When Does Query-Aware Quantization Help? From Redundancy In-Distribution to Gains in Multimodal Retrieval

Anonymous authors
Paper under double-blind review

Abstract

Embedding-based retrieval at scale depends on vector quantization to hold large databases in memory. A natural idea is to compress vectors so that similarity to *queries*, rather than the vectors themselves, is preserved. Methods built on this idea, however, often yield only modest gains over strong query-agnostic baselines on standard benchmarks, and the field has lacked a principled account of when query awareness should help. In this paper, we find that on standard benchmarks the query and database distributions are so closely aligned that a rotation optimized on the data alone already captures the query structure, leaving query-aware objectives little to exploit. The redundancy persists even locally, where genuine query structure does exist: we show it adds nothing beyond local data statistics. On multimodal text-to-image retrieval (Yandex Text2Image), where the two distributions differ by construction, the situation reverses: at moderate compression rates, our proposed spectral method delivers 7.1 percentage points more Recall@10 than the standard OPQ+IVFPQ pipeline (a 26% relative improvement) and 1.7 points more than the strongest baseline we test, at equal memory and search cost; every query-aware and locally adaptive variant we test also improves on OPQ+IVFPQ in this regime. We distill both findings into a diagnostic that, from a sample of calibration queries alone, determines in minutes whether a given workload will benefit from query-aware quantization before any method is built.

1 Introduction

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) grounds large language models in external knowledge via approximate nearest neighbor (ANN) search (Johnson et al., 2019) over large stores of dense embeddings, and remains an active area of systems research (Zhao et al., 2024; Shen et al., 2024), where memory and bandwidth dominate retrieval cost. Vector quantization, most commonly product quantization (PQ) (Jégou et al., 2011) and its rotated variant, optimized product quantization (OPQ) (Ge et al., 2014), is the standard remedy: conventional quantizers minimize isotropic reconstruction error $\mathbb{E}\|x - Q(x)\|^2$ (for database vector x and compressed approximation $Q(x)$), treating all directions equally.

Retrieval quality, however, depends on inner-product similarity against *queries*: distortion along directions that queries probe degrades ranking, while distortion elsewhere is free. This has motivated *query-aware* quantization: score-aware losses (Guo et al., 2020; Zhang et al., 2022; 2023), query-distribution-aware codebook training (Jaiswal et al., 2023), and query-aware VQ for attention-score approximation (He et al., 2025). Yet a puzzle persists: against a well-tuned OPQ baseline, the gains from query awareness on *in-distribution* benchmarks, where queries and the database come from the same encoder and domain such as MS-MARCO, are often modest, while *out-of-distribution* benchmarks, where they come from different encoders or modalities such as text-to-image search, are comparatively underexplored. Is the theory wrong, or is the signal it targets simply absent? (We define both terms precisely in Section 2.)

This paper. This paper gives an answer for the workloads we study, stated up front: *query-aware quantization helps only when the query distribution differs from the database distribution. This condition fails on standard benchmarks, so query awareness cannot help there, and we show why; it*

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.

holds in multimodal retrieval (Yandex Text2Image), where our proposed spectral method gains 7.1 *percentage points of Recall@10 over OPQ combined with an inverted file (Ge et al., 2014; Jégou et al., 2011), the standard pipeline, and remains the strongest method we test at that budget; and it is checkable in advance with a simple diagnostic.* We reach this answer through a sequence of controlled analyses that, to our knowledge, are each the first of their kind for this problem (detailed below), and use them to design methods that target the signal where it actually resides. Our contributions:

(1) Exact task-aware objective (Section 2). We formalize the query-aware objective via $H = \mathbb{E}_q[qq^\top]$: the expected squared similarity error of any quantizer equals the H-weighted reconstruction error (Theorem 1), with a ranking-stability corollary. Unlike prior score-aware losses (Guo et al., 2020; Zhang et al., 2023), which weight errors by a chosen or learned function, this identity is exact and grounds all that follows.

(2) In-distribution: query awareness is redundant at every granularity (Sections 3, 5). We give, to our knowledge, the first direct measurement of query–database covariance alignment on ANN benchmarks (top-20 alignment 0.77–0.96): query-agnostic OPQ is already implicitly query-aware, and global instantiations of the H-weighted objective fail to beat it. Locally, we borrow two statistical controls not previously used in retrieval research, shuffled nulls and split-half reliability, and find that genuine local query structure does exist on MS-MARCO. Yet it is *redundant*: we propose LQ-PQ, to our knowledge the first locally query-aware product quantizer, and compare it to an LOPQ-style (Kalantidis and Avrithis, 2014) control that shares its exact code path but substitutes local *data* covariance. The two are statistically indistinguishable, while both add $\sim$3.3 Recall@10 points over OPQ+IVFPQ, about half of which an unweighted per-group control already recovers.

(3) Out-of-distribution: query awareness pays, but only globally (Section 6). On Yandex Text2Image, where query and database are misaligned by construction (alignment 0.50), our proposed task-aware quantization (TAQ), whose query-spectrum bit allocation is the first of its kind for ANN search, retrieves 7.1 Recall@10 points more than the strongest standard pipeline (0.343 vs. 0.272, 26% relative), ahead of the best local method, confirmed by an exhaustive-search control, and our local LQ-PQ is within 0.2 points of it while beating its data-covariance control by $+1.5$. At 4 bits/subspace a bit-allocation floor reverses the ranking toward local data codebooks for both query-aware methods: on this benchmark, query awareness pays at both granularities, given bit headroom.

