ULBS: UNIFORM LOCAL BEAM SURVIVAL FOR SEMANTIC-ID GENERATIVE RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative recommendation formulates recommendation as conditional generation, offering an alternative to conventional discriminative retrieval and ranking paradigms. However, SID-based generative recommendation relies on constrained beam search at inference, while standard teacher-forcing training fails to account for competition among candidate prefixes. Under this discrepancy, target items with competitive scores can still be pruned prematurely during prefix search. Recent methods partially address this train–inference mismatch but do not directly supervise the cross-parent beam boundary that determines whether a target prefix survives. To bridge this gap, we propose **Uniform Local Beam Survival (ULBS)**, a beam-aligned training objective that optimizes intermediate target-prefix survival against this boundary. At each semantic step, ULBS obtains the beam cutoff from a detached constrained-beam trace and applies a smooth margin loss between the cumulative target-prefix score and this cutoff. This provides uniform supervision across intermediate survival events with localized credit assignment, without adding model parameters or inference-time overhead. Extensive experiments on Amazon benchmarks with LLM backbones demonstrate that ULBS outperforms strong baselines, effectively alleviating the beam admission bottleneck while revealing within-beam ranking as a critical frontier for generative recommendation.

1 INTRODUCTION

Generative recommendation departs from traditional large-scale item classification by framing recommendation as autoregressive generation over semantic identifiers (SIDs) (Rajput et al., 2023; Kong et al., 2025). This formulation significantly compresses the target vocabulary and facilitates structured generalization across items.

However, a fundamental mismatch undermines its deployment: standard training relies on teacher forcing to optimize next-token likelihoods, whereas inference employs constrained beam search over an item prefix tree. At each decoding step, beam search retains only the B highest-scoring partial sequences and expands them to form the candidates for the next step. Since beam search irrevocably removes branches outside the top-B set at each step, a target item can have a competitive full-sequence score yet be prematurely pruned if an intermediate prefix falls outside the beam.

Several recent methods attempt to mitigate this train-inference mismatch, but differ in how directly they account for beam-search competition. For instance, MSL evaluates each gold token against a restricted set of legal tokens derived from an item trie to prevent invalid tokens, without explicitly modeling beam competition (Wang et al., 2025). Targeting beam-search behavior, BEAR optimizes each gold token against a prefix-local top-B boundary within the same parent (Yang et al., 2026b), while APAO adaptively reweights prefix-level objectives to emphasize high-loss prefixes (Yu et al., 2026). Crucially, none of these objectives directly supervises the cross-parent beam-admission boundary that arises during inference, which is determined by the B-th highest-scoring legal prefix after competing branches from all parent paths are expanded (Figure 1). Beam admission and within-beam ranking are therefore distinct optimization targets. A target must first survive this cross-parent competition before its relative ranking among surviving candidates becomes relevant.

This raises a key question: **can training directly supervise the cross-parent admission boundary induced by beam search, without differentiating through the search process or altering inference?** We answer this question with Uniform Local Beam Survival (ULBS), a beam-aligned training

objective that directly supervises intermediate prefix survival using the width-B beam boundary. At each semantic step, ULBS identifies the beam cutoff from a detached constrained-beam trace and applies a softplus margin loss between the cumulative gold-prefix score and this boundary. For local credit assignment, ULBS assigns local credit strictly to the current decision position by stopping gradients through earlier gold-prefix states and then weights all boundary events uniformly across semantic steps. Moreover, ULBS introduces neither additional model parameters nor inference-time decoding overhead.

Across three Qwen backbones, ULBS leads all reported metrics on Industrial and Office and all HR metrics on Musical Instruments. Furthermore, prefix-level and hit/miss analyses show that, while boundary supervision rescues items pruned by baselines and improves beam admission, it can also perturb the relative ordering of candidates already inside the beam. By isolating the admission bottleneck from within-beam ranking, our findings clarify the scope of boundary-level supervision and identify within-beam ranking as a critical frontier for beam-aligned recommendation.

Our main contributions are:

1. We formulate the local beam survival problem for trie-constrained semantic-ID generation and propose ULBS, a smooth boundary objective with **local credit assignment** and **uniform reduction** that anchors training supervision to the realized **cross-parent beam cutoff**, validated against cumulative-credit and adaptive-weighting alternatives.
2. We show that ULBS consistently improves retrieval over strong baselines across multiple LLM backbones on canonical benchmarks, without adding any inference-time decoding mechanism.
3. We disentangle **candidate admission** from **within-beam ranking** through prefix-level diagnostic analyses that explain both the empirical gains and the limits of boundary-level supervision.

2 Background and Related Work

Semantic-ID generation. Generative recommendation models items as discrete semantic identifiers (SIDs) rather than atomic item IDs. TIGER (Rajput et al., 2023) constructs hierarchical SID tuples using residual vector quantization (RQ-VAE) (Lee et al., 2022), framing recommendation as autoregressive generation over semantic code tokens. MiniOneRec (Kong et al., 2025) extends this paradigm to pretrained LLM backbones, combining SID-aware supervised fine-tuning (SFT) and reinforcement-learning post-training with constrained decoding. Following MiniOneRec's SID configuration, we represent each item with a three-token SID. Exact SID retrieval is strictly constrained by a legal item trie during both training and inference in this work.

Search-aware learning and beam objectives. Incorporating search dynamics into sequence training has a long history in NLP. Beam-Search Optimization (BSO) (Wiseman & Rush, 2016) constructs training updates from beam-search states, while continuous relaxations make the otherwise discrete pruning operations differentiable for end-to-end optimization (Goyal et al., 2017). Both require maintaining or differentiating through multiple candidate trajectories at every decoding step. These compute and memory costs scale with vocabulary size and become expensive for large autoregressive models. Lexically constrained decoding constitutes a separate algorithmic layer. Grid Beam Search (Hokamp & Liu, 2017), for instance, modifies beam search to enforce predefined output constraints but does not provide a search-aware training objective.

