training fixes root probabilities, since shared parameters can still change them indirectly. The parent marginal aligned by $\mathcal{L}_J$ is induced by retained legal L2 mass and need not equal the bare L1 distribution (Appendix B.3).

3.2.4 Core Objective and Deployment

In the temporal study, core JP-BND adds joint-prefix distillation to the H-BND objective:

$$\boxed{\mathcal{L}_{\text{JP-BND}} = \mathcal{L}_{\text{H-BND}} + \lambda_J \mathcal{L}_J}. \tag{19}$$

An optional safety-augmented variant adds $\lambda_B \mathcal{L}_B$ to target the beam boundary and $\lambda_A \mathcal{L}_A$ for root stability (Appendices B.4 and B.5). Inference uses only the trained student with the original mapping, trie, decoder, and item-expansion policy. No behavioral graph or teacher cache is needed at inference.

4 Experiments

4.1 Protocol and Metrics

Temporal benchmark. We train on 900,774 instances from anonymized video interactions: 810,697 next-item instances and 90,077 content-denotation replay instances used to preserve the shared content-to-SID capability during behavioral training. Content embeddings are mapped to a fixed three-level 4096^3 ContentSID space by a greedy-margin variant of balanced RQ-KMeans (Zhou et al., 2025). Qwen3-0.6B (Yang et al., 2025) first undergoes shared SID-token alignment and content-to-SID denotation training. Our behavioral recommendation arms, referred to as Stage 3, start independently from the same resulting checkpoint and share instance identities and order, optimizer, global batch size 256, the 3,519-step budget, SID mapping, trie, and decoder (Appendices C.1–C.5). Specifically, A0 uses AR, A1 adds BPA, and A2 adds dense BND. H-BND gates only BND on the 21.55% of instances whose target L1 lies outside the frozen A0 legal top-B frontier, with $B = 500$ for mask construction. By default, $\lambda_{\mathrm{BPA}} = 0.10$, $\lambda_{\mathrm{BND}} = 0.20$, $\lambda_J = 0.10$, and $\lambda_B = \lambda_A = 0.05$.

Because deployed recommenders are refreshed periodically while user behavior evolves, we evaluate temporally separated populations rather than an IID split. *Delayed* contains 5,000 instances from a non-overlapping evaluation period three days after the training window, measuring near-term temporal generalization. *Dynamics* contains 10,000 instances drawn from a broader behavioral window, with users disjoint from the training population, providing a view of sensitivity to behavioral change. *Retention* contains 10,090 replay instances measuring the preservation of content denotation. Reporting them separately reveals gains and regressions that could be hidden by an aggregate score. A separate 10,000-instance Delayed held-out sample is evaluated only after the H-BND, conditional control ($+\mathcal{L}_{\mathrm{cond}}$), core JP-BND ($+\mathcal{L}_J$), and safety-augmented JP-BND ($+\mathcal{L}_J+\mathcal{L}_B+\mathcal{L}_A$)

Table 1: **Primary temporal comparison.** Percentages are reported on the same 5,000 Delayed development instances under $B = 500$. L1, L1L2, and Full-SID are Hit@500 metrics. Expected Item is secondary. Prior-method baselines are matched adaptations, not official reproductions.

Method	L1	L1L2	Full-SID	Exp. Item	Inference change
Base behavioral supervision					
A0: AR	15.50	2.20	1.52	0.78	None
A1: AR+BPA	16.42	2.38	1.64	0.80	None
A2: AR+BPA+BND	33.02	7.92	4.28	1.92	None
H-BND	30.86	6.48	3.82	2.14	None
Internal controls					
Constant-total early-token weighting	12.82	1.88	1.32	0.38	None
Gold-prefix ranking	30.58	6.48	3.56	2.02	None
Route-score control (AGCR-OPD)	30.44	6.14	3.78	2.24	None
Prior-method adaptations					
SmartGR-style (Zhang et al., 2026)	23.32	4.32	2.14	0.34	None
BEAR-style (Yang et al., 2026)	29.00	5.78	3.66	1.98	None
ISD-style (Wang et al., 2026a)	31.82	9.10	4.96	2.51	External support
Joint-prefix behavioral distillation					
Conditional control ($+\mathcal{L}_{\text{cond}}$)	37.26	16.20	8.60	2.06	None
Core JP-BND (ours) ($+\mathcal{L}_J$)	39.26	**18.60**	**10.66**	**3.90**	None
Safety-augmented JP-BND	**39.42**	18.54	10.62	3.84	None

checkpoints for all three training seeds are frozen. This comparison measures generalization to unseen instances from the same Delayed temporal population rather than to a new temporal distribution.

Metrics and statistics. Primary evaluation uses $B = 500$ and item budget $K_{\text{item}} = 500$. Full-SID Hit@500 is primary, with L1/L1L2 final-path coverage and collision-aware Expected Item as additional endpoints (Appendices C.6 and C.7). Beam sensitivity changes only inference B. We report scheduled-final checkpoints without evaluation-based selection. Exact sample-ID joins and 10,000 paired bootstrap draws quantify evaluation uncertainty, separately from sample SD across training seeds. Percentage-point differences are computed before rounding.

4.2 Primary Temporal Results

Table 1 lists the internal controls and prior-method adaptations evaluated alongside our method. Core JP-BND reaches 10.66% Full-SID and 3.90% Expected Item. Relative to H-BND, the gains are $+6.84$ points (95% CI $[6.00, 7.66]$) and $+1.76$ points (95% CI $[1.18, 2.34]$), respectively. The smaller item-level gain shows that improved SID-path recovery is only partly converted into item credit under the fixed SID mapping and expansion policy. Improving SID construction or within-SID item resolution is outside the scope of this work. Neither the internal controls nor the prior-method adaptations match this gain under unchanged inference. ISD-style improves retrieval but requires external inference support. Appendix D.5 details the control implementations and paired diagnostics.

Training cost. The accuracy gain comes with higher training cost. On the same 8-H20 GPU setup, core JP-BND takes 46.84 hours versus 2.35 hours for H-BND ($19.92\times$). This overhead is training-only and arises from multi-parent scoring. Reference training and teacher preparation add one-time costs, described in Appendix A.5. Inference remains unchanged.

4.3 Robustness and Delayed Held-Out Evaluation

Across three training seeds, core JP-BND improves Full-SID over H-BND by 7.10 ± 0.27 points on the Delayed development population. On the separate 10,000-instance Delayed held-out sample from the same population, mean Full-SID increases from 4.57% to 11.74%, with a mean paired gain of 7.17 ± 0.32 points. Complete seed-wise results for both evaluations are reported in Appendix D.1.

Table 2 evaluates whether JP-BND's retrieval gain persists when the inference beam width is varied. Core JP-BND retains positive Full-SID gains over H-BND from $B = 20$ to 500. These runs change intermediate pruning and therefore differ from truncating the fixed $B = 500$ output at rank K. In that fixed list, core JP-BND is below H-BND at $K = 20/50$ but above it from $K = 100$ onward,

Table 2: **Beam-width sensitivity on Delayed.** Full-SID hit rate (%) under actual decoding with beam width B. Δ is core JP-BND minus H-BND in percentage points, with paired 95% CIs.

B	20	50	100	200	500
H-BND	1.28	1.80	2.46	2.90	3.82
Core JP-BND	1.94	3.54	5.22	7.26	10.66
Δ	+0.66	+1.74	+2.76	+4.36	+6.84
95% CI	[0.24, 1.06]	[1.22, 2.26]	[2.12, 3.40]	[3.66, 5.06]	[6.00, 7.66]

Table 3: **Public transfer across Amazon Books and Home.** All values are percentages under the budget-matched sampled student-training protocol and fixed $B = 500$ evaluator.

	Amazon Books				Amazon Home			
Method	L1	L1L2	Full-SID	Exp. Item	L1	L1L2	Full-SID	Exp. Item
A0: AR	25.047	3.170	2.360	0.969	25.840	5.243	3.997	2.778
A1: AR+BPA	25.523	3.129	2.332	0.998	24.865	4.895	3.754	2.523
A2: AR+BPA+BND	28.513	3.706	2.739	1.136	26.031	5.280	4.016	2.861
Core JP-BND (ours)	**29.862**	**3.921**	**2.891**	**1.213**	26.100	**5.415**	**4.236**	2.809
Safety-augmented JP-BND	29.003	3.702	2.766	1.205	**26.348**	5.331	4.180	**2.962**

indicating that its gain is concentrated more in deeper returned ranks than at the very top of the list. It remains above the matched conditional control at every reported K (see Table 10 in Appendix D.2).

4.4 Public Transfer across Amazon Domains

Table 3 reports budget-matched public transfer on Amazon Books and Home. Both domains use a TIGER-style T5 recommender with independently constructed 4096^3 ContentSID hierarchies and no industrial Stage-1/Stage-2 pretraining. Within each domain, all arms share training samples, initialization, optimizer and budget. Unlike the temporal benchmark's H-BND base, public-domain JP-BND adds joint distillation to dense A2. Frozen A2 supplies the reference frontier. On Books, core JP-BND improves all four endpoints over A2, including $+0.153$ Full-SID and $+0.077$ Expected Item points. On Home, core improves Full-SID by $+0.220$ points but decreases Expected Item by 0.052 points, while the safety-augmented variant improves both over A2 by $+0.164$ and $+0.101$ points. Thus SID-path gains transfer across both domains, while item-level realization depends on domain and objective. Protocols and paired intervals appear in Appendices F.1–F.3.

5 Mechanism and Ablation Analysis

5.1 Component attribution

We write H for the H-BND base objective and omit coefficients in configuration labels. The matched $H + \mathcal{L}_{\text{cond}}$ and $H + \mathcal{L}_J$ arms isolate joint normalization beyond multi-parent behavioral supervision.

Table 4: **Matched conditional and joint intervention.** Final-path Hit@500 (%) at the primary seed. The first four metrics use 5,000 Delayed development instances, with Expected Item as fractional credit. Delayed held-out, Dynamics, and Retention report Full-SID on 10,000, 10,000, and 10,090 instances.

	Delayed development				Delayed held-out	Dynamics	Retention
Training objective	L1	L1L2	Full-SID	Exp. Item	Full-SID	Full-SID	Full-SID
H-BND (H)	30.86	6.48	3.82	2.14	4.77	1.98	21.33
Conditional control ($H + \mathcal{L}_{\text{cond}}$)	37.26	16.20	8.60	2.06	9.33	1.41	21.36
Core JP-BND ($H + \mathcal{L}_J$)	39.26	18.60	10.66	3.90	11.63	1.25	21.44

Matched conditional versus joint supervision. In Table 4, conditional supervision raises Delayed Full-SID from 3.82% to 8.60%, a 4.78-point gain. Joint normalization further improves Full-SID by 2.06 points and Expected Item by 1.84 points. The same Full-SID progression appears on Delayed held-out: 4.77% $\rightarrow$ 9.33% $\rightarrow$ 11.63%. Yet conditional supervision's large path-level gain does not

improve Expected Item. Appendix E.4 shows that this gap is driven mainly by earlier-ranked SID groups exhausting the fixed item budget, leaving many recovered Full-SIDs without item credit. The Delayed gains also incur a temporal trade-off: Dynamics Full-SID decreases from 1.98% for H to 1.41% for the conditional control and 1.25% for core JP-BND, while Retention remains similar.

