SOPQ: SIMPLE OPTIMIZED PRODUCT QUANTIZATION VIA CORRECTED REVERSE WATER FILLING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents SOPQ. It is a provably correct and one-shot optimization algorithm for optimizing Product Quantization (PQ). The main combinatorial challenge is that of optimally assigning coordinates to *variable-width* PQ segments. OPQ (Ge et al., 2014) suggests a good fixed-segment-width heuristic. SOPQ, on the other hand, provably approximates the optimal solution. The crux of the algorithm is a novel Corrected Reverse Water Filling (CRWF) formula that accounts for dimension-dependent effects in k-means clustering. Thereby, converting the coupled combinatorial problem to standard bin-packing of coordinates to segments by their CRWF allocated bits. We then prove that the SVD rotation globally minimizes the CRWF waterline while simultaneously decorrelating the coordinates.[1] SOPQ outperforms EDEN, E-RaBitQ, TurboQuant, and PQ at low bit rates across five VQ-bench datasets. It matches OPQ's compression quality while offering strong guarantees, faster and predictable training, and a smaller memory footprint.

1 INTRODUCTION AND PROBLEM FORMULATION

Vector Quantization (VQ) (Shannon, 1948; Shannon et al., 1959; Gersho & Gray, 1991) is a cornerstone of modern AI infrastructure. Vector databases (e.g. Pinecone **?**) rely on Product Quantization (PQ) and Optimized Product Quantization (OPQ) for billion-scale approximate nearest neighbor (ANN) search (Jégou et al., 2011; Ge et al., 2013; 2014; Johnson et al., 2019; Simhadri et al., 2022), while large language models (LLMs) quantize model weights (Frantar et al., 2023; Lin et al., 2024; Dettmers et al., 2023; Chee et al., 2023; Tseng et al., 2024) and key-value (KV) caches (Liu et al., 2024; Hooper et al., 2024; Kang et al., 2024) down to 0.5-4 bits per coordinate. See (Padaki et al., 2026) for a comprehensive survey and comparative benchmarks.

Recent fast quantizers (Vargaftik et al., 2022; Gao & Long, 2024; Gao et al., 2025; Zandieh et al., 2025; Ashkboos et al., 2024) precondition data with a random rotation (Johnson & Lindenstrauss, 1984; Ailon & Chazelle, 2009; Ailon & Liberty, 2009) so that all coordinates share identical marginal variance and a single scalar codebook. This is optimal *only* if every coordinate is artificially forced to use the exact same bit budget. But, without that constraint, random rotation almost *maximizes* reconstruction error.

Empirical quality standards are set by methods that precondition data with carefully chosen rotations. Among multi-codebook methods, OPQ (Ge et al., 2013; 2014) sets the gold standard. Its design, however, is fundamentally heuristic. Non-Parametric OPQ alternates between full $d \times d$ Procrustes SVD updates and k-means re-clustering. Parametric OPQ forces every subspace to have the same fixed width d/m and greedily buckets eigenvalues without accounting for low-bit space-filling penalties or zero-bit coordinates.

Our goal in this paper is to derive a provable, fast algorithm for minimizing PQ reconstruction error that matches the quality of OPQ.

The data model: the dataset $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$ is assumed drawn from a d-dimensional mean-zero Gaussian distribution $\mathcal{N}(0, \Sigma)$. Assuming zero mean loses no generality because the

[1] Random rotation, a common pre-conditioner in quantization, is shown to effectively *maximize* mean squared error.

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.

algorithm explicitly centers the data ($\mu \leftarrow \frac{1}{n}\sum_{i=1}^{n} x_i$) prior to rotation. Modeling the centered marginals as Gaussian is not just a mathematical convenience. Among all distributions with a given covariance matrix Σ, the Gaussian distribution maximizes differential entropy and the Shannon rate-distortion function $R(D)$ (Cover & Thomas, 2006). Minimizing PQ distortion for a Gaussian source therefore minimizes an *upper bound* on distortion across arbitrary distributions with covariance Σ.

Objective Function: A generalized PQ is specified by two design choices:

1. An isometric rotation matrix $R \in O(d)$: the input vectors $x \in \mathbb{R}^d$ are rotated to $y = Rx$.
2. A grouping $G = \{g_1, \ldots, g_m\}$: each subset of coordinates $g_r \subset [d]$ $(r = 1, \ldots, m)$ is quantized independently by running $|g_r|$-dimensional k-means with k centroids, plus potentially an unassigned set $g_0 = [d] \setminus \bigcup_{r=1}^{m} g_r$ which is unquantized and reconstructed by the value 0.

Letting $e_i^2 = \mathbb{E}[(y_i - \hat{y}_i)^2]$ denote the resulting Mean Squared Error (MSE) contributed by reconstructing coordinate i, our objective is to minimize total MSE jointly over R and G:

$$\min_{R \in O(d)} \min_{G} \sum_{i=1}^{d} e_i^2(R, G). \tag{1}$$

2 Optimal Grouping Without Rotation

We first solve the inner grouping problem in (1) assuming the coordinates $y_1, \ldots, y_d$ are mutually independent mean-zero Gaussians with variances $s_1^2 \geq \cdots \geq s_d^2$. Section 4 will explain why requiring mutual independence is not limiting.

The structure of the section is as follows. First, we show that, for given segment widths, any grouping that achieves $e_i^2 = \min(\theta^2, s_i^2)$ for some parameter θ is optimal. Here, $e_i^2 = s_i^2$ iff $i \in g_0$, which is the unquantized inactive set. For all other $i \leq d'$ coordinates, $e_i^2 = \theta^2$. Within each active cluster, $e_i^2 = e_j^2$ "for free" by k-means. However, getting $e_i^2 = e_j^2$ across groups is combinatorially hard. We therefore devise a Corrected Reverse Water Filling (CRWF) formula. Using this formula we assign $b_1^\star, \ldots, b_{d'}^\star$, and show that if two groups have the same total bit allocation, their per-coordinate k-means error is almost identical. Then, simply bin-packing $b_i^\star$ into m bins produces a near-optimal grouping.

Observe that any coordinate $i \in [d]$ can be reconstructed trivially by its mean 0 using zero bits and incurring $e_i^2 = \mathbb{E}[(y_i - 0)^2] = s_i^2$. Consequently, including coordinate i in a quantized segment g is beneficial if and only if the resulting error satisfies $e_i^2 < s_i^2$; otherwise coordinate i is best left unquantized ($i \in g_0, \hat{y}_i = 0$) with $e_i^2 = s_i^2$. Going forward, we only group active coordinates $i \leq d'$ such that $s_i^2 > \theta^2$. We explain how to compute θ^2 and d' in Section 3.

Theorem 1 (Zador-Gersho Gaussian Vector Quantization Theorem Zador, 1982; Gersho, 1979; Gray & Neuhoff, 1998)**.** *Let $g \subset [d]$ be a segment of dimension $d_g = |g|$ consisting of independent Gaussian coordinates $\{y_j : j \in g\}$ with variances $\{s_j^2 : j \in g\}$. Whenever all coordinates in g are actively quantized an optimal k-means codebook in $\mathbb{R}^{d_g}$ equates the per-coordinate MSE across all coordinates $i \in g$. The common Zador-Gersho segment error formula is:*

$$e_i^2 \;=\; c_{d_g} \left(\prod_{j \in g} s_j^2 \right)^{1/d_g} k^{-2/d_g} \qquad \forall\, i \in g, \tag{2}$$

where c_d is the Gaussian Zador–Gersho factor (Zador, 1982; Gersho, 1979), which decreases from $c_1 = \sqrt{3}\pi/2 \approx 2.72$ to 1 as $d \to \infty$ (values in Appendix A).

