VecDB @ VLDB 2026 · The 2nd Workshop on Vector Databases · Poster Session 1 · Galleria · 4 Sep 2026, Boston

SPI: Query-Depth-Adaptive Indexing for Streaming RAG in Vector Databases

Dong Liu[1] Yanxuan Yu[2] Shinan Liu[3] [1]Yale University [2]Columbia University [3]University of Hong Kong
dong.liu@aya.yale.edu yy3523@columbia.edu shinan6@hku.hk

Code & configs
github.com/FastLM/SPI_VecDB

Motivation & Semantic Pyramid Indexing

Streaming RAG, uniform ANN

Workload $\mathcal{W} = \{(q_t, t)\} \cup \{(d, t_d^{\text{arr}})\}$ on VecDBs (Milvus, Qdrant, Weaviate). A uniform index stores $\mathbf{e}_d \in \mathbb{R}^D$ and answers every q by

$$\hat{d}(q) = \arg\max_{d \in \mathcal{D}} \text{sim}(\mathbf{q}, \mathbf{e}_d) \quad \text{or ANN} \quad \widetilde{G}_K(q) \approx G_K(q).$$

"Resolution" is *semantic discriminability* (separating d^+ from d^-), not the retrieval unit (always documents). Broad q ("Who discovered gravity?") needs only a shallow high-entropy $\mathbf{e}^{(1)}$; fine q ("Equation in Newton's third-law paper") needs $\mathbf{e}^{(L)}$ that splits near-duplicates. Forcing $\ell \equiv L$ wastes compute when $\ell^*(q) \ll L$ and may leave $\Delta_L(q) < \tau$. Inserts either block (rebuild) or degrade (lazy append).

Similarity margin and $\ell^*(q)$

$$\Delta_\ell(q) = \mathbb{E}_{d^+}\big[\text{sim}(\mathbf{q}^{(\ell)}, \mathbf{e}_{d^+}^{(\ell)})\big] - \mathbb{E}_{d^-}\big[\text{sim}(\mathbf{q}^{(\ell)}, \mathbf{e}_{d^-}^{(\ell)})\big],$$

$\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / (\|\mathbf{u}\| \, \|\mathbf{v}\|)$. The optimal depth is the first separable level:

$$\ell^*(q) = \min\big\{\ell \;\big|\; \Delta_\ell(q) \geq \tau\big\}.$$

Evaluating $\ell^*(q)$ is infeasible online; SPI approximates it from $\mathbf{q}^{(1)}$ via a controller. Expected cost under a well-calibrated router:

$$\mathbb{E}[T_{\text{SPI}}(q)] \leq \sum_{\ell=1}^{\ell^*(q)} p_\ell \cdot O\big(|C_\ell| \log |C_\ell|\big) + \mathcal{O}(d_q),$$

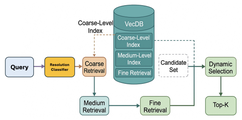
where $p_\ell = \Pr[\ell_{\text{final}} = \ell]$ and $\mathcal{O}(d_q)$ is the 0.8 ms controller.

SPI $\neq$ hierarchical ANN

HNSW / SPANN organise *geometry* with fixed traversal π_{geo}. SPI organises *meaning*: the same $\mathcal{D}$ is stored as

$$\mathbf{e}_d = \big(\mathbf{e}_d^{(1)}, \ldots, \mathbf{e}_d^{(L)}\big), \quad \mathbf{e}_d^{(\ell)} \in \mathbb{R}^{D_\ell}, \quad D_1 \geq \cdots \geq D_L,$$

and routes $q \mapsto \ell_{\text{final}}(q) \approx \ell^*(q)$. Each $\mathcal{I}^{(\ell)} = \text{Index}(\{\mathbf{e}_d^{(\ell)}\}_{d \in \mathcal{D}})$ is FAISS IVF–PQ or Qdrant HNSW—orthogonal to CAGRA / DiskANN.



Enter at ℓ=1; descend only while $\Delta_\ell(q) < \tau$.