Beyond the specific methods, the diagnostic protocol itself (global alignment, shuffled nulls, split-half reliability) is of independent value: from calibration queries alone, it correctly separates the two regimes across all four datasets in this study.

1.1 Related Work

PQ, OPQ, and IVF. Product quantization (Jégou et al., 2011) decomposes $\mathbb{R}^d$ into m subspaces quantized independently for fast lookup-table distance computation. OPQ (Ge et al., 2014) prepends an orthogonal rotation minimizing PQ reconstruction error, preserving inner products and IVF Voronoi geometry. LOPQ (Kalantidis and Avrithis, 2014) adapts rotations and codebooks *locally* per coarse cell using local *data* statistics. Our LQ-PQ inherits LOPQ's locality but substitutes local *query* statistics for LOPQ's local *data* statistics (Section 4 details the controlled design): an informative negative result in-distribution, where query information adds nothing beyond what local data statistics already capture, whereas out-of-distribution the same substitution pays at sufficient bit budget (Section 6). IVF with residual encoding (Jégou et al., 2011; Johnson et al., 2019) is the standard scalable pipeline; we leave it unmodified.

Query-aware quantization. ScaNN (Guo et al., 2020) introduced score-aware quantization, penalizing the residual parallel to the database vector under a uniform query assumption; Zhang et al. extended this to additive quantization (Zhang et al., 2022) and to weighting learned within a general framework (Zhang et al., 2023). Two works target the OOD setting directly: OOD-DiskANN (Jaiswal et al., 2023) trains PQ pivots on base-and-query mixtures, and A^{2}ATS (He et al., 2025) optimizes an inner-product objective for KV-cache retrieval. Each weights errors by a chosen or learned function; our objective (Theorem 1) is exact, which is what lets us pinpoint *when* it pays off, a question this line of work has not previously answered.

Adaptive bit allocation. BAPQ (Chen et al., 2015) allocates bits across PQ subspaces from data variance, using the water-filling machinery of rate–distortion theory (Gray and Neuhoff, 1998). We derive the query-covariance analog (Section 2.1), to our knowledge the first bit allocation for ANN driven by the query second-moment spectrum, and use it to drive TAQ's leading result out-of-distribution (Section 6). We also pin down exactly where it stops mattering, in-distribution and at very low bit budgets, turning two potential failure modes into a precise operating envelope.

Out-of-distribution retrieval. The Big-ANN benchmark (Simhadri et al., 2024) established OOD query workloads, which prior work treats as a setting to engineer around. We instead treat the divide as a measurable property of the workload, query–data alignment, and show it governs whether query-aware quantization helps at all.

2 Problem Setup and Task-Aware Objective

Setup. Let $\mathcal{D} = \{x_i\}_{i=1}^N \subset \mathbb{R}^d$ be database embeddings and $q \sim \mathcal{Q}$ a query. Similarity is the inner product $s(q,x) := q^\top x$, retrieval returns $\mathrm{TopK}(q) = \arg\max_{x\in\mathcal{D}} s(q,x)$, and a quantizer $Q : \mathbb{R}^d \to \mathcal{C}$ replaces x by $\hat{x} := Q(x)$ at search time. Quantization perturbs each similarity score by $\epsilon(q,x) := s(q,x) - s(q,\hat{x}) = q^\top(x - \hat{x})$, the central error quantity of our analysis. We write $H := \mathbb{E}_{q\sim\mathcal{Q}}[qq^\top]$ for the query second-moment matrix and $\Sigma := \mathbb{E}_{x\sim\mathcal{D}}[xx^\top]$ for the analogous database second-moment matrix; Theorem 1 below shows why H is the right object for quantization error, and Section 3 compares it to Σ directly. H is the uncentered second moment, as Theorem 1 requires; calibration queries are never centered.

In-distribution vs. out-of-distribution queries. In *in-distribution* retrieval, queries and database vectors come from the same encoder and domain (e.g., MS-MARCO questions and passages), so queries statistically resemble the data. In *out-of-distribution* (OOD) retrieval, queries originate from a different modality or encoder and differ *by construction*: in text-to-image search, a multimodal (cross-modal) setting, text queries occupy a different, far more concentrated region of the space than the image database (Jaiswal et al., 2023; Simhadri et al., 2024). In second-moment terms, in-distribution workloads have $H \approx \Sigma$ (Finding 1), while on Yandex Text2Image we measure query effective rank 21.6 vs. data 116.2 (of d=200) and top-20 alignment only 0.50 (Section 3). This distinction determines whether the task-aware objective carries any information beyond standard reconstruction.

Theorem 1 (Similarity Error Equivalence)**.** *Let $\mathbb{E}[\|q\|^2] < \infty$ and $H := \mathbb{E}_{q\sim\mathcal{Q}}[qq^\top]$. For any x and any quantizer Q independent of q,*

$$\mathbb{E}_q\big[\epsilon(q,x)^2\big] = (x - \hat{x})^\top H\,(x - \hat{x}). \tag{1}$$

Proof. $\epsilon(q,x)^2 = (x-\hat{x})^\top qq^\top (x-\hat{x})$; take expectations and use linearity. Full details in Appendix A. □

Minimizing the expected squared similarity error is therefore equivalent to solving the H-weighted (Mahalanobis) reconstruction problem

$$\min_Q\ \mathbb{E}_{x\sim\mathcal{D}}\big[(x - Q(x))^\top H\,(x - Q(x))\big]. \tag{2}$$

With $H = I$ (the identity matrix), Equation (2) reduces to standard vector quantization. This objective is the population-level statement underlying prior score-aware and query-aware losses (Guo et al., 2020; Zhang et al., 2023; He et al., 2025); we treat it as the shared premise of the family, not as a novel objective. It is an exact identity for expected squared score error, which serves as a surrogate for top-k ranking quality; prior score-aware losses instead reweight errors toward the top-k margin.