In Appendix C (Table 5) we confirm this empirically across $d_g \in \{1, 2, 4, 8, 16, 32, 64\}$. Theorem 1 holds accurately at finite PQ codebook size $k = 256$ even across a $256\times$ within-segment variance ratio, k-means equalizes e_i^2 across coordinates within 0.8%-3.1% while matching Equation (2).

Using Theorem 1, we now characterize the optimal grouping across all m segments for given segment widths.

Theorem 2 (Across-Segment Error Equalization and the Waterline θ^2)**.** *Let $s_1^2 \geq \cdots \geq s_d^2 > 0$ and fix segment widths $d_{g_1}, \ldots, d_{g_m}$ with $\sum_{r=1}^m d_{g_r} = d'$. Every partition $G = \{g_1, \ldots, g_m\}$ of the active coordinates $[d']$ into segments of these widths (with unquantized set $g_0 = [d] \setminus [d']$) satisfies $\sum_{i=1}^{d'} e_i^2 \geq d'\theta^2$, where*

$$\theta^2 \;:=\; \Big(k^{-2m} \prod_{r=1}^{m} c_{d_{g_r}}^{d_{g_r}} \prod_{j=1}^{d'} s_j^2\Big)^{1/d'} \tag{3}$$

*depends on the widths but not on which coordinates are assigned to which segment. Equality holds **if and only if** all m segments achieve the same per-coordinate error, which is then θ^2. Hence any equal-error partition is optimal among partitions with these widths, and if $s_{d'+1}^2 \leq \theta^2 < s_{d'}^2$, its total MSE is $\sum_{i=1}^{d} \min(\theta^2, s_i^2)$.*

Proof. Raising Equation (2) to the power d_{g_r} gives $\big(e^2(g_r)\big)^{d_{g_r}} = c_{d_{g_r}}^{d_{g_r}} (\prod_{j \in g_r} s_j^2) k^{-2}$. Multiplying across $r = 1, \ldots, m$ cancels the $1/d_{g_r}$ exponents on the variances and on k:

$$\prod_{r=1}^{m} \big(e^2(g_r)\big)^{d_{g_r}} \;=\; \left(\prod_{r=1}^{m} c_{d_{g_r}}^{d_{g_r}}\right) \left(\prod_{j=1}^{d'} s_j^2\right) k^{-2m} \;=\; \big(\theta^2\big)^{d'}. \tag{4}$$

For fixed widths, the right-hand side does not depend on how the coordinates are assigned to segments. By the weighted AM–GM inequality, the total active error satisfies

$$\frac{1}{d'} \sum_{r=1}^{m} d_{g_r}\, e^2(g_r) \;\geq\; \left(\prod_{r=1}^{m} \big(e^2(g_r)\big)^{d_{g_r}}\right)^{1/d'} \;=\; \theta^2, \tag{5}$$

with equality if and only if $e^2(g_r) = \theta^2$ for all $r \in [m]$. □

The widths enter (3) only through $\prod_r c_{d_{g_r}}^{d_{g_r}}$. Since $d \ln c_d$ grows slowly with d (from 1.05 at $d = 1$ to 3.0 at $d = 64$; Table 4), different width profiles yield slightly different waterlines; under CRWF, this width dependence reduces to the small Jensen gaps of Theorem 3.

2.1 The Combinatorial Challenge and Why OPQ Is Heuristic

By Theorem 2, our goal is to partition the active coordinates $[d']$ into m disjoint groups $g_1, \ldots, g_m$ such that every group has the same per-coordinate k-means error, Equation (2). Directly searching for a variable-width partition is a challenging combinatorial problem because both the space-filling constant c_{d_g} and the exponent $-2/d_g$ depend non-linearly on the cardinality $d_g = |g|$ of the group.

Parametric OPQ (Ge et al., 2014) circumvents this coupling heuristically by *forcing all segments to have the exact same fixed width $d_g = d/m$*. While holding $d_g = d/m$ constant factors $c_{d/m} k^{-2m/d}$ out of (2) (reducing OPQ's bucketing rule to greedily balancing $\prod_{j \in g_r} s_j^2$), this restriction incurs two severe penalties in the low-bit regime: (1) it cannot discard zero-budget coordinates below the waterline ($d' < d$), wasting rotation flops and codebook capacity on trailing noise dimensions, and (2) it forces leading high-variance components into fixed d/m-width subspaces, requiring secondary within-subspace rotations or expensive iterative Procrustes–Lloyd optimization.

3 The Solution: Corrected Reverse Water Filling (CRWF)

We now show that the non-linear, dimension-dependent condition $e^2(g) = \theta^2$ in (2) can be approximately decoupled into an additive per-coordinate bit-allocation rule: Corrected Reverse Water Filling (CRWF). We define the effective per-coordinate distortion coefficient $c_{\text{eff}} : \mathbb{R}_{\geq 0} \to [1, c_1]$ as the pullback of the space-filling penalty function c_d evaluated at dimension $d = C/b$:

$$c_{\text{eff}}(b) \;:=\; c_{C/b} \;\text{ for any }\; b > 0, \tag{6}$$

Here $C = \log_2(k)$ ($C = 8$ bits by default).

Using $c_{\text{eff}}(b)$, we define the CRWF coordinate bit allocation $b_j^\star(\theta)$ at waterline $\theta^2 > 0$ as the solution to

$$c_{\text{eff}}\big(b_j^\star\big)\, s_j^2\, 2^{-2b_j^\star} \;=\; \theta^2 \quad\Longleftrightarrow\quad b_j^\star(\theta) \;=\; \left(\underbrace{\frac{1}{2}\log_2 \frac{s_j^2}{\theta^2}}_{b_j^{\text{rwf}}(\theta)} \;+\; \underbrace{\frac{1}{2}\log_2 c_{\text{eff}}\big(b_j^\star(\theta)\big)}_{\text{space-filling correction } \Delta b_j \geq 0} \right)^+, \tag{7}$$

where $\theta > 0$ is chosen so that $\sum_{j=1}^d b_j^\star(\theta) = mC$. Because $c_{\text{eff}}(0) = 1$, all coordinates with $s_j^2 \leq \theta^2$ receive $b_j^\star(\theta) = 0$ bits and are set aside into the unquantized set g_0 (retaining $d' = |\{j : s_j^2 > \theta^2\}|$ active coordinates).

The following theorem is the crux of the paper: it proves that under CRWF (7), *any* segment $g \subset [d']$ whose allocated bit budgets $b_j^\star(\theta)$ sum to $C = \log_2 k$ achieves approximate per-coordinate k-means error $e^2(g) \approx \theta^2$ independent of the segment's dimension d_g!

