VENICE: UNIVERSAL SPECTRAL RATE-DISTORTION QUANTIZATION FRAMEWORK FOR VECTOR SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional vector embeddings impose severe memory and latency constraints in large-scale maximum inner product search. Vector quantization has emerged as the standard compression paradigm, where existing state-of-the-art methods perform effectively under classical assumptions of uncorrelated dimensions and isotropic energy distributions. Most real-world embeddings, however, exhibit a different geometry characterized by coordinate cross-correlations and anisotropic spectral decay, which we prove randomized Hadamard transforms redistribute rather than remove. To harness the benefits of non-uniform bit budgets, remove these cross-correlations, and correct inner product distortion, we present VENICE, a universal spectral framework that wraps arbitrary base quantizers as plug-and-play black boxes. Grounded in Shannon rate-distortion theory, VENICE transforms anisotropic embeddings into decoupled sub-vectors of uniform sensitivity through closed-form spectral decorrelation, water-filling rate allocation, and sensitivity equalization. This establishes the geometry downstream codecs were designed for, while norm restoration corrects the shrinkage of their output. We not only mathematically prove the foundational properties of VENICE, but also empirically validate its benefits across five benchmarks, wrapping six codecs of diverse nature. Across 4-, 2-, and 1-bit regimes, VENICE consistently improves both training-free and learned codecs on anisotropic spectra, cutting recall loss by up to 93% and rescuing ultra-low-bit quantizers from performance collapse.

1 INTRODUCTION

Vector search—returning the Top-k items of a collection most similar to a query under inner product or Euclidean distance—is the retrieval primitive behind semantic search, recommendation and retrieval-augmented generation (Johnson et al., 2019; Guo et al., 2020; Malkov & Yashunin, 2018; Wang et al., 2021). Such collections routinely run to billions of high-dimensional vectors and cannot be held in full precision, so they are quantized and the scores are estimated from the codes. We focus on maximum inner product search (MIPS); the same geometry arises in the quantization of LLM weights and KV caches (Chee et al., 2023; Ashkboos et al., 2024), which we do not address.

Quantizers are designed to minimize reconstruction error, yet what a retrieval system loses is the distortion of the scores it ranks by: for an encoding error $\boldsymbol{e}$, that loss is $\mathcal{L}(\boldsymbol{e}) = \boldsymbol{e}^\top \boldsymbol{H} \boldsymbol{e}$, with $\boldsymbol{H}$ the second moment of the query distribution (Section 3.1). This loss is separable only in the eigenbasis of $\boldsymbol{H}$, where each coordinate enters with an effective weight $\lambda_k \sigma_k^2$: its variance, scaled by its sensitivity λ_k, the cost the loss charges for an error along it. We show that a uniform budget then inflates the distortion by the ratio of the arithmetic to the geometric mean of these weights, which equals one only if the weights are equal (Section 3.2). The two standard remedies prepare the data before coding, in opposite directions. *Dimension balancing* applies a random orthogonal rotation so all coordinates carry comparable variance, which is what justifies giving each the same number of bits (Gao et al., 2025; Zandieh et al., 2026); *adaptive allocation* keeps them unequal and varies the budget accordingly, granting more bits to the coordinates that carry more energy and few or none to the rest (Guo et al., 2016a; Paparrizos et al., 2022). We prove that a random rotation redistributes the energy of $\boldsymbol{H}$ rather than removing it from the off-diagonal, and provably equalizes the diagonal only at high effective rank, which learned embeddings lack (Mu et al., 2017; Gao et al., 2019) (Section 3.3). Recent work combines the two, splitting PCA-ordered coordinates into

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.

consecutive groups, each rotated and given its own budget (Li et al., 2025), but sets that budget per segment rather than per coordinate, and by variance alone.

VENICE does not choose between the two. It makes coordinates equal in *sensitivity* and applies the randomized rotation only inside blocks it has first made isotropic. It rests on four pillars, each with a formal guarantee (Section 4). *Spectral decorrelation* rotates into the eigenbasis of $\boldsymbol{H}$ and subtracts the rotated mean, which cancels the cross-talk exactly and leaves every coordinate with its own weight $\lambda_k \sigma_k^2$. *Reverse water-filling* distributes the rate over those weights in closed form, down to a level below which a coordinate earns no bits at all: pruning is part of the optimum. *Sensitivity equalization* multiplies each surviving coordinate by $\lambda_k^{1/2}$, which turns the Hessian block of every tier—the coordinates that share a bit-width—into the identity, fulfilling expectations of base codecs; only then is a randomized Hadamard transform applied inside the tier, spreading its energy at no cost. Both orderings matter: the budget has to be decided while the weights are still unequal, since equalizing first would leave nothing to allocate on, and within a tier the rescaling has to precede the rotation, which would otherwise make the block non-diagonal again and restore the cross-talk. *Norm restoration* finally overwrites the decoded length with a stored f-bit norm, undoing the shrinkage that decoding introduces, which differs from one vector to the next and reorders candidates whose scores are close. What each base quantizer receives is a zero-mean sub-vector whose coordinates are equally weighted, of comparable variance, and with energy spread across them: the regime it was built for. No stage is learned—the four reduce to an eigendecomposition, a bisection on a scalar, a diagonal rescaling followed by a random rotation, and a per-vector norm ratio.

VENICE is thus designed as a framework rather than a standalone codec: because its four pillars act on the geometry and never on the code, it leaves the base quantizer's encoder and decoder untouched and wraps it as it is, adding one eigenbasis shared by the collection and one f-bit scalar per vector. We wrap six codecs, training-free and learned alike, on five collections of 100 to 3072 dimensions, at 4, 2 and 1 bits per dimension. Recall loss drops by up to 93% across every codec, collection, and rate but one, with VENICE reducing recall loss by up to 50% over SAQ (Li et al., 2025). The sole exception is the null case the theory predicts: on GLOVE, whose spectrum is nearly flat, a water-filling returns a uniform allocation, and VENICE reduces to a change of basis, yielding no statistically significant difference. A leave-one-out ablation finds no redundant pillar: removing spectral decorrelation or reverse water-filling costs up to 18.0 recall points at 1 bit.

Our contributions are the following.

- **A retrieval-aware analysis of vector quantization (Section 3).** From the $\boldsymbol{H}$-weighted retrieval loss we derive three design requirements—removing coordinate cross-talk, allocating rate to the weights $\lambda_k \sigma_k^2$, and correcting heterogeneous norm contraction—and prove that randomized Hadamard transforms retain most of the off-diagonal energy of $\boldsymbol{H}$ and guarantee a uniform diagonal only at high effective rank, which most embeddings lack.
- **VENICE, a closed-form, codec-agnostic framework (Section 4).** VENICE separates data geometry from the code to wrap any base quantizer untouched via four pillars with formal guarantees: spectral decorrelation in the eigenbasis of $\boldsymbol{H}$, reverse water-filling with built-in pruning, sensitivity equalization by $\lambda_k^{1/2}$ into tiers with identity Hessian blocks followed by a localized randomized Hadamard transform—an exact isometry of the retrieval loss—and per-vector norm restoration.
- **Empirical validation (Section 5.2).** Across six training-free and learned codecs, five collections and three bit budgets, VENICE delivers consistent recall gains outside the predicted null case, surpasses SAQ overall when wrapping SAQ's own base codec (CAQ), and has no redundant pillar under a leave-one-out ablation.

2 RELATED WORK

Learned codebooks. PQ (Jegou et al., 2011) splits a vector into subvectors and learns one codebook per subspace by k-means; additive schemes (Babenko & Lempitsky, 2014) drop the split and represent a vector as a sum of codewords drawn from multiple full-dimensional codebooks. OPQ (Ge et al., 2013) learns the rotation applied before the split, in its parametric form choosing it so that variance is balanced across subspaces. A parallel line replaces Euclidean reconstruction with an objective aligned to retrieval: Guo et al. (2016b) learns codebooks that minimize inner-product error, and AVQ (Guo et al., 2020) weights residual components parallel to each datapoint more

heavily than orthogonal ones. Across all of them the bit budget is fixed in advance and identical across codebooks.

Training-free codecs. A second family (Vargaftik et al., 2022; Gao & Long, 2024; Gao et al., 2025; Ann et al., 2026; Zandieh et al., 2026; Lenhardt et al., 2026) involves no learning phase: the quantizer is built from a randomized orthogonal transform followed by a fixed rule applied at the same rate throughout, so its construction uses no training data and carries guarantees stated independently of the collection. Several of these codecs store the norm of a vector separately from its code.

Adaptive codes. A third family varies the budget: across PQ subspaces (Guo et al., 2016a), across subspace codebooks sized by explained variance (Paparrizos et al., 2022), and across segments of PCA-projected coordinates by dynamic programming on quantization error (Li et al., 2025); LVQ (Aguerrebere et al., 2024) instead adapts the scale of each vector at a fixed bit-width. In the three schemes that allocate budget, rate follows the variance of the data and the unit of allocation is a group of coordinates. VENICE departs on both counts, allocating one budget per coordinate and weighting that coordinate's variance by how much the retrieval objective charges an error along it. SAQ (Li et al., 2025) is closest to VENICE, making it a natural comparison on retrieval accuracy (Section 5): it alone allocates bit-widths rather than dictionary sizes, and also rotates each group before quantizing it. The groups differ, though—in SAQ contiguous segments found by dynamic programming, in VENICE the coordinates that received the same budget—and VENICE rescales them before rotating, so the rotation acts on coordinates of equal retrieval weight (Section 4.3).