A related line of work has emphasized the importance of intermediate prefixes in identifier generation. In generative retrieval, RIPOR (Zeng et al., 2023) introduces prefix-oriented optimization, recognizing that accurate relevance at the complete-identifier level alone is insufficient under sequential decoding. More recently, generative recommendation methods have explicitly targeted the mismatch induced by beam pruning. BEAR (Yang et al., 2026b) optimizes each gold token against a prefix-local top-B boundary within the same parent, while APAO (Yu et al., 2026) introduces prefix-level objectives and adaptively emphasizes high-loss prefixes through worst-prefix reweighting. Despite accounting for beam-search behavior, these objectives still rely on local or prefix-level surrogate signals rather than the realized cross-parent beam cutoff. Semantic-ID recommendation makes such direct supervision particularly tractable because its output sequences are short and constrained by a finite item trie.

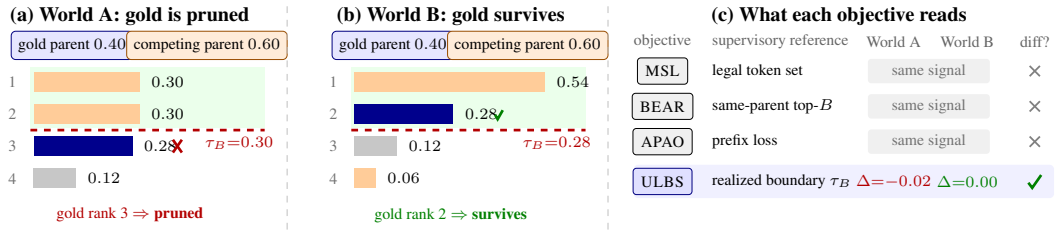

Figure 1: A two-world counterexample with beam width B=2. Holding the gold parent, gold extension score, and same-parent ranking fixed while changing only the rival parent's child scores shifts the realized beam cutoff and flips gold-prefix survival. ULBS directly observes this cutoff, unlike objectives whose supervision is not derived from the live beam state.

ULBS exploits this structure by using beam search only to obtain the detached width-B boundary, rather than differentiating through competing search paths. The resulting boundary serves as a direct training reference for target-prefix survival without changing inference-time decoding.

A reference-axis taxonomy. The fundamental distinction among existing training objectives lies in the supervisory reference against which the gold target is evaluated (summarized in Table 1). Standard masked likelihood (MSL) normalizes over valid trie tokens to exclude invalid continuations from the softmax denominator but remains agnostic to beam membership (Wang et al., 2025). BEAR introduces beam awareness but bounds predictions only against sibling tokens under the same parent (Yang et al., 2026b). APAO shifts supervision to the prefix level through pointwise likelihood and pairwise objectives with sampled negative prefixes (Yu et al., 2026). Crucially, none of these baselines reads the realized beam. In contrast, ULBS anchors intermediate target-prefix scores directly to the realized cross-parent beam cutoff across all active candidate branches.

Table 1: Supervisory references used by the compared objectives.

Method	Object	Reference	Granularity	Beam trace?
MSL	gold-token score	legal token set	token	No
BEAR	gold-token rank	same-parent top-B	token	No
APAO	prefix likelihood	pointwise / sampled negatives	prefix	No
ULBS	prefix cumulative score	cross-parent beam cutoff	prefix	Yes

A minimal counterexample: why cross-parent boundaries matter. To illustrate why parent-local references are insufficient, Figure 1 presents a synthetic two-world counterexample (width B=2, one expansion step). In both World A and World B, the gold prefix retains an identical cumulative score (0.28), an identical gold parent score (0.40), and ranks first among its same-parent siblings. However, varying the distribution of child scores under the rival parent changes the global survival threshold: the realized beam boundary shifts from $\tau_B = 0.30$ in World A to $\tau_B = 0.28$ in World B. Consequently, the gold prefix is pruned in World A but admitted into the beam in World B.

MSL and BEAR receive identical local references in the two worlds by construction. APAO does not observe the realized beam boundary either, because its pointwise variant depends only on the positive-prefix likelihood, while its pairwise variant compares against sampled negatives rather than the live beam cutoff. Thus, none of these objectives is directly conditioned on the survival transition illustrated here. In contrast, ULBS directly tracks the realized boundary τ_B, yielding different boundary gaps when the same gold prefix is pruned in World A ($\Delta = -0.02$) and admitted at the cutoff in World B ($\Delta = 0.00$). Although synthetic, this mismatch also occurs in trained models during inference (Section 4, Table 3).

Decoupling admission from within-beam ranking. The ultimate output of a constrained generative recommender is a ranked set, not an isolated token. Notably, theoretical results show that even a Bayes-optimal generative retriever can suffer poor top-k recall under constrained beam search (Wu et al., 2025), indicating that decoding itself can impose a retrieval bottleneck beyond model estimation error. This observation motivates us to decompose beam-search outcomes into two distinct

components, namely *candidate admission* (whether the complete target SID survives into the top-B beam) and *within-beam ranking* (its relative position among surviving candidates). This distinction is critical. An objective that improves intermediate prefix survival can increase target retrieval (HR@K), while changes in the relative scores of already-admitted items can still affect their within-beam ordering and hence NDCG@K. This decomposition provides the analytical foundation for our diagnostic study in Section 4.

3 Uniform local beam survival

3.1 Problem Setup and Formulation

For each recommendation domain d, let $\mathcal{Y}_d$ denote the catalog of legal item SIDs. Let $x \in \mathcal{X}_d$ denote a recommendation context in domain d, instantiated in our experiments as a chronologically ordered sequence of SIDs for the user's historical interactions. A candidate item is represented by a legal SID $y = (y_1, \ldots, y_T) \in \mathcal{Y}_d$, and the gold target item is denoted by $y^* = (y_1^*, \ldots, y_T^*) \in \mathcal{Y}_d$. For every position $t \in \{1, \ldots, T\}$ and legal prefix $y_{<t}$, the domain-specific item trie defines the legal continuation set $\mathcal{V}_d(y_{<t}) = \{v : \ y_{<t} \| v \text{ is a prefix of some } y' \in \mathcal{Y}_d\}$. During inference, constrained beam search of width B retains the B highest-scoring valid prefixes at each step. For any $y_{\leq t}$, we define its cumulative log-probability as $S_t(x, y) = \sum_{i=1}^{t} \log P_\theta(y_i \mid x, y_{<i})$. Let $\tau_t(x)$ denote the score of the B-th highest-scoring legal prefix after expanding all prefixes in the live beam at depth $t-1$. When fewer than B legal extensions exist, we use the lowest extension score as the boundary by convention.