Proposition 1 (Ranking Stability)**.** *If $|\epsilon(q,x)| \le \delta$ for all $x \in \mathcal{D}$, then quantized retrieval preserves the order of any pair (x^+, x^-) whose margin $\Delta(q) = s(q,x^+) - s(q,x^-)$ exceeds 2δ. (Proof in Appendix A.)*

Proposition 1 is a worst-case statement; Theorem 1 controls the *expected* squared error. The two connect through a standard tail bound: writing $e(x) := (x-\hat{x})^\top H(x-\hat{x})$ for the H-weighted error of Theorem 1, Markov's inequality applied to the squared perturbation gives, for a fixed margin threshold Δ, a per-pair rank-inversion probability of at most $2\big(e(x^+) + e(x^-)\big)/\Delta^2 \le 4\max_x e(x)/\Delta^2$

for margin Δ, so reducing the H-weighted objective directly reduces expected rank inversions on low-margin pairs.

Spectral form. Since H is symmetric, it admits an eigendecomposition $H = U\Lambda U^\top$, where $U = [u_1, \ldots, u_d]$ is orthogonal (its columns are H's eigenvectors, the principal directions of query variation) and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ holds the corresponding eigenvalues. Writing the residual projections onto these directions as $\tilde{r}_j = u_j^\top (x - \hat{x})$,

$$(x - \hat{x})^\top H (x - \hat{x}) = \sum_{j=1}^{d} \lambda_j \, \tilde{r}_j^2. \tag{3}$$

Directions with large λ_j are where quantization error is expensive. This suggests two global mechanisms, which together define our proposed method, task-aware quantization (TAQ): (i) transform the space by $L = \Lambda^{1/2} U^\top$ and quantize isotropically (whitening); (ii) allocate bits across subspaces by water-filling on λ_j, worked out next. As implemented, the allocation weights subspaces by the raw eigenvalues in the whitened coordinates, so the two mechanisms compound and the realized objective weights subspace error by roughly λ_s^2 rather than λ_s; we report TAQ as implemented. We first derive mechanism (ii) in full, then, in Section 3.3, test both mechanisms together: they fail in practice on standard benchmarks, and Finding 1 explains why. Section 6 shows the same two mechanisms, unchanged, become the strongest method once the in-distribution premise no longer holds.

2.1 Spectral Rate Allocation

We now derive mechanism (ii) above: a query-spectrum analog of classical bit allocation. Partition dimensions into m equal subspaces with budgets b_s, $\sum_s b_s \le B$. Under the high-rate model $D_s(b_s) \approx c_s 2^{-2b_s/d_s}$ (Gray and Neuhoff, 1998), weighting each subspace's distortion by the mean eigenvalue $\bar{\lambda}_s$ of H on its dimensions and solving the Lagrangian relaxation yields the water-filling allocation

$$b_s^* = \Big[\frac{B}{m} + \frac{d_s}{2} \log_2 \frac{\bar{\lambda}_s c_s}{\big(\prod_t \bar{\lambda}_t c_t\big)^{1/m}} \Big]_+, \tag{4}$$

with $[\cdot]_+$ denoting clipping to $b_s \ge 0$ and budget re-distribution. In practice we take c_s equal across subspaces, floor each $\bar{\lambda}_s$ at 10^{-8}, restrict $4 \le b_s \le 12$, and round to integers by largest remainder; at 4 bits/subspace the floor binds everywhere and the allocation is uniform. Two caveats: the high-rate approximation is loose at very low b_s (the mechanism behind the 4-bit crossover of Section 6), and in-distribution the spectrum of H nearly coincides with the data spectrum, collapsing (4) to its variance-driven counterpart (Chen et al., 2015). Whether this collapse actually happens on real data, and what survives it, is the question we turn to next.

3 Diagnosis: Where Is the Query-Aware Signal?

We now ask the question that determines whether (2) can beat query-agnostic baselines: *does the query distribution differ from the data distribution, and at what granularity?*

Protocol. For each dataset (MS-MARCO-1M, d=384, bi-encoder embeddings; SIFT1M, d=128; GloVe-1.2M, d=100) we estimate the query second-moment H from held-out calibration queries and the data covariance Σ from a 100K database sample, and compare: effective (participation-ratio) rank, spectra, and principal angles between the leading k eigenvectors of each (k=20 for the global comparison of Finding 1, k=5 for the local comparison of Finding 2, matching each analysis's granularity). Alignment is the mean singular value of $U_q^\top U_x$ for the two leading-k eigenvector matrices; Σ is centered, H is not. MS-MARCO embeddings are from msmarco-MiniLM-L6-cos-v5 on the first 1M passages; calibration and evaluation queries are disjoint halves of each public query set. Yandex Text2Image, our OOD benchmark, is analyzed with the same protocol in Section 6.