Each of the above is a codec: a geometry and a code designed together. VENICE separates them, determining only the geometry and leaving the code to any wrapped quantizer. This separation lets one construction be applied, unchanged, to learned and training-free codecs alike (Section 5).

3 Problem Formulation and Design Requirements

3.1 Retrieval Distortion under Vector Quantization

Given a query $\boldsymbol{q} \in \mathbb{R}^d$ and a finite collection of vectors $\boldsymbol{X} \subset \mathbb{R}^d$, MIPS retrieves the vector $\boldsymbol{x}^* \in \arg\max_{\boldsymbol{x}\in\boldsymbol{X}} \langle \boldsymbol{q}, \boldsymbol{x}\rangle$. When collection cardinality $|\boldsymbol{X}|$ and dimensionality d are large, storing full-precision vectors creates severe memory and bandwidth bottlenecks. This motivates lossy vector quantization $\widehat{\boldsymbol{x}} = \mathcal{Q}(\boldsymbol{x})$ at low bit-widths $b \ll 32$ bits per dimension (typically $b \leq 8$) to compress storage footprint and accelerate query throughput (Jegou et al., 2011; Johnson et al., 2019).

Replacing $\boldsymbol{x}$ by $\widehat{\boldsymbol{x}}$ induces an inner product error $\Delta_{\mathrm{IP}} = \langle \boldsymbol{q}, \boldsymbol{e}\rangle$, where $\boldsymbol{e} = \boldsymbol{x} - \widehat{\boldsymbol{x}}$ is the quantization residual. Its squared error expands as $\Delta_{\mathrm{IP}}^2 = \boldsymbol{e}^\top(\boldsymbol{q}\boldsymbol{q}^\top)\boldsymbol{e}$. Since $\boldsymbol{e}$ does not depend on $\boldsymbol{q}$, taking the expectation over the query distribution $\boldsymbol{q} \sim \mathcal{D}_{\boldsymbol{q}}$ gives

$$\mathcal{L}(\boldsymbol{e}) \triangleq \mathbb{E}_{\boldsymbol{q}}\left[\Delta_{\mathrm{IP}}^2\right] = \boldsymbol{e}^\top \boldsymbol{H} \boldsymbol{e}, \qquad \boldsymbol{H} \triangleq \mathbb{E}_{\boldsymbol{q}}\left[\boldsymbol{q}\boldsymbol{q}^\top\right], \tag{1}$$

so retrieval distortion is the $\boldsymbol{H}$-weighted quadratic form of the residual, with $\boldsymbol{H}$ the query second-moment matrix (referred to as the Hessian in the quantization literature). Being symmetric positive semi-definite, $\boldsymbol{H}$ admits the eigendecomposition

$$\boldsymbol{H} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top = \sum_{k=1}^{d} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k^\top, \;\; \boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d], \;\; \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \;\; \lambda_1 \geq \cdots \geq \lambda_d \geq 0. \tag{2}$$

Following standard practice we assume that queries and datapoints share the same second-moment structure (Guo et al., 2016b) and estimate $\boldsymbol{H}$ by the empirical second moment $\frac{1}{|\boldsymbol{X}|}\sum_{\boldsymbol{x}\in\boldsymbol{X}} \boldsymbol{x}\boldsymbol{x}^\top$. In contrast to isotropic Euclidean error, equation 1 penalizes residual directions unequally: an error along $\boldsymbol{v}_k$ costs λ_k per unit of squared magnitude, and plain mean squared error is recovered only when $\boldsymbol{H} \propto \boldsymbol{I}$. Learned embeddings, however, exhibit sharply decaying spectra with $\lambda_1 \gg \lambda_d$ (Mu et al., 2017; Gao et al., 2019), so the two objectives are not equivalent.

3.2 Design Requirements for Retrieval-Aware Quantization

We isolate three structural properties of equation 1 that bear directly on quantizer design.

CROSS-TALK BETWEEN COORDINATES. In the canonical basis the loss splits into a diagonal and an off-diagonal part,

$$\mathcal{L}(\boldsymbol{e}) = \sum_{j=1}^{d} H_{jj} e_j^2 + \sum_{j \neq k} H_{jk} e_j e_k. \tag{3}$$

Encoders acting independently on coordinates, or on disjoint blocks of them—separable encoders—can only control the first sum; this class includes, for instance, scalar quantization (Lloyd, 1982), PQ (Jegou et al., 2011), and the rotation-plus-coordinate-wise codecs of recent MIPS systems (Gao & Long, 2024; Zandieh et al., 2026; Lenhardt et al., 2026). The loss and its diagonal part agree for every $\boldsymbol{e}$ if and only if $\boldsymbol{H}$ is diagonal, respectively block-diagonal, and the neglected energy $\mathrm{od}(\boldsymbol{H}) \triangleq \sum_{j \neq k} H_{jk}^2 = \|\boldsymbol{H}\|_F^2 - \sum_j H_{jj}^2$ is large precisely when the eigenbasis $\boldsymbol{V}$ of equation 2 is misaligned with the coordinate axes. This energy is inaccessible to any encoder in this class.

RATE ALLOCATION ACROSS COORDINATES. In the eigenbasis of $\boldsymbol{H}$ the loss is separable, but the bit-rate must still be distributed across coordinates. The high-resolution distortion-rate law makes the squared error of coordinate j at bit-width b_j proportional to $\sigma_j^2 2^{-2b_j}$, with σ_j^2 the source variance along $\boldsymbol{v}_j$ (Gersho & Gray, 2012); its contribution to equation 1 is therefore proportional to $w_j 2^{-2b_j}$, with effective weight $w_j = \lambda_j \sigma_j^2 > 0$. Standard pipelines enforce a uniform bit allocation $b_j \equiv R$. The following classical result (Goyal, 2001) quantifies the cost of ignoring the heterogeneity of $\boldsymbol{w} = (w_1, \ldots, w_d)$; we prove it in Appendix A.1.

Proposition 1 (Transform coding gain). *Let $w_1, \ldots, w_d > 0$, and let $D(\boldsymbol{b}) = \sum_j w_j 2^{-2b_j}$ be the total distortion of the bit allocation $\boldsymbol{b} = (b_1, \ldots, b_d)$ subject to $b_j \geq 0$ and the budget $\frac{1}{d}\sum_j b_j = R$. Write $\mathrm{AM}(\boldsymbol{w})$ and $\mathrm{GM}(\boldsymbol{w})$ for the arithmetic and geometric means of $\boldsymbol{w}$. If $R \geq \max_j \frac{1}{2}\log_2\big(\mathrm{GM}(\boldsymbol{w})/w_j\big)$, the uniform allocation $b_j \equiv R$ satisfies*

$$\frac{D(\textit{uniform})}{\min_{\boldsymbol{b}} D(\boldsymbol{b})} = \frac{\mathrm{AM}(\boldsymbol{w})}{\mathrm{GM}(\boldsymbol{w})} \geq 1, \tag{4}$$

with equality iff $w_1 = \cdots = w_d$; the minimum is attained at $b_j^\star = R + \frac{1}{2}\log_2\big(w_j/\mathrm{GM}(\boldsymbol{w})\big)$.

The penalty is the arithmetic-to-geometric mean ratio of the effective weights: a uniform budget is optimal only for a flat profile, and the ratio grows without bound as $\boldsymbol{w}$ becomes skewed. Under the estimator of Section 3.1, σ_j^2 equals λ_j up to a squared mean component, so $\boldsymbol{w}$ tracks the squared spectrum, whose sharp decay is precisely the regime where uniform allocation is most costly.

NORM CONTRACTION UNDER MMSE RECONSTRUCTION. Any quantizer reconstructing by the conditional mean of its cell satisfies the orthogonality principle $\mathbb{E}[\langle \widehat{\boldsymbol{x}}, \boldsymbol{e}\rangle] = 0$ (Gersho & Gray, 2012), hence $\mathbb{E}[\|\widehat{\boldsymbol{x}}\|_2^2] = \mathbb{E}[\|\boldsymbol{x}\|_2^2] - \mathbb{E}[\|\boldsymbol{e}\|_2^2] \leq \mathbb{E}[\|\boldsymbol{x}\|_2^2]$: reconstructions are systematically shorter than their originals and inner products are biased toward zero. A *common* multiplicative factor would be harmless, since MIPS rankings are invariant to a global positive rescaling of all scores; what matters is that the shrinkage is *heterogeneous* across datapoints, which reorders candidates with nearby scores; this is the regime that determines Top-k recall. These three properties are simultaneously addressed by quantizing in the eigenbasis of $\boldsymbol{H}$, allocating rate to the spectrum, and restoring norms after decoding. A natural objection is that a cheap randomized rotation already secures the first, a possibility we examine in Section 3.3.