Theorem 3 (CRWF). *Let $b_1^\star(\theta), \ldots, b_{d'}^\star(\theta) > 0$ be the active CRWF bit allocations defined by* (7)*, and let $\psi(b) := \ln c_{\text{eff}}(b)$. For any subset of coordinates $g \subset [d']$ of width $d_g = |g|$, if the CRWF bit budgets in g sum to the segment capacity,*

$$\sum_{j \in g} b_j^\star(\theta) \;=\; C \;=\; \log_2 k, \tag{8}$$

then the d_g-dimensional k-means per-coordinate error (2) *of segment g is*

$$e^2(g) \;=\; \theta^2 \exp(J_g), \qquad \textit{where} \qquad J_g \;:=\; \psi\Big(\frac{1}{d_g}\sum_{j \in g} b_j^\star\Big) \;-\; \frac{1}{d_g}\sum_{j \in g} \psi(b_j^\star) \tag{9}$$

is the Jensen gap of ψ over the allocations in g. In particular, $e^2(g) = \theta^2$ exactly whenever all coordinates in g receive the same allocation.

Proof. By (7), every $j \in g$ satisfies $s_j^2 = \theta^2\, 2^{2b_j^\star} / c_{\text{eff}}(b_j^\star)$. Substituting this and $k = 2^C = 2^{\sum_{j \in g} b_j^\star}$ into (2), the factors $2^{2b_j^\star}$ cancel:

$$e^2(g) \;=\; c_{d_g} \left(\prod_{j \in g} \frac{\theta^2\, 2^{2b_j^\star}}{c_{\text{eff}}(b_j^\star)} \right)^{1/d_g} 2^{-\frac{2}{d_g}\sum_{j \in g} b_j^\star} \;=\; \theta^2\, \frac{c_{d_g}}{\Big(\prod_{j \in g} c_{\text{eff}}(b_j^\star)\Big)^{1/d_g}}.$$

Since $d_g = C/\bar{b}_g$, definition (6) gives $c_{d_g} = c_{\text{eff}}(\bar{b}_g)$. Taking logarithms yields $\ln\big(e^2(g)/\theta^2\big) = \psi(\bar{b}_g) - \frac{1}{d_g}\sum_{j \in g} \psi(b_j^\star) = J_g$. □

The Jensen gap J_g is small because ψ is very flat: over the entire low-bit range $b \in [1/8, 4]$, $\psi(b)$ increases only from 0.05 to 0.72 (Table 4). Consequently, $e^2(g)$ is within about 2% of θ^2 even in extreme cases. For example, a segment pairing one 4-bit coordinate with thirty-two $1/8$-bit coordinates ($d_g = 33$) has $J_g = 0.016$, and the largest gap among all 30 segments of the five bucketings in Appendix D is $J_g = 0.027$.

3.1 Segment allocation by bin packing

Theorem 3 transforms the coupled multi-dimensional grouping problem (2) into an elementary one dimensional bin packing problem. Namely, group the d' active coordinates into m bins of equal total bit budget $\sum_{j \in g_r} b_j^\star(\theta) \approx C$.

A standard greedy approach is Longest Processing Time (`LPT` / `Most-Empty-Bin`). Assign coordinates in decreasing order $b_1^\star \geq \cdots \geq b_{d'}^\star$ to the segment with the smallest current bit sum. In Appendix D (Table 8), we verify Theorem 3 on a controlled 126-variable, 6-tier benchmark ($d_g \in [2, 64]$), confirming that under CRWF, all five segment bucketings (`FFD`, `LPT`, Complementary Fold, Adjacent Pairs, Equal-Width 21D) achieve $e_i^2 = 1.00 \pm 0.04$ across all tiers and identical Total MSE within 0.23% (126.29–126.58 vs. theoretical 126.00), whereas Standard RWF incurs a $1.96\times$ distortion staircase across tiers.

Table 1: Spectral statistics of PCA-rotated sources ($40{,}000$ rows each) and the resulting water-filling distortion gain over isotropic random-rotation allocation, with the active truncated dimension d' at $\bar{b} = 1$ bit/dim in parentheses.

Source	d_{eff}	λ_1/λ_d	AM/GM	coding gain	rel. spread	gain @1b (d')	gain @2b
Gaussian, Zipf(2) scales	1024	$3.0{\cdot}10^5$	183.6	3.76 b/dim	0.62	183.6× (962)	183.6×
GloVe-100	100	6.3	1.04	0.03 b/dim	0.05	1.04× (100)	1.04×
ImageNet-CLIP-512	511	$1.9{\cdot}10^4$	2.55	0.67 b/dim	0.16	2.45× (440)	2.52×
ArXiv-nomic-768	767	$1.1{\cdot}10^4$	4.80	1.13 b/dim	0.17	4.07× (513)	4.78×

4 Optimizing the Preconditioning Rotation $R \in O(d)$

Having shown that CRWF + LPT achieves total distortion $\sum_{i=1}^{d} \min(\theta^2(s^2), s_i^2)$ for any uncorrelated variance profile $s^2 = (s_1^2, \ldots, s_d^2)$, we now optimize the orthogonal preconditioner $R \in O(d)$. For the analysis we use the per-coordinate model $D_{\text{crwf}}(s^2; B) := \min_{\sum b_j \le B} \sum_{j=1}^{d} s_j^2 c_{\text{eff}}(b_j) 2^{-2b_j}$. Its minimizer equalizes $s_j^2 c_{\text{eff}}(b_j) 2^{-2b_j} (2\ln 2 - \psi'(b_j))$ rather than $s_j^2 c_{\text{eff}}(b_j) 2^{-2b_j}$ as in (7), so D_{crwf} is slightly below the CRWF distortion; the gap is under 0.2% for Zipf, exponential, and flat spectra at 0.25–2 bits per coordinate.

Recall from the Schur–Horn theorem (Horn, 1954; Marshall et al., 2011) that as R ranges over $O(d)$, the diagonal entries $s^2(R) = \text{diag}(R\Sigma R^\top)$ form the majorization polytope $\{s^2 \in \mathbb{R}^d_{\ge 0} : (\bar{\lambda}, \ldots, \bar{\lambda}) \prec s^2 \prec (\lambda_1, \ldots, \lambda_d)\}$, whose majorization maximum $s^2 = \lambda$ (maximal variance skew) is achieved by the SVD rotation $R = U^\top$, and whose minimum $s^2 = (\bar{\lambda}, \ldots, \bar{\lambda})$ (equal variances) is achieved by variance-equalizing/random rotations.