Two-phase architecture

Phase 1 (offline). Learn $\{f_\ell, g_\ell\}_{\ell=2}^L$ by InfoNCE $\mathcal{L}_{\text{retr}}^{(\ell)} = -\log \frac{e^{\text{sim}(\mathbf{q}^{(\ell)}, \mathbf{e}_{d^+}^{(\ell)})/\kappa}}{\sum_{d'} e^{\text{sim}(\mathbf{q}^{(\ell)}, \mathbf{e}_{d'}^{(\ell)})/\kappa}}$.

Phase 2 (online). Serve $\{\mathcal{I}^{(\ell)}\}$, append-only inserts, LSH over N nodes. Decoupled: any compatible encoder (e.g. MRL) can plug into Phase 2.

Contributions

1. **Index.** Query-adaptive pyramid $\mathcal{I} = \{\mathcal{I}^{(\ell)}\}_{\ell=1}^L$ with $C_1 \supseteq C_2 \supseteq \cdots \supseteq C_{\ell_{\text{final}}}$ and $|C_{\ell_{\text{final}}}| = K$.
2. **Guarantee.** If $\gamma(q, \ell') > 2\delta_\ell$, then $G_K(q, \ell) = G_K(q, \ell')$ (Thm. 1): early exit is exact.
3. **Systems.** 1.4–2.3× latency at matched Recall@10; $S(8) = 6.3\times$ (78.6% efficiency); FAISS/Qdrant plugin.

Method: Pyramid, Control, Streaming

Residual pyramid

Level 1 is frozen Contriever, $\mathbf{e}_d^{(1)} \in \mathbb{R}^{768}$. For $\ell \geq 2$, a patch-Transformer f_ℓ ($\approx$ 5M) is trained on pairs with $\Delta_{\ell-1}(q) < \tau$:

$$\mathbf{e}_d^{(\ell)} = f_\ell\big(\mathbf{e}_d^{(\ell-1)}\big) + \mathbf{W}_\ell \, \mathbf{e}_d^{(1)}, \qquad \mathbf{W}_\ell \in \mathbb{R}^{D_\ell \times D_1}.$$

Default $(D_1, D_2, D_3) = (768, 512, 256)$. Split $\mathbf{e} \in \mathbb{R}^{D_{\ell-1}}$ into K patches of size $p = D_{\ell-1}/K$ (ℓ=2: K=12, p=64; ℓ=3: K=8, p=64), prepend `[CLS]`, two blocks (H=4). Query side is dual-encoder, unshared: $\mathbf{q}^{(\ell)} = g_\ell(\mathbf{q}^{(\ell-1)})$, fired only if the controller proceeds. Encoding cost $T_{\text{enc}}^{q,\ell} \in \{8.1, 3.9, 1.8\}$ ms, so $\Sigma_{\ell=1}^{\ell_{\text{final}}} T_{\text{enc}}^{q,\ell} \in [15.1, 24.9]$ ms. Documents encode asynchronously and never block queries. Each level is independently partitioned: $\mathcal{I}^{(\ell)} = \cup_{i=1}^N \text{Index}(\{\mathbf{e}_d^{(\ell)}\}_{d \in \mathcal{D}_i})$.
InfoNCE at level ℓ plus consistency:

$$\mathcal{L}_{\text{total}} = \sum_{\ell=1}^{L} \alpha_\ell \, \mathcal{L}_{\text{retr}}^{(\ell)} + \beta \, \mathcal{L}_{\text{cons}} + \gamma \, \mathcal{L}_{\text{reg}},$$

$$\mathcal{L}_{\text{cons}} = \sum_{\ell=1}^{L-1} \big\| \mathbf{e}_d^{(\ell)} - \text{proj}_\ell(\mathbf{e}_d^{(\ell+1)}) \big\|_2^2, \quad \alpha_\ell = \tfrac{1}{L}, \; \beta = 0.5, \; \gamma = 0.01.$$

After training, $\Delta_\ell(q) \geq \Delta_{\ell-1}(q)$ on queries routed to ℓ.