3.3 RANDOMIZED ROTATIONS UNDER SPECTRAL ANISOTROPY

A Randomized Hadamard Transform (RHT) (Ailon & Chazelle, 2006) is the standard data-oblivious remedy for coordinate heterogeneity and the preprocessing step of the rotation-based codecs cited above. Let $\boldsymbol{R} = \boldsymbol{W}\boldsymbol{D}$, where $\boldsymbol{W}$ is the normalized Walsh-Hadamard matrix, so that $|W_{ij}| = d^{-1/2}$, and $\boldsymbol{D} = \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_d)$ has i.i.d. Rademacher entries; write $\widetilde{\boldsymbol{H}} = \boldsymbol{R}\boldsymbol{H}\boldsymbol{R}^\top$. We measure how far $\widetilde{\boldsymbol{H}}$ is from the isotropic target $(\mathrm{Tr}\,\boldsymbol{H}/d)\,\boldsymbol{I}$, which is the relevant target because these codecs minimize an unweighted squared error, optimal precisely when $\boldsymbol{H} \propto \boldsymbol{I}$: the question is whether the rotation makes that implicit assumption true. Proofs are deferred to Appendix A.2.

Theorem 1 (Diagonal equalization). *$\forall i$, $\mathbb{E}_{\boldsymbol{D}}[\widetilde{H}_{ii}] = \mathrm{Tr}(\boldsymbol{H})/d$ and*

$$\mathrm{Var}_{\boldsymbol{D}}\big(\widetilde{H}_{ii}\big) = \frac{2\,\mathrm{od}(\boldsymbol{H})}{d^2}. \tag{5}$$

Algorithm 1 VENICE: universal spectral rate-distortion framework wrapping a base quantizer

Require: Vectors $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, with rows indexed by $i \in \{1,\dots,N\}$ and coordinates by $k \in \{1,\dots,d\}$; target rate $R \in \mathbb{R}_{>0}$ in b/dim; bit tiers $\mathcal{B} = \{0\} \cup \mathcal{B}^+$, with active tiers $\mathcal{B}^+ \subset \mathbb{Z}_{\geq 1}$; norm precision f in bits; iterations $T_{\max}$; base quantizers $(\mathcal{Q}_b, \mathcal{Q}_b^{-1})_{b\in\mathcal{B}^+}$, applied row-wise.

Ensure: Per index: basis $\boldsymbol{V}$, rotated mean $\boldsymbol{\nu}$, allocation $\boldsymbol{b}^\star$, scales $\boldsymbol{s}$, tier rotations $\{\boldsymbol{R}_b\}$. Per vector: codes $\big(c_i^{(b)}\big)_{b\in\mathcal{B}^+}$ and quantized norm n_i.

$\boldsymbol{H} \leftarrow \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$; $\boldsymbol{H} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top, \lambda_1 \geq \cdots \geq \lambda_d \geq 0$ // **Pillar I**: decorrelation

$\boldsymbol{\mu} \leftarrow \frac{1}{N}\boldsymbol{X}^\top\mathbf{1}_N$; $\boldsymbol{\nu} \leftarrow \boldsymbol{V}^\top\boldsymbol{\mu}$; $\boldsymbol{Z} \leftarrow \boldsymbol{X}\boldsymbol{V} - \mathbf{1}_N\boldsymbol{\nu}^\top$ // rotate, then center

$w_k \leftarrow \lambda_k\sigma_k^2$ with $\sigma_k^2 = \frac{1}{N}\sum_i Z_{ik}^2, \forall k$

$[\theta_{\min}, \theta_{\max}] \leftarrow \big[\, 2^{-2\max\mathcal{B}}\min_k w_k,\ \max_k w_k \,\big]$ // **Pillar II**: reverse water-filling

for $t = 1,\dots,T_{\max}$ **do**

  $\theta \leftarrow \frac{1}{2}(\theta_{\min}+\theta_{\max})$; **if** $\sum_k \big\lfloor \frac{1}{2}\log_2(w_k/\theta)\big\rfloor_{\mathcal{B}} > Rd$ **then** $\theta_{\min} \leftarrow \theta$ **else** $\theta_{\max} \leftarrow \theta$

$b_k^\star \leftarrow \big\lfloor \frac{1}{2}\log_2(w_k/\theta_{\max})\big\rfloor_{\mathcal{B}}, \forall k$; $\mathcal{I}_b \leftarrow \{k : b_k^\star = b\}$, $d_b \leftarrow |\mathcal{I}_b|$, $\forall b \in \mathcal{B}$; $\mathcal{I}_+ \leftarrow \bigcup_{b\in\mathcal{B}^+}\mathcal{I}_b$

$s_k \leftarrow \lambda_k^{-1/2}$ for $k \in \mathcal{I}_+$ and 1 otherwise; $\boldsymbol{U} \leftarrow \boldsymbol{Z}\,\mathrm{diag}(\boldsymbol{s})^{-1}$ // **Pillar III**: $\lambda_k s_k^2 = 1$ on $\mathcal{I}_+$

for all $b \in \mathcal{B}^+$: $\boldsymbol{C}_b \leftarrow \mathcal{Q}_b\big(\boldsymbol{U}_{[:,\mathcal{I}_b]}\,\boldsymbol{R}_b^\top\big)$ // localized RHT

$n_i \leftarrow \big[\, \|\boldsymbol{z}_{i,\mathcal{I}_+}\| \,\big]_f, \forall i$ // **Pillar IV**: norm of the active coordinates

Decode. For every active tier $b \in \mathcal{B}^+$ set $\widehat{\boldsymbol{U}}_{[:,\mathcal{I}_b]} \leftarrow \mathcal{Q}_b^{-1}(\boldsymbol{C}_b)\,\boldsymbol{R}_b$, so that $\widehat{\boldsymbol{u}}_i$ is supported on $\mathcal{I}_+$. Undo the scaling, $\widetilde{\boldsymbol{z}}_i \leftarrow \mathrm{diag}(\boldsymbol{s})\,\widehat{\boldsymbol{u}}_i$; rescale the resulting direction to the stored length, $\widehat{\boldsymbol{z}}_i \leftarrow n_i \cdot \big(\widetilde{\boldsymbol{z}}_i/\|\widetilde{\boldsymbol{z}}_i\|\big)$, or $\mathbf{0}$ if $\widetilde{\boldsymbol{z}}_i = \mathbf{0}$; rotate back, $\widehat{\boldsymbol{x}}_i \leftarrow \boldsymbol{V}(\widehat{\boldsymbol{z}}_i + \boldsymbol{\nu})$.

Notation. $\boldsymbol{z}_i$ and $\boldsymbol{u}_i$ are the i-th rows of $\boldsymbol{Z}$ and $\boldsymbol{U}$, $\boldsymbol{U}_{[:,\mathcal{I}]}$ is the submatrix of columns indexed by $\mathcal{I}$, and $\boldsymbol{z}_{i,\mathcal{I}}$ is the restriction of $\boldsymbol{z}_i$ to the coordinates in $\mathcal{I}$; $\mathcal{I}_+$ is the set of active coordinates and $\mathcal{I}_0$ its complement; $\boldsymbol{C}_b$ stacks the tier-b codes $c_i^{(b)}$, one per vector; $[\,a\,]_f$ is the scalar a quantized to f bits; $\lfloor r\rfloor_{\mathcal{B}} = \max\{b \in \mathcal{B} : b \leq \max(r,0)\}$ rounds a continuous rate down to an admissible tier; $\boldsymbol{R}_b = \boldsymbol{W}_{d_b}\boldsymbol{D}_b$ is the RHT of Section 3.3 restricted to the tier-b block, applied block-wise so that $\boldsymbol{R}_b$ is orthogonal for every d_b.

The transform equalizes the diagonal in expectation, but its fluctuation is governed by $\mathrm{od}(\boldsymbol{H})$, the same quantity that Section 3.2 identified as inaccessible to separable encoders, and vanishes only when $\boldsymbol{H}$ is already diagonal. The proof uses only $|W_{ij}| = d^{-1/2}$.

Corollary 1 (Off-diagonal retention)**.** *Let* $r(\boldsymbol{H}) \triangleq (\mathrm{Tr}\,\boldsymbol{H})^2/\|\boldsymbol{H}\|_F^2 \in [1,d]$. *Then*

$$\frac{\mathbb{E}_{\boldsymbol{D}}\big[\mathrm{od}(\widetilde{\boldsymbol{H}})\big]}{\|\boldsymbol{H}\|_F^2} = 1 - \frac{r(\boldsymbol{H})}{d} - \frac{2\,\mathrm{od}(\boldsymbol{H})}{d\,\|\boldsymbol{H}\|_F^2} \;\geq\; 1 - \frac{r(\boldsymbol{H})+2}{d}. \tag{6}$$

The quantity $r(\boldsymbol{H})$ is the effective number of significant eigenvalues: $r(\boldsymbol{H}) = d$ for a flat spectrum and $r(\boldsymbol{H}) = 1$ when $\boldsymbol{H}$ has rank one. Since $\|\widetilde{\boldsymbol{H}}\|_F = \|\boldsymbol{H}\|_F$ for every rotation, equation 6 says that a random rotation redistributes Frobenius energy rather than removing it, leaving, in expectation, a fraction at least $1 - (r(\boldsymbol{H})+2)/d$ off the diagonal.