Theorem 4 (SVD Globally Minimizes CRWF Distortion; Random Rotation Maximizes It)**.** *For every bit budget $B > 0$ and any smooth distortion coefficient $c_{\text{eff}}(b)$ such that $\phi(b) := c_{\text{eff}}(b) 2^{-2b}$ is strictly decreasing and convex on $\mathbb{R}_{\ge 0}$ (which holds for the measured c_{eff} of Table 4), $D_{\text{crwf}}(s^2; B)$ is strictly Schur-concave in the active variance profile s^2. Consequently:*

1. ***Global Minimum (SVD Rotation):*** $\min_{R \in O(d)} D_{\text{crwf}}(s^2(R); B) = D_{\text{crwf}}(\lambda; B)$, *achieved simultaneously for every bit budget $B > 0$ by the SVD rotation $R = U^\top$.*
2. ***Global Maximum (Equal-Variance / Random Rotation):*** $\max_{R \in O(d)} D_{\text{crwf}}(s^2(R); B) = d\bar{\lambda}\, c_{\text{eff}}(B/d)\, 2^{-2B/d}$, *achieved whenever R equalizes the coordinate variances $s_1^2 = \cdots = s_d^2 = \bar{\lambda}$. For Haar-random $R \in O(d)$, $\max_j |s_j^2(R)/\bar{\lambda} - 1| = O(\sqrt{\rho \log d / d})$ w.h.p., driving distortion directly to the global worst case (proofs in Appendix B).*

The Dual Purpose of the SVD Rotation. Theorem 4 highlights that $R = U^\top$ serves two essential, simultaneous roles in SOPQ: (1) because $U^\top \Sigma U = \Lambda$ is strictly diagonal, $R = U^\top$ **decorrelates the coordinates** (making $y_1, \ldots, y_d$ mutually independent for Gaussian data, which is the exact precondition required for Within-Segment Equalization in Section 2), and (2) among all rotations, $R = U^\top$ **maximizes the variance skew** $s_j^2 = \lambda_j$, globally minimizing the waterline θ^2. Table 1 quantifies this spectral skew across four representative sources.

5 The Complete SOPQ Algorithm

Combining centering, SVD rotation, CRWF bit allocation with zero-bit coordinate truncation ($d \to d'$), LPT variable-width bin packing, and independent per-segment k-means yields **SOPQ** (Algorithm 1):

Proposition 5 (Lossless Truncation and Asymptotic Complexity of SOPQ)**.** *For active dimension $d' = |\{j : b_j^\star(\theta) > 0\}| \le d$, m segments of widths $\sum_{r=1}^{m} d_{g_r} = d'$, and codebook size k (proof in Appendix B):*

1. ***Lossless $d \to d'$ Truncation:*** *Trimming $U \in \mathbb{R}^{d \times d}$ to $U_{d'} \in \mathbb{R}^{d \times d'}$ produces the exact same reconstructions $\hat{x} = \mu + U_{d'}\hat{y}$ and inner products $\langle q, \hat{x} - \mu \rangle = \langle U_{d'}^\top q, \hat{y} \rangle$ as the full $d \times d$ rotation while reducing rotation memory and flops by a factor of d'/d.*

Algorithm 1: The SOPQ algorithm.

```
fit(X, b, k=256):
  mu ← mean(X);  Sigma ← cov(X - mu);  U, s2 ← eig(Sigma)    # 1. Center & SVD
  C  ← log2(k);  B ← d * b;  m ← round(B / C)
  ceff ← interp(C / d, c_d)                                  # ceff(b) = c_{C/b}, Table 4
  theta, b_star, d' ← solve_crwf(s2, B, ceff)                # 2. CRWF bit allocation, Eq. (7)
  U_d' ← U[:, 1:d'];  Y ← U_d'^T (X - mu)                    # 3. Trim zero-bit coords (d' ≤ d)
  scalar_encode({j ≤ d' : b_star[j] > C});  update B, m      # 1D encode if b_j^\star > C
  G ← bit_pack_LPT(b_star[1:d'], m)                          # 4. LPT variable-width segments
  for each segment g_r in G:  C_r ← kmeans(Y[g_r], k)        # 5. Per-segment k-means
  return model = (mu, U_d', G, {C_r})

encode(x):
  y ← U_d'^T (x - mu)                                        # O(d * d') truncated projection
  return concat_{r=1..m}( argmin_{c in C_r} ||y[g_r] - c||_2 )

score(q):
  qt ← U_d'^T q;  for r=1..m: LUT_r[c] ← <qt[g_r], C_r[c]>   # O(d*d' + d'*k) once per query
  for each candidate code:  score ← sum_{r=1}^m LUT_r[code_r] # O(m) byte-aligned SIMD lookups
```

2. ***Zero Asymptotic Dictionary Memory Overhead:*** *Total model memory is* $d + d \cdot d' + k\sum_{r=1}^{m} d_{g_r} + d'\lceil \log_2 m \rceil = d \cdot d' + d'k + O(d' \log m)$*, which is* $\Theta(d \cdot d')$ *whenever* $k = O(d)$*.*

3. ***LPT Spillage Bound:*** *LPT produces exact* m *segments with* $\max_{1 \le r \le m} \sum_{j \in g_r} b_j^\star - C \le \frac{m-1}{m} b_{j^\star}^\star < 1$ *bit whenever the last placed coordinate satisfies* $j^\star \ge B$*.*

6 Experimental Evaluation

We evaluate SOPQ inside VQ-bench (Padaki et al., 2026) on five embedding benchmarks from VIBE (Jääsaari et al., 2025): Llama ($n = 257$K, $d = 128$, inner product), Yahoo ($n = 677$K, $d = 384$), ImageNet ($n = 1.28$M, $d = 512$), COCO ($n = 282$K, $d = 768$), and CCNews ($n = 495$K, $d = 768$). Every quantizer is fit on $25{,}000$ database vectors, encodes the full database, and is evaluated on $1{,}000$ held-out queries. Recall@k is the overlap between the approximate and true top-k when each query's 100 true nearest neighbors are re-ranked by approximate score. Reported bit rates include model size (rotation $U_{d'}$, codebooks $\{C_r\}$, metadata, etc.).

We compare against 15 quantizers in VQ-bench, including TurboQuant (Zandieh et al., 2025), E-RaBitQ (Gao et al., 2025), EDEN (Vargaftik et al., 2022), PQ (Jégou et al., 2011), and OPQ (Ge et al., 2013). SOPQ uses $k = 256$ and the c_{eff} values in Appendix A. All timing measurements were conducted on a GCP Compute Engine virtual machine equipped with a 64-vCPU AMD EPYC 7B13 processor with clock speed 2.0 GHz and 120 GB of ECC RAM. The encoding phase for each quantizer used all 64 threads.

End-to-End Recall on VQ-bench. Figure 1 plots Recall@10 against total bits/dim on Yahoo, ImageNet, and COCO; Table 2 reports all five datasets near 2 and 4 bits/dim (Llama and CCNews curves in Appendix E, Figure 2). SOPQ matches OPQ to within 0.013 Recall@10 on every dataset and rate (e.g., 0.756 vs. 0.753 on ImageNet and 0.685 vs. 0.681 on COCO near 2 b/dim), and outperforms the rotation-plus-scalar quantizers EDEN, E-RaBitQ, and TurboQuant at ≈ 2 b/dim on all five datasets, by up to 0.30 on COCO (0.685 vs. 0.382 for EDEN). Table 3 compares time and memory metrics on ImageNet ($n = 1.28$M): SOPQ roughly enjoys the same time and memory savings as OPQ-P (e.g., over $10\times$ faster fit compared to OPQ, due to no iterative loop).