3.1 Finding 1: Global Query and Data Covariance Nearly Coincide

Finding 1 (Global redundancy of query awareness)**.** *On all three benchmarks, the top-20 eigenspaces of H and Σ reach mean subspace alignment of* 0.77–0.96 *and their spectra closely coincide (Figure 1 (top)). Consequently, a rotation or allocation computed from data statistics (OPQ,*

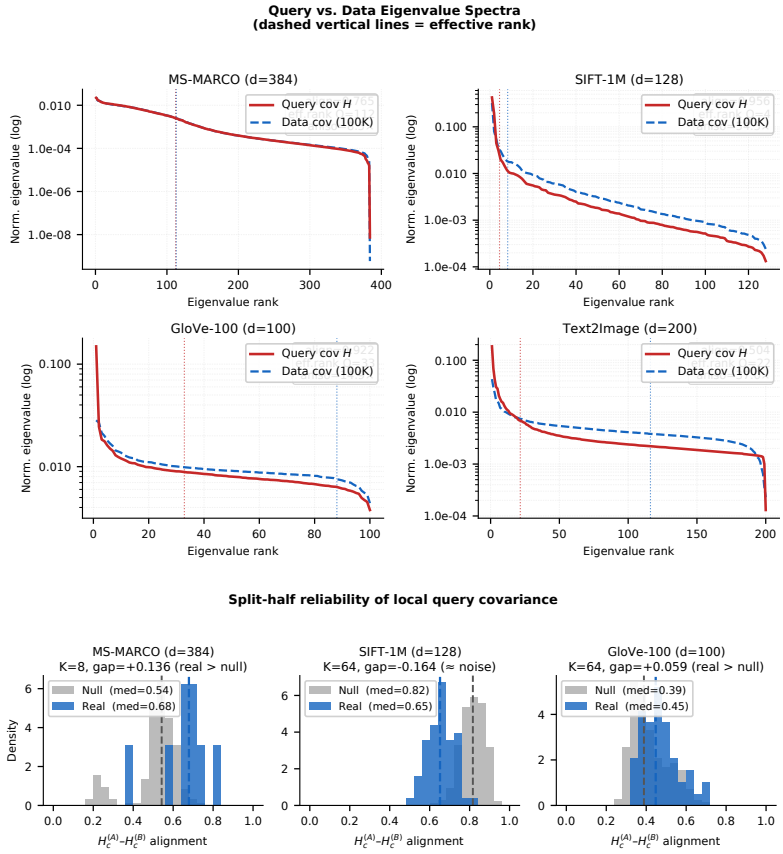


Figure 1: **Top:** query (H) vs. data (Σ) eigenvalue spectra; alignment is 0.77–0.96 in-distribution vs. 0.50 on Text2Image. **Bottom:** split-half reliability of local query covariance vs. shuffled null: real structure on MS-MARCO ($+0.136$), noise on SIFT-1M (-0.164), borderline on GloVe ($+0.059$); full panels in the supplement.

BAPQ) is numerically almost identical to one computed from query statistics, and global query-aware methods can offer at most marginal gains in-distribution.

This finding predicts that *any* global instantiation of (2) with a single H approximately reduces to its query-agnostic counterpart in-distribution, consistent with the modest margins of score-aware losses over OPQ (Guo et al., 2020) and with query-aware training paying off mainly out-of-distribution (Jaiswal et al., 2023).

3.2 Finding 2: Local Query Structure Is Real on RAG Workloads, but Is Noise on SIFT

Globally aligned distributions can still differ *conditionally*: the queries routed to one region of space may concentrate on directions unrelated to that region's local data spread. We test this with routing by IVF assignment, followed by two noise controls: (i) a *shuffled null* that assigns queries to cells uniformly at random (preserving sample sizes, destroying structure), and (ii) *split-half reliability*: the top-5 subspace alignment between two independent halves of each cell's queries (Figure 1, bottom).

Finding 2 (Local query structure)**.** *With sufficient per-cell query counts ($M_c \approx 900$, $K{=}8$ coarse regions), MS-MARCO exhibits reliable local query structure: median principal angle* $49.6°$ *vs. null* $72.2°$*; split-half alignment* 0.679 *vs. null* 0.542*; local effective rank* 63 *vs. null* 102*. On SIFT1M the split-half test* inverts *(*0.652 *vs. null* 0.816*): within-cell query subspaces are less self-consistent than random global draws, i.e., the apparent local structure is sampling noise. GloVe is borderline (*$+0.059$ *reliability gap).*

Table 1: Best in-distribution Recall@10 (10K calibration queries, best over nprobe$\leq$1024; SIFT rows use non-residual encoding for all methods). †: sole cell where a query-aware method beats OPQ+IVFPQ (see text).

Dataset	b/vec	OPQ+IVFPQ	TAQ+IVF	OPQ-Q+IVF
MS-MARCO (d=384)	192	.546	.269	.580†
	384	.783	.625	.708
SIFT-1M (d=128)	64	.364	.233	.231
	128	.584	.455	.456
GloVe-100 (d=100)	40	.101	.077	.082
	80	.243	.198	.196

Two implications follow. First, the exploitable query-aware signal on RAG-representative data is *local*: queries routed to a region concentrate on a low-rank subspace ($63 \ll 112$ global) that is misaligned with the region's data covariance, a structure that no global method, and no data-driven local method (LOPQ), can capture. Second, the signal exists at *coarse* granularity: it was validated at K=8 regions with $\sim$900 queries each; per-fine-cell estimates (K=256, $M_c \approx 13$) are noise-dominated. A sound method must therefore estimate local query statistics coarsely and regularize.