Theorem 2 (Concentration)**.** *Let* $\rho(\boldsymbol{H}) \triangleq \mathrm{Tr}(\boldsymbol{H})/\|\boldsymbol{H}\|_2$. *There is an absolute constant* $c > 0$ *such that, for every* $\epsilon \in (0,1)$,

$$\mathbb{P}_{\boldsymbol{D}}\Big(\max_{i\leq d}\big|\widetilde{H}_{ii} - \mathrm{Tr}(\boldsymbol{H})/d\big| > \epsilon\ \mathrm{Tr}(\boldsymbol{H})/d\Big) \leq 2d\exp\big(-c\min\{\epsilon^2 r(\boldsymbol{H}),\ \epsilon\rho(\boldsymbol{H})\}\big). \tag{7}$$

The right-hand side falls below one only if $r(\boldsymbol{H}) \gtrsim \epsilon^{-2}\log(2d)$ and $\rho(\boldsymbol{H}) \gtrsim \epsilon^{-1}\log(2d)$. Both quantities are at most d, with equality only for a flat spectrum, so at any moderate accuracy ϵ the guarantee requires a high effective rank: precisely the situation excluded in Section 3.1. Theorem 1 also locates the regime in which a randomized transform is exactly benign: a vanishing equation 5 is not enough, since a diagonal but non-isotropic $\boldsymbol{H}$ has no fluctuation and is still rotated into a non-diagonal $\boldsymbol{R}\boldsymbol{H}\boldsymbol{R}^\top$, whereas on a block equal to $\lambda\boldsymbol{I}$ the transform acts trivially, $\boldsymbol{O}(\lambda\boldsymbol{I})\boldsymbol{O}^\top = \lambda\boldsymbol{I}$ for every orthogonal $\boldsymbol{O}$. A randomized rotation can thus reshape the source distribution at no cost to equation 1, when applied within the eigenbasis of $\boldsymbol{H}$ rather than in place of it (Section 4).

4 THE VENICE FRAMEWORK

VENICE realizes, in a single index-construction stage, the three remedies of Section 3.2 together with the rotation that Section 3.3 shows to be exactly benign inside the eigenbasis. *Spectral decorrelation* (Section 4.1) rotates into the eigenbasis of $\boldsymbol{H}$ and subtracts the rotated mean, turning the Hessian into $\boldsymbol{\Lambda}$ and cancelling the cross-talk term of equation 3. *Reverse water-filling* (Section 4.2) distributes the target rate across coordinates according to the weights $\boldsymbol{w}$ of Section 3.2, realizing the gain of equation 4 over a uniform budget; bit-widths come from the tier set $\mathcal{B}$, and the coordinates that do not earn an active tier are pruned. *Sensitivity equalization* (Section 4.3) multiplies each active coordinate by $\lambda_k^{1/2}$, which makes the Hessian block of every tier the identity, the condition required by Section 3.3, and only then applies a randomized rotation inside the block, reshaping the sub-vector without changing equation 1. *Norm restoration* (Section 4.4) counteracts the heterogeneous shrinkage of Section 3.2 with one factor per datapoint.

Algorithm 1 states the procedure, with $\boldsymbol{X}$ now denoting the $N \times d$ matrix stacking the collection. Every active tier thus receives a sub-vector whose coordinates are equally important, its Hessian block being the identity, and whose energy the in-block rotation spreads uniformly: the regime a base quantizer is built for. VENICE asks nothing of $\mathcal{Q}_b$ beyond an encoder and a decoder, which is what lets one construction wrap codecs of different nature; it is also free of learned parameters, its four stages reducing to an eigendecomposition, a bisection on a scalar, a rescaling followed by a random rotation, and a per-vector norm ratio.

4.1 PILLAR I: SPECTRAL DECORRELATION

VENICE encodes in the eigenbasis $\boldsymbol{V}$ of equation 2, computed once from the index. Writing $\boldsymbol{\mu}$ for the sample mean of $\boldsymbol{X}$ and $\boldsymbol{\nu} \triangleq \boldsymbol{V}^\top \boldsymbol{\mu}$ for its coordinates in that basis, each vector is encoded as $\boldsymbol{z}_i = \boldsymbol{V}^\top \boldsymbol{x}_i - \boldsymbol{\nu}$ and reconstructed as $\widehat{\boldsymbol{x}}_i = \boldsymbol{V}(\widehat{\boldsymbol{z}}_i + \boldsymbol{\nu})$, with $\boldsymbol{V}$ and $\boldsymbol{\nu}$ shared by the whole index. Since $\boldsymbol{V}$ is orthogonal and $\boldsymbol{\nu}$ is re-added exactly, the residual is $\boldsymbol{e}_i = \boldsymbol{V}(\boldsymbol{z}_i - \widehat{\boldsymbol{z}}_i)$, so equation 1 evaluates to

$$\mathcal{L}(\boldsymbol{e}_i) = (\boldsymbol{z}_i - \widehat{\boldsymbol{z}}_i)^\top \big(\boldsymbol{V}^\top \boldsymbol{H} \boldsymbol{V}\big)(\boldsymbol{z}_i - \widehat{\boldsymbol{z}}_i) = \sum_{k=1}^{d} \lambda_k \big(z_{ik} - \widehat{z}_{ik}\big)^2, \tag{8}$$

because $\boldsymbol{V}^\top \boldsymbol{H} \boldsymbol{V} = \boldsymbol{\Lambda}$ by equation 2. This settles the first requirement of Section 3.2 exactly: $\mathrm{od}(\boldsymbol{V}^\top \boldsymbol{H} \boldsymbol{V}) = 0$, so the off-diagonal sum of equation 3 is removed, the diagonal objective available to a separable encoder coincides with the retrieval loss for every residual, and the energy declared there to be inaccessible no longer exists. No other basis achieves this, since $\mathrm{od}(\boldsymbol{U}^\top \boldsymbol{H} \boldsymbol{U}) = 0$ forces the columns of $\boldsymbol{U}$ to be eigenvectors of $\boldsymbol{H}$, whereas a data-oblivious rotation leaves, in expectation, a fraction at least $1 - (r(\boldsymbol{H}) + 2)/d$ of that energy off the diagonal by equation 6. Finally, the coordinates z_{ik} have zero mean and variance $\sigma_k^2 = \lambda_k - \nu_k^2$, since $\boldsymbol{V}^\top \boldsymbol{\Sigma} \boldsymbol{V} = \boldsymbol{\Lambda} - \boldsymbol{\nu}\boldsymbol{\nu}^\top$ for the covariance $\boldsymbol{\Sigma} = \boldsymbol{H} - \boldsymbol{\mu}\boldsymbol{\mu}^\top$; they form the weights $w_k = \lambda_k \sigma_k^2$ used by Section 4.2.

4.2 PILLAR II: REVERSE WATER-FILLING

The off-diagonal term of equation 3 couples the coordinates, so before Section 4.1 the distortion was not a sum of per-coordinate ones and no rate could be assigned to a coordinate in isolation. It now is: Section 3.2 charges coordinate k the distortion $w_k 2^{-2b_k}$ at bit-width b_k, so an allocation costs $D(\boldsymbol{b}) = \sum_k w_k 2^{-2b_k}$, and the weights $w_k = \lambda_k \sigma_k^2$ decay steeply (Section 3.1), which is the regime in which the uniform $b_k \equiv R$ is most wasteful, by equation 4. We therefore allocate by reverse water-filling (Shannon et al., 1959; Cover et al., 1991), derived for $D(\boldsymbol{b})$ from the Karush-Kuhn-Tucker conditions in Appendix B.1.

Proposition 2 (Reverse water-filling)**.** *Let $w_1, \ldots, w_d > 0$ and $R > 0$. The problem $\min_{\boldsymbol{b}} D(\boldsymbol{b})$ subject to $b_k \geq 0$ and $\frac{1}{d}\sum_k b_k = R$ has a unique minimizer,*

$$b_k^\star = \Big(\tfrac{1}{2}\log_2 \frac{w_k}{\theta}\Big)_+, \qquad D(\boldsymbol{b}^\star) = \sum_{k=1}^{d} \min(\theta, w_k), \tag{9}$$

where $(\cdot)_+ = \max(\cdot, 0)$ and $\theta > 0$ is the unique level at which $\sum_k b_k^\star = Rd$.

The level θ is a distortion threshold determined by the rate budget. Every coordinate with $w_k > \theta$ receives rate until its distortion equals θ; every coordinate with $w_k \leq \theta$ already costs no more than

that, so the same rate reduces $D(\boldsymbol{b})$ more elsewhere and it is left unencoded, $b_k^\star = 0$. Pruning is therefore part of the optimum rather than a criterion beside it, and it concentrates the Rd bits on the $m = |\mathcal{I}_+| \le d$ active coordinates, an average of $Rd/m \ge R$ bits each.