Auxiliary variants. The auxiliary terms offer limited stabilization of core JP-BND. On Dynamics, these variants reach at most 1.39% Full-SID versus 1.25% for core JP-BND, while Delayed remains between 10.62% and 10.84% versus 10.66% for core. Retention changes little. Removing the joint objective while retaining both auxiliaries returns Delayed Full-SID to 3.78%, showing that they do not explain the core gain. Continuation reweighting instead shifts the Delayed–Dynamics trade-off to 8.36%/1.68%. Appendices D.3–D.4 report the complete results and paired comparisons.

5.2 From Distribution Alignment to Search Recovery

Fixed-candidate alignment. To separate score changes from candidate-set changes, all three checkpoints rescore the same frozen H-BND L1L2 candidate set against the same teacher. Conditional training reduces within-parent KL from 5.159 to 1.933. Joint normalization leaves this term unchanged at 1.933 while further reducing parent-marginal KL from 2.632 to 2.285 (Figure 2(a)). This matches the additional parent-marginal component in Equation 17 and distinguishes joint normalization from merely exposing behavioral supervision across multiple parents.

Constrained beam-search audit. We rerun constrained beam search on all 5,000 Delayed instances using each checkpoint's frontier. Among common target-L1 survivors, conditional supervision improves within-parent L2 rank and raises the cutoff margin, whereas joint leaves within-parent rank largely unchanged but improves global joint-prefix rank. Despite near-zero mean margin change, joint yields more immediate rescues than regressions and increases final-path coverage (Figure 2(b, c)). Together with the fixed-candidate analysis, this supports selective cross-parent reordering rather than a uniform margin lift. Appendix E reports full statistics, including final-path intersections in E.4.

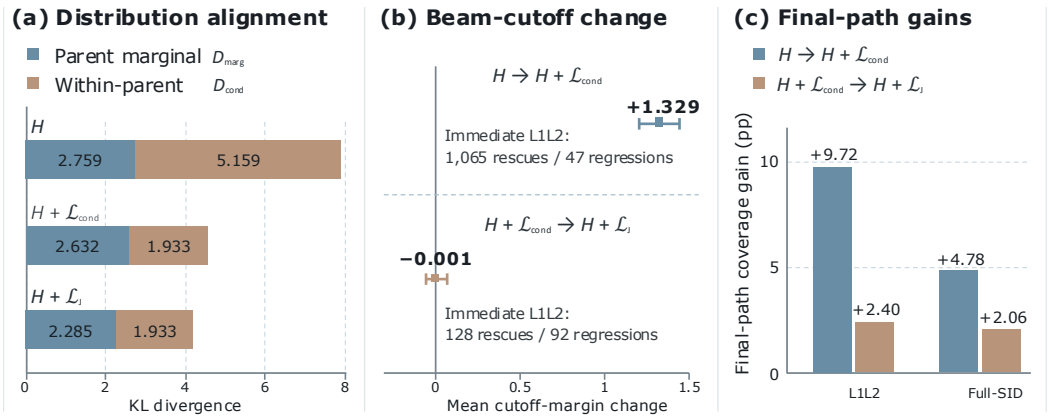

Figure 2: **Distributional and search effects of matched interventions.** (a) KL decomposition on a fixed H-BND candidate set. (b) Paired cutoff-margin changes with 95% CIs and immediate L1L2 rescue/regression events among common-L1 survivors. (c) L1L2 and Full-SID final-path coverage gains on all 5,000 Delayed instances. KL divergences and cutoff margins use natural-log scores.

6 Discussion and Conclusion

JP-BND improves navigation over fixed SID spaces by aligning behavioral supervision with joint prefixes that compete during beam search. Core JP-BND isolates joint-prefix navigation, while optional auxiliaries trade off SID-path recovery and downstream item realization. Matched controls and mechanism audits separate the multi-parent conditional gain from the additional joint benefit and link the latter to reduced parent-marginal mismatch and selective cross-parent reordering. Gains are stable across seeds and Delayed held-out evaluation, persist under narrower beams, and transfer at the SID-path level to Amazon Books and Home without changing inference. Improving training efficiency and adapting supervision under faster distribution shifts remain important next steps.

REPRODUCIBILITY STATEMENT

The main text and Appendices A–F specify the objectives, shared initializations, training protocols, and evaluation rules, including actual beam-width sensitivity. Appendix G lists planned code, fixed SID assets, checkpoints, and reconstruction recipes. Amazon Books and Home provide the public retraining and evaluation paths. Industrial code and permitted artifacts support implementation checks, while private interactions and raw content are not redistributed.

AI USE STATEMENT

AI tools assisted literature discovery, research ideation and experimental design, mathematical derivations and proof checking, manuscript drafting and editing, and consistency checks against experimental records. All experiments were executed by the authors, and all numerical results, mathematical claims, citations, and final manuscript content were reviewed and verified by the authors. The authors take full responsibility for the submission and its contents.

REFERENCES

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

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Fake Lin, Binbin Hu, Zhi Zheng, Xi Zhu, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, and Tong Xu. Token-level collaborative alignment for LLM-based generative recommendation. *arXiv preprint arXiv:2601.18457*, 2026.

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*, 2023.

Junting Wang, Xinrui He, Yunzhe Li, and Hari Sundaram. Understanding semantic IDs: From item representation to item selection in generative recommendation. *arXiv preprint arXiv:2607.24995*, 2026a.

Puji Wang, Yingchen Zhang, Ruqing Zhang, Jiafeng Guo, and Xueqi Cheng. One-pass decoding for generative recommendation with WFST-constrained A* search. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026b.

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 ACM International Conference on Information and Knowledge Management*, 2024a.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024b.

Sam Wiseman and Alexander M. Rush. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1296–1306, 2016.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Weiqin Yang, Bohao Wang, Zhenxiang Xu, Jiawei Chen, Shengjia Zhang, Jingbang Chen, Canghong Jin, and Can Wang. BEAR: Towards beam-search-aware optimization for recommendation with large language models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026.

Xin Yu, Stephen Li, Sina Aghaei, Zifan Zhu, Jiamu Bai, Guanjie Huang, Bo Peng, Yiyao Liu, and Lingzhou Xue. Difficulty-aware semantic-ID optimization for generative recommendation. *arXiv preprint arXiv:2608.20611*, 2026.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. Scalable and effective generative information retrieval. *arXiv preprint arXiv:2311.09134*, 2023.

Ziheng Zhang, Yu Cui, Bohao Wang, Yong He, Chao Yu, Chuan Yuan, Wujie Sun, Can Wang, and Jiawei Chen. SmartGR: Hierarchy and beam-aware knowledge distillation for generative recommendation. *arXiv preprint arXiv:2608.02048*, 2026.

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 *IEEE International Conference on Data Engineering*, 2024.

Guorui Zhou, Honghui Bao, Jiaming Huang, Jiaxin Deng, Jinghao Zhang, Junda She, Kuo Cai, Lejian Ren, Lu Ren, Qiang Luo, et al. Openonerec technical report. *arXiv preprint arXiv:2512.24762*, 2025.

APPENDIX

A BEHAVIORAL TEACHERS AND TRAINING OBJECTIVES

A.1 HISTORY-TO-TARGET AFFINITY GRAPH

For each training instance, we deduplicate historical source items, remove self pairs, and connect each eligible historical source x to the training target v. Users are split into two disjoint halves before aggregation. Let w_{xv}^A and w_{xv}^B denote the corresponding counts. Stable edges occur in both halves, with

$$w_{xv}^{\text{stable}} = \min(w_{xv}^A, w_{xv}^B). \tag{20}$$

Define

$$o_x = \sum_v w_{xv}^{\text{stable}}, \qquad t_v = \sum_x w_{xv}^{\text{stable}}, \qquad \alpha(x,v) = \frac{w_{xv}^{\text{stable}}}{\sqrt{o_x t_v}}. \tag{21}$$

The industrial graph contains 27,913,004 directed pairs before stable filtering and 484,234 afterward, spanning 86,441 stable nodes. The edges represent predictive history-to-target affinity and do not assume a first-order transition process.

Evaluation identities, evaluation-period interactions, and post-target evidence from contributing instances are excluded. The graph is fitted over the training partition rather than leave-one-example-out, so partition safety and per-instance as-of safety are distinct contracts.

Item-level evidence is $w_i(v) = \sum_{x \in H_i} \alpha(x,v)$, as in Equation 3. The L1 and within-parent L2 teachers are defined in Equation 4. Instance–depth pairs with empty teachers are excluded from BND.

A.2 BASE TRAINING OBJECTIVES AND HARD-EXAMPLE BND

Let $\mathcal{B}$ denote a training batch. For each instance i, the framed target SID sequence contains five supervised tokens,

$$\mathbf{y}_i = (y_{i,0}, \ldots, y_{i,4}) = (\text{SID-begin}, \text{L1}, \text{L2}, \text{L3}, \text{SID-end}).$$

The autoregressive objective is

$$\mathcal{L}_{\text{AR}} = -\frac{1}{5|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{t=0}^{4} \log p_\theta \left(y_{i,t} \mid H_i, y_{i,<t} \right). \tag{22}$$

All five target positions contribute equally.

Behavioral Prefix Alignment (BPA) acts at the L1 and L2 prediction positions. Let $h_{i,d}$ be the hidden state predicting token $c_{i,d}^\star$, and let

$$e_c = \text{sg}(W_{\text{head}}[c])$$

be a detached row of the current output head. With cosine similarity s, temperature $T = 0.1$, and distinct batch negatives $\mathcal{N}_{i,d}$,

$$\ell_{i,d}^{\text{BPA}} = -\log \frac{\exp(s(h_{i,d}, e_{c_{i,d}^\star})/T)}{\exp(s(h_{i,d}, e_{c_{i,d}^\star})/T) + \sum_{c \in \mathcal{N}_{i,d}} \exp(s(h_{i,d}, e_c)/T)}. \tag{23}$$

The observed token is excluded from the negative set and repeated negative IDs are deduplicated.

For each instance, let

$$\mathcal{D}_i = \{d \in \{\text{L1}, \text{L2}\} : |\mathcal{N}_{i,d}| > 0\}, \qquad \mathcal{E} = \{i \in \mathcal{B} : \mathcal{D}_i \neq \emptyset\}.$$

The BPA loss is

$$\mathcal{L}_{\text{BPA}} = \frac{\sum_{i \in \mathcal{E}} \sum_{d \in \mathcal{D}_i} \ell_{i,d}^{\text{BPA}}}{2|\mathcal{E}|}. \tag{24}$$

When $\mathcal{E}$ is empty, the loss is zero. If only one depth has available negatives for an instance, the other depth contributes zero while the fixed two-depth normalization is retained.