7 Summary and Conclusion

We presented SOPQ, a one-shot algorithm for optimizing Product Quantization derived from first principles. Its analysis rests on three results: (1) k-means equalizes the per-coordinate error within

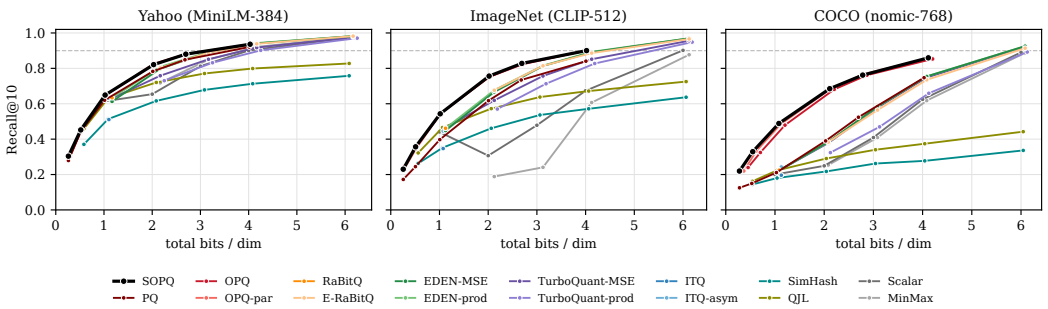

Figure 1: Recall@10 versus total bits per dimension (including model overhead) on Yahoo, ImageNet, and COCO for all VQ-bench quantizers. Llama and CCNews are in Appendix E, Figure 2.

Table 2: Recall@10 near 2 and 4 bits/dim across five VQ-bench datasets. Each method uses its configuration closest to the target rate; best per column in bold.

	Method	Yahoo	ImageNet	COCO	CCNews	Llama
≈ 2 b/dim	PQ	2.01: 0.783	2.01: 0.618	2.03: 0.390	2.02: 0.791	2.03: 0.225
	OPQ	2.05: 0.824	2.03: 0.753	2.20: 0.681	2.12: 0.885	2.06: **0.369**
	OPQ-P	2.03: **0.826**	2.02: 0.752	2.12: 0.682	2.07: 0.887	2.05: 0.365
	E-RaBitQ	2.17: 0.803	2.13: 0.676	2.08: 0.383	2.08: 0.799	2.50: 0.222
	EDEN-prod	2.17: 0.808	2.13: 0.677	2.08: 0.382	2.08: 0.801	1.50: 0.161
	TurboQuant-prod	2.25: 0.730	2.19: 0.570	2.13: 0.324	2.13: 0.747	2.75: 0.176
	SOPQ	2.03: 0.821	2.02: **0.756**	2.11: **0.685**	2.06: **0.887**	2.05: 0.359
≈ 4 b/dim	PQ	4.01: 0.921	4.01: 0.841	4.03: 0.747	4.02: 0.924	4.03: 0.528
	OPQ	4.05: 0.937	4.03: **0.902**	4.20: 0.853	4.12: 0.950	4.06: **0.658**
	OPQ-P	4.03: 0.936	4.02: 0.900	4.12: 0.858	4.07: **0.951**	4.05: 0.645
	E-RaBitQ	4.17: 0.939	4.13: 0.886	4.08: 0.735	4.08: 0.933	3.50: 0.320
	EDEN-prod	4.17: **0.943**	4.13: 0.893	4.08: 0.755	4.08: 0.935	3.50: 0.359
	TurboQuant-prod	4.25: 0.902	4.19: 0.827	4.13: 0.659	4.13: 0.903	3.75: 0.242
	SOPQ	4.03: 0.936	4.02: 0.900	4.12: **0.860**	4.07: 0.949	4.05: 0.645

each segment, and for given segment widths, total MSE is minimized exactly when all segments share the same error θ^2 (Theorems 1 and 2); (2) under CRWF, any segment whose allocations sum to $C = \log_2 k$ has error $\theta^2 \exp(J_g)$, where the Jensen gap J_g of the flat function $\psi = \ln c_{\text{eff}}$ keeps the error within about 2% of θ^2 even in extreme cases, reducing the combinatorial grouping problem to bin packing (Theorem 3); and (3) the SVD rotation decorrelates the coordinates and maximizes their variance skew, globally minimizing the CRWF distortion, whereas random rotations maximize it (Theorem 4).

Empirically, SOPQ consistently matches the recall of OPQ on all five VQ-bench benchmarks and often exceeds it; it is also faster to train and query than OPQ. SOPQ is thus a fast, predictable, and theoretically grounded alternative to OPQ for modern AI workloads.

7.1 AI USE STATEMENT

The guiding theoretical questions of this paper originated from the authors, but AI tools were used to refine the theoretical model and analysis. AI tools were also used to assist the authors for formulating claims and drafting proof sketches. AI tools were used to write the experimental code for this work. AI was not used to generate synthetic datasets, propose or refine hypotheses, provide feedback on methodology, assist with translation, clean or reformat datasets, support qualitative data analysis, or interpret results.

Table 3: System cost on ImageNet ($n = 1{,}281{,}167$, $d = 512$), the largest of the five datasets, near 2 bits/dim. Fit uses $25{,}000$ vectors; encode covers the full database; score latency is the median per-query time to score its 100 candidates.

Method	Bits/dim	Model b/dim	Fit	Encode	Score (p50)	Recall@10
PQ	2.01	0.006	9.1 s	30.2 s	523 μs	0.618
OPQ	2.03	0.032	115 s	31.9 s	962 μs	0.753
OPQ-P	2.02	0.019	9.6 s	31.3 s	708 μs	0.752
E-RaBitQ	2.13	0.000	0.1 s	2.5 s	233 μs	0.676
EDEN-prod	2.13	0.000	0.1 s	1.9 s	237 μs	0.677
TurboQuant-prod	2.19	0.000	0.4 s	2.6 s	446 μs	0.570
SOPQ	2.02	0.019	10.0 s	31.2 s	700 μs	**0.756**

REFERENCES

Nir Ailon and Bernard Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. Quip: 2-bit quantization of large language models with guarantees. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 4396–4429. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0df38cd13520747e1e64e5b123a78ef8-Paper-Conference.pdf.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. Rate–distortion theory, Ch. 10.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. Introduces the NF4 data type.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Jianyang Gao and Cheng Long. Rabitq: Quantizing high-dimensional vectors with a theoretical error bound for approximate nearest neighbor search. *Proc. ACM Manag. Data*, 2(3), May 2024.

Jianyang Gao, Yutong Gou, Yuexuan Xu, Yongyi Yang, Cheng Long, and Raymond Chi-Wing Wong. Practical and asymptotically optimal quantization of high-dimensional vectors in euclidean space for approximate nearest neighbor search. *Proc. ACM Manag. Data*, 2025. doi: 10.1145/3725413. URL https://doi.org/10.1145/3725413. Extended RaBitQ (SIGMOD 2025); B bits/dimension.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2946–2953, 2013.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):744–755, 2014. doi: 10.1109/TPAMI.2013.240.

Allen Gersho. Asymptotically optimal block quantization. *IEEE Transactions on Information Theory*, 25(4):373–380, 1979. doi: 10.1109/TIT.1979.1056067.

Allen Gersho and Robert M. Gray. Vector quantization and signal compression. In *The Kluwer International Series in Engineering and Computer Science*, 1991.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998. doi: 10.1109/18.720541.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Alfred Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76(3):620–630, 1954.

Elias Jääsaari, Ville Hyvönen, Matteo Ceccarello, Teemu Roos, and Martin Aumüller. Vibe: Vector index benchmark for embeddings, 2025.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. arXiv:1702.08734, 2017.