VENICE evaluates equation 9 over the tier set $\mathcal{B}$, replacing $(\cdot)_+$ by $\lfloor\cdot\rfloor_{\mathcal{B}}$ and locating θ by bisection on $[\theta_{\min}, \max_k w_k]$ (Algorithm 1, lines 4–6), which converges because raising θ lowers every $b_k^\star$; rounding down keeps the realized rate at or below R, and restricting rates to $\mathcal{B}$ is the usual Lagrangian treatment of a discrete quantizer family (Shoham & Gersho, 1988). The law $w_k 2^{-2b_k}$ is derived in the limit of large b_k, whereas the allocation is computed at low rate. It remains exact at $b_k = 0$, and any constant factor it misses is absorbed into θ and leaves $\boldsymbol{b}^\star$ unchanged, so the approximation affects only how rate is divided, never the ranking by w_k (Appendix B.1).

Coordinates in a tier share a bit-width but not a sensitivity: the λ_k inside it still differ, so equation 8 charges their errors unequally, while the base quantizer minimizes an unweighted one. Section 4.3 removes the mismatch.

4.3 PILLAR III: SENSITIVITY EQUALIZATION

A tier groups coordinates of the same bit-width but of different sensitivity, and $\mathcal{Q}_b$ never sees λ_k. VENICE moves the weights out of the metric and into the data,

$$u_{ik} = \lambda_k^{1/2} z_{ik} \quad (k \in \mathcal{I}_+), \qquad \text{so that} \qquad \boldsymbol{\Lambda}_b \mapsto \boldsymbol{\Lambda}_b^{-1/2} \boldsymbol{\Lambda}_b \boldsymbol{\Lambda}_b^{-1/2} = \boldsymbol{I}_{d_b} \quad \forall b \in \mathcal{B}^+, \tag{10}$$

making every tier block $\boldsymbol{\Lambda}_b$ of $\boldsymbol{\Lambda}$ the identity, so that unweighted squared error on $\boldsymbol{u}$ is exactly what those coordinates contribute to equation 8. The same rescaling gives u_{ik} variance w_k, and since Section 4.2 assigns the same bit-width to coordinates of comparable w_k, a tier is also a group of comparable variance. An identity block is moreover the condition under which Section 3.3 shows an in-block rotation to be harmless, so each tier is then rotated by the RHT $\boldsymbol{R}_b$ of that section, $\boldsymbol{y}_{i,b} = \boldsymbol{R}_b \boldsymbol{u}_{i,b}$, where $\boldsymbol{u}_{i,b}$ is the tier-b block of $\boldsymbol{u}_i$; the rotation spreads its energy at no cost to the loss, and $\boldsymbol{y}_{i,b}$ is the sub-vector handed to $\mathcal{Q}_b$. The order matters: rotating first would leave $\boldsymbol{R}_b \boldsymbol{\Lambda}_b \boldsymbol{R}_b^\top$, no longer diagonal, and the cross-talk removed in Section 4.1 would return. Both steps are isometries of equation 8, proved in Appendix B.2.

Proposition 3 (Tier isometry)**.** *Let $\widehat{\boldsymbol{y}}_{i,b}$ be the reconstruction of $\boldsymbol{y}_{i,b}$ returned by $\mathcal{Q}_b$, and let $\widetilde{\boldsymbol{z}}_i$ have blocks $\boldsymbol{\Lambda}_b^{-1/2} \boldsymbol{R}_b^\top \widehat{\boldsymbol{y}}_{i,b}$ and zero on $\mathcal{I}_0$. For every orthogonal $\boldsymbol{R}_b$,*

$$\mathcal{L}\big(\boldsymbol{V}(\boldsymbol{z}_i - \widetilde{\boldsymbol{z}}_i)\big) = \sum_{b \in \mathcal{B}^+} \big\|\boldsymbol{y}_{i,b} - \widehat{\boldsymbol{y}}_{i,b}\big\|_2^2 + \sum_{k \in \mathcal{I}_0} \lambda_k z_{ik}^2. \tag{11}$$

Every tier therefore reaches its quantizer as a zero-mean sub-vector whose coordinates are equally weighted, of comparable variance, and with energy spread across them: the regime a base quantizer is built for. Its squared error is the first term of equation 11, charged to equation 1 without reweighting, so any reduction of it lowers the retrieval loss by the same amount, whatever $\mathcal{Q}_b$ optimizes internally; the second term is the pruning cost already fixed by Section 4.2. All components are attributed for: the shrinkage of Section 3.2 survives, and Section 4.4 corrects it.

4.4 PILLAR IV: NORM RESTORATION

The first three pillars shape the input of $\mathcal{Q}_b$; the last corrects its output. On average a decoder returns a reconstruction shorter than the vector it encodes (Lemma 4, Appendix B.3), and the amount differs from one vector to the next. A common factor would be harmless, since MIPS orderings are invariant to a global positive rescaling, but unequal factors reorder candidates whose scores are close—the regime that decides top-k recall (Section 3.2). The correction is therefore made one datapoint at a time: VENICE stores the norm of its active coordinates to f bits, $n_i = [\,\|\boldsymbol{z}_{i,\mathcal{I}_+}\|\,]_f$, and rescales the decoded direction to that length,

$$\widehat{\boldsymbol{z}}_i = n_i \frac{\widetilde{\boldsymbol{z}}_i}{\|\widetilde{\boldsymbol{z}}_i\|}, \qquad \text{so that} \qquad \|\widehat{\boldsymbol{z}}_i\| = n_i \quad \forall i. \tag{12}$$

Every contraction factor is set to one, and what remains is an error in direction. The step reads only $\widetilde{\boldsymbol{z}}_i$ and overwrites its length, so it assumes nothing about how the code was produced: a quantizer that already restores its own norm, as in Lenhardt et al. (2026) or Zandieh et al. (2026), is left as it was up to the f-bit precision, since the stored length replaces the one it returns, while a quantizer that does not is corrected.

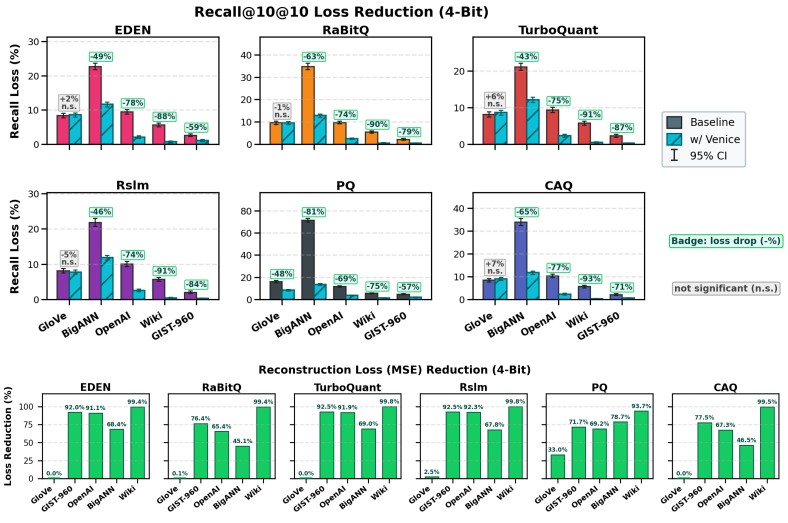

Figure 1: Recall loss ($1 - \text{Recall@10@10}$, top) and reduction of the reconstruction error $\mathcal{L}_{\text{MSE}}$ (bottom) at 4 bits per dimension, with and without VENICE. Badges: relative loss drop; n.s.: not significant (95% CI).

5 EXPERIMENTAL EVALUATION

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate on five standard benchmarks spanning modalities, data geometries, and dimensional regimes: GLOVE (d=100, 1M vectors) (Pennington et al., 2014), BIGANN (d=128, 100M vectors) (Jegou et al., 2011), GIST-960 (d=960, 1M vectors) (Jegou et al., 2011), OPENAI (d=1536, 2M vectors) (Neelakantan et al., 2022; Simhadri et al., 2022), and WIKI (d=3072, 2M vectors) (Wikimedia Foundation, 2026; Gemini Team, Google, 2026); Appendix C.1 describes each. For each we evaluate 500 randomly sampled test queries against exact full-precision ground truth.

Base codecs. We consider six codecs of different nature: EDEN (Vargaftik et al., 2022), EXTENDED-RABITQ (hereafter RABITQ) (Gao et al., 2025), TURBOQUANT (Zandieh et al., 2026), RSLM (Lenhardt et al., 2026), PQ (Jegou et al., 2011), and CAQ (Li et al., 2025); Appendix C.2 describes each. Every codec is evaluated standalone and wrapped, unmodified, by VENICE at 4-, 2- and 1-bit per dimension on average.