At test time, target retrieval depends on a sequence of stepwise admission events: the partial gold prefix $y^*_{\leq t}$ must remain in the width-B beam at every intermediate step $t < T$. The hard pruning indicator at position t is given by $v_t = \mathbb{1}[\Delta_t < 0]$, where

$$\Delta_t(x) = S_t(x, y^*) - \tau_t(x) \tag{1}$$

denotes the beam-admission gap between the cumulative gold-prefix score and the realized beam cutoff. We omit the dependence on x when clear from context. Because v_t is discontinuous and has zero gradient almost everywhere with respect to the model scores, it is unsuitable for direct gradient-based optimization. To address this, we introduce the following smooth and margin-based survival surrogate:

$$\ell_t^{\text{surv}} = \text{softplus}\left(\frac{m - \Delta_t}{\gamma}\right) \tag{2}$$

where $m > 0$ specifies a target safety margin above the cutoff and $\gamma > 0$ is the softplus temperature (fixed to $\gamma = 1$ in all experiments). For any $m > 0$, the pruning indicator is upper-bounded by the scaled surrogate:

$$v_t \ \leq \ \frac{\gamma}{m} \text{softplus}\left(\frac{m - \Delta_t}{\gamma}\right) \tag{3}$$

Thus, Eq. 2 is a smooth and margin-aware convex surrogate in the scalar gap Δ_t. After scaling by γ/m, it upper-bounds the discrete pruning indicator.

Admission versus within-beam ranking. Let $A(x) = \mathbb{1}[y^* \in \text{Beam}_T]$ denote whether the complete gold SID is retained in the final width-B beam, and, conditional on $A(x) = 1$, let $R(x) \in \{1, \ldots, B\}$ denote the rank of y^* within the final beam. With a single gold item per context, Hit Rate depends only on final beam admission, $\text{HR@}B = \mathbb{E}_x[A(x)]$, whereas ranking metrics such as $\text{NDCG@}B$ depend jointly on admission and the target's within-beam rank, $\text{NDCG@}B = \mathbb{E}_x\left[\frac{A(x)}{\log_2(R(x)+1)}\right]$. Eq. 2 targets intermediate beam-survival events and therefore provides admission-oriented supervision, while leaving within-beam ranking implicit.

3.2 Local Credit and Uniform Reduction

While the forward boundary gap Δ_t in Eq. 1 depends on the cumulative gold-prefix score $S_t(x, y^*)$, allowing gradients to backpropagate recursively through all preceding prefix tokens $y^*_{<t}$ would repeatedly penalize early tokens for downstream survival failures. To ensure localized credit assignment,

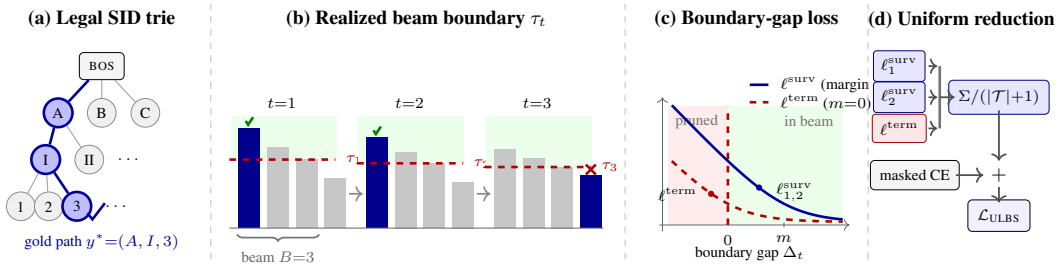

Figure 2: Overview of ULBS on a width-3 case study. A detached constrained-beam trace provides the realized cutoff τ_t along the gold path. Intermediate positions use margin-based survival losses, the complete SID uses a zero-margin terminal loss, and all boundary events are uniformly averaged with masked cross-entropy.

ULBS stops gradients through the cumulative score of the preceding gold prefix when evaluating position t:

$$S_t^{\text{local}}(x, y^*) = \text{sg}[S_{t-1}(x, y^*)] + \log P_\theta(y_t^* \mid x, y_{<t}^*), \tag{4}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator. Consequently, the survival event at position t contributes gradient only through the current conditional log-probability, rather than through earlier token-level score terms.

In addition to intermediate survival events, we incorporate a terminal calibration term ℓ^{term} at the final sequence position T. Because no further SID token remains after step T, this term does not represent an intermediate survival decision but rather calibrates the complete gold SID against the final beam boundary with zero margin: $\ell^{\text{term}} = \text{softplus}(-\Delta_T/\gamma)$.

Averaging all valid boundary events uniformly and combining them with standard masked cross-entropy yields the full ULBS training objective:

$$\mathcal{L}_{\text{ULBS}} = \mathcal{L}_{\text{CE}} + \frac{\alpha}{|\mathcal{T}(y^*)| + 1} \left(\sum_{t \in \mathcal{T}(y^*)} \ell_t^{\text{surv}} + \ell^{\text{term}} \right). \tag{5}$$

where $\mathcal{T}(y^*) \subseteq \{1, \ldots, T-1\}$ denotes the set of intermediate semantic positions at which the gold prefix admits at least two legal continuations in the domain-specific trie, and $\alpha \geq 0$ controls the contribution of the beam-survival objective. ULBS operates only during training, introducing no additional model parameters and leaving inference-time decoding unchanged. Figure 2 illustrates the complete workflow.

3.3 Target-Conditioned Trace and Gold Reset

Computing Eq. 2 requires obtaining the realized beam cutoff $\tau_t(x)$ during training. We obtain this boundary using a detached constrained-beam trace that does not construct gradient paths through competing beam candidates. However, a technical challenge arises if the gold prefix $y_{<t}^*$ is pruned from the live beam at an intermediate depth $t < T$: in an unconstrained trace, the gold path would vanish, preventing the evaluation of subsequent survival losses at depths $> t$.

To retain a well-defined training signal at later semantic positions, ULBS employs a training-only gold reset within the training trace. If the gold prefix falls below the width-B threshold at step t, it receives a temporary selection-only boost that keeps the gold prefix in the training beam for the next conditional expansion. The boost affects selection only and is removed from the carried cumulative score before the next boundary gap is evaluated.