William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into hilbert space. *Contemporary mathematics*, 26:189–206, 1984.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. GEAR: An efficient KV cache compression recipe for near-lossless generative inference of LLM. *arXiv preprint arXiv:2403.05527*, 2024.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2-bit quantization for KV cache. In *International Conference on Machine Learning (ICML)*, 2024.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2 edition, 2011.

Ashwin Padaki, Amir Ingber, and Edo Liberty. Vq-bench: A composable vector quantization framework, 2026. URL https://arxiv.org/abs/2608.11240.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

Harsha Vardhan Simhadri, George Williams, Martin Aumüller, Matthijs Douze, Artem Babenko, Dmitry Baranchuk, Qi Chen, Lucas Hosseini, Ravishankar Krishnaswamy, Gopal Srinivasa, Suhas Jayaram Subramanya, and Jingdong Wang. Results of the NeurIPS'21 challenge on billion-scale approximate nearest neighbor search. In *Proceedings of the NeurIPS Competition Track*, 2022.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning (ICML)*, 2024.

Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben Itzhak, and Michael Mitzenmacher. EDEN: Communication-efficient and robust distributed mean estimation for federated learning. In *International Conference on Machine Learning (ICML)*, PMLR 162, pp. 21984–22014, 2022. arXiv:2108.08842.

Paul L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982. doi: 10.1109/TIT.1982.1056490.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. TurboQuant: Online vector quantization with near-optimal distortion rate. *arXiv preprint arXiv:2504.19874*, 2025.

A Calculating c_d and $c_{\text{eff}}(b)$ in Practice

Modeling each d-dimensional Voronoi cell by a Euclidean ball of equal volume yields Zador's closed-form sphere formula for the Gaussian distortion factor c_d (Zador, 1982; Gersho, 1979):

$$c_d \;=\; \frac{2}{d+2}\,\Gamma\!\left(1+\frac{d}{2}\right)^{2/d}\left(1+\frac{2}{d}\right)^{\frac{d+2}{2}} \qquad \left(\text{with } c_1=\frac{\sqrt{3}\pi}{2}\approx 2.721 \text{ and } \lim_{d\to\infty} c_d = 1\right). \tag{10}$$

In practice, since k-means codebook size is fixed at $k = 256$ ($C = 8$ bits), the simplest and most accurate engineering implementation is a direct lookup/interpolation table of $c_{\text{eff}}(b) = c_{8/b}$ (Table 4).

Table 4: Lookup table of c_d and $c_{\text{eff}}(b) = c_{8/b}$ ($k = 256$) from (10) vs. measured k-means on $\mathcal{N}(0, I_d)$, and $\psi(b) = \ln c_{\text{eff}}(b)$ used in Theorem 3.

Segment Dim d	1	2	3	4	6	8	12	16	24	32	64
Bit Rate $b = 8/d$	8.00	4.00	2.67	2.00	1.33	1.00	0.67	0.50	0.33	0.25	0.125
Sphere formula c_d (10)	2.721	2.000	1.734	1.591	1.436	1.351	1.258	1.207	1.152	1.121	1.070
Measured k-means (k=256)	2.868	2.058	1.781	1.591	1.405	1.315	1.216	1.169	1.118	1.089	1.048
$\psi(8/d) = \ln c_d$ (measured)	1.054	0.722	0.577	0.464	0.340	0.274	0.196	0.156	0.112	0.086	0.047

B Complete Mathematical Proofs and Query-Aware Extension

B.1 Proof of Theorem 4 (SVD Global Minimum and Random-Rotation Global Maximum)

Schur–Horn Majorization (Horn, 1954). Write $\Sigma = U\Lambda U^\top$ and $Q = RU \in O(d)$. Then $s_j^2(R) = (Q\Lambda Q^\top)_{jj} = \sum_{i=1}^d Q_{ji}^2\lambda_i = (P\lambda)_j$, where $P_{ji} = Q_{ji}^2 \geq 0$ is doubly stochastic ($\sum_i P_{ji} = \sum_j P_{ji} = 1$). By the Hardy–Littlewood–Pólya and Horn theorems (Marshall et al., 2011; Horn, 1954), $\{s^2(R) : R \in O(d)\} = \{s^2 \in \mathbb{R}^d_{\geq 0} : (\bar\lambda, \ldots, \bar\lambda) \prec s^2 \prec (\lambda_1, \ldots, \lambda_d)\}$.

Strict Schur-Concavity of $D_{\text{crwf}}(s^2; B)$. Writing $\phi(b) = \exp\big(\psi(b) - 2b\ln 2\big)$ with $\psi = \ln c_{\text{eff}}$ gives $\phi' = -(2\ln 2 - \psi')\,\phi$ and $\phi'' = \big((2\ln 2 - \psi')^2 + \psi''\big)\,\phi$, so ϕ is strictly decreasing and convex whenever $\psi' < 2\ln 2$ and $\psi'' > -(2\ln 2 - \psi')^2$. The measured values in Table 4 (with $\psi(0) = 0$) satisfy both with a wide margin: finite differences give $0 < \psi' \leq 0.4$ and $\psi'' \geq -0.6$, whereas $(2\ln 2 - 0.4)^2 > 0.97$. For each fixed bit vector $b \in \mathbb{R}^d_{\geq 0}$, $\Phi(s^2; b) = \sum_{j=1}^d s_j^2\phi(b_j)$ is affine in s^2, so its pointwise minimum $D_{\text{crwf}}(s^2; B) = \min_{\sum b_j \leq B, b_j \geq 0} \Phi(s^2; b)$ is globally concave and symmetric on $\mathbb{R}^d_{>0}$. Let $\tilde b(s^2)$ denote the minimizer, which is unique because ϕ is strictly convex. By the Envelope Theorem,

$$\frac{\partial D_{\text{crwf}}(s^2; B)}{\partial (s_j^2)} \;=\; \phi\big(\tilde b_j(s^2)\big) \;=\; c_{\text{eff}}\big(\tilde b_j(s^2)\big)\, 2^{-2\tilde b_j(s^2)}. \tag{11}$$

Stationarity gives $s_j^2\phi'(\tilde b_j) = -\lambda$ on the active set, and ϕ' is increasing, so $\tilde b_j(s^2)$ is strictly increasing in s_j^2. Since ϕ is strictly decreasing, any $s_i^2 > s_j^2$ with $\tilde b_i > 0$ satisfies $(s_i^2 - s_j^2)\big(\frac{\partial D_{\text{crwf}}}{\partial s_i^2} -$

$\frac{\partial D_{\mathrm{crwf}}}{\partial s_j^2}\big) < 0$. By the Schur–Ostrowski criterion, $D_{\mathrm{crwf}}(\cdot; B)$ is strictly Schur-concave, placing $R = U^\top$ ($s^2 = \lambda$) at the global minimum and equal-variance rotations ($s^2 = \bar{\lambda}\mathbf{1}$) at the global maximum.