Metrics. $\text{Recall@}X\text{@}Y$ is the fraction of the exact Top-X neighbors appearing in the Top-Y returned; we report the *recall loss* $1 - \text{Recall@}X\text{@}Y$, with X=Y=10 as our primary metric. We also report the *normalized reconstruction error* $\mathcal{L}_{\text{MSE}} = \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|^2 / \|\boldsymbol{x}\|^2$.

Configuration. VENICE uses a single configuration across all datasets and base codecs, with no per-dataset tuning: $\boldsymbol{H}$ estimated from 10^4 vectors, bit-widths drawn from $\mathcal{B} = \{0, 1, 2, 4, 8\}$, 60 bisection steps for θ, and norms quantized to f=16 bits.

Storage and cost. Reported rates are the bits per dimension of the code. Apart from PQ, which stores codes only, every evaluated method stores one additional scalar per vector, a norm or a scale factor, at f bits. Beyond this, VENICE adds only index-level state: the shared eigenbasis $\boldsymbol{V}$ and tier rotations $\{\boldsymbol{R}_b\}$ incur an $\mathcal{O}(d^2)$ overhead amortized over N points. At serving time, the primary computational step is the query projection $\boldsymbol{V}^\top \boldsymbol{q}$, of $\mathcal{O}(d^2)$ complexity, applied once per query since datapoints are pre-rotated offline. Because $\boldsymbol{V}$ is shared across queries, it remains resident in modern CPU L3 cache and is applied with SIMD. Baseline codecs originally incurred comparable complexity to rotate queries, though several adopted the faster $\mathcal{O}(d \log d)$ FWHT in practice.

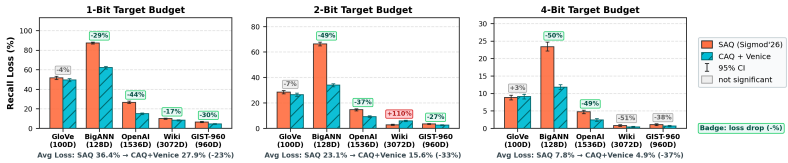

Figure 2: Recall@10@10 loss of SAQ and of CAQ wrapped by VENICE. Badges: relative drop.

5.2 EXPERIMENTAL RESULTS

Wrapping base codecs. Figure 1 reports recall loss with and without VENICE at 4 bits per dimension. Our framework reduces the loss for every codec on every dataset except GLOVE, by 43 to 93%. The pattern holds as the budget tightens: at 1 bit VENICE still removes more than 60% of the loss on OPENAI and WIKI for all six codecs which is significant as it makes these ultra low-bit rates suddenly useful in practice (Appendix D.1; see also Appendix D.2 for Recall@1@1 showing similar pattern). BIGANN is the dataset with the highest recall loss, with 85% of its variance in 64 of its 128 dimensions, and VENICE removes 43 to 81% of the loss there. In contrast, wrapping GLOVE at 4 bits provides no benefits for majority of the codecs. The spectrum of GLOVE is nearly flat, so VENICE even after water-filling keeps all dimensions to be 4-bit and the difference is not significant, except on PQ, which benefits from the rotation and norm restoration. Reconstruction error $\mathcal{L}_{\text{MSE}}$ improves as well, falling by 45.1 to 99.8% on all datasets except GLOVE.

Comparison with SAQ. SAQ is CAQ together with its segmentation and bit allocation (Section 2), so wrapping CAQ with VENICE keeps the base quantizer fixed and replaces the construction around it: bits follow the weights $\lambda_k \sigma_k^2$ rather than the variances σ_k^2 alone (Li et al., 2025, Eq. 17), and the coordinates sharing a budget need not form a contiguous block. Figure 2 shows macro-averaged Recall loss improving by 23%, 33% and 37% at 1, 2 and 4 bits in favor of VENICE. The margin is concentrated where SAQ's segmentation is most constrained: its segments are multiples of 64 dimensions, so BIGANN (d=128) admits at most two of them and VENICE reduces the recall loss by 50% at 4 bits. VENICE is also ahead on OPENAI at all bit rates, while SAQ is ahead only on WIKI at 2 bits, where 48 segments give SAQ fine-grained per-segment scale factors advantage.

Ablation. We remove one pillar at a time on PQ and RABITQ (Appendix D.3), contrasting a learned codebook against a training-free codec with built-in rotation and norm scaling. First, contribution of *spectral decorrelation* and *reverse water-filling* grow sharply as the budget tightens: at 1 bit they contribute up to 16.3 and 18.0 points of Recall@10@10 on PQ (16.4 and 17.1 on RABITQ), versus 3.3 and 3.0 points at 4 bits (5.5 and 4.3 on RABITQ), while remaining statistically neutral at 4 bits on the near-flat spectrum of GLOVE. Second, *sensitivity equalization* matters little on RABITQ, which already rotates its input (at most 1.7 points), but yields up to 7.8 points on PQ (GIST-960, 2 bits) by balancing coordinates across subspaces. Third, *norm restoration* (Section 4.4) corrects centroid norm shrinkage on PQ across all datasets, whereas on norm-preserving RABITQ it is neutral when full-dimensional norms are retained (GLOVE across all bit-widths; WIKI and GIST-960 at 4 bits) and only activates when water-filling truncates 0-bit tail dimensions.

6 CONCLUSION

Most vector quantizers assume that coordinates are uncorrelated and equally important, yet real-world embeddings are strongly correlated and skewed across directions. We showed that a global randomized rotation alone does not fix this mismatch on anisotropic data: it redistributes coordinate correlations rather than removing them. Instead of designing a new quantizer, VENICE fixes the data geometry around any existing one in closed form. It decorrelates coordinates in the eigenbasis of $\boldsymbol{H}$, allocates more bits to directions that matter most for retrieval while pruning the rest, equalizes and locally rotates coordinates within each bit tier, and restores vector norms after decoding. Wrapping six training-free and learned codecs without modification, VENICE cuts recall loss by up to 93% and reconstruction error up to 99.8% on anisotropic embeddings. This sharp drop in reconstruction error is particularly valuable for applications that rely on faithful vector recovery, opening the door to integrating VENICE with IVF and graph indices, estimating $\boldsymbol{H}$ from live query logs, and compressing LLM weights and KV caches.

AI USE STATEMENT

In this work, we used generative AI tools for the early reviewer feedback to check phrasing, mathematical content and references, including suggestions of related work to compare against. We also used them to assist in the writing of proofs, by checking derivations, pointing out errors and suggesting corrections. We have not used generative AI tools for creating new datasets. Additionally, we used generative AI tools to generate some of the evaluation and plotting code. We have reviewed all AI-assisted work by at least two authors and cross-checked the numbers from the independent evaluations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All theoretical results are stated with their assumptions and proved in Appendix A and Appendix B. VENICE is described in detail in Algorithm 1, and Section 5.1 gives its single configuration, used unchanged across datasets and codecs, together with the evaluation protocol. Appendix C describes the five datasets and the six base codecs. The results are reported in Section 5.2 and Appendix D.

REFERENCES

Cecilia Aguerrebere, Mark Hildebrand, Ishwar Singh Bhati, Theodore Willke, and Mariano Tepper. Locally-adaptive quantization for streaming vector search, 2024. URL https://arxiv.org/abs/2402.02044.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pp. 557–563, 2006.

Heesang Ann, Joongkyu Lee, and Min-hwan Oh. Block-sphere vector quantization. *arXiv preprint arXiv:2605.19972*, 2026.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Artem Babenko and Victor Lempitsky. Additive quantization for extreme vector compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 931–938, 2014.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in neural information processing systems*, 36:4396–4429, 2023.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

Jianyang Gao and Cheng Long. Rabitq: Quantizing high-dimensional vectors with a theoretical error bound for approximate nearest neighbor search. *Proceedings of the ACM on Management of Data*, 2(3):1–27, 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. *Proceedings of the ACM on Management of Data*, 3(3):1–26, 2025.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):744–755, 2013.

Gemini Team, Google. Gemini embedding 2: A native multimodal embedding model from gemini, 2026. URL https://arxiv.org/abs/2605.27295.

Allen Gersho and Robert M Gray. *Vector quantization and signal compression*. Springer Science & Business Media, 2012.

Vivek K Goyal. Theoretical foundations of transform coding. *IEEE Signal processing magazine*, 18 (5):9–21, 2001.

Qin-Zhen Guo, Zhi Zeng, Shuwu Zhang, Guixuan Zhang, and Yuan Zhang. Adaptive bit allocation product quantization. *Neurocomputing*, 171:866–877, 2016a.

Ruiqi Guo, Sanjiv Kumar, Krzysztof Choromanski, and David Simcha. Quantization based fast inner product search. In *Artificial intelligence and statistics*, pp. 482–490. PMLR, 2016b.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, pp. 3887–3896. PMLR, 2020.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):117–128, January 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.57.

Hervé Jegou, Romain Tavenard, Matthijs Douze, and Laurent Amsaleg. Searching in one billion vectors: Re-evaluating the state of the art. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1817–1824, 2011.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547, 2019.