We emphasize that the reset is used only within the training trace: it does not modify the deployed beam-search procedure, nor does it imply that a prefix pruned during inference can later re-enter the beam. Events after the first falloff are therefore *conditional counterfactuals*, in that they evaluate target-prefix survival under the target-conditioned training trace rather than under the unmodified deployed beam trajectory. While this introduces an exposure-bias limitation analogous to scheduled sampling (Bengio et al., 2015) and DAgger oracle trajectories (Ross et al., 2011), it ensures that every non-trivial semantic token receives well-defined gradient supervision.

Algorithm 1 Per-example ULBS training

Require: Context $x \in \mathcal{X}_d$, gold SID $y^* = (y_1^*, \ldots, y_T^*) \in \mathcal{Y}_d$, domain-specific item trie, beam width B, margin m, temperature γ, loss weight α
Ensure: Total loss $\mathcal{L}_{\text{ULBS}}$
1: Run a detached and target-conditioned constrained-beam trace to obtain $\{\tau_t(x)\}_{t=1}^{T}$, applying the training-only gold reset after any intermediate falloff.
2: **for** each $t \in \mathcal{T}(y^*)$ **do**
3: Compute $S_t^{\text{local}}(x, y^*)$ via Eq. 4, $\Delta_t \leftarrow S_t^{\text{local}}(x, y^*) - \tau_t(x)$, and $\ell_t^{\text{surv}} \leftarrow \text{softplus}\Big(\frac{m-\Delta_t}{\gamma}\Big)$.
4: **end for**
5: Compute $S_T^{\text{local}}(x, y^*)$ via Eq. 4, $\Delta_T \leftarrow S_T^{\text{local}}(x, y^*) - \tau_T(x)$, and $\ell^{\text{term}} \leftarrow \text{softplus}\Big(-\frac{\Delta_T}{\gamma}\Big)$.
6: Compute $\mathcal{L}_{\text{ULBS}} \leftarrow \mathcal{L}_{\text{CE}} + \frac{\alpha}{|\mathcal{T}(y^*)|+1}\Big(\sum_{t\in\mathcal{T}(y^*)} \ell_t^{\text{surv}} + \ell^{\text{term}}\Big)$.
7: **return** $\mathcal{L}_{\text{ULBS}}$

3.4 Why Uniform Local Credit?

The design of ULBS is guided by two optimization choices: localized credit assignment and uniform event weighting. We motivate both choices from an optimization perspective:

Local versus cumulative credit. Under cumulative credit assignment, the survival loss at depth t backpropagates through all token-level log-probability terms contributing to the cumulative gold-prefix score. Consequently, earlier SID decisions accumulate gradient contributions from multiple downstream survival events. This induces gradient concentration on the coarsest semantic tokens, often distorting early representations even when a boundary violation occurs at a later semantic position. By truncating the backward path through earlier cumulative score terms via Eq. 4, local credit assigns each survival event to the current semantic decision and avoids repeated downstream credit on earlier SID positions.

Uniform versus adaptive weighting. APAO (Yu et al., 2026) adaptively assigns greater emphasis to high-loss prefixes through worst-prefix reweighting. Our mechanism controls (Section 4.4) show that adaptive weighting underperforms uniform weighting across both credit schemes and domains. This coupling can over-emphasize high-loss prefixes and underweight already-safe ones. Uniform reduction ($\frac{1}{|\mathcal{T}(y^*)|+1}$) treats each competitive boundary event equally, providing a simple and non-adaptive operating point that avoids coupling event weights to their instantaneous losses.

3.5 Relation to Search-Aware Objectives

ULBS differs fundamentally from both classical and contemporary search-aware methods:

- **Versus Beam-Search Optimization (BSO) (Wiseman & Rush, 2016):** BSO constructs margin violations between the gold prefix and a competing beam candidate and backpropagates through both paths. In contrast, ULBS computes the beam boundary τ_t in a detached forward trace without backpropagating through rival beams. This yields a one-sided margin in which the realized beam cutoff acts as a detached reference, so gradients flow only through the gold-side score rather than through competing beam paths.
- **Versus BEAR (Yang et al., 2026b):** For comparison, we adapt BEAR, originally formulated for autoregressive generation of natural-language item descriptions, to SID continuations. Implementation details are provided in Section 4. Under this adaptation, BEAR compares each gold token against a prefix-local top-B threshold over legal continuations under the same gold prefix. As illustrated in Figure 1, such parent-local constraints do not capture changes in the global cutoff caused by cross-parent competition. To our knowledge, ULBS is the first objective to anchor supervision to the realized cross-parent beam cutoff τ_t obtained from the detached beam trace.

Algorithm 1 summarizes the complete per-example training procedure of ULBS.

Table 2: Five-seed mean recommendation performance ($\times 100$) across three Amazon domains and three LLM backbones. Best results within each backbone-domain-metric setting are bolded.

	Industrial				Office				Musical			
Method	NDCG @5	NDCG @10	HR @5	HR @10	NDCG @5	NDCG @10	HR @5	HR @10	NDCG @5	NDCG @10	HR @5	HR @10
Qwen2.5-7B												
MSL	5.3716	6.2111	6.8784	9.4992	5.5851	6.3251	7.5134	9.5192	5.3582	5.7735	6.4785	7.7838
BEAR-flat	5.0996	5.8847	6.5652	8.9962	5.3675	6.0694	7.2421	9.4287	**5.8236**	**6.1548**	6.6266	7.8839
APAO	4.8940	5.6663	6.1549	8.5903	5.6430	6.3136	7.5545	9.6383	5.6801	6.0949	6.7187	8.0040
ULBS	**6.2735**	**6.9602**	**7.6594**	**9.8125**	**5.8930**	**6.4737**	**7.8504**	**9.6547**	5.4100	5.8641	**6.8428**	**8.0320**
Qwen3-1.7B												
MSL	6.4774	7.4071	8.1697	11.0670	6.0658	7.1015	8.8437	12.0701	5.3231	5.8600	6.9603	8.6420
BEAR-flat	6.3491	7.2540	7.9418	10.7729	6.2521	7.2841	9.1930	11.8578	**5.8050**	**6.3223**	7.1939	8.7955
APAO	6.2841	7.1921	7.8535	10.6699	6.1930	7.1630	9.0286	12.0427	5.6861	6.2497	7.0137	8.7621
ULBS	**6.8173**	**7.7473**	**8.5080**	**11.3905**	**6.5304**	**7.3881**	**9.5287**	**12.7346**	5.7787	6.2868	**7.2139**	**8.8021**
Qwen3.5-2B												
MSL	6.3157	7.1734	8.4197	11.0891	6.0889	7.0121	8.7752	11.6454	5.7716	6.2613	7.2873	8.8288
BEAR-flat	5.9138	6.7253	7.7653	10.3169	6.4107	7.3270	9.0766	11.9194	6.2057	6.7088	7.2272	8.7888
APAO	6.3033	7.3021	8.2874	11.4126	6.5166	7.4437	9.1588	11.9948	**6.7138**	**7.2187**	7.6009	9.1625
ULBS	**6.7189**	**7.6790**	**8.6256**	**11.6038**	**6.7725**	**7.5724**	**9.5082**	**12.0222**	6.5802	7.1169	**7.7010**	**9.3760**