Random Rotation Concentration. For Haar-distributed $R \in O(d)$, each row $r_j \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ satisfies $\mathbb{E}[s_j^2(R)] = \bar{\lambda}$ and $\mathrm{Var}[s_j^2(R)] = \frac{2(d\,\mathrm{tr}\Sigma^2 - (\mathrm{tr}\Sigma)^2)}{d^2(d+2)} = \frac{2(\rho-1)}{d+2}\bar{\lambda}^2$ with $\rho = d\,\mathrm{tr}\Sigma^2/(\mathrm{tr}\Sigma)^2 \in [1, d]$. By Hanson–Wright concentration on $\mathbb{S}^{d-1}$ and a union bound over $j \in [d]$, $\max_{1\le j\le d} |s_j^2(R)/\bar{\lambda} - 1| = O(\sqrt{\rho \log d/d})$ w.h.p. □

B.2 Proof of Proposition 5 (LPT Spillage and Complexity Bounds)

Lossless $d \to d'$ Truncation and Memory. Every coordinate $j > d'$ receives $b_j^\star = 0$ bits and is reconstructed by 0, so $\hat{x} - \mu = \sum_{j=1}^{d'} u_j \hat{y}_j + \sum_{j=d'+1}^{d} u_j \cdot 0 = U_{d'}\hat{y}$. Segment r stores k centroids in $\mathbb{R}^{d_{g_r}}$ (kd_{g_r} floats), summing across $r = 1, \ldots, m$ to $k\sum_{r=1}^m d_{g_r} = d'k$ floats.

LPT Spillage Bound. Let g_{r^+} be the segment with the maximum accumulated budget $B_{\max} = \sum_{j\in g_{r^+}} b_j^\star$. If $|g_{r^+}| = 1$, then $B_{\max} \le b_1^\star \le C$. If $|g_{r^+}| \ge 2$, let $j^\star \ge m+1$ be the last coordinate placed into g_{r^+}. Just before $j^\star$ was added, g_{r^+} had the smallest load among all m segments, so $B_{\max} - b_{j^\star}^\star \le \frac{1}{m}(B - b_{j^\star}^\star)$, giving $B_{\max} - C \le \frac{m-1}{m} b_{j^\star}^\star$. Since $b_1^\star \ge \cdots \ge b_{j^\star}^\star > 0$ sum to B, Markov's inequality gives $b_{j^\star}^\star \le B/j^\star \le \frac{B}{m+1}$ unconditionally (and $b_{j^\star}^\star \le 1 \implies B_{\max} < C + 1$ whenever $j^\star \ge B$). □

B.3 Query-Aware Inner-Product Score Allocation (SOPQ + WHITEN)

For ANN inner-product search where queries q have second-moment matrix $M = \mathbb{E}[qq^\top]$, expected score error is $\mathbb{E}_q[\langle q, x - \hat{x}\rangle^2] = \mathrm{tr}((RMR^\top)\mathbb{E}[ee^\top]) = \sum_{j=1}^d m_j e_j^2$, where $m_j = (RMR^\top)_{jj}$. Replacing s_j^2 in (7) with the query-weighted variance $w_j = m_j s_j^2$ yields score-optimal CRWF allocation ($w_j = \lambda_j^2$ when $M \propto \Sigma$; when calibration queries with $M \ne \Sigma$ are available, whitening by $M^{1/2}$ and taking the SVD of $M^{1/2}\Sigma M^{1/2}$ jointly diagonalizes both matrices, labeled `SOPQ + whiten` in Table 2).

C Per-Coordinate k-Means Error Equalization Experiment

To verify Theorem 1—that d_g-dimensional k-means equalizes per-coordinate MSE across all actively quantized dimensions ($e_1^2 \approx \cdots \approx e_{d_g}^2 \approx e^2(g)$) and matches $c_{d_g}(\prod_{j\in g} s_j^2)^{1/d_g} k^{-2/d_g}$—we evaluate $d_g \in \{1, 2, 4, 8, 16, 32, 64\}$ at fixed codebook size $k = 256$ ($C = 8.0$ bits). For each $d_g \ge 2$, we assign skewed target rates $b_1 > b_2 > \cdots > b_{d_g} > 0$ spaced linearly from $1.5(8/d_g)$ down to $0.5(8/d_g)$ ($\sum_{j=1}^{d_g} b_j = 8.0$ bits, all strictly active) and set $s_j^2 = 2^{2b_j}$, so that $(\prod_{j=1}^{d_g} s_j^2)^{1/d_g} k^{-2/d_g} \equiv 1.0000$ and the theoretical per-coordinate MSE is c_{d_g} (or $c_{\mathrm{eff}}(8/d_g)$). Table 5 confirms that even at $d_g = 2$ where $s_1^2/s_2^2 = 256\times$, k-means equalizes $e_1^2 = 1.878$ and $e_2^2 = 1.997$ (CV = 3.1%), and across $d_g \in [2, 64]$ the across-coordinate CV is 0.8%–3.1% while tracking c_{d_g} and $c_{\mathrm{eff}}(8/d_g)$.

D Controlled 6-Tier Verification: CRWF vs. Standard RWF Across Segment Bucketings

Motivation and Setup. To verify Theorem 3—that under CRWF, segments with equal total bit allocation $\sum_{j\in g} b_j^\star = C = 8.0$ bits achieve the same k-means per-coordinate error $\theta^2 = 1.00$ independent of segment dimension d_g and composition—we construct a 126-dimensional Gaussian source partitioned into 6 geometric bit-budget tiers (Table 6), where each tier sums to 8.0 bits ($B = 48.0$ bits across $m = 6$ segments, $k = 256$). We compare setting s_j^2 at target waterline $\theta^2 = 1.0$

Table 5: Experimental verification of per-coordinate k-means error equalization ($k = 256$, $n_{\text{train}} = 45{,}000$, $n_{\text{test}} = 15{,}000$, 3 seeds) across segment dimensions $d_g \in \{1, 2, 4, 8, 16, 32, 64\}$ with skewed variances $s_1^2 > \cdots > s_{d_g}^2$ satisfying $(\prod_{j=1}^{d_g} s_j^2)^{1/d_g} k^{-2/d_g} = 1.0000$.

Seg. Dim. d_g	**Mean Rate** $\bar{b}_g = 8/d_g$	**Variance Range** $[s_{d_g}^2,\ s_1^2]$	**Skew** $s_1^2/s_{d_g}^2$	**Zador Pred.** c_{d_g}	**CRWF Pred.** $c_{\text{eff}}(8/d_g)$	**Meas. Mean** $\frac{1}{d_g}\sum_j e_j^2$	**Max / Min Var** $(e_1^2,\ e_{d_g}^2)$	**Across-Coord CV** (sd/mean)
1	8.000b	[65536, 65536]	1.0×	2.7207	2.3947	**2.8678**	(2.868, 2.868)	**0.0**%
2	4.000b	[16.00, 4096.0]	256.0×	2.0000	1.9717	**1.9373**	(1.878, 1.997)	**3.1**%
4	2.000b	[4.00, 64.00]	16.0×	1.5910	1.5855	**1.5250**	(1.470, 1.578)	**2.7**%
8	1.000b	[2.00, 8.00]	4.0×	1.3509	1.3231	**1.3041**	(1.293, 1.328)	**1.3**%
16	0.500b	[1.41, 2.83]	2.0×	1.2073	1.1699	**1.1630**	(1.158, 1.170)	**0.8**%
32	0.250b	[1.19, 1.68]	1.41×	1.1212	1.0872	**1.0882**	(1.096, 1.065)	**1.3**%
64	0.125b	[1.09, 1.30]	1.19×	1.0700	1.0441	**1.0480**	(1.061, 1.001)	**2.1**%

(theoretical Total MSE = 126.00) via **CRWF** ($s_j^2 = 2^{2b_j}/c_{\text{eff}}(b_j)$) vs. **Standard RWF** ($s_j^2 = 2^{2b_j}$) across the five zero-spillage bucketing configurations detailed in Table 7.