3.3 Negative Results for Global Instantiations

Consistent with Finding 1, our controlled comparison of global instantiations of (2) against OPQ+IVF on MS-MARCO-1M, SIFT1M, and GloVe shows: (i) whitening by $L = \Lambda^{1/2} U^\top$ with allocation (TAQ+IVF; at 4 bits the floor makes it whitening-only) *underperforms* OPQ: the non-orthogonal transform distorts the IVF Voronoi geometry and PQ subspace balance, costing more than query weighting recovers; (ii) query-covariance water-filling on top of the OPQ rotation (OPQ-Q+IVF) recovers most but not all of the gap; (iii) neither beats plain OPQ at moderate-to-high bit budgets. Table 1 consolidates the comparison. TAQ+IVF underperforms OPQ+IVFPQ on all six dataset/budget combinations, by 0.02–0.28 Recall@10: on in-distribution data, the data-optimized OPQ rotation dominates the query-covariance whitening. OPQ-Q+IVF falls behind on five of six rows; the single exception (MS-MARCO at 192 bits/vec, $+0.034$) occurs precisely where the water-filling floor collapses bit allocation to uniform, making OPQ-Q algorithmically identical to OPQ+IVFPQ; run inside our own harness, OPQ+IVFPQ reaches 0.579 there, so the gap is a FAISS-versus-harness difference, not a query-aware mechanism.

These negative results are the empirical face of Finding 1, and they shape the method we propose next.

4 LQ-PQ: Locally Query-Aware Product Quantization

LQ-PQ operationalizes Finding 2 under the constraints imposed by Section 3.3. The design principle: *query awareness may only enter components that cannot distort search geometry.* Rotation, IVF partition, residual encoding, code format, and the ADC search loop are all standard; only PQ codebook *training* is reweighted, locally. Pseudocode for the full offline pipeline appears in Appendix B.

Step 1: OPQ rotation. Learn the orthogonal R by standard OPQ on the database; rotate database and queries. Orthogonality preserves inner products and IVF geometry.

Step 2: IVF with residual encoding. Build a K-cell IVF index on the rotated database (K=1024); encode residuals $r_i = x'_i - c_{a(i)}$, matching standard IVF-PQ semantics.

Step 3: Coarse query regions with shrinkage. Cluster the K IVF centroids into G groups ($G \in \{8, 16, 32\}$; diagnostics support coarse granularity). Route calibration queries to cells (nprobe $= 1$), aggregate per group g, and estimate the local second moment with shrinkage toward the global:

$$H_g \;=\; (1 - \alpha_g)\, H \;+\; \alpha_g\, \widehat{H}_g, \qquad \alpha_g = \frac{M_g}{M_g + \tau}, \tag{5}$$

where M_g is the group's query count and τ a shrinkage constant. Groups with few queries fall back to the global H; groups with many queries exploit their local structure. This directly addresses the small-sample failure mode identified in Finding 2.

Step 4: H_g-weighted codebook training. For each group g and subspace s, let $B_{g,s}$ be the (ridge-regularized) block of H_g on the dimensions of s. Train the codebook by k-means in the metric induced by $B_{g,s}$: equivalently, standard k-means on $B_{g,s}^{1/2}$-transformed residual sub-vectors, mapping centroids back by $B_{g,s}^{-1/2}$. With $H_g = I$ this reduces exactly to plain PQ. The transform lives entirely inside training; encoded vectors, code sizes, and lookup-table search are unchanged.

Step 5: Search. Standard IVF-PQ ADC. The query's group is known from its nearest coarse centroid; one LUT is built per group encountered among the probed cells (at most G), each exactly as in standard PQ. The query rotation is shared with OPQ+IVFPQ, so that is the only extra online cost.

An exact test of local query information. LQ-PQ isolates one question with no confounds: does local *query* covariance improve codebook adaptation beyond local *data* covariance, as in LOPQ (Kalantidis and Avrithis, 2014)? We hold every design choice fixed to LQ-PQ's and change only the covariance source, giving an LOPQ-style control (canonical LOPQ additionally learns per-cell rotations, which we omit for both methods) (Σ_g in place of H_g) that differs from LQ-PQ in this one respect alone, to our knowledge the first such controlled test in the retrieval literature. The question is not rhetorical: Finding 2 shows the two covariances differ by $\sim 50^\circ$–70° of principal angle locally on MS-MARCO.

5 Experiments

Setup. Datasets: MS-MARCO-1M (d=384, inner product; RAG-representative) and SIFT1M (d=128); GloVe-1.2M (d=100) enters only the global-method comparison of Table 1. Calibration: the large train query split used in Section 3, disjoint from evaluation queries. Metric: Recall@10 at fixed bits per vector, best over nprobe $\in \{1, \ldots, 128\}$. All methods use identical total bit budgets, IVF K=1024, residual encoding.