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Datasets and backbone. We evaluate on Amazon Reviews 2018 (Ni et al., 2019), following the data processing and SID construction protocol of MiniOneRec (Kong et al., 2025). Our main evaluation uses the *Industrial* and *Office* domains, with *Musical Instruments* included as an additional stress-test domain.

We evaluate all methods with three LLM backbones: QWEN2.5-7B (Qwen Team, 2024), QWEN3-1.7B, and QWEN3.5-2B. This multi-backbone evaluation tests whether the effect of ULBS generalizes across model scales and generations. Across backbones, all methods are fine-tuned using rank-8 LoRA (Hu et al., 2022) for three epochs with a global batch size of 64 (micro-batch size 4). We maintain the same SID assignments, item trie, and trie-constrained beam search decoder across all methods. Checkpoints are evaluated on the held-out test set with a deployed beam width of $B = 10$. Unless otherwise specified, we use $B = 10$, $\alpha = 0.3$, $m = 0.2$, and softplus temperature $\gamma = 1$ throughout the experiments. All ablation and efficiency analyses are conducted with the QWEN2.5-7B backbone unless otherwise specified.

Compared baselines. We benchmark ULBS against three representative training objectives under matched model and decoding constraints: (1) **MSL** (Wang et al., 2025): masked softmax likelihood restricted to legal next-token continuations; (2) **BEAR-flat** (Yang et al., 2026b): our semantic-ID adaptation of BEAR, enforcing its prefix-local top-B token constraint over legal SID continuations. The SID-adapted BEAR retains the released sigmoid surrogate together with its temperature and regularization-weight settings; (3) **APAO** (Yu et al., 2026): prefix-level optimization with adaptive worst-prefix reweighting.

Metrics. We report Hit Rate (HR@k) and Normalized Discounted Cumulative Gain (NDCG@k) at cutoffs $k \in \{5, 10\}$ under the standard single-target next-item evaluation protocol. For a pair (x, y^*), HR@k indicates whether the gold SID appears in the top-k legal items returned by constrained beam search, while NDCG@k additionally discounts its rank within the returned list. To account for training stochasticity, we run each main-comparison configuration with five random seeds, $\mathcal{S} = \{7, 42, 123, 2026, 31415\}$.

4.2 PRIMARY RESULTS

Table 2 reports five-seed mean performance across three LLM backbones on the two main domains, *Industrial* and *Office*, together with the additional *Musical Instruments* stress-test domain.

Across all three backbones, ULBS achieves the highest mean HR@5, HR@10, NDCG@5, and NDCG@10 on both *Industrial* and *Office*, leading all 24 backbone-domain-metric comparisons on the two main domains. This consistency indicates that the gains are not specific to a particular Qwen backbone or retrieval cutoff. On *Musical Instruments*, the pattern is different but consistent across backbones: ULBS achieves the highest HR@5 and HR@10 with all three backbones, while it does not attain the best NDCG@5 or NDCG@10.

4.3 Local Admission vs. Realized Survival

To examine whether the failure mode illustrated in Figure 1 occurs in practice, we conduct a validation-only pilot on 50 fixed queries per domain using seed-42 checkpoints (Table 3). Across all model-domain pairs, we observe cases in which the gold token passes the same-parent top-B criterion but the corresponding gold prefix still falls below the realized cross-parent beam cutoff τ_t under $B = 10$. These local-pass/global-falloff events account for 1.4%-6.7% of eligible decisions, showing that the local-global mismatch arises in actual decoding traces. Using the raw cumulative gold-prefix score alone as a score-only falloff predictor yields AUCs of 0.773-0.894, indicating that prefix likelihood is informative but does not encode the realized cross-parent boundary.

Table 3: Validation pilot for local-global beam divergence under $B = 10$. Eligible events are natural pre-reset semantic decisions whose gold parent has at least 10 legal continuations. Div. denotes local-pass/global-falloff events.

	Industrial		Office	
Method	Div.	NLL AUC	Div.	NLL AUC
MSL	2/77 (2.6%)	.881	3/71 (4.2%)	.781
BEAR-flat	2/78 (2.6%)	.866	1/72 (1.4%)	.773
APAO	3/71 (4.2%)	.836	5/75 (6.7%)	.813
ULBS	4/73 (5.5%)	.894	1/66 (1.5%)	.843

4.4 Mechanism Controls and Component Ablations

Credit assignment and weighting schemes. ULBS combines localized credit assignment with uniform event reduction. To decouple these two design choices, we evaluate the 2×2 factorial space of credit schemes (Local vs. Cumulative) and weighting policies (Uniform vs. Adaptive) (Table 5). Two findings emerge:

- **Consistent benefit of uniform weighting:** Uniform weighting outperforms its adaptive counterpart under both credit schemes and on both domains. Under local credit, it improves mean NDCG@10 by 0.500 points on Industrial and 0.452 points on Office.
- **Benefit of local credit:** Local credit improves over cumulative credit under both weighting policies, with a particularly large gain on Office. Under uniform weighting, it raises mean NDCG@10 by 0.156 points on Industrial and 0.795 points on Office.

The two controls are therefore complementary: uniform weighting provides consistent gains across credit schemes, while local credit yields additional improvements, particularly on Office. Their combination achieves the highest HR@10 and NDCG@10 on both domains.