Let $s_{i,d}$ indicate a nonempty BND teacher and z_i an exposure mask. The BND reduction is

$$\mathcal{L}_{\mathrm{BND}}(z) = \frac{\sum_{i\in\mathcal{B}} \sum_{d\in\{\mathrm{L1,L2}\}} z_i s_{i,d}\, \mathrm{KL}(q^{\mathrm{BND}}_{i,d} \| p_{\theta,i,d})}{\sum_{i\in\mathcal{B}} \sum_{d\in\{\mathrm{L1,L2}\}} z_i s_{i,d}}, \tag{25}$$

with zero loss when the denominator is zero. For the temporal recommender, the student distribution is normalized over the full language-model vocabulary at each local prediction position.

The temporal hard-example mask is constructed once using a frozen AR-only (A0) reference:

$$z^H_i = \mathbf{1}[\mathrm{rank}_{\mathrm{A0}}(a^\star_i \mid H_i) > B]\,, \tag{26}$$

where rank is taken over legal L1 tokens, and $B = 500$ is fixed when constructing the temporal hard-example mask. Thus, $z^H_i = 1$ marks instances whose target L1 is not reachable within the A0 legal top-B frontier. The mask gates only the BND term. The superscript H identifies the H-BND exposure mask, not the model used to compute that rank. Its objective is

$$\mathcal{L}_{\text{H-BND}} = \mathcal{L}_{\mathrm{AR}} + \lambda_{\mathrm{BPA}}\mathcal{L}_{\mathrm{BPA}} + \lambda_{\mathrm{BND}}\mathcal{L}_{\mathrm{BND}}(z^H). \tag{27}$$

Only the BND term is gated by z^H. AR and BPA retain their original exposure. The temporal H-BND mask activates on 194,135 of 900,774 training instances, or 21.55%.

A.3 JP-BND Support and Candidate-Set Contract

Equations 7–14 specify the joint teacher and joint loss. Frozen A0 determines the exposure mask, whereas a separately trained frozen H-BND checkpoint supplies the offline width-500 frontier. Neither is the student initialization. The student recomputes its own width-500 frontier and expands all legal L2 children. Teacher-only parents are not inserted, and zero-teacher-mass legal rivals remain in the denominator.

Of the final 900,774 training instances, 760,490 (84.43%) have nonzero cached joint support. Supported instances contain on average 190.68 positive prefixes across 116.26 parents. A fixed 5,000-instance training audit finds active-instance mean retained mass rising from 0.1340 at initialization to 0.9833 at the final checkpoint, with 83.18% active instances at the latter. At student initialization, 35 instances retain a single teacher-supported prefix and have $Z_i < 10^{-2}$. No such cases are observed at audited checkpoints from step 880 onward.

At the final checkpoint, 9.99% of current student frontier parents lie outside the frozen reference frontier, carrying 2.51% of legal-L1 probability mass. These statistics quantify frontier overlap, not behavioral relevance.

A.4 Information Access and Computation

Table 5: **Information-access contract for the temporal protocol.**

Component	Construction	Current-instance target access
SID mapping	Content-side, fixed before recommendation	No behavioral target access
Behavior graph	Training-partition history–target pairs	Training targets used in graph fitting
Frozen references	A0 mask and H-BND joint support	Target L1 read for A0 exposure only
Offline JP teacher	History, graph, mapping, H-BND frontier	No target-prefix lookup or insertion
Candidate set / joint loss	Student frontier and all legal L2 children	No gold/teacher parent insertion
Optional auxiliaries	Frozen root anchor, current boundary scores	Target lookup allowed for boundary gate
AR / BPA / local BND	Supervised objectives	Stated target supervision
Inference	Trained student and original decoder	No graph, labels, cache, or reference

Joint KL averages uniformly over $\mathcal{B}_J = \{i \in \mathcal{B} : Z_i > 0\}$, without retained-mass thresholds or Z_i weighting. Both optional auxiliaries average over the complete local batch. Candidate membership, parent selection, beam cutoff, and boundary gate are detached, while candidate scores remain

differentiable. Forward/backward checks verify the domains, reductions, and finite nonzero gradients. Table 5 summarizes target access by component and Algorithm 1 gives the procedure.

Algorithm 1 Core JP-BND training for the temporal protocol

Require: Training instances $\mathcal{D}$, fixed SID mapping g, legal trie $\mathcal{T}$, behavioral graph $\mathcal{G}$, frozen H-BND reference θ_H, frozen exposure mask z^H, training beam width $B = 500$
Require: Shared Stage-2 initialization θ_0
Ensure: Trained student θ

Offline preparation
for each training instance i **do**
  Compute frozen H-BND L1 frontier $\mathcal{F}_i^H$
  Aggregate behavioral mass $W_i(a, b)$ through g
  Cache q_i^{JP} on $\mathcal{S}_i^H$ (zero mass if empty)
end for
Student training
Initialize $\theta \leftarrow \theta_0$
for each training batch $\mathcal{B}$ **do**
  for each instance $i \in \mathcal{B}$ **do**
    Compute legal top-B L1 frontier $\mathcal{F}_i^\theta$
    Expand legal L2 children to form Ω_i^θ
    Compute $S_\theta(a, b \mid H_i)$ and normalize to obtain P_i^θ
    Compute retained teacher mass $Z_i = \sum_{(a,b)\in\Omega_i^\theta} q_i^{\mathrm{JP}}(a, b)$
    if $Z_i > 0$ **then**
      Restrict and renormalize q_i^{JP} to obtain q_i
    end if
  end for
  Compute $\mathcal{L}_J$ over $\mathcal{B}_J = \{i \in \mathcal{B} : Z_i > 0\}$ (zero if empty)
  Optimize $\mathcal{L}_{\text{H-BND}} + \lambda_J \mathcal{L}_J$
end for
return θ

A.5 Training-Time and Memory Cost

Training time and memory. On eight NVIDIA H20-3e GPUs (BF16, ZeRO-2, global batch 256, seed 20260805), 3,519 updates take 2.35 hours for H-BND and 46.84 hours for core JP-BND ($19.92\times$), excluding offline preparation. Inference remains unchanged.

Implementation overhead. Core JP-BND constructs current student frontiers instance by instance. The differentiable scorer prefills each prompt once and reuses the KV cache across checkpointed parent chunks. Frontier construction uses a separate prefill. Instance-wise processing, dense per-parent vocabulary logits, and backward recomputation add overhead beyond multi-parent scoring.

One-time preparation. The reported 2.35-hour H-BND run supplies the frozen reference. Behavior-only teacher rewriting and cache writing take approximately 72 minutes for two deterministic CPU-only passes over an existing source view. This timing excludes initial graph, frontier, support, and root-anchor construction. References and caches are reusable across student seeds.

B Joint-Prefix Theory and Auxiliary Objectives

This section gives the full derivations behind the joint-prefix objective and the formal definitions of the training-only auxiliary terms. Unless stated otherwise, all statements below condition on a fixed candidate set Ω_i^θ. Training does not differentiate through the discrete selection of this set.

B.1 Joint KL Decomposition and Proof

For one active instance i with $Z_i > 0$, let $q_i(a, b)$ and $P_i^\theta(a, b)$ be distributions over the same fixed candidate set Ω_i^θ. Since P_i^θ is obtained by a softmax over Ω_i^θ, every candidate has strictly positive probability. Hence $q_i \ll P_i^\theta$.

Their L1 marginals are

$$q_{i,\mathrm{L1}}(a) = \sum_{b:(a,b)\in\Omega_i^\theta} q_i(a,b), \qquad P_{i,\mathrm{L1}}^\theta(a) = \sum_{b:(a,b)\in\Omega_i^\theta} P_i^\theta(a,b). \tag{28}$$

For any parent a with $q_{i,\mathrm{L1}}(a) > 0$, define

$$q_i(b \mid a) = \frac{q_i(a,b)}{q_{i,\mathrm{L1}}(a)}, \qquad P_i^\theta(b \mid a) = \frac{P_i^\theta(a,b)}{P_{i,\mathrm{L1}}^\theta(a)}. \tag{29}$$

Because $q_{i,\mathrm{L1}}(a) > 0$ implies that at least one candidate with parent a belongs to Ω_i^θ, the strict positivity of the softmax also implies $P_{i,\mathrm{L1}}^\theta(a) > 0$.

Hence,

$$q_i(a,b) = q_{i,\mathrm{L1}}(a) q_i(b \mid a), \qquad P_i^\theta(a,b) = P_{i,\mathrm{L1}}^\theta(a) P_i^\theta(b \mid a). \tag{30}$$

Substituting these factorizations into the KL definition gives

$$\begin{aligned}
\mathrm{KL}(q_i \| P_i^\theta) &= \sum_{a,b} q_i(a,b) \log \frac{q_i(a,b)}{P_i^\theta(a,b)} && (31)\\
&= \sum_{a,b} q_{i,\mathrm{L1}}(a) q_i(b \mid a) \log \frac{q_{i,\mathrm{L1}}(a) q_i(b \mid a)}{P_{i,\mathrm{L1}}^\theta(a) P_i^\theta(b \mid a)} && (32)\\
&= \sum_{a} q_{i,\mathrm{L1}}(a) \log \frac{q_{i,\mathrm{L1}}(a)}{P_{i,\mathrm{L1}}^\theta(a)} \\
&\quad + \sum_{a:q_{i,\mathrm{L1}}(a)>0} q_{i,\mathrm{L1}}(a) \sum_b q_i(b \mid a) \log \frac{q_i(b \mid a)}{P_i^\theta(b \mid a)} && (33)\\
&= \mathrm{KL}(q_{i,\mathrm{L1}} \| P_{i,\mathrm{L1}}^\theta) + \sum_{a:q_{i,\mathrm{L1}}(a)>0} q_{i,\mathrm{L1}}(a)\, \mathrm{KL}\big(q_i(\cdot \mid a) \| P_i^\theta(\cdot \mid a)\big). && (34)
\end{aligned}$$

The second term is exactly the per-instance objective used by the matched multi-parent conditional control. The additional term introduced by joint normalization is therefore

$$\mathcal{L}_{\mathrm{marg,i}} = \mathrm{KL}(q_{i,\mathrm{L1}} \| P_{i,\mathrm{L1}}^\theta). \tag{35}$$

With the same active-instance set $\mathcal{B}_J = \{i \in \mathcal{B} : Z_i > 0\}$ and the same reduction,

$$\mathcal{L}_J = \mathcal{L}_{\mathrm{cond}} + \mathcal{L}_{\mathrm{marg}}, \qquad \mathcal{L}_{\mathrm{marg}} = \frac{1}{|\mathcal{B}_J|} \sum_{i\in\mathcal{B}_J} \mathrm{KL}(q_{i,\mathrm{L1}} \| P_{i,\mathrm{L1}}^\theta). \tag{36}$$

All three losses are zero when $\mathcal{B}_J$ is empty. Equation 34 is the standard chain rule for KL divergence. Its role here is to isolate the parent-marginal component tested by the matched conditional intervention.