Baselines. FAISS IVFPQ (Jégou et al., 2011; Johnson et al., 2019); OPQ+IVFPQ (primary baseline) (Ge et al., 2014); an LOPQ-style (Kalantidis and Avrithis, 2014) local-data control (Σ_g for H_g); a global-H ablation of our own LQ-PQ ($\alpha_g \equiv 0$); plus, for completeness, our own global query-aware instantiations from Section 3.3 (TAQ+IVF, OPQ-Q+IVF); and an unweighted per-group control ($B_{g,s}$=I) that isolates locality from covariance weighting. FAISS-IVFPQ and OPQ+IVFPQ are library configurations; the local methods share one implementation whose scoring matches `faiss.IndexIVFPQ` exactly under shared codebooks, and whose OPQ+IVFPQ re-implementation lands within about 1 point of FAISS on the inner-product datasets. Codebooks are trained and residuals assigned in the $B_{g,s}^{1/2}$-transformed space; queries are routed to cells by inner product.

Main result: locality pays; the covariance source does not. Table 2 reports the decisive MS-MARCO run at 8 bits per subspace (384 bits/vector) with 100K calibration queries (G=32, τ=20000, nprobe = 128, $M_g \geq 2000$ per region). Both local methods beat OPQ+IVFPQ (0.7650) by about $+3.3$ points (paired bootstrap, 10K resamples, $p < 10^{-4}$). The unweighted per-group control already gains $+1.5$, so roughly half of the local benefit is locality alone and the remainder is covariance weighting. But LQ-PQ and the LOPQ-style control differ by only $+0.12$ points ($p = 0.20$); an earlier 3×4 sweep over (G, τ) reached the same conclusion in every cell. The covariance *source* does not matter.

SIFT1M: noise hurts, as diagnosed. On SIFT1M, where Finding 2 identified the local query signal as sampling noise, query covariance actively *hurts* codebook training: the LOPQ-style control reaches 0.4018/0.6284 Recall@10 at 4/8 bits per subspace versus 0.3475/0.5847 for LQ-PQ (paired bootstrap, $p < 10^{-4}$ at both budgets). LOPQ-style still beats OPQ+IVFPQ (0.3720/0.5898 in the same run) while LQ-PQ falls below it: fitting codebooks to a noise-dominated query covariance costs more than locality gains.

Ablations. Finer granularity helps both methods equally (G=32 best), and the unweighted control shows that per-region specialization alone provides about half of the gain, with covariance weighting

Table 2: Decisive MS-MARCO run (100K cal. queries, 8 bits/sub, G=32, nprobe=128). Gains are vs. OPQ+IVFPQ.

Method	R@10	gain
FAISS-IVFPQ	.5816	
OPQ+IVFPQ	.7650	
Unweighted local (B=I)	.7801	+1.5
LOPQ-style	.7975	+3.3
LQ-PQ (ours)	.7987	+3.4

Table 3: Yandex T2I (OOD): Recall@10 at nprobe=128, equal bits/vec; local rows at G=32, τ=5000. All query-aware and local methods exceed OPQ+IVFPQ significantly (p_{bonf}<.001).

Method	100 b/vec	200 b/vec
FAISS-IVFPQ	.1098	.2543
OPQ+IVFPQ	.1191	.2721
TAQ+IVF (ours)	.1258	**.3427**
OPQ-Q+IVF (ours)	.1364	.2890
LQ-PQ (ours)	.1384	.3405
LOPQ-style	**.1449**	.3257

supplying the rest. With small calibration sets (1.7K queries), heavy shrinkage and even pure-global weighting outperformed local estimates; the 100K-query run removes this sample-size bottleneck without changing the conclusion.

Why is real local structure redundant? A plausible mechanism: k-means centroids sit where the *data* mass is, so reweighting the metric by H_g only redistributes error among directions the local residuals already occupy; if query-heavy directions carry little residual variance, the reweighting buys nothing. Learned codebooks or re-ranking may fare differently.

6 The Out-of-Distribution Regime

Findings 1–2 and the controlled comparison of Section 5 establish that in-distribution, the task-aware objective carries no information beyond the data distribution. Its premise, $H \neq \Sigma$, holds by construction in cross-modal retrieval. We test the full method suite on the Yandex Text2Image benchmark (Simhadri et al., 2024): 1M image embeddings (d=200, inner product) searched by text-query embeddings from a separate encoder tower; 50K calibration / 10K evaluation queries, disjoint; exact ground truth by brute force.

The diagnostic confirms this is the payable regime. Applying the protocol of Section 3 to T2I, before running any method: query effective rank 21.6 vs. data 116.2 (of 200); top-20 subspace alignment 0.50, against 0.77–0.96 in-distribution; per-cell alignment 2–3× the shuffled null, with split-half reliability up to 0.97. Both global and local query structure are real and large here, exactly the signature Finding 1 identifies as necessary for query awareness to pay off.

Setup. All methods at equal bits/vector (m=25; 4 and 8 bits/subspace), shared IVF (K=1024, residual encoding for all IVF variants), nprobe swept to 128 for every method, paired bootstrap with Bonferroni correction, plus an exhaustive-search arm with no routing.

Result 1: our proposed methods lead at 8 bits. TAQ (H-whitening plus query-spectrum water-filling allocation, Section 2) reaches 0.3427 Recall@10, +7.1 points over OPQ+IVFPQ and +1.7 over the strongest non-proposed method, LOPQ-style local codebooks (0.3257); our local LQ-PQ follows within 0.2 points at 0.3405 (Table 3). An exhaustive-search control (all methods scoring all 1M vectors, raw encoding: PQ 0.222, OPQ+PQ 0.251, TAQ 0.265) shows TAQ's quantizer beats OPQ's by +1.4 points ($p_{\text{bonf}} < 0.001$) even without routing; the larger gain under IVF reflects the combination with residual encoding and coarse partitioning, which we do not decompose further. Notably, the whitening transform that *hurt* in-distribution (distorting IVF geometry against a near-isotropic H) helps here: under severe anisotropy (query anisotropy 37×), normalizing the query spectrum before quantization outweighs the geometric cost.