Ablation of auxiliary loss terms. We isolate the intermediate survival loss ℓ^{surv} and terminal calibration loss ℓ^{term} through component-off controls (Table 6). On Industrial, removing either term substantially reduces NDCG@10: the survival-only (5.9722) and terminal-only (5.4784) variants both fall below MSL (6.2416), whereas full ULBS reaches 7.3301. On Office, the terminal-only variant achieves the highest HR@10 (10.2548), while full ULBS achieves the highest NDCG@10 (6.7926), showing that combining both terms yields the strongest ranking performance. Because the auxiliary losses are renormalized over the active terms, these controls should not be interpreted as a linear additive decomposition.

Decoder width and hyperparameter sensitivity. Using frozen seed-42 adapters, we evaluate the off-default inference widths $B \in \{5, 20\}$ without retraining, while retaining $B = 10$ as the default protocol (Table 7). At both off-default widths, ULBS achieves the highest NDCG on both Industrial and Office, indicating that its gains are not tied to a single decoding width. We additionally evaluate a 3×3 sensitivity grid over the survival-loss weight $\alpha \in \{0.15, 0.30, 0.60\}$ and margin $m \in \{0.10, 0.20, 0.40\}$ on Industrial (Table 8), where NDCG@10 ranges from 5.729 to 7.535.

Following the fixed experimental protocol, we retain the pre-specified defaults $\alpha = 0.3$ and $m = 0.2$ rather than selecting the best-performing cell post hoc.

4.5 Admission-Ranking Divergence on Musical Instruments

To test the boundary of our core claim, we evaluate all models on the *Musical Instruments* dataset (Table 2).

Unlike on the two main domains, the gains on *Musical Instruments* separate cleanly along the admission-ranking axis. Across all three backbones, ULBS achieves the highest mean HR@5 and HR@10, yet does not attain the best NDCG@5 or NDCG@10. This divergence is consistent with the distinction between target admission and within-beam ranking:

- **Boundary supervision improves beam admission:** ULBS achieves the highest mean HR@10, indicating that intermediate margin enforcement effectively rescues target prefixes from being pruned out of the beam.
- **Within-beam ranking remains distinct:** Despite its consistent HR advantage, ULBS does not achieve the highest NDCG at either cutoff on Musical Instruments for any of the three backbones, showing that improved target admission does not necessarily translate into better ordering among admitted items. On Musical, a higher proportion of target items naturally survive the beam boundary. Consequently, boundary supervision yields diminishing returns for admission while perturbing the delicate log-probability ranking among already-admitted candidates.

These results clarify the scope of the current objective: ULBS directly targets beam admission, whereas improvements in within-beam ordering are not universally guaranteed. Boundary supervision is effective in regimes where beam pruning is the primary bottleneck, but requires complementary ranking constraints in scenarios dominated by intra-beam ordering. Incorporating an explicit ranking-oriented objective therefore represents a natural direction for complementary supervision.

4.6 Computational and Memory Efficiency

A key design advantage of ULBS is that it modifies only the training objective, leaving the inference-time decoder and trie unchanged. We benchmark efficiency on Qwen2.5-7B using 256 fixed test queries with batch size 4 (Table 9):

- **Inference decoding:** ULBS exhibits nearly identical latency and throughput to MSL on both domains, with unchanged peak memory usage (15.1 GiB). Thus, it introduces no additional inference-time mechanism and negligible measured decoding overhead.
- **Training cost:** The detached beam trace increases Industrial training time from 3.5 h to 6.4 h ($1.83\times$). A batch-64 single-step probe (55 vs. 32 s) shows the same peak VRAM, indicating that the additional cost is primarily computational rather than memory-capacity overhead. This overhead is confined to offline training and does not carry over to deployment.

5 Conclusion

In this paper, we proposed Uniform Local Beam Survival (ULBS), a beam-aligned training objective for semantic-ID generative recommendation. By tracking the realized cross-parent beam cutoff through a detached constrained-beam trace and enforcing localized survival margins, ULBS provides beam-survival supervision without differentiating through competing beam paths or modifying the deployed decoder. Across canonical MiniOneRec benchmarks, ULBS achieves the strongest mean performance on the two main domains and remains competitive on the additional Musical Instruments stress test. Mechanism analyses further show that ULBS directly targets this admission mismatch, while within-beam ranking remains a distinct source of error. By demonstrating negligible measured deployment-time overhead alongside higher offline training cost, ULBS establishes beam-aligned training in semantic-ID recommendation, and identifies within-beam ranking as a complementary objective for future work.

Reproducibility statement. To ensure reproducibility, our code, configuration scripts, and evaluation logs are archived and will be publicly released upon publication. Implementation details like hyperparameters are documented in the Appendix.

Ethics statement. This work uses the public Amazon Reviews benchmark and does not collect any new personal data. While our research is purely algorithmic, practitioners deploying generative recommenders should remain mindful of societal risks such as popularity bias amplification and fairness disparities.

AI use statement. Large language models were used to polish text for clarity and assist with literature review for related work. The authors verified all retrieved references, finalized all content, and take full responsibility for the paper.

REFERENCES

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1506.03099.

Junchen Fu, Xuri Ge, Alexandros Karatzoglou, Ioannis Arapakis, Suzan Verberne, Joemon M Jose, and Zhaochun Ren. Differentiable semantic id for generative recommendation. *arXiv preprint arXiv:2601.19711*, 2026.

Kartik Goyal, Graham Neubig, Chris Dyer, and Taylor Berg-Kirkpatrick. A continuous relaxation of beam search for end-to-end training of neural sequence models. *arXiv preprint arXiv:1708.00111*, 2017. URL https://arxiv.org/abs/1708.00111.

Chengcheng Guo, Kuo Cai, Yu Zhou, Qiang Luo, Ruiming Tang, Han Li, Kun Gai, and Guorui Zhou. Promise: Process reward models unlock test-time scaling laws in generative recommendations. *arXiv preprint arXiv:2601.04674*, 2026. URL https://arxiv.org/abs/2601.04674.

Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1535–1546. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1141. URL https://aclanthology.org/P17-1141/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Yanchen Jiang, Zhe Feng, Christopher P Mah, Aranyak Mehta, and Di Wang. One model, two markets: Bid-aware generative recommendation. *arXiv preprint arXiv:2603.22231*, 2026.

Xiaoyu Kong, Leheng Sheng, Junfei Tan, Yuxin Chen, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Minionerec: An open-source framework for scaling generative recommendation. *arXiv preprint arXiv:2510.24431*, 2025. URL https://arxiv.org/abs/2510.24431.