B.2 Score-Gradient Interpretation

The student joint distribution is

$$P_i^\theta(a,b) = \frac{\exp S_\theta(a,b \mid H_i)}{\sum_{(a',b')\in\Omega_i^\theta} \exp S_\theta(a',b' \mid H_i)}. \tag{37}$$

Since the entropy of q_i is constant with respect to the student scores, the joint KL is a softmax cross-entropy up to an additive constant:

$$\mathrm{KL}(q_i \| P_i^\theta) = -\sum_{a,b} q_i(a,b) \log P_i^\theta(a,b) + \mathrm{const}. \tag{38}$$

Differentiating with respect to one cumulative score gives

$$\frac{\partial \mathrm{KL}(q_i \| P_i^\theta)}{\partial S_\theta(a, b \mid H_i)} = P_i^\theta(a, b) - q_i(a, b). \tag{39}$$

Because

$$S_\theta(a, b \mid H_i) = \log p_\theta(a \mid H_i) + \log p_\theta(b \mid H_i, a),$$

the direct derivative with respect to the root log-probability coordinate is

$$\frac{\partial \mathrm{KL}(q_i \| P_i^\theta)}{\partial \log p_\theta(a \mid H_i)} = \sum_{b:(a,b)\in\Omega_i^\theta} \left[P_i^\theta(a, b) - q_i(a, b)\right] \tag{40}$$

$$= P_{i,\mathrm{L1}}^\theta(a) - q_{i,\mathrm{L1}}(a). \tag{41}$$

Thus the joint objective contains a direct score-coordinate signal that can increase or decrease the aggregate probability mass assigned to an L1 parent.

For comparison, within one fixed parent a, the conditional student distribution used by the matched $\mathcal{L}_{\mathrm{cond}}$ control is

$$P_i^\theta(b \mid a) = \frac{\exp S_\theta(a, b \mid H_i)}{\sum_{b':(a,b')\in\Omega_i^\theta} \exp S_\theta(a, b' \mid H_i)} \tag{42}$$

$$= \frac{p_\theta(b \mid H_i, a)}{\sum_{b':(a,b')\in\Omega_i^\theta} p_\theta(b' \mid H_i, a)}. \tag{43}$$

The common factor $p_\theta(a \mid H_i)$ cancels exactly. Therefore, for the matched conditional objective,

$$\frac{\partial \mathcal{L}_{\mathrm{cond}}}{\partial \log p_\theta(a \mid H_i)} = 0 \tag{44}$$

as a direct score-coordinate derivative on a fixed candidate set.

This does not imply that conditional supervision leaves root probabilities unchanged after a parameter update. Shared model parameters can change $p_\theta(a \mid H_i)$ indirectly. The distinction is that $\mathcal{L}_{\mathrm{cond}}$ has no direct root-coordinate term, whereas $\mathcal{L}_J$ does through Equation 41. The original BND objective also contains separate L1 supervision.

B.3 Joint Parent Marginal

For parent a, define

$$\Omega_i^\theta(a) = \left\{b : (a, b) \in \Omega_i^\theta\right\}. \tag{45}$$

From the cumulative score

$$S_\theta(a, b \mid H_i) = \log p_\theta(a \mid H_i) + \log p_\theta(b \mid H_i, a),$$

we have

$$\sum_{b\in\Omega_i^\theta(a)} \exp S_\theta(a, b \mid H_i) = p_\theta(a \mid H_i) \sum_{b\in\Omega_i^\theta(a)} p_\theta(b \mid H_i, a) \tag{46}$$

$$= p_\theta(a \mid H_i)\mu_{i,\mathrm{L2}}^{\mathrm{legal}}(a), \tag{47}$$

where

$$\mu_{i,\mathrm{L2}}^{\mathrm{legal}}(a) = \sum_{b\in\Omega_i^\theta(a)} p_\theta(b \mid H_i, a). \tag{48}$$

Consequently,

$$P_{i,\mathrm{L1}}^\theta(a) = \frac{p_\theta(a \mid H_i)\mu_{i,\mathrm{L2}}^{\mathrm{legal}}(a)}{\sum_{a'\in\mathcal{F}_i^\theta} p_\theta(a' \mid H_i)\mu_{i,\mathrm{L2}}^{\mathrm{legal}}(a')}. \tag{49}$$

If the L2 distribution were normalized only over the legal children in $\Omega_i^\theta(a)$, then $\mu_{i,\mathrm{L2}}^{\mathrm{legal}}(a) = 1$. In the temporal implementation, however, $p_\theta(b \mid H_i, a)$ is the raw language-model probability over the full vocabulary, while $\Omega_i^\theta(a)$ contains only legal trie children. Hence $\mu_{i,\mathrm{L2}}^{\mathrm{legal}}(a)$ measures the raw conditional probability mass assigned to those legal children. It can vary across parents because their legal child sets differ and probability mass may fall on tokens outside those sets.

The parent marginal induced by joint-prefix normalization is therefore not generally equal to the bare L1 distribution $p_\theta(a \mid H_i)$. It depends both on the L1 probability and on the raw L2 probability mass retained by the legal children of that parent. This is the quantity aligned by $\mathcal{L}_{\mathrm{marg}}$.

B.4 Beam-Boundary Auxiliary

Let

$$S_i^\star = S_\theta(a_i^\star, b_i^\star \mid H_i) \tag{50}$$

be the target joint-prefix score. At depth 2, let $\mathcal{A}_{i,2}^\theta$ denote the legal expansions of the surviving L1 frontier. The decoder retains up to B prefixes with highest cumulative scores under its deterministic tie-breaking rule. Let $\tau_{i,2,B}$ denote the score of the last retained prefix, and let $F_{i,2}$ indicate whether the target L1L2 prefix belongs to the retained set.

Actual prefix survival is defined by $F_{i,2}$ rather than by the sign of a score margin. The cutoff margin

$$m_{i,2} = S_i^\star - \tau_{i,2,B} \tag{51}$$

is used as a score diagnostic. For a target in the candidate set, $m_{i,2} > 0$ implies survival and $m_{i,2} < 0$ implies pruning. At $m_{i,2} = 0$, the actual retained set determines survival, including the target ranked exactly Bth.

Let $\mathcal{B}$ denote the local training batch. The optional boundary auxiliary is

$$\mathcal{L}_B = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} g_i \,\mathrm{softplus}\left(\mathrm{sg}(\tau_{i,2,B}) - S_i^\star\right), \tag{52}$$

where g_i is active only when the target L1 survives, the target L1L2 score is at or below the cutoff, and the behavioral teacher assigns the target prefix more probability than the current student.

Candidate membership, the cutoff, and the gate are detached. Candidate scores remain differentiable. On an active instance, with the detached gate and cutoff held fixed, the softplus term is smooth and satisfies

$$\mathrm{softplus}\left(\tau_{i,2,B} - S_i^\star\right) \geq \left[\tau_{i,2,B} - S_i^\star\right]_+. \tag{53}$$

It therefore provides a smooth surrogate for the pruning violation on the active region. Because g_i is a detached discrete gate, $\mathcal{L}_B$ as a whole is not globally smooth across changes in gate membership.

With the cutoff detached, the derivative on an active instance is proportional to

$$-\sigma(\tau_{i,2,B} - S_i^\star). \tag{54}$$

The auxiliary therefore pushes an active target prefix upward relative to the observed cutoff without differentiating through which candidate defines that cutoff.

B.5 Root-Anchor Auxiliary

Using the same local batch $\mathcal{B}$, the optional root-stability auxiliary is

$$\mathcal{L}_A = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathrm{KL}\Big(p_{H,i,\mathrm{L1}}^{\mathrm{legal}} \| p_{\theta,i,\mathrm{L1}}^{\mathrm{legal}}\Big). \tag{55}$$

For one instance, write

$$\mathcal{L}_{A,i} = \mathrm{KL}\Big(p_{H,i,\mathrm{L1}}^{\mathrm{legal}} \| p_{\theta,i,\mathrm{L1}}^{\mathrm{legal}}\Big). \tag{56}$$

Pinsker's inequality gives

$$\left\| p_{H,i,\mathrm{L1}}^{\mathrm{legal}} - p_{\theta,i,\mathrm{L1}}^{\mathrm{legal}} \right\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\mathcal{L}_{A,i}}. \tag{57}$$

Therefore, for any subset $\mathcal{R}$ of legal L1 tokens,

$$\left|p^{\text{legal}}_{\theta,i,\text{L1}}(\mathcal{R}) - p^{\text{legal}}_{H,i,\text{L1}}(\mathcal{R})\right| \le \sqrt{\frac{1}{2}\mathcal{L}_{A,i}}. \tag{58}$$

The anchor controls probability-mass drift over arbitrary legal root regions. It does not fix individual token ranks and does not guarantee preservation of the top-B parent set.

B.6 Continuation Reweighting Identity

The CA-JP-BND variant uses the same behavioral joint-prefix support as JP-BND. Let p_H denote the frozen H-BND reference. For a behavior-supported item v with $g(v) = (a_v, b_v, c_v)$, define its suffix-completion score

$$\rho_H(v \mid H_i, a_v, b_v) = p_H(c_v \mid H_i, a_v, b_v) p_H(\text{SID-end} \mid H_i, a_v, b_v, c_v). \tag{59}$$

The continuation-weighted teacher is normalized over the same nonempty $\mathcal{S}_i^H$ and is zero outside it. On this support,

$$q_i^{\text{CA}}(a, b) \propto \sum_{v:(a_v,b_v)=(a,b)} w_i(v)\rho_H(v \mid H_i, a, b). \tag{60}$$

For any supported prefix (a, b), define its behavior-weighted mean continuation score

$$\bar{\rho}_i(a, b) = \frac{\sum_{v:(a_v,b_v)=(a,b)} w_i(v)\rho_H(v \mid H_i, a, b)}{W_i(a, b)}, \qquad W_i(a, b) > 0. \tag{61}$$

Using the JP teacher definition,

$$\frac{q_i^{\text{CA}}(a, b)}{q_i^{\text{JP}}(a, b)} = \frac{\bar{\rho}_i(a, b)}{\mathbb{E}_{(a',b')\sim q_i^{\text{JP}}}[\bar{\rho}_i(a', b')]}. \tag{62}$$

Continuation weighting therefore introduces no new behavioral support. It only redistributes probability mass within the existing JP support according to the frozen reference's suffix-completion confidence.

C Primary Temporal Benchmark and Training Pipeline

C.1 Behavioral Data Construction

We use a fixed extract of anonymized video-interaction sequences containing item IDs, second-level timestamps, engagement signals, and play ratio. Eligible targets have either an explicit positive-engagement signal or a play ratio ≥ 0.5, while other observed events may remain in history. We remove malformed and unmapped events, sort by time, and remove consecutive duplicates. Cutoffs require five prior interactions, with at most ten cutoffs per user under a frozen selection rule. We first freeze 1,000,000 canonical anchors with sample/user identities, history, target, time, and prompt identity, then join the SID mapping.

Excluding evaluation/development identities and ineligible instances leaves 810,697 next-item and 90,077 content-denotation replay instances (900,774 total). Replay provides Full-SID hard-CE supervision without history to preserve content denotation. Every Stage-3 arm shares this instance set and order.