Result 2: the OOD regime rewards query awareness broadly, but TAQ leads it. Every query-aware and local method we test beats OPQ+IVFPQ at both budgets (Table 3), by +0.7 to +7.1 Recall@10 points, all significant, confirming the diagnostic's alignment score separates the regimes exactly as the theory predicts. Among them, TAQ's margin is the largest at 8 bits; we examine when this ordering changes next.

Result 3: local query awareness pays here too, given bit headroom. At 8 bits, LQ-PQ beats its local-data control by +1.5 points (95% CI $[+1.2, +1.8]$), the reverse of the in-distribution tie: where

the diagnostic finds real local query structure and the budget can express it, local query covariance is no longer redundant. At 4 bits/subspace both query-aware methods lose to local data codebooks: a floor constraint collapses TAQ's allocation (4) to whitening-only (0.1258 vs. 0.1449), and LQ-PQ trails LOPQ-style by 0.6 points. Query awareness on this benchmark thus pays at both granularities, but needs bit headroom, not extreme compression.

Throughput note. TAQ's ADC performs the same per-candidate LUT lookups as standard PQ plus one $d\times d$ whitening per query ($\approx 4\,\mu$s). The $\sim 100\times$ QPS gap of our prototype is mostly a Python vs. C++ artifact; a C++ kernel would narrow it substantially, though variable-width codes preclude FastScan-style uniform-nibble SIMD lookups, so part of the gap is structural.

7 Conclusion

We asked when query-aware quantization helps, and obtained a complete, two-regime answer. In-distribution, it does not help at any granularity: the query and data covariances coincide globally, making query-agnostic rotations implicitly query-aware, and the real local query structure we detect is redundant for codebook training: local codebooks provide a $\sim$3-point gain whichever covariance drives them. Out-of-distribution, where the premise $H \neq \Sigma$ holds by construction, the conclusion reverses: every query-aware and local method beats OPQ+IVFPQ on text-to-image retrieval, our global TAQ and local LQ-PQ are strongest at moderate bit budgets ($+7.1$ Recall@10 over OPQ+IVFPQ, quantizer-level advantage confirmed by exhaustive-search controls), and a bit-allocation floor explains the crossover to local codebooks at extreme compression. Local *query* weighting ties local *data* weighting in-distribution, hurts where the local signal is noise, and pays out-of-distribution at sufficient bit budget: the diagnostic's local test predicts all three outcomes.

The diagnostic protocol separates these regimes from calibration queries alone, in minutes, across all four datasets studied, and doubles as a pre-deployment test: when alignment is high, the actionable result is simpler than query awareness (local codebooks add $\sim$3 Recall@10 points over OPQ+IVFPQ at equal bits); when alignment is low and the bit budget permits, spectral query-aware quantization is worth its complexity. Our conclusions cover k-means product quantization with IVF; other quantizer families may exploit query structure differently. The split-half protocol needs substantial calibration queries per region, the high-rate model of Section 2.1 is loose at very low bits, and broader workload coverage (recommendation, code search, multilingual) would strengthen generality.

References

P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui. Retrieval-augmented generation for AI-generated content: a survey. *arXiv:2402.19473*, 2024.

M. Shen, M. Umar, K. Maeng, G. E. Suh, and U. Gupta. Towards understanding systems trade-offs in retrieval-augmented generation model inference. *arXiv:2412.11854*, 2024.

P. Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 33(1):117–128, 2011.

T. Ge, K. He, Q. Ke, and J. Sun. Optimized product quantization. *IEEE TPAMI*, 36(4):744–755, 2014.

Y. Kalantidis and Y. Avrithis. Locally optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2014.

J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019.

R. Guo, P. Sun, E. Lindgren, Q. Geng, D. Simcha, F. Chern, and S. Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *ICML*, PMLR 119:3887–3896, 2020.

J. Zhang, Q. Liu, D. Lian, Z. Liu, L. Wu, and E. Chen. Anisotropic additive quantization for fast inner product search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4354–4362, 2022.

J. Zhang, D. Lian, H. Zhang, B. Wang, and E. Chen. Query-aware quantization for maximum inner product search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4875–4883, 2023.

S. Jaiswal, R. Krishnaswamy, A. Garg, H. V. Simhadri, and S. Agrawal. OOD-DiskANN: Efficient and scalable graph ANNS for out-of-distribution queries. In *WWW*, 2023.

H. V. Simhadri et al. Results of the Big ANN: NeurIPS'23 competition. *arXiv:2409.17424*, 2024.

J. He, J. Xing, N. Wang, R. Xu, S. Wu, P. Zhou, Q. Liu, C. J. Xue, and Q. Li. A^{2}ATS: Retrieval-based KV cache reduction via windowed rotary position embedding and query-aware vector quantization. *arXiv:2502.12665*, 2025.

Z. Chen et al. Adaptive bit allocation product quantization for approximate nearest neighbor search. *Neurocomputing*, 2015.