Table 6: Setup of the 126-variable, 6-tier geometric source ($\theta^2 = 1.0$, total budget $B = 48.0$ bits across $m = 6$ segments of 8.0 bits each).

Parameter	**Tier 1**	**Tier 2**	**Tier 3**	**Tier 4**	**Tier 5**	**Tier 6**	**Total**
Number of coordinates	2	4	8	16	32	64	$d = 126$
Bit budget b_j (bits/coord)	4.0	2.0	1.0	0.5	0.25	0.125	$B = 48.0$
Tier bit sum (bits)	8.0	8.0	8.0	8.0	8.0	8.0	$m = 6$ bins
s_j^2(RWF) $= 2^{2b_j}$	256.00	16.00	4.00	2.00	1.41	1.19	761.36
s_j^2(CRWF) $= 2^{2b_j}/c_{\text{eff}}(b_j)$	129.84	10.09	3.02	1.71	1.30	1.14	466.09

E Additional Experimental Plots and Ablations

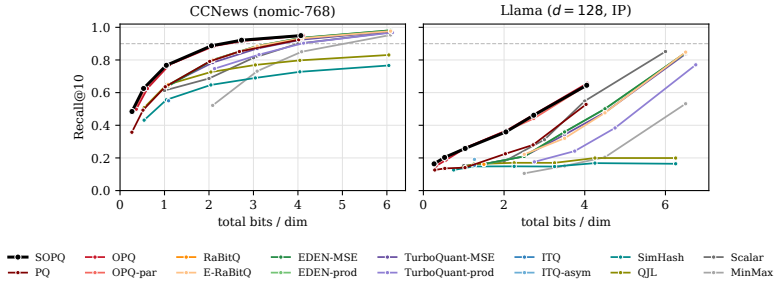

Figure 2: Recall@10 versus total bits per dimension (including model overhead) on CCNews and Llama for all VQ-bench quantizers, complementing Figure 1.

Table 7: Segment-by-segment composition of the five bucketing configurations ($m = 6$ segments per configuration, $B_s = 8.0$ bits per segment). Each row reports the number of coordinates drawn from each tier in Table 6 and the resulting segment dimension d_g.

Configuration	s	Coordinate Count from Each Tier in Segment s						Dim. d_g	Bits B_s
		Tier 1 (4.0b)	Tier 2 (2.0b)	Tier 3 (1.0b)	Tier 4 (0.5b)	Tier 5 (0.25b)	Tier 6 (0.125b)		
1. Homogeneous FFD	1	2	0	0	0	0	0	2	8.0
	2	0	4	0	0	0	0	4	8.0
	3	0	0	8	0	0	0	8	8.0
	4	0	0	0	16	0	0	16	8.0
	5	0	0	0	0	32	0	32	8.0
	6	0	0	0	0	0	64	64	8.0
2. Canonical LPT	1	1	0	0	3	5	10	19	8.0
	2	1	0	0	3	5	10	19	8.0
	3	0	1	2	3	5	10	21	8.0
	4	0	1	2	3	5	10	21	8.0
	5	0	1	2	2	6	12	23	8.0
	6	0	1	2	2	6	12	23	8.0
3. Complementary Fold	1	1	0	0	0	0	32	33	8.0
	2	1	0	0	0	0	32	33	8.0
	3	0	2	0	0	16	0	18	8.0
	4	0	2	0	0	16	0	18	8.0
	5	0	0	4	8	0	0	12	8.0
	6	0	0	4	8	0	0	12	8.0
4. Adjacent Pairs	1	1	2	0	0	0	0	3	8.0
	2	1	2	0	0	0	0	3	8.0
	3	0	0	4	8	0	0	12	8.0
	4	0	0	4	8	0	0	12	8.0
	5	0	0	0	0	16	32	48	8.0
	6	0	0	0	0	16	32	48	8.0
5. Equal-Width	1	1	0	0	2	6	12	21	8.0
	2	1	0	0	2	6	12	21	8.0
	3	0	1	2	3	5	10	21	8.0
	4	0	1	2	3	5	10	21	8.0
	5	0	1	2	3	5	10	21	8.0
	6	0	1	2	3	5	10	21	8.0

Table 8: **126-Variable, 6-Tier Geometric Benchmark ($m = 6$ segments of 8.0 bits, $k = 256$): CRWF vs. Standard RWF.** Under **CRWF**, per-coordinate MSE is equalized to 1.00 ± 0.04 across all 6 tiers (max/min spread $1.04\times$–$1.07\times$), and Total MSE is invariant ($\mathbf{126.29}$–$\mathbf{126.58}$ vs. theoretical 126.00) across all five bucketings. Under **Standard RWF**, homogeneous FFD suffers a $\mathbf{1.96}\times$ water-level staircase ($1.048 \to 2.052$), which LPT collapses back to $\mathbf{1.06}\times$.

Bucketing Strategy (m=6, B_s=8.0b)	Total MSE (mean ± std)	Measured Per-Coordinate MSE by Tier						Max / Min
		4b	2b	1b	0.5b	0.25b	0.125b	
Corrected CRWF Variances ($s_j^2 = 2^{2b}/c_{\text{eff}}(b)$; theoretical Target MSE = 1.000/coord, Total = $\mathbf{126.00}$)								
1. Homogeneous FFD	$\mathbf{126.29 \pm 0.14}$	1.040	0.997	0.989	0.996	1.003	1.004	$\mathbf{1.05}\times$
2. Canonical LPT	$\mathbf{126.54 \pm 0.07}$	0.964	0.976	0.996	1.012	1.019	0.999	$\mathbf{1.06}\times$
3. Complementary Fold	$\mathbf{126.51 \pm 0.07}$	0.971	0.970	0.992	0.994	1.005	1.011	$\mathbf{1.04}\times$
4. Adjacent Pairs	$\mathbf{126.32 \pm 0.07}$	0.965	1.029	0.989	0.996	1.023	0.995	$\mathbf{1.07}\times$
5. Equal-Width	$\mathbf{126.58 \pm 0.09}$	0.958	0.985	0.998	1.015	1.019	0.999	$\mathbf{1.06}\times$
Standard RWF Variances ($s_j^2 = 2^{2b}$, constant c_0)								
1. Homogeneous FFD	141.55 ± 0.17	$\mathbf{2.052}$	$\mathbf{1.581}$	$\mathbf{1.309}$	1.165	1.091	1.048	$\mathbf{1.96}\times$
2. Canonical LPT	$\mathbf{140.06 \pm 0.10}$	$\mathbf{1.077}$	$\mathbf{1.097}$	$\mathbf{1.111}$	1.133	1.142	1.093	$\mathbf{1.06}\times$
3. Complementary Fold	140.49 ± 0.10	1.030	1.092	1.204	1.213	1.135	1.073	$1.18\times$
4. Adjacent Pairs	141.31 ± 0.13	1.657	1.741	1.201	1.212	1.086	1.051	$1.66\times$
5. Equal-Width	140.12 ± 0.15	1.067	1.097	1.122	1.135	1.140	1.094	$\mathbf{1.07}\times$