C.2 Fixed ContentSID Space

The ContentSID mapping is constructed before recommendation training from 128-dimensional item embeddings produced by a fine-tuned InternVideo2 model (Wang et al., 2024b). We use a three-layer greedy-margin variant of balanced RQ-KMeans (Zhou et al., 2025), with 4,096 codes per layer, producing a 4096^3 SID space. The frozen catalog contains 14,102,490 distinct complete SIDs, so multiple items may share the same identifier.

The quantizer is constructed once using the greedy-margin fitting procedure. Capacity constraints apply to sampled codebook fitting, not to full-catalog encoding. The resulting centroids, item-to-SID mapping, and legal trie are frozen before recommender training and shared by all temporal arms. JP-BND never updates these assets. Fitting configurations and inference code belong to the planned release in Appendix G.1.

C.3 Shared Content-Side Initialization

Qwen3-0.6B (Yang et al., 2025) undergoes two shared content-side stages, performed once before all Stage-3 comparisons.

Stage 1: SID-token alignment. The input pairs a complete SID with an item description from training-side canonical data (250,000 pairs). Only newly introduced SID embedding rows are trainable. Existing vocabulary rows and non-embedding parameters are frozen, with zero weight decay to prevent drift. The budget is 2,000 steps, selecting checkpoint 1,500 by validation loss. A tensor audit verifies frozen parameters.

Stage 2: content-to-SID denotation. The model generates a complete SID from its item description using 483,910 training and 10,090 validation examples. Hard autoregressive prediction is combined with a 0.2-weighted frozen-quantizer geometry signal, constructed from the item's residual-to-centroid distances under the true parent prefix. This shared content objective is not JP-BND. Training starts from Stage 1 and runs 3,000 steps, selecting checkpoint 3,000 by validation hard CE. No Delayed, Dynamics, Retention, or Delayed held-out targets train either content stage.

C.4 Stage-3 Behavioral Recommendation Training

Next-item instances use ContentSID histories to predict the next eligible complete SID. All arms start independently from the same Stage-2 checkpoint, not from another Stage-3 endpoint. A0 uses AR, A1 adds BPA, and A2 adds dense BND. H-BND gates only BND when the target L1 falls outside the legal top-B frontier of frozen A0, with $B = 500$ for mask construction (194,135/900,774 instances, 21.55%). Frozen H-BND provides joint support and optional reference scores. Core JP-BND adds $\lambda_J \mathcal{L}_J$.

Within a seed, arms share instance identities/order, optimizer, global batch 256, and 3,519 updates, along with SID mapping, trie, and decoder. The industrial temporal protocol processes 900,864 training exposures ($3{,}519 \times 256$), including 90 padding exposures beyond the 900,774 training instances. Shared coefficients are $\lambda_{\mathrm{BPA}} = 0.10$, $\lambda_{\mathrm{BND}} = 0.20$, $\lambda_J = 0.10$, and $\lambda_B = \lambda_A = 0.05$.

C.5 Temporal Populations and Evaluation Protocol

The three development populations contain 5,000 *Delayed*, 10,000 *Dynamics*, and 10,090 *Retention* instances. Delayed uses a non-overlapping evaluation period three days after the training window and measures near-term temporal generalization. Dynamics is drawn from a broader behavioral window with users disjoint from the training population, providing a view of sensitivity to behavioral change. Retention uses replay instances to measure preservation of content denotation. These populations were used during method development and are not treated as untouched test data.

A separate 10,000-instance Delayed held-out sample from the same canonical Delayed population evaluates H-BND, the conditional control ($+\mathcal{L}_{\mathrm{cond}}$), core JP-BND ($+\mathcal{L}_J$), and safety-augmented JP-BND ($+\mathcal{L}_J + \mathcal{L}_B + \mathcal{L}_A$) across all three training seeds after the corresponding checkpoints are frozen. It uses the same temporal setting as Delayed rather than a new temporal distribution.

Primary temporal endpoint evaluations use constrained path beam $B = 500$ and a fixed item-expansion budget $K_{\mathrm{item}} = 500$. Beam-width sensitivity reuses the same trained checkpoints and varies only inference B. The primary training seed is 20260805, with 20260831 and 20260901 used for training-seed robustness.

C.6 Final-Path Coverage and Search Survival

For returned complete paths $\mathcal{P}_B(H_i)$ under beam width B, let $\mathbf{s}$ denote a complete SID path and define

$$C_{i,d} = \mathbf{1}[\exists\, \mathbf{s} \in \mathcal{P}_B(H_i) : \mathbf{s}_{1:d} = g_{1:d}(v_i^\star)], \qquad \mathrm{Hit}_d = \frac{100}{n} \sum_i C_{i,d}. \tag{63}$$

For the primary $B = 500$ evaluation, we report $d = 1$, 2, and 3 as L1, L1L2, and Full-SID Hit@500. Beam-width audits use the same coverage definition with the corresponding value of B. These measure final-path coverage rather than immediate search survival.

For immediate survival, let $F_{i,d}$ indicate whether the target prefix remains in the frontier immediately after depth-d pruning. In a tree beam, $C_{i,d} \le F_{i,d}$ because a surviving prefix can later lose all matching descendants. Consequently, endpoint ratios such as $\mathrm{Hit}_2 \,/\, \mathrm{Hit}_1$ do not identify a causal propagation probability across depths.

For reference and comparison models, an immediate rescue is $(F_{i,d}^{\mathrm{ref}}, F_{i,d}^{\mathrm{cmp}}) = (0, 1)$, and a regression is $(1, 0)$. Final-path transitions replace F with C. These are distinct events, so immediate-to-final intersections require the same instance to satisfy both changes.

C.7 Collision-Aware Expected Item

Let

$$V(\mathbf{s}) = \{v : g(v) = \mathbf{s}\}, \qquad M_{\mathbf{s}} = |V(\mathbf{s})|$$

be the items and size of the collision group associated with complete SID $\mathbf{s}$.

The decoder preserves the model-score order of complete SID paths and deduplicates repeated complete SIDs by retaining their first occurrence. Let

$$(\mathbf{s}_{i,1}, \ldots, \mathbf{s}_{i,K_i})$$

denote the resulting ordered SID list for instance i. Item expansion traverses this list in order under a fixed budget $K_{\mathrm{item}} = 500$. Define the remaining item budget before SID $\mathbf{s}_{i,j}$ as

$$R_{i,j} = \left[K_{\mathrm{item}} - \sum_{\ell<j} k_{i,\mathbf{s}_{i,\ell}} \right]_+, \tag{64}$$

and allocate

$$k_{i,\mathbf{s}_{i,j}} = \min\left(M_{\mathbf{s}_{i,j}}, R_{i,j}\right). \tag{65}$$

Thus a collision group is fully included when sufficient budget remains and is truncated only when it crosses the fixed item budget.

The model provides no score or ordering among individual items sharing the same SID. We therefore assign uniform expected credit within each collision group. The per-instance credit is

$$\mathrm{EHit}_i = \sum_{j=1}^{K_i} \mathbf{1}[v_i^\star \in V(\mathbf{s}_{i,j})] \, \frac{k_{i,\mathbf{s}_{i,j}}}{M_{\mathbf{s}_{i,j}}}, \qquad \text{Expected Item} = \frac{100}{n} \sum_i \mathrm{EHit}_i. \tag{66}$$

Expected Item is therefore a fractional collision-aware endpoint rather than a binary hit rate or a model-defined item-ranking metric. A Full-SID hit can receive zero Expected Item credit if earlier ranked SID groups exhaust the fixed item budget before the target SID group is reached. We do not report item-level Hit@K or NDCG because the decoder assigns no scores or ranks to individual items within a shared SID.

C.8 Uncertainty and Checkpoint Selection

Pairwise confidence intervals are computed from exact sample-ID joins using 10,000 paired bootstrap draws. Reported deltas are absolute percentage-point differences computed before rounding. Expected Item uses the same procedure on its fractional per-instance credits. These intervals quantify evaluation-sample uncertainty, while training-seed variability is reported separately.

Only scheduled final Stage-3 checkpoints are formal endpoint results. Intermediate checkpoints and fixed diagnostic windows are used only for mechanism audits and are not model-selection pools.

D Robustness and Ablations

This appendix follows the method groups in Table 1. A0 uses AR, A1 adds BPA, and A2 adds dense BND. H-BND gates only the BND term using the frozen AR-only reference. Their objectives and training protocol are specified in Appendices A.2 and C.4.

For compact objective labels, let $H \equiv \mathcal{L}_{\text{H-BND}}$. The matched conditional control is $H + \mathcal{L}_{\text{cond}}$, core JP-BND is $H + \mathcal{L}_J$, and safety-augmented JP-BND is $H + \mathcal{L}_J + \mathcal{L}_B + \mathcal{L}_A$. These labels omit the fixed coefficients in Section 4.1.

D.1 Training-Seed Stability and Delayed Held-Out Evaluation

Table 6: **Training-seed stability of Full-SID retrieval.** Full-SID Hit@500 (%) across three fixed training seeds on Delayed development and the 10,000-instance Delayed held-out evaluation. SD is the sample standard deviation across training seeds.

	Delayed development				Delayed held-out			
Method	0805	0831	0901	Mean ± SD	0805	0831	0901	Mean ± SD
H-BND	3.82	3.26	3.58	3.55 ± 0.28	4.77	4.27	4.67	4.57 ± 0.26
Conditional control	8.60	8.60	8.50	8.57 ± 0.06	9.33	9.32	9.37	9.34 ± 0.03
Core JP-BND	10.66	10.64	10.66	10.65 ± 0.01	11.63	11.76	11.84	11.74 ± 0.11
Safety-augmented JP-BND	10.62	10.64	10.60	10.62 ± 0.02	11.68	11.81	11.77	11.75 ± 0.07

Table 7: **Complete Delayed held-out results.** All four endpoints are percentages on the same 10,000 Delayed held-out instances from the canonical Delayed population. Rows differ by training seed, not by evaluation sample.

Seed	Method	L1	L1L2	Full-SID	Exp. Item
0805	H-BND	31.85	7.25	4.77	2.640
	Conditional control	38.82	17.13	9.33	2.290
	Core JP-BND	40.97	19.42	11.63	4.283
	Safety-augmented JP-BND	41.18	19.40	11.68	4.300
0831	H-BND	30.94	7.06	4.27	2.040
	Conditional control	38.58	17.05	9.32	2.280
	Core JP-BND	40.99	19.45	11.76	4.420
	Safety-augmented JP-BND	41.25	19.44	11.81	4.340
0901	H-BND	31.35	7.06	4.67	2.490
	Conditional control	38.68	17.10	9.37	2.290
	Core JP-BND	41.06	19.51	11.84	4.423
	Safety-augmented JP-BND	41.08	19.48	11.77	4.377

Tables 6–7 report the fixed training seeds. The matched conditional control is stable on Delayed, reaching 8.60/8.60/8.50% Full-SID across the three seeds (8.57 ± 0.06). Its three-seed Full-SID mean ± SD is 1.41 ± 0.02 on Dynamics and 21.37 ± 0.01 on Retention.