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. URL https://arxiv.org/abs/2203.01941.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 188–197, 2019.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.05065.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://arxiv.org/abs/1011.0686.

Bohao Wang, Feng Liu, Jiawei Chen, Xingyu Lou, Changwang Zhang, Jun Wang, Yuegang Sun, Yan Feng, Chun Chen, and Can Wang. Msl: Not all tokens are what you need for tuning llm as a recommender. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025. arXiv:2504.04178.

You Wang, Zhao Liu, Guoping Tang, Yiqing Yang, Shuo Su, Jing Liu, Naifu Zhou, Xiaoyou Zhou, Wei Jiang, Jian Liang, Xiao Lv, Ruiming Tang, Liyin Hong, and Wenwu Ou. Multi-decoder onerec: Controllable generative retrieval for multi-objective industrial recommendation. *arXiv preprint arXiv:2607.26500*, 2026. URL https://arxiv.org/abs/2607.26500.

Sam Wiseman and Alexander M. Rush. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1296–1306. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1137. URL https://aclanthology.org/D16-1137/.

Shiguang Wu, Zhaochun Ren, Xin Xin, Jiyuan Yang, Mengqi Zhang, Zhumin Chen, Maarten de Rijke, and Pengjie Ren. Constrained auto-regressive decoding constrains generative retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2025. doi: 10.1145/3726302.3729934.

Ruochen Yang, Yusheng Huang, Youfeng Zheng, Shuang Wen, Liangliang Chen, Pengbo Xu, Xiaoyu Zhang, Shijun Wang, Shuang Yang, Zhaojie Liu, Lantao Hu, Wenwu Ou, Jiawei Sheng, and Tingwen Liu. Reward guided decoding for generative recommendation. *arXiv preprint arXiv:2607.25344*, 2026a. URL https://arxiv.org/abs/2607.25344.

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*, pp. 2196–2207. ACM, 2026b. doi: 10.1145/3805712.3809533.

Yuanqing Yu, Yifan Wang, Weizhi Ma, Zhiqiang Guo, and Min Zhang. Apao: Bridging the training-inference gap in generative recommendation via adaptive prefix-aware optimization. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2026. doi: 10.1145/3770855.3818052.

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. URL https://arxiv.org/abs/2311.09134.

Yuanbo Zhao, Ruochen Liu, Senzhang Wang, Jun Yin, Yuxin Dong, Huan Gong, Hao Chen, Shirui Pan, and Chengqi Zhang. From token generation to item ranking: Direct generative recommendation with semantic ids. *arXiv preprint arXiv:2602.07847*, 2026. URL https://arxiv.org/abs/2602.07847.

A Appendix

A.1 Extended Related Work

Recent work in generative recommendation addresses decoding and ranking limitations through complementary inference-time and architectural mechanisms:

- **Inference-time search guidance:** RGD (Yang et al., 2026a) derives a KL-regularized reward-guided decoding distribution that injects reward signals at each generation step without retraining the base generator. Similarly, PROMISE (Guo et al., 2026) guides beam expansions using an auxiliary process reward model, explicitly trading inference compute for generation quality. GEM-Rec (Jiang et al., 2026) similarly modifies token scores only at decoding time through prefix-conditional bid signals. Such inference-time transformations redefine the realized beam cutoff, implying that a boundary-aware training trace should apply the same score transformation when combined with these controllers.
- **Architecture routing and multi-task decoders:** Multi-Decoder OneRec (Wang et al., 2026) combines shared representations with objective-specific parameter-efficient decoder experts and coordinates their outputs through explicit route quotas during decoding.
- **Direct item-level representation:** DiGR (Zhao et al., 2026) uses SIDs as item representations rather than autoregressive generation targets and performs direct item-level scoring, thereby avoiding the token-mediated decoding pipeline analyzed in their ranking formulation.
- **Trainable semantic-ID interfaces:** DIGER (Fu et al., 2026) jointly optimizes the item-to-SID assignment with the downstream objective, requiring explicit quantization and reconstruction anchors for stable training. In contrast, ULBS keeps the SID interface fixed and reads the realized beam cutoff through a detached trace rather than differentiating through discrete search decisions.

ULBS is complementary to these inference-side and architectural approaches. It changes only the training objective and introduces no additional inference-time mechanism. In principle, ULBS could be combined with inference-time reward guidance or multi-decoder architectures because these methods intervene at different stages of the recommendation pipeline.

A.2 Mechanistic Diagnostics

This section provides granular diagnostic analyses to unpack the structural tension between candidate admission and within-beam ranking.

Table 4: Five-seed diagnostic by prefix depth and local trie branching. Rescue and regress are mean counts of MSL-miss→ULBS-hit and MSL-hit→ULBS-miss transitions, respectively.

Dataset	Depth	Branch	Rows	Rescue	Regress	ΔNDCG (pp)
Industrial	0	5+	4533	144.4	130.2	+0.749
Industrial	1	5+	4533	144.4	130.2	+0.749
Industrial	2	1	1785	37.8	11.6	+1.822
Industrial	2	2–4	1603	32.4	23.8	+0.397
Industrial	2	5+	1145	74.2	94.8	-0.431
Office	0	5+	4866	139.2	147.2	+0.149
Office	1	2–4	17	0.0	0.0	+0.000
Office	1	5+	4849	139.2	147.2	+0.149
Office	2	1	2334	45.2	32.0	+0.767
Office	2	2–4	2000	68.6	82.6	-0.684
Office	2	5+	532	25.4	32.6	+0.565
Musical	0	5+	4995	100.2	87.8	+0.091
Musical	1	1	2	0.0	0.0	+0.000
Musical	1	2–4	260	3.0	3.6	-0.387
Musical	1	5+	4733	97.2	84.2	+0.117
Musical	2	1	1785	20.2	15.0	+0.163
Musical	2	2–4	2672	55.8	49.2	+0.252
Musical	2	5+	538	24.2	23.6	-0.952

Table 5: Factorial ablation of credit assignment and event weighting. Results are mean over seeds 123 and 2026 ($\times 100$).

	Industrial		Office	
Control	HR@10	NDCG@10	HR@10	NDCG@10
Cumulative + uniform	10.2581	7.2776	9.2581	6.1186
Cumulative + adaptive	9.9051	6.8334	9.1451	5.8571
Local + uniform	10.5670	7.4333	10.3165	6.9139
Local + adaptive	10.0265	6.9335	10.2343	6.4615

Table 6: Seed-42 ablation of intermediate survival and terminal calibration losses ($\times 100$).