Precision, pools, memory

$N_\ell = \big\lfloor N_1 \, \Pi_{k<\ell} \, r_k \big\rfloor$, N_1=1000, $(r_1, r_2) = (0.20, 0.05)$. $b_\ell = \min(32, \, 8 \cdot 2^{\ell-1}) \Rightarrow (8, 16, 32)$. Quantization $\delta_\ell^{\text{q}} \leq \|\mathbf{e}\|_\infty / 2^{b_\ell - 1}$ satisfies $\delta_\ell^{\text{q}} \cdot \Delta_\ell(q) \ll \tau$.

$$M_{\text{store}} = |\mathcal{D}| \sum_\ell D_\ell \tfrac{b_\ell}{8}, \quad M_{\text{active}}(q) = \sum_{\ell \leq \ell_{\text{final}}} N_\ell D_\ell \tfrac{b_\ell}{8}.$$

Progressive retrieval and LSH

$$C_{1,i} = \text{Top}_{N_1/N}\big(\text{sim}(\mathbf{q}^{(1)}, \{\mathbf{e}_d^{(1)}\}_{d \in \mathcal{D}_i})\big), \; C_1 = \cup_i C_{1,i}.$$

Deeper levels re-score the retained set:

$$C_{\ell,i} = \text{Top}_{N_\ell/N}\big(\text{sim}(\mathbf{q}^{(\ell)}, \{\mathbf{e}_d^{(\ell)}\}_{d \in C_{\ell-1} \cap \mathcal{D}_i})\big).$$

Always $|C_{\ell_{\text{final}}}| = K$. Local top-$(k_\ell + \delta)$ with slack $\delta = \lceil k_\ell / \sqrt{N} \rceil$. LSH cell $h_{\text{LSH}}(\mathbf{e}) = \arg\min_i \|\mathbf{e} - \mathbf{c}_i\|$; leakage $O(\varepsilon_\ell / \delta_{\text{LSH}}^2) < 2\%$. EMA $L_i(t) = \alpha L_i(t-1) + (1-\alpha) Q_i(t)$, α=0.9; $i^* = \arg\min_i L_i(t)$ over async gRPC.

Controller

$$p_k = \frac{|q_k^{(1)}|}{\Sigma_j \, |q_j^{(1)}|}, \quad H(\mathbf{q}^{(1)}) = -\sum_{k=1}^{D_1} p_k \log p_k, \quad (\hat{\ell}, \sigma_\ell) = g_{\text{ctrl}}(\mathbf{q}^{(1)}, H).$$

$$\ell_{\text{final}} = \begin{cases} \hat{\ell}, & \sigma_\ell \leq \theta, \\ \min(\hat{\ell} + 1, \, L), & \sigma_\ell > \theta, \end{cases} \qquad \theta = 0.35.$$

Oracle fold (disjoint 80/10/10): $\ell_q^* = \min\{\ell \mid \text{R@100}^{(\ell)} \geq 0.99 \cdot \text{R@100}^{(L)}\}$, CE + temperature T=1.2. In-domain 87.5% ($45/32/23\%$ at $\ell = 1, 2, 3$); OOD (BEIR Quora/FiQA) 82.1%, still $\geq 3.8\times$ vs. full depth.

Streaming and Thm. 1

Steady state: $\mathcal{D}^{(\ell)}(t) = \mathcal{D}$. Online: $\mathcal{D}^{(\ell)}(t) = \{d \mid t_d^{\text{arr}} + T_{\text{ins}}^{(\ell-1)} \leq t\}$, so $\mathcal{D}^{(1)}(t) \supseteq \cdots \supseteq \mathcal{D}^{(L)}(t)$. Level 1 is synchronous ($T_{\text{enc}}^{(1)} \approx 8$ ms); $\ell \geq 2$ is async, $T_{\text{ins}}^{(\ell)} = T_{\text{enc}}^{(1)} + O(D_\ell / B_{\text{enc}})$. Re-anchor $\hat{\mathbf{e}}_d^{(\ell)} = \text{proj}_\ell(\mathbf{e}_d^{(\ell)}) + \mathbf{W}_\ell \mathbf{e}_d^{(1)}$.