On Delayed held-out, the conditional control reaches 9.33/9.32/9.37% Full-SID (9.34 ± 0.03). Relative to H-BND, its Full-SID gains are $+4.56$ $[4.01, 5.12]$, $+5.05$ $[4.50, 5.60]$, and $+4.70$ $[4.14, 5.26]$ points. Core JP-BND adds a further $+2.30$ $[1.91, 2.69]$, $+2.44$ $[2.06, 2.83]$, and $+2.47$ $[2.08, 2.86]$ points. Both objective contrasts improve Full-SID at every training seed. The matched conditional control has no resolved Expected Item difference from H-BND, whereas core JP-BND improves all four endpoints over the matched conditional control at every seed.

Every Delayed held-out seed uses the same 10,000 instances. The paired bootstrap uses 10,000 draws with seed 20260822, separately from training randomness. Core JP-BND improves all four endpoints over H-BND at every seed. Safety-augmented JP-BND has no resolved Full-SID difference from core JP-BND at any seed. No model or checkpoint is selected from these outcomes.

D.2 Actual Beam Width and Fixed-List Rank

B is the decoding beam width. K is a post-decoding SID-rank cutoff, distinct from the item-expansion budget K_{item}. We reuse the primary-seed checkpoints without retraining. Decoding is rerun for $B \in \{10, 20, 50, 100, 200\}$, while $B = 500$ reuses the primary outputs. The item budget remains $K_{\text{item}} = 500$ throughout. Tables 8 and 9 report endpoints and paired Full-SID differences. Core JP-BND improves over both H-BND and the matched conditional control with positive 95% CIs at every tested $B \geq 20$. At $B = 10$, both differences are unresolved.

Table 8: **Actual beam-width endpoints on Delayed.** Full-SID and Expected Item (%) on the same 5,000 instances. Decoding is rerun for $B < 500$ with checkpoints fixed. Expected Item retains $K_{\text{item}} = 500$.

	Full-SID			Expected Item		
B	H	$H + \mathcal{L}_{\text{cond}}$	$H + \mathcal{L}_J$	H	$H + \mathcal{L}_{\text{cond}}$	$H + \mathcal{L}_J$
10	0.78	0.98	1.06	0.78	0.98	1.06
20	1.28	1.60	1.94	1.28	1.60	1.94
50	1.80	2.60	3.54	1.76	2.54	3.54
100	2.46	4.10	5.22	2.36	3.92	5.10
200	2.90	5.62	7.26	2.52	4.22	6.14
500	3.82	8.60	10.66	2.14	2.06	3.90

Table 9: **Paired Full-SID gains under actual beam decoding.** Values are percentage-point differences with 95% paired-bootstrap confidence intervals.

B	Core JP-BND − H-BND	Core JP-BND − Conditional control
10	$+0.28$ $[-0.04, 0.60]$	$+0.08$ $[-0.10, 0.26]$
20	$+0.66$ $[0.24, 1.06]$	$+0.34$ $[0.10, 0.60]$
50	$+1.74$ $[1.22, 2.26]$	$+0.94$ $[0.60, 1.30]$
100	$+2.76$ $[2.12, 3.40]$	$+1.12$ $[0.72, 1.54]$
200	$+4.36$ $[3.66, 5.06]$	$+1.64$ $[1.20, 2.10]$
500	$+6.84$ $[6.00, 7.66]$	$+2.06$ $[1.56, 2.56]$

Table 10: **Target-SID rank concentration within a fixed $B = 500$ output.** Values are target-SID Hit@K (%). K is a post-decoding rank cutoff, not a beam width. The decoder remains fixed at $B = 500$.

Method	$K = 10$	$K = 20$	$K = 50$	$K = 100$	$K = 200$	$K = 500$
H-BND	1.36	1.62	2.00	2.48	2.96	3.82
Conditional control	0.16	0.36	1.02	2.28	3.96	8.60
Core JP-BND	0.34	0.64	1.28	3.22	5.60	10.66

Table 11: **SID-level ranking within the fixed $B = 500$ output.** MRR and NDCG are computed from target-SID ranks and reported on the original $[0, 1]$ scale, not as percentages or item-level metrics.

Method	MRR@500	NDCG@10	NDCG@20
H-BND	0.006239	0.007633	0.008293
Conditional control	0.001172	0.000696	0.001193
Core JP-BND	0.001990	0.001539	0.002307

Table 10 truncates the fixed $B = 500$ list without rerunning search. Core JP-BND is below H-BND at $K = 10/20/50$ but above it at the reported cutoffs from $K = 100$ onward. Of 411 newly retrieved targets, 219 (53.3%) appear at ranks 201–500 and 314 (76.4%) at ranks 101–500. The gain therefore reflects broader final-path coverage rather than stronger head ranking. Relative to the matched conditional control, core JP-BND improves every reported fixed-list Hit@K, with positive paired MRR@500 and SID-level NDCG@10/20 gains. These ranking metrics nevertheless remain below H-BND (Table 11). Fixed-list ranking and actual beam-width sensitivity measure different effects because changing B also changes intermediate survivor sets.

D.3 Component Ablations and Safety Controls

The matched conditional control replaces $\mathcal{L}_J$ with $\mathcal{L}_{\text{cond}}$ while preserving the offline behavioral support, candidate-generation rule, initialization, and training budget. It supervises L2 separately within each supported parent and weights each conditional loss by the teacher parent marginal. Core JP-BND instead jointly normalizes over competing L1L2 prefixes. The independently trained models can induce different student frontiers. Safety-augmented JP-BND adds the optional beam-boundary loss $\mathcal{L}_B$ and root-stability loss $\mathcal{L}_A$ to core JP-BND. Table 12 includes both standalone auxiliaries and their deletions from the safety-augmented objective.

Table 12: **Component endpoints at the primary seed.** Values are Full-SID Hit@500 (%).

Training objective	Delayed	Dynamics	Retention
H	3.82	1.98	21.33
$H + \mathcal{L}_A$	3.80	1.96	21.38
$H + \mathcal{L}_B$	3.86	1.93	21.37
$H + \mathcal{L}_B + \mathcal{L}_A$ $(-\mathcal{L}_J)$	3.78	1.96	21.26
$H + \mathcal{L}_{\text{cond}}$	8.60	1.41	21.36
$H + \mathcal{L}_J$	10.66	1.25	21.44
$H + \mathcal{L}_J + \mathcal{L}_A$ $(-\mathcal{L}_B)$	10.84	1.39	21.41
$H + \mathcal{L}_J + \mathcal{L}_B$ $(-\mathcal{L}_A)$	10.68	1.26	21.37
$H + \mathcal{L}_J + \mathcal{L}_B + \mathcal{L}_A$	10.62	1.37	21.43
CA-JP-BND†	8.36	1.68	21.37

† CA-JP-BND uses the continuation-weighted teacher q^{CA} in the joint and boundary terms while retaining the same joint-prefix support.

Relative to safety-augmented JP-BND, core JP-BND changes Delayed Full-SID by $+0.04$ points $[-0.18, 0.26]$ and Expected Item by $+0.056$ $[-0.140, 0.256]$. The anchor-only and boundary-only arms both remain near H-BND. For $H + \mathcal{L}_B$, Delayed L1L2 improves by $+0.14$ points $[0.02, 0.28]$, but its Full-SID changes relative to H-BND are $+0.04$ $[-0.06, 0.14]$ on Delayed, -0.05 $[-0.11, 0.01]$ on Dynamics, and $+0.040$ $[-0.089, 0.159]$ on Retention. It has no resolved Expected Item gain on any population. Removing the boundary term from safety-augmented JP-BND raises Delayed Full-SID to 10.84% ($+0.22$, $[0.02, 0.42]$), while removing the anchor gives 10.68% ($+0.06$, $[-0.16, 0.28]$). Thus neither auxiliary explains the main Delayed gain.

On Dynamics, core JP-BND reaches 1.25% Full-SID and safety-augmented JP-BND reaches 1.37%. Deleting the anchor gives 1.26%, while deleting the boundary term gives 1.39%. Relative to safety-augmented JP-BND, core JP-BND changes Full-SID by -0.12 points $[-0.20, -0.05]$ on Dynamics and $+0.010$ $[-0.139, 0.159]$ on Retention. A fixed 128-instance Dynamics audit compares root distributions with H-BND. Root KL, total variation, and top-500 overlap are 0.0546, 0.1260, and 89.11% for core JP-BND, versus 0.0456, 0.1153, and 89.84% for safety-augmented JP-BND.

Safety-matched conditional control. Replacing $\mathcal{L}_J$ with $\mathcal{L}_{\text{cond}}$ while retaining both auxiliaries yields 37.70/16.26/8.78/2.42% for L1/L1L2/Full-SID/Expected Item on Delayed, versus 39.42/18.54/10.62/3.84% for safety-augmented JP-BND. The latter has a $+1.84$-point Full-SID gain $[1.34, 2.34]$, consistent with the additional joint gain over the matched conditional control without auxiliaries.

D.4 Continuation Weighting

CA-JP-BND uses the same objective structure as safety-augmented JP-BND but substitutes the continuation-weighted teacher q^{CA} in the joint and boundary terms. Frozen-reference suffix-completion confidence reweights behavioral mass without changing the joint-prefix support (Appendix B.6). Thus CA-JP-BND changes the teacher, not the number of auxiliary losses.

On Delayed, CA-JP-BND reaches 37.76/11.40/8.36/3.62% for L1/L1L2/Full-SID/Expected Item, compared with 39.42/18.54/10.62/3.84% for safety-augmented JP-BND.

Removing continuation from this safety-matched pair improves Delayed Full-SID by $+2.26$ points $[1.58, 2.94]$ but changes Dynamics by -0.31 $[-0.50, -0.12]$. Retention has no resolved difference. On fixed 128-instance teacher windows, continuation reduces median effective support among

nonempty teachers from 109.95 to 9.71 on Delayed and 83.37 to 8.08 on Dynamics. It changes the Delayed–Dynamics trade-off rather than explaining the core JP-BND gain.

D.5 Internal Controls and Prior-Method Adaptations

Following Table 1, we separate internal controls from prior-method adaptations. The trained controls start independently from the shared Stage-2 checkpoint and use the same training instances, order, and 3,519-update budget. Frozen reference models supply supervision, not student initialization. ISD-style instead reuses the A2 checkpoint and changes only decoding. All methods retain the fixed SID mapping. Unless stated otherwise, results below use the same 5,000 Delayed development instances with $B = 500$ and $K_{\text{item}} = 500$.

D.5.1 Internal Controls

Constant-total early-token weighting. This control tests whether emphasizing early SID tokens alone improves retrieval. It modifies only AR, without BPA or BND, assigning SID-begin/L1/L2/L3/SID-end weights $[2/3, 1.5, 1.5, 2/3, 2/3]$. Their sum remains 5, preserving the total token weight while shifting weight to L1 and L2. It reaches 1.32% Full-SID and 0.38% Expected Item. Relative to A0, the Full-SID difference is -0.20 points with 95% CI $[-0.38, -0.04]$. Positional reweighting alone does not reproduce the joint-prefix retrieval gain.