R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Trans. Inf. Theory*, 44(6):2325–2383, 1998.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

Large language model assistants (Claude) were used substantially throughout this project's research and writing process, and their use is disclosed here in the interest of transparency; the authors take full responsibility for the submission's content, correctness, and originality.

Research and experimentation. An LLM assistant helped formalize the task-aware objective (Theorem 1) from the authors' informal research idea, and helped writing the proofs in Appendix A, which the authors verified, helped design the noise-controlled diagnostic protocol (shuffled nulls, split-half reliability). A separate LLM coding agent (Claude Code) implemented the experimental codebase, ran all reported experiments, and was used interactively to debug the implementation. All experimental results were reviewed by the authors, and key comparisons were independently verified against library reference implementations (e.g., exact scoring parity with `faiss.IndexIVFPQ`).

Writing. LLM assistance was used substantially in drafting, restructuring, and copyediting this manuscript, including tightening prose, checking internal consistency between the text, tables, and figures, and verifying that mathematical notation was defined before use. The authors reviewed, edited, and take responsibility for all text, claims, and conclusions in this paper.

A COMPLETE PROOFS

Proof of Theorem 1. Fix $x \in \mathbb{R}^d$ and let $r := x - \hat{x}$. Since the quantizer Q is trained without access to the evaluation query q, the residual r is a fixed (non-random) vector with respect to the distribution of q. By the definition of ϵ,

$$\epsilon(q,x)^2 \;=\; \big(q^\top r\big)^2 \;=\; \big(q^\top r\big)\big(r^\top q\big) \;=\; r^\top\big(qq^\top\big)\, r.$$

Taking expectations over $q \sim \mathcal{Q}$ and using linearity of expectation entrywise (each entry of $qq^\top$ has finite expectation because $\mathbb{E}\|q\|^2 < \infty$ bounds all second moments via Cauchy–Schwarz),

$$\mathbb{E}_q\big[\epsilon(q,x)^2\big] \;=\; r^\top \mathbb{E}_q\big[qq^\top\big]\, r \;=\; r^\top H\, r \;=\; (x-\hat{x})^\top H\,(x-\hat{x}). \qquad \square$$

Proof of Proposition 1. Let (x^+, x^-) be any pair with margin $\Delta(q) = s(q,x^+) - s(q,x^-) > 2\delta$, and suppose $|\epsilon(q,x)| \le \delta$ for all $x \in \mathcal{D}$. The quantized score gap is

$$s(q,\hat{x}^+) - s(q,\hat{x}^-) \;=\; \big[s(q,x^+) - \epsilon(q,x^+)\big] - \big[s(q,x^-) - \epsilon(q,x^-)\big] \;=\; \Delta(q) - \epsilon(q,x^+) + \epsilon(q,x^-).$$

By the triangle inequality, $\big|\epsilon(q,x^+) - \epsilon(q,x^-)\big| \le 2\delta < \Delta(q)$, hence $s(q,\hat{x}^+) - s(q,\hat{x}^-) \ge \Delta(q) - 2\delta > 0$: the quantized scores preserve the original order of the pair. $\square$

Derivation of the tail bound in Section 2. Let $Z := \epsilon(q, x^+) - \epsilon(q, x^-)$. For a fixed threshold $\Delta > 0$, a rank inversion requires $Z \geq \Delta$, so Markov's inequality applied to Z^2 gives $\Pr[Z \geq \Delta] \leq \mathbb{E}[Z^2]/\Delta^2$. Since $Z^2 \leq 2\big(\epsilon(q, x^+)^2 + \epsilon(q, x^-)^2\big)$ and $\mathbb{E}_q[\epsilon(q, x)^2] = e(x)$ by Theorem 1, $\Pr[Z \geq \Delta] \leq 2\big(e(x^+) + e(x^-)\big)/\Delta^2 \leq 4 \max_x e(x)/\Delta^2$. No zero-mean assumption is needed. □

B LQ-PQ ALGORITHM

Algorithm 1 LQ-PQ (offline)

Require: database $\{x_i\}$, calibration queries $\{q_j\}_{j=1}^{M}$, subspaces m, cells K, groups G, shrinkage τ, ridge η, bits b

$R \leftarrow \mathrm{OPQ}(\{x_i\}, m)$; $x'_i \leftarrow Rx_i$; $q'_j \leftarrow Rq_j$

IVF: cluster $\{x'_i\}$ into K cells; residuals $r_i \leftarrow x'_i - c_{a(i)}$

Group cells: k-means on centroids $\rightarrow$ groups $1..G$

Route queries to cells (nprobe=1); aggregate per group

$H \leftarrow \frac{1}{M} \sum_j q'_j q'^{\top}_j$; $\widehat{H}_g \leftarrow \frac{1}{M_g} \sum_{j \in g} q'_j q'^{\top}_j$

$H_g \leftarrow \big(\tau H + \sum_{j \in g} q'_j q'^{\top}_j\big)/(\tau + M_g)$ ▷ equals $(1 - \alpha_g)H + \alpha_g \widehat{H}_g$; well-defined at $M_g{=}0$

for each group g, subspace s **do**

  $B_{g,s} \leftarrow H_g[\mathrm{dims}[s], \mathrm{dims}[s]] + \eta I$

  k-means on $\{B_{g,s}^{1/2} r_{i,s} : a(i) \in g\}$ with 2^b centroids; store $B_{g,s}^{-1/2} \times$centroids

end for

Encode all residuals with their group's codebooks