Theorem 1. $\mathbb{E}_d \|\mathbf{e}_d^{(\ell)} - \text{proj}_\ell(\mathbf{e}_d^{(\ell')})\|_2^2 \leq \varepsilon_\ell$ gives $\delta_\ell = \sqrt{\varepsilon_\ell} < 0.18$ and $\big|\text{sim}^{(\ell)} - \text{sim}^{(\ell')}\big| \leq 2\delta_\ell$. If $\gamma(q, \ell') = \min_{d^* \in G_K, \, d' \notin G_K} [\text{sim}_{d^*} - \text{sim}_{d'}] > 2\delta_\ell$, then $G_K(q, \ell) = G_K(q, \ell')$.

Experiments & Takeaways

Setup (Q1)

Single node: RTX 4090 24 GB, i9-13900K, FAISS 1.7.4. End-to-end latency (no LLM) $T = T_{\text{enc}} + T_{\text{ANN}} + T_{\text{ctrl}}$, K=10.

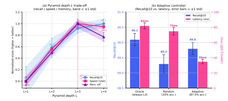
$$\text{R@}K = \tfrac{|G_K \cap \text{Rel}|}{\min(K, |\text{Rel}|)}, \quad \text{QPS} = 10^3 / T_{\text{ms}}.$$

6980 MS MARCO + 3610 NQ; default L=3, IVF–PQ.

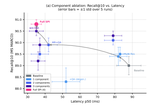
Method	R@10	NDCG	ms	GB
Contriever	85.2	76.4	120	7.2
ColBERTv2	89.3	80.1	108	5.8
SPANN	86.8	77.0	115	5.9
HNSW (flat)	87.1	77.4	122	6.2
SPI	**90.1**	**81.2**	**35**	**4.2**

Macro-average. Vs. SPANN: $2.8\times$ faster, $+3.3$ R@10. QPS $8.1 \to 28.6$.

Depth L and controller



Knee at L=3: $(\text{R@10}, T, M) = (90.1, \, 35\,\text{ms}, \, 4.2\,\text{GB})$. Adaptive: R@10: $90.1 \to 89.8$ for T: $82 \to 35$ ms, $\Pr[\ell_{\text{final}}$=$1, 2, 3] = (0.45, 0.32, 0.23)$.



Ablation: Multi-Res alone leaves $T \approx T_{\text{full}}$; g_{ctrl} is required for early exit. Full SPI: $\text{R@10} = 90.8$ at $T = 35$ ms on MS MARCO.

Streaming (Q2) and scaling (Q3)

Burst $|\mathcal{D}_{\text{new}}| = 10^3$ at 100 QPS: $T_{\text{ins}}^{(2,3)} = 47$ ms p95, $\Delta\text{R@10} \leq 0.6$. Complementary to IP-DiskANN / Quake. Corpus $|\mathcal{D}| \propto N$ (1M→16M); LSH on $\mathbf{e}_d^{(1)}$.

$$S(N) = \frac{T(1)}{T(N)}, \qquad \eta(N) = \frac{S(N)}{N}, \qquad T_{\text{dist}} \leq \tfrac{T_{\text{seq}}}{N} + T_{\text{net}} + T_{\text{sync}}.$$

Nodes N	Docs	ms	QPS	$S(N)$
1	1M	22	45.5	$1.0\times$
4	4M	26	153.8	$3.4\times$
8	8M	28	285.7	$\mathbf{6.3}\times$
16	16M	33	483.4	$10.6\times$

$\eta(8) = 78.6\%$. $\eta(16) = 66.4\%$ ($T_{\text{net}}/T \approx 0.20$).

Takeaways

- Put $\ell_{\text{final}}(q)$ in the VecDB index, not in the RAG pipeline config.
- Early exit is exact whenever $\gamma(q, \ell') > 2\delta_\ell$ (Thm. 1).
- Streaming freshness and geometric repair are composable: $\mathcal{I}^{(1)}$ can sit on IP-DiskANN / Quake.
- Operating point: L=3, $b_\ell \in \{8, 16, 32\}$, $D_\ell \in \{768, 512, 256\}$.