Gold-prefix ranking. Here gold denotes the observed target prefix. This control adds a 0.10-weighted cumulative L1L2 ranking term to H-BND. The current student first constructs its target-free top-B L1 frontier with $B = 500$, then expands all legal L2 children under those parents to form the L1L2 candidate set. The ranking term promotes the target joint prefix against competing prefixes only when the target is already present in this candidate set. No target parent or target prefix is inserted. Unlike JP-BND, this term supervises the observed target route rather than a multi-target behavioral distribution. It reaches 3.56% Full-SID and 2.02% Expected Item. In the paired comparison with A2, Full-SID changes by -0.72 points $[-1.08, -0.36]$. Its teacher-forced L2 negative log-likelihood is higher by 0.2867 and target-versus-best-sibling margin lower by 0.0725 than A2.

Route-score control (AGCR-OPD). This control adds a Smooth-L1 route-score loss to H-BND, weighted by 0.05 and gated by positive teacher advantage. A frozen A2 teacher supplies cumulative route-score targets at L1 and L1L2, weighted 0.5 and 1.0 within the auxiliary term. The intervention fits route scores rather than matching a distribution over competing joint prefixes. It reaches 3.78% Full-SID and 2.24% Expected Item. Relative to H-BND, the corresponding differences are -0.04 points $[-0.22, 0.12]$ and $+0.10$ points $[-0.02, 0.24]$. Neither endpoint has a resolved gain under this regression control.

D.5.2 Prior-Method Adaptations

These controls adapt published objectives or decoding mechanisms to the fixed-SID experimental setting rather than reproducing their original end-to-end systems.

SmartGR-style. Our matched-teacher adaptation (Zhang et al., 2026) uses a frozen H-BND model of the same size as the student. It adds hierarchical token-distribution distillation and beam-prefix ordering distillation to H-BND with weights 0.60 and 0.30, respectively. The ordering targets come from the teacher, not the observed target prefix used by Gold-prefix ranking. The teacher is used only during training. This is a same-size matched comparison rather than the original larger-teacher compression setting. It reaches 2.14% Full-SID and 0.34% Expected Item. Relative to H-BND, these change by -1.68 points $[-2.06, -1.32]$ and -1.80 points $[-2.18, -1.44]$. The matched hierarchical and beam-prefix distillation therefore does not reproduce the JP-BND gain.

BEAR-style. This local-survival adaptation (Yang et al., 2026) adds $0.25\mathcal{L}_{\text{BEAR}}$ to H-BND. For SID depth d, let $\mathcal{C}_{i,d}^{\text{legal}}$ be the legal children under the teacher-forced target prefix, $y_{i,d}^{\star}$ the target token, and $\beta_{i,d}^{B}$ the Bth-largest legal-token probability. The additive objective is

$$\mathcal{L}_{\text{BEAR}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{\substack{d \in \{1,2,3\} \\ |\mathcal{C}_{i,d}^{\text{legal}}| > B}} \log \sigma \left(\frac{\log \text{sg}(\beta_{i,d}^{B}) - \log p_\theta(y_{i,d}^{\star} \mid H_i, y_{i,<d}^{\star})}{\xi} \right), \tag{67}$$

with $B = 500$, $\xi = 2.75$, and token-probability temperature 1. Here σ is the logistic sigmoid and sg denotes stop-gradient. The loss sums active depths within each instance and averages over the complete batch. Branches with at most B children contribute zero. It reuses the AR forward pass without constructing a student beam or joint teacher. Unlike JP-BND, it supervises local token survival under a supplied parent rather than cumulative-score competition across parents. This additive-objective adaptation differs from the released detached-CE-reweighting implementation.

Table 13: **BEAR-style adaptation at the primary seed.** Endpoints are percentages. The last column gives BEAR-style minus H-BND Full-SID in percentage points with a 95% paired-bootstrap interval.

Population	L1	L1L2	Full-SID	Exp. Item	Δ Full-SID vs. H-BND (95% CI)
Delayed	29.00	5.78	3.66	1.980	-0.16 $[-0.36, 0.04]$
Dynamics	36.66	5.04	1.95	0.799	-0.03 $[-0.12, 0.06]$
Retention	86.501	36.987	21.477	17.086	$+0.149$ $[-0.030, 0.337]$

The nontrivial branching condition holds at 100.0% of L1, 97.1% of L2, and 0.03% of L3 positions. Formal runtime is $1.113\times$ H-BND. On Delayed, BEAR-style reaches 3.66% Full-SID and 1.980% Expected Item. Its Full-SID difference from H-BND is unresolved, while Expected Item changes by -0.160 points $[-0.320, -0.020]$. Dynamics L1 rises by $+0.68$ points $[0.35, 1.00]$ without resolved Full-SID or Expected Item gains. Table 13 reports all three development populations.

ISD-style. Our supported-decoding adaptation (Wang et al., 2026a) reuses the frozen A2 checkpoint without additional training. It combines model prefix rankings with training-history-derived external support using reciprocal-rank fusion with constant 60. The SID mapping and item-expansion policy remain fixed, but prefix selection now depends on external support at inference. It reaches 4.96% Full-SID and 2.51% Expected Item. Relative to unmodified A2 decoding, the gains are $+0.68$ points $[0.42, 0.96]$ and $+0.591$ points $[0.373, 0.822]$, respectively. The deeper-path gain therefore comes with an inference change, unlike the training-only JP-BND intervention.

E Mechanism Audits

Configuration labels use $H = \mathcal{L}_{\text{H-BND}}$ and omit fixed coefficients. These audits distinguish fixed-domain alignment from realized search behavior.

E.1 Matched Conditional Intervention

Table 4 reports the three checkpoints' endpoints. Table 14 gives the paired core-minus-conditional Delayed differences. The same conditional checkpoint obtains 31.52/3.93/1.41/0.363% on Dynamics and 86.72/36.76/21.36/17.03% on Retention for L1/L1L2/Full-SID/Expected Item. The fixed-budget item-credit explanation is in Appendix E.4.

Table 14: **Paired core-versus-conditional differences on Delayed.** Deltas are core JP-BND minus $H + \mathcal{L}_{\text{cond}}$ on the same 5,000 instances. Deltas and confidence intervals are in percentage points.

Endpoint	Δ	95% CI
L1	+2.00	$[1.26, 2.74]$
L1L2	+2.40	$[1.80, 2.98]$
Full-SID	+2.06	$[1.56, 2.56]$
Expected Item	+1.84	$[1.42, 2.26]$

E.2 Fixed-Candidate KL Decomposition

We freeze H-BND's legal L1 frontier and all its legal L2 children on a 128-instance Delayed window. The same 107 instances have nonzero restricted teacher mass for all rescoring comparisons. Each checkpoint scores this identical domain against the identical teacher. D_{joint} is joint KL, D_{marg} its parent-marginal term, and D_{cond} its teacher-weighted within-parent term. They satisfy

$D_{\text{joint}} = D_{\text{marg}} + D_{\text{cond}}$ (Appendix B.1). Table 15 retains fuller precision than Figure 2(a). This diagnostic does not freeze the candidates used by the separate native-search audit.

Table 15: **Fixed-candidate KL decomposition.** KL values are means in nats over the same 107 active instances of a fixed 128-instance window. All models rescore identical frozen L1L2 candidates.

Training objective	D_{joint}	D_{marg}	D_{cond}
H	7.9182	2.7592	5.1590
$H + \mathcal{L}_{\text{cond}}$	4.5658	2.6324	1.9334
$H + \mathcal{L}_J$	4.2181	2.2848	1.9333

E.3 Constrained-Beam Cutoff Geometry

We use the cutoff margin defined in Equation 51. Actual survival uses the retained-set indicator $F_{i,2}$, not a margin-sign surrogate (Appendix B.4). We trace the same non-chunked $B = 500$ decoder, trie, and 5,000 ordered Delayed instances as primary evaluation. Observer-on/off equality of complete ordered outputs is verified on the first eight instances per model.

Rank, margin, and immediate-transition statistics use common-L1 subsets of 3,428 and 3,434 instances, with all paired fields present. Table 16 reports mean margin changes and 10,000-draw sample-ID bootstrap CIs on these subsets. For $H \to H + \mathcal{L}_{\text{cond}}$, within-parent L2 rank improves

Table 16: **Constrained L1L2 cutoff geometry on Delayed.** Immediate L1L2 transitions, rank statistics, and margin statistics condition on target L1 survival under both models. Mean margin changes and their 95% paired-bootstrap intervals are in nats.

Transition	Common L1	Immediate L1L2 R/Reg.	Mean Δm_2	95% CI
$H \to H + \mathcal{L}_{\text{cond}}$	3,428	1,065/47	+1.3289	$[1.2503, 1.4066]$
$H + \mathcal{L}_{\text{cond}} \to H + \mathcal{L}_J$	3,434	128/92	-0.0010	$[-0.0196, 0.0177]$

from mean/median 209.14/63 to 84.61/3. Global joint rank improves from 83,044.60/13,438.5 to 37,739.30/523. For $H + \mathcal{L}_{\text{cond}} \to H + \mathcal{L}_J$, within-parent paired rank change has mean -0.231 and median 0, while global rank improves by 2,326.27 on average, with median paired improvement 5. Paired medians are computed from per-instance differences, not differences of endpoint medians. Cutoff sign crossings match immediate rescue/regression events in both comparisons. The joint comparison has net rescues without an average cutoff-margin lift.

E.4 Final-Path and Item-Credit Realization

Table 17 traces final-path transitions on all 5,000 instances, rather than the common-L1 subset. R and Reg. denote false-to-true rescues and true-to-false regressions under the $C_{i,d}$ indicators.

Table 17: **Final-path realization on Delayed.** Counts on all 5,000 instances. The last column is the intersection of immediate L1L2 rescues and final Full-SID rescues.

Transition	L1L2 R	L1L2 Reg.	Full-SID R	Full-SID Reg.	Immediate L1L2 R $\cap$ Full-SID R
$H \to H + \mathcal{L}_{\text{cond}}$	552	66	314	75	207
$H + \mathcal{L}_{\text{cond}} \to H + \mathcal{L}_J$	170	50	131	28	13

Among immediate L1L2 rescues, 380 and 27 also remain final-path L1L2 rescues, while 207 and 13 also satisfy Full-SID rescue. These are intersections of two transition events on the same instance, not causal conversion rates. Full-SID hit counts are 191, 430, and 533 for H, $H + \mathcal{L}_{\text{cond}}$, and $H + \mathcal{L}_J$, consistent with the $314 - 75$ and $131 - 28$ net increases.

For $H \to H + \mathcal{L}_{\text{cond}}$, 264 of 314 Full-SID rescues receive zero Expected Item credit because earlier SID groups exhaust the fixed item budget. Target-SID collision sizes are small (median 1, P90 2). Recovered and lost Full-SIDs contribute $+50$ and -44 credits, while realization changes among common hits contribute -10, yielding a net -4 credits over 5,000 instances.