	Industrial		Office	
Condition	HR@10	NDCG@10	HR@10	NDCG@10
MSL	9.5081	6.2416	9.5561	6.1194
Survival only	8.8904	5.9722	8.8985	5.7839
Terminal only	9.2654	5.4784	**10.2548**	6.4589
ULBS	**10.1037**	**7.3301**	9.8233	**6.7926**

Prefix-depth and trie branching strata. We further stratify the aligned five-seed predictions by prefix depth and local legal branching factor (Table 4). The effect of ULBS is heterogeneous across prefix depths and local trie geometry:

- **Sparse branching:** At depth 2, single-child prefixes show positive ΔNDCG in all three domains, with the largest gain on Industrial ($+1.82$ points). This pattern is consistent with boundary supervision being more favorable when the continuation set is locally unambiguous.
- **Branching-dependent heterogeneity:** At depth 2, larger continuation sets do not exhibit a monotonic pattern. The $5+$ stratum is negative on Industrial (-0.43) and Musical (-0.95), but remains positive on Office ($+0.57$); conversely, the Office 2-4 stratum decreases by -0.68 points. These mixed outcomes suggest that local branching alone does not determine the benefit of boundary supervision, and that within-beam ranking remains relevant in more competitive subtrees.

Depth-0 rows reproduce the aggregate result, while some depth-1 rows do likewise when nearly all prefixes fall into the same branching stratum.

Why canonical and Musical differ. The contrast between the canonical categories (Industrial/Office) and the stress test (Musical) is consistent with the structural heterogeneity observed above. Although both settings share identical LLM backbones, LoRA configurations, and training budgets, they differ substantially in their local trie structure and retrieval behavior. On Musical, a larger fraction of target items naturally survive the beam boundary under standard likelihood training. Together, these diagnostics suggest that the balance between admission gains and within-beam ranking regressions is domain dependent.

A.3 Mechanism Controls and Ablations

Credit assignment and weighting controls. To decouple the credit assignment mechanism from the event reduction policy, we evaluate all four combinations of local versus cumulative credit and uniform versus adaptive weighting across seeds 123 and 2026 on Industrial and Office (Table 5).

Across both domains and credit schemes, uniform weighting outperforms the corresponding adaptive variant. The largest NDCG@10 gap is 0.500 points under local credit on Industrial. Local-uniform credit provides localized gradient assignment by stopping backward flow through earlier cumulative score terms, while uniform weighting avoids coupling event weights to instantaneous losses.

Component-off loss ablations. Table 6 examines the contribution of the intermediate survival loss ℓ^{surv} and the terminal calibration loss ℓ^{term} on seed 42.

Table 7: Decoder-width sensitivity with frozen seed-42 adapters ($\times 100$). Metrics are reported at $k = 5$ for $B = 5$ and $k = 10$ for $B = 20$.

	Industrial				Office			
	B=5		B=20		B=5		B=20	
Method	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10
MSL	6.7284	5.3828	7.7653	5.5028	**7.4394**	5.2833	7.7682	5.0155
BEAR-flat	6.9932	4.9069	8.3830	5.2086	6.6790	4.6871	7.3572	4.5556
APAO	6.5078	4.2394	5.8240	3.5429	7.0900	5.0652	8.0148	5.0602
ULBS	**8.0079**	**6.6533**	**9.9051**	**7.1481**	7.3572	**5.5661**	**8.8779**	**6.0077**

Table 8: Seed-42 sensitivity to α and m on Industrial with local credit and uniform weighting ($\times 100$).

α	m	HR@10	NDCG@10
0.15	0.10	9.7287	6.2220
0.15	0.20	8.8683	5.7285
0.15	0.40	9.2433	6.1056
0.30	0.10	9.5963	6.8918
0.30	0.20	10.1037	7.3301
0.30	0.40	8.8904	6.4615
0.60	0.10	10.0154	7.0261
0.60	0.20	10.4787	7.5353
0.60	0.40	10.3463	7.3304

Because our implementation dynamically renormalizes active terms, these experiments serve as component-off controls rather than a linear additive decomposition. Neither term in isolation matches the complete ULBS objective: omitting terminal calibration causes an immediate drop in NDCG@10 (falling from 7.33 to 5.97 on Industrial), showing that neither auxiliary term alone reproduces the performance of the full objective.

Decoder width and hyperparameter sensitivity. We evaluate sensitivity to decoding beam width $B \in \{5, 20\}$ on frozen seed-42 checkpoints (Table 7). ULBS demonstrates robust retrieval across beam budgets. Furthermore, Table 8 presents the full 3×3 sensitivity grid over the survival weight $\alpha \in \{0.15, 0.30, 0.60\}$ and margin $m \in \{0.10, 0.20, 0.40\}$ on Industrial. We fix $\gamma = 1$ throughout. With γ held fixed, varying m is equivalent to varying the normalized margin m/γ while keeping the gap scale Δ_t/γ unchanged.

A.4 Hardware Profiling and Training Efficiency

To ensure complete computational transparency, we profile the runtime and memory overhead of ULBS relative to MSL (Table 9).

All efficiency experiments use NVIDIA RTX 4090 GPUs and a dedicated batch-size-4 probe over 256 aligned test queries following eight warm-up batches. At inference time, ULBS introduces no additional decoding mechanism and exhibits negligible measured overhead: median latency remains close to that of MSL (303.13 vs. 299.13,ms on Industrial), while peak VRAM is unchanged at 15.1,GiB. During training, the detached beam trace increases Industrial runtime from 3.5,h to 6.4,h ($1.83\times$ based on the reported wall-clock values). Single-step profiling with NVML likewise shows unchanged peak VRAM, indicating that the additional training cost is primarily computational rather than due to increased parameter or memory capacity.

Table 9: Inference efficiency on 256 fixed queries with Qwen2.5-7B and beam width $B = 10$. Latency is reported per query with batch size 4.

	Industrial			Office		
Method	p50 (ms)	p95 (ms)	Throughput (q/s)	p50 (ms)	p95 (ms)	Throughput (q/s)
MSL	299.13	329.10	3.340	306.49	335.32	3.283
ULBS	303.13	333.41	3.298	307.39	335.85	3.266