Rastislav Lenhardt, Teodora Dobos, Thomas Vecchiato, Jiri Isa, and Igor Ginzburg. Rslm: Training-free vector quantization for approximate nearest neighbor search, 2026. URL https://arxiv.org/abs/2608.30384.

Hui Li, Shiyuan Deng, Xiao Yan, Xiangyu Zhi, and James Cheng. Saq: Pushing the limits of vector quantization through code adjustment and dimension segmentation. *Proceedings of the ACM on Management of Data*, 3(6), December 2025. doi: 10.1145/3769824. URL https://doi.org/10.1145/3769824.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training, 2022. URL https://arxiv.org/abs/2201.10005.

John Paparrizos, Ikraduya Edian, Chunwei Liu, Aaron J Elmore, and Michael J Franklin. Fast adaptive similarity search through variance-aware quantization. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 2969–2983. IEEE, 2022.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration, 2013. URL https://arxiv.org/abs/1306.2872.

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

Yair Shoham and Allen Gersho. Efficient bit allocation for an arbitrary set of quantizers (speech coding). *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(9):1445–1453, 1988.

Harsha Vardhan Simhadri, George Williams, Martin Aumüller, Matthijs Douze, Artem Babenko, Dmitry Baranchuk, Qi Chen, Lucas Hosseini, Ravishankar Krishnaswamny, Gopal Srinivasa, et al. Results of the neurips'21 challenge on billion-scale approximate nearest neighbor search. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 177–189. PMLR, 2022.

Shay Vargaftik, Ran Ben-Basat, Amit Portnoy, Gal Mendelson, and Yaniv Ben-Itzhak. Eden: Communication-efficient and robust distributed mean estimation for federated learning. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21984–22014, 2022. URL https://proceedings.mlr.press/v162/vargaftik22a/vargaftik22a.pdf.

Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search. *arXiv preprint arXiv:2101.12631*, 2021.

Wikimedia Foundation. Wikimedia downloads: Wikipedia database dumps. https://dumps.wikimedia.org/, 2026. English Wikipedia snapshot; accessed June 2026.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. Turboquant: Online vector quantization with near-optimal distortion rate. In *International Conference on Learning Representations*, volume 2026, pp. 56418–56439, 2026.

A Proofs for Section 3

A.1 Transform Coding Gain

Proposition 4 (Proposition 1 restated). *Let $w_1, \dots, w_d > 0$ and $R > 0$, let $D(\boldsymbol{b}) = \sum_{j=1}^{d} w_j 2^{-2b_j}$, and consider*

$$\min_{\boldsymbol{b}\in\mathbb{R}^d} D(\boldsymbol{b}) \qquad \textit{subject to} \qquad b_j \geq 0 \ \ \forall j, \qquad \tfrac{1}{d}\textstyle\sum_{j=1}^{d} b_j = R. \tag{13}$$

Write $\mathrm{AM}(\boldsymbol{w}) = \frac{1}{d}\sum_j w_j$ and $\mathrm{GM}(\boldsymbol{w}) = \big(\prod_j w_j\big)^{1/d}$. If $R \geq \max_j \frac{1}{2}\log_2\big(\mathrm{GM}(\boldsymbol{w})/w_j\big)$, then equation 13 is solved uniquely by $b_j^\star = R + \frac{1}{2}\log_2\big(w_j/\mathrm{GM}(\boldsymbol{w})\big)$ and

$$\frac{D(\textit{uniform})}{\min_{\boldsymbol{b}} D(\boldsymbol{b})} = \frac{\mathrm{AM}(\boldsymbol{w})}{\mathrm{GM}(\boldsymbol{w})} \ \geq \ 1, \tag{14}$$

with equality if and only if $w_1 = \cdots = w_d$.

Proof. Step 1 (relaxation and its unique solution). Drop the constraints $b_j \geq 0$ and consider

$$\min_{\boldsymbol{b}\in\mathbb{R}^d} D(\boldsymbol{b}) \qquad \text{subject to} \qquad \textstyle\sum_j b_j = Rd. \tag{15}$$

Each map $b \mapsto w_j 2^{-2b}$ has second derivative $4(\ln 2)^2 w_j 2^{-2b} > 0$, so D is strictly convex on $\mathbb{R}^d$; the feasible set of equation 15 is a nonempty affine subspace. A strictly convex differentiable function admits at most one minimizer on a convex set, and D is coercive along the subspace,[1] so the minimizer exists and is unique and is characterized by stationarity of $L(\boldsymbol{b}, \mu) = D(\boldsymbol{b}) + \mu\big(\sum_j b_j - Rd\big)$:

$$\frac{\partial L}{\partial b_j} = -2\ln 2\ w_j 2^{-2b_j} + \mu = 0 \qquad \Longrightarrow \qquad w_j 2^{-2b_j} = \theta \triangleq \frac{\mu}{2\ln 2} \quad \forall j, \tag{16}$$

so at the optimum every coordinate carries the *same* distortion θ. Note $\theta > 0$, since $w_j 2^{-2b_j} > 0$.

Step 2 (value of θ). From equation 16, $b_j = \frac{1}{2}\log_2(w_j/\theta)$. Imposing the budget,

$$\sum_{j=1}^{d} \tfrac{1}{2}\log_2\frac{w_j}{\theta} = Rd \iff \tfrac{1}{2}\Big(\sum_j \log_2 w_j - d\log_2\theta\Big) = Rd \iff \log_2\theta = \frac{1}{d}\sum_j \log_2 w_j - 2R, \tag{17}$$

that is $\theta = \mathrm{GM}(\boldsymbol{w})\, 2^{-2R}$, and therefore

$$b_j^\star = \tfrac{1}{2}\log_2\frac{w_j}{\mathrm{GM}(\boldsymbol{w})2^{-2R}} = R + \tfrac{1}{2}\log_2\frac{w_j}{\mathrm{GM}(\boldsymbol{w})}, \qquad D(\boldsymbol{b}^\star) = \sum_{j=1}^{d}\theta = d\,\mathrm{GM}(\boldsymbol{w})\,2^{-2R}. \tag{18}$$

Step 3 (the hypothesis makes the relaxation tight). We have $b_j^\star \geq 0$ for all j if and only if $R \geq \frac{1}{2}\log_2\big(\mathrm{GM}(\boldsymbol{w})/w_j\big)$ for all j, i.e. if and only if $R \geq \max_j \frac{1}{2}\log_2\big(\mathrm{GM}(\boldsymbol{w})/w_j\big)$, which is the hypothesis. Hence $\boldsymbol{b}^\star$ is feasible for equation 13. Since the feasible set of equation 13 is contained in that of equation 15, a feasible point attaining the relaxed minimum attains the constrained one as well, and uniqueness is inherited from Step 1.

Step 4 (uniform allocation). The point $b_j \equiv R$ is feasible for equation 13, since $R > 0$ and $\frac{1}{d}\sum_j R = R$, and

$$D(\text{uniform}) = \sum_{j=1}^{d} w_j 2^{-2R} = d\,\mathrm{AM}(\boldsymbol{w})\,2^{-2R}. \tag{19}$$

Step 5 (ratio). Dividing by equation 18,

$$\frac{D(\text{uniform})}{D(\boldsymbol{b}^\star)} = \frac{d\,\mathrm{AM}(\boldsymbol{w})\,2^{-2R}}{d\,\mathrm{GM}(\boldsymbol{w})\,2^{-2R}} = \frac{\mathrm{AM}(\boldsymbol{w})}{\mathrm{GM}(\boldsymbol{w})}. \tag{20}$$

[1] If $\|\boldsymbol{b}\| \to \infty$ with $\sum_j b_j = Rd$, some $b_j \to -\infty$, hence $D(\boldsymbol{b}) \to +\infty$.

Finally, $\log$ is strictly concave on $(0,\infty)$, so Jensen's inequality gives $\frac{1}{d}\sum_j \log w_j \le \log\big(\frac{1}{d}\sum_j w_j\big)$, i.e. $\log \mathrm{GM}(\boldsymbol{w}) \le \log \mathrm{AM}(\boldsymbol{w})$, with equality if and only if $w_1 = \cdots = w_d$. Exponentiating yields equation 14. □

Remark 1. The hypothesis on R states exactly that no coordinate is pruned at the optimum. When it fails, the constrained minimum exceeds the relaxed one, so

$$\frac{D(\text{uniform})}{\min_{\boldsymbol{b}} D(\boldsymbol{b})} \;\le\; \frac{\mathrm{AM}(\boldsymbol{w})}{\mathrm{GM}(\boldsymbol{w})}, \tag{21}$$

and equation 14 becomes an upper bound on the attainable gain. The pruning regime is the one treated by Proposition 2, whose water level θ is the same quantity as in equation 16.