F PUBLIC-DOMAIN TRANSFER

F.1 SHARED PROTOCOL AND DOMAIN-SPECIFIC DATA

Amazon Books and Home follow the same public-data pipeline, with separately constructed catalogs, ContentSID mappings, and recommender initializations. Public item metadata are deterministically serialized and encoded using Qwen3-Embedding-8B's 512-dimensional Matryoshka representation. Within each domain, we use the same quantizer construction procedure as Appendix C.2, but fit separate codebooks on that domain's public-content embeddings, with 4,096 codes per layer. The embedding model is used only for ContentSID construction, not recommender training or inference. Table 18 summarizes the domain-specific assets and matched instance and update budgets.

Table 18: **Public-domain assets and evaluation populations.** The two domains use separate assets and samples but equal counts of training instances, student updates, and evaluation instances.

Quantity	Books	Home
Valid item embeddings	4,448,123	3,735,555
Excluded empty-text items	58	29
Quantizer fitting items	3,539,053	2,617,352
Sampled training instances	3,374,080	3,374,080
Global batch size	1,024	1,024
Student updates	3,295	3,295
Full validation instances	1,046,262	6,137,587
Evaluated validation instances	1,046,262	1,046,262
Validation selection	Complete set	Fixed subsample

Both domains use a TIGER-style T5 encoder–decoder with codebook-initialized SID embeddings and per-level 4,096-code heads, without industrial Stage-1/Stage-2 content pretraining. A target-independent hash ranking of sample IDs selects 3,374,080 training instances per domain. For Books, the selector uses SHA-256 over global sample IDs from the full 16,874,820-instance training pool. All arms within a domain share the selected instances, source order, initialization, optimizer, training seed, and 3,295-update budget, corresponding to 3,295 complete global batches of 1,024.

Evaluation uses exact constrained beam search with $B = 500$ and the fixed item-expansion budget $K_{\text{item}} = 500$. Books uses its complete validation set. Home instead uses a target-independent subsample drawn from its 6,137,587-instance validation set using seed 20260805. This subsample contains 1,046,262 instances, matching the Books evaluation size, and its membership and order are fixed across all Home arms. Every reported Home arm evaluates the complete selected sample, with no short-beam or unknown-target cases.

F.2 BEHAVIORAL SUPERVISION AND JOINT-TEACHER CONSTRUCTION

In both domains, A0 uses AR, A1 adds BPA, and A2 adds dense BND. The public experiments do not use the industrial H-BND hard-example mask. We therefore write

$$\mathcal{L}_{\text{Core}} = \mathcal{L}_{\text{A2}} + 0.10\mathcal{L}_J,$$

and

$$\mathcal{L}_{\text{Safety}} = \mathcal{L}_{\text{A2}} + 0.10\mathcal{L}_J + 0.05\mathcal{L}_B + 0.05\mathcal{L}_A.$$

Core and safety-augmented JP-BND start independently from the same domain-specific initialization as A2, rather than from the trained A2 checkpoint.

Behavior graphs are constructed separately from each domain's sampled training instances using the same construction procedure. A BLAKE2b hash splits users into disjoint halves. Repeated history–target pairs are deduplicated within instances, and self-transitions are removed. An edge is retained when it appears in both halves and is supported by at least two distinct users. Its stable weight is the smaller half-count. We apply the degree correction in Equation 21, with degrees summing stable weights, and no PMI, neighbor-count pruning, recency weighting, temperature, or additional frequency threshold.

Within each domain, the frozen sampled-A2 checkpoint after 3,295 updates provides the target-free width-500 L1 reference frontier $\mathcal{F}_i^{\text{A2}}$. The offline joint teacher aggregates positive behavioral

mass over legal L1L2 prefixes below that frontier. A2 itself is trained without a joint teacher and supplies the reference frontier, not the JP-BND student's initialization. During training, the student recomputes its own width-500 L1 frontier and expands all legal L2 children. The teacher is restricted and renormalized on this candidate set. Legal rivals with zero teacher mass remain in the student normalization, and parents outside the student frontier are not inserted from the teacher or target.

The Books construction audit records 38,759,081 unique directed pairs, 358,587 stable pairs, and 113,282 active nodes. Local L1 and target-parent L2 teacher coverage are 65.51% and 19.78%, respectively. Joint support is nonempty on 2,105,535 training instances (62.40%). Among supported instances, the represented-parent mean/median is 48.69/14 and the positive-prefix mean/median is 117.52/18. Singleton support occurs on 12.59% of supported instances.

A2 reference training and offline teacher materialization are pipeline prerequisites in both domains. The matched budget counts student updates, not total pipeline computation. Inference uses only the trained student with the fixed mapping, trie, and item-expansion policy.

F.3 Cross-Domain Transfer Estimates and Diagnostics

Table 3 reports absolute endpoints for both domains. Table 19 reports matched paired comparisons on the same 1,046,262 evaluation instances within each domain, using exact sample-ID joins and 10,000 paired-bootstrap draws. Differences are computed before rounding.

Table 19: **Public-domain paired comparisons.** Differences are percentage points with 95% paired-bootstrap confidence intervals.

Domain	Comparison	Full-SID	Expected Item
Books	Core JP-BND − A2	$+0.153$ $[0.131, 0.174]$	$+0.077$ $[0.062, 0.091]$
Books	Safety-augmented JP-BND − A2	$+0.027$ $[0.006, 0.048]$	$+0.069$ $[0.056, 0.083]$
Books	Safety-augmented JP-BND − Core JP-BND	-0.126 $[-0.146, -0.106]$	-0.007 $[-0.021, 0.007]$
Home	Core JP-BND − A2	$+0.220$ $[0.196, 0.243]$	-0.052 $[-0.072, -0.032]$
Home	Safety-augmented JP-BND − A2	$+0.164$ $[0.142, 0.186]$	$+0.101$ $[0.083, 0.119]$
Home	Safety-augmented JP-BND − Core JP-BND	-0.056 $[-0.079, -0.032]$	$+0.153$ $[0.134, 0.173]$

Core JP-BND improves Full-SID over A2 in both domains. On Books, it also improves Expected Item, while the safety-augmented variant gives back 0.126 Full-SID points relative to core with no resolved Expected Item difference. Home shows a different path–item trade-off: core gains 0.220 Full-SID points but loses 0.052 Expected Item points, whereas the safety-augmented variant remains above A2 on both endpoints. Relative to core, it gives back 0.056 Full-SID points while gaining 0.153 Expected Item points. Thus the auxiliaries do not increase the core path-recovery gain in either domain, but can shift its realization into item-level credit under the fixed expansion policy.

The final-path transitions show the same path-level pattern. Core JP-BND produces 9,734 L1L2 rescues versus 7,487 regressions on Books and 10,152 versus 8,735 on Home. At Full-SID, the corresponding counts are 7,081 versus 5,481 and 9,174 versus 6,877. These counts measure coverage changes among returned paths rather than immediate beam-survival transitions. On Books, 6,514 L1L2 rescues are also Full-SID rescues, and 885 newly recovered Full-SIDs receive positive Expected Item credit. On Home, Expected Item credit increases on 5,415 instances and decreases on 5,966, consistent with the negative mean item-level change of core.

Additional Books diagnostics support the same conclusion under alternative resampling and search views. A user-cluster bootstrap gives a core-minus-A2 Full-SID gain of $+0.153$ points with 95% CI $[0.132, 0.175]$. An independent 5,000-instance beam-search audit contains 2,598 common-L1 instances, with 48 cutoff-sign rescues versus 27 regressions, no zero-margin ties, and a mean cutoff-margin change of $+0.0628$ natural-log units.

The public experiments test transfer of the joint-prefix objective but do not include the matched conditional control, so marginal and conditional effects are not separately identified in these domains. User-sequence indices are ordinal rather than wall-clock timestamps, and the public evaluations do not replicate the industrial temporal-generalization setting.

Table 20: **Planned release artifacts and their roles.** The public scope includes Books and Home.

Artifact	Role
Books and Home data recipes	Reconstruct domain-specific interaction data and frozen training/evaluation selection procedures and sample-ID manifests.
Fixed SID assets	Separate item-to-SID mappings, legal tries, and collision-group sizes for each domain's training and evaluation search space.
Initialization assets	Domain-specific T5 configurations, codebook initialization code and states for Books and Home, together with industrial SID tokenizer files and Stage-2 checkpoint where available.
Training code and configurations	Stage-1/Stage-2 recipes and the reported AR, BPA, BND, H-BND, matched conditional, BEAR-style, core JP-BND, and optional objectives.
References and teacher builders	Industrial A0 exposure-mask and H-BND support builders, separate Books/Home sampled-A2 frontier references, and behavioral-graph and local/joint-teacher materialization code.
Endpoint checkpoints	Reported A0, H-BND, sampled-A2, and core JP-BND checkpoints where available, with endpoint/reference roles distinguished from shared initializations.
Upstream content assets and recipes	Item embeddings, metadata serialization and embedding recipes, and quantizer fitting/inference code, configurations, and centroids.
Evaluation and provenance	Exact constrained $B = 500$ decoding, item expansion, paired statistics, objective-contract tests, diagnostics and figure data, split/sample manifests, environment versions, seeds, and artifact hashes.

F.4 Minimal Public Reconstruction

Reconstruction in either domain requires its interaction data, fixed ContentSID mapping, legal trie, and frozen training and evaluation manifests. Initialize T5 from the domain-specific codebook state and train dense A2 under the matched budget. Freeze that checkpoint to materialize the joint teacher. Core and safety-augmented JP-BND start independently from the same initial state under the objectives in Appendix F.2, rather than continuing from trained A2 parameters.

Evaluate on the full Books validation set or the frozen Home subsample using constrained $B = 500$ decoding and $K_{\text{item}} = 500$ expansion. Loading fixed SID assets avoids rerunning the embedding model or quantizer. Plans for releasing these assets or their upstream reconstruction recipes are described in Appendix G.1.

G Reproducibility and Release Plan

G.1 Frozen Contracts and Planned Release

Appendix A.4 specifies information access. Appendices C and F document the industrial and public training and evaluation protocols. We plan to release code and reproducibility artifacts after packaging and verification. Table 20 summarizes the intended scope.

G.2 Industrial Reproducibility Scope

For the industrial setting, we plan to release the implementation, model configurations, non-sensitive manifests, aggregate results, and available model and item-side assets needed to inspect the training and evaluation pipeline. A released Stage-2 checkpoint would provide the shared initialization used by all Stage-3 arms, while Stage-3 A0 remains an endpoint/reference artifact rather than the shared initialization.

The industrial benchmark itself depends on proprietary interaction histories and replay populations. We therefore use Books and Home as the public retraining path: both domains reconstruct the full JP-BND training pipeline from public interactions with independently built ContentSID assets, matched training budgets, and the same constrained decoding procedure, without industrial Stage-1/Stage-2 pretraining.