A.2 Randomized Rotations

Setup and conventions. Throughout this subsection $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ is symmetric positive semi-definite with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$; $\boldsymbol{W}$ is the normalized Walsh–Hadamard matrix, so that $\boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}_d$ and $|W_{ij}| = d^{-1/2}$ for all i, j; $\boldsymbol{D} = \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_d)$ has i.i.d. Rademacher entries, $\mathbb{P}[\varepsilon_j = \pm 1] = 1/2$; $\boldsymbol{R} = \boldsymbol{W}\boldsymbol{D}$, which is orthogonal for every realization of $\boldsymbol{D}$; and $\widetilde{\boldsymbol{H}} = \boldsymbol{R}\boldsymbol{H}\boldsymbol{R}^\top = \boldsymbol{W}\boldsymbol{D}\boldsymbol{H}\boldsymbol{D}\boldsymbol{W}^\top$. All expectations are over $\boldsymbol{D}$. We use repeatedly

$$\mathbb{E}[\varepsilon_j] = 0, \qquad \varepsilon_j^2 = 1 \;\text{ a.s.}, \qquad \mathbb{E}\Big[\prod_{j\in J} \varepsilon_j\Big] = 0 \tag{22}$$

whenever some index of the multiset J has odd multiplicity.

Lemma 1 (Spectral ratios)**.** *Let $\boldsymbol{H} \ne \boldsymbol{0}$ be symmetric positive semi-definite and set $r(\boldsymbol{H}) = (\mathrm{Tr}\,\boldsymbol{H})^2/\|\boldsymbol{H}\|_F^2$ and $\rho(\boldsymbol{H}) = \mathrm{Tr}\,\boldsymbol{H}/\|\boldsymbol{H}\|_2$. Then*

$$1 \le r(\boldsymbol{H}) \le d, \qquad 1 \le \rho(\boldsymbol{H}) \le d, \tag{23}$$

and either upper bound is attained if and only if $\lambda_1 = \cdots = \lambda_d$, either lower bound if and only if $\mathrm{rank}(\boldsymbol{H}) = 1$.

Proof. Write $\mathrm{Tr}\,\boldsymbol{H} = \sum_i \lambda_i$, $\|\boldsymbol{H}\|_F^2 = \sum_i \lambda_i^2$ and $\|\boldsymbol{H}\|_2 = \lambda_1$, all with $\lambda_i \ge 0$.

Upper bounds. Cauchy–Schwarz applied to $(\lambda_1, \ldots, \lambda_d)$ and $(1, \ldots, 1)$ gives $\big(\sum_i \lambda_i\big)^2 \le d\sum_i \lambda_i^2$, i.e. $r(\boldsymbol{H}) \le d$, with equality if and only if the two vectors are proportional, i.e. all λ_i are equal. Similarly $\sum_i \lambda_i \le d\lambda_1$ gives $\rho(\boldsymbol{H}) \le d$, with equality if and only if $\lambda_i = \lambda_1$ for all i.

Lower bounds. Since $\lambda_i \ge 0$, $\big(\sum_i \lambda_i\big)^2 = \sum_i \lambda_i^2 + \sum_{i\ne j} \lambda_i\lambda_j \ge \sum_i \lambda_i^2$, so $r(\boldsymbol{H}) \ge 1$, with equality if and only if $\lambda_i\lambda_j = 0$ for all $i \ne j$, i.e. at most one eigenvalue is nonzero. Likewise $\sum_i \lambda_i \ge \lambda_1$ gives $\rho(\boldsymbol{H}) \ge 1$, with equality if and only if $\lambda_i = 0$ for $i \ge 2$. □

Theorem 3 (Diagonal equalization; Theorem 1 restated)**.** *For every $i \in \{1, \ldots, d\}$,*

$$\mathbb{E}_{\boldsymbol{D}}\big[\widetilde{H}_{ii}\big] = \frac{\mathrm{Tr}\,\boldsymbol{H}}{d}, \qquad \mathrm{Var}_{\boldsymbol{D}}\big(\widetilde{H}_{ii}\big) = \frac{2\,\mathrm{od}(\boldsymbol{H})}{d^2}. \tag{24}$$

Proof. Fix i. Expanding the matrix product,

$$\widetilde{H}_{ii} = \big(\boldsymbol{W}\boldsymbol{D}\boldsymbol{H}\boldsymbol{D}\boldsymbol{W}^\top\big)_{ii} = \sum_{j=1}^{d}\sum_{k=1}^{d} W_{ij}\,\varepsilon_j\,H_{jk}\,\varepsilon_k\,W_{ik}. \tag{25}$$

Split the double sum into its diagonal and off-diagonal parts,

$$\widetilde{H}_{ii} = \underbrace{\sum_{j=1}^{d} W_{ij}^2\varepsilon_j^2 H_{jj}}_{=:\,M_i} \;+\; \underbrace{\sum_{j\ne k} W_{ij}W_{ik}H_{jk}\,\varepsilon_j\varepsilon_k}_{=:\,T_i}. \tag{26}$$

Deterministic part. Since $\varepsilon_j^2 = 1$ a.s. and $W_{ij}^2 = 1/d$,

$$M_i = \frac{1}{d}\sum_{j=1}^{d} H_{jj} = \frac{\operatorname{Tr} \boldsymbol{H}}{d} \qquad \text{almost surely.} \tag{27}$$

Mean. For $j \neq k$ the product $\varepsilon_j\varepsilon_k$ has two indices of multiplicity one, so $\mathbb{E}[\varepsilon_j\varepsilon_k] = 0$ by equation 22; hence $\mathbb{E}[T_i] = 0$ and $\mathbb{E}[\widetilde{H}_{ii}] = M_i = \operatorname{Tr} \boldsymbol{H}/d$, the first claim.

Variance. Since $\operatorname{Var}(\widetilde{H}_{ii}) = \operatorname{Var}(T_i) = \mathbb{E}[T_i^2]$,

$$\mathbb{E}\big[T_i^2\big] = \sum_{j\neq k}\sum_{p\neq q} W_{ij}W_{ik}W_{ip}W_{iq}\, H_{jk}H_{pq}\, \mathbb{E}\big[\varepsilon_j\varepsilon_k\varepsilon_p\varepsilon_q\big]. \tag{28}$$

By equation 22 the expectation vanishes unless every index occurs an even number of times; with $j \neq k$ and $p \neq q$ this happens exactly for $(p,q) = (j,k)$ and $(p,q) = (k,j)$, in which case the expectation equals 1. Both surviving terms contribute $W_{ij}^2 W_{ik}^2 H_{jk}^2$, using $H_{jk} = H_{kj}$, so

$$\mathbb{E}\big[T_i^2\big] = 2\sum_{j\neq k} W_{ij}^2 W_{ik}^2 H_{jk}^2 = 2\sum_{j\neq k}\frac{1}{d}\cdot\frac{1}{d}\,H_{jk}^2 = \frac{2}{d^2}\sum_{j\neq k} H_{jk}^2 = \frac{2\,\mathrm{od}(\boldsymbol{H})}{d^2}. \tag{29}$$

□

Remark 2. The proof uses only $|W_{ij}| = d^{-1/2}$; orthogonality of $\boldsymbol{W}$ is never invoked. The conclusion therefore holds for any flat-entry matrix, which is why equation 24 is a property of the *randomization*, not of the Hadamard structure. In particular $\operatorname{Var}(\widetilde{H}_{ii}) = 0$ if and only if $\mathrm{od}(\boldsymbol{H}) = 0$: the diagonal is equalized deterministically exactly when there is nothing off the diagonal to begin with.

Corollary 2 (Off-diagonal retention; Corollary 1 restated)**.** *With* $r(\boldsymbol{H}) = (\operatorname{Tr}\boldsymbol{H})^2/\|\boldsymbol{H}\|_F^2$,

$$\frac{\mathbb{E}_{\boldsymbol{D}}\big[\mathrm{od}(\widetilde{\boldsymbol{H}})\big]}{\|\boldsymbol{H}\|_F^2} = 1 - \frac{r(\boldsymbol{H})}{d} - \frac{2\,\mathrm{od}(\boldsymbol{H})}{d\,\|\boldsymbol{H}\|_F^2} \;\geq\; 1 - \frac{r(\boldsymbol{H})+2}{d}. \tag{30}$$

Proof. Since $\boldsymbol{R}$ is orthogonal, $\|\widetilde{\boldsymbol{H}}\|_F^2 = \|\boldsymbol{R}\boldsymbol{H}\boldsymbol{R}^\top\|_F^2 = \|\boldsymbol{H}\|_F^2$ for every realization of $\boldsymbol{D}$. By the definition of $\mathrm{od}(\cdot)$,

$$\mathrm{od}\big(\widetilde{\boldsymbol{H}}\big) = \|\widetilde{\boldsymbol{H}}\|_F^2 - \sum_{i=1}^{d}\widetilde{H}_{ii}^2 = \|\boldsymbol{H}\|_F^2 - \sum_{i=1}^{d}\widetilde{H}_{ii}^2. \tag{31}$$

Taking expectations and using $\mathbb{E}[X^2] = (\mathbb{E}[X])^2 + \operatorname{Var}(X)$ together with Theorem 1,

$$\mathbb{E}\Big[\sum_{i=1}^{d}\widetilde{H}_{ii}^2\Big] = \sum_{i=1}^{d}\left[\Big(\frac{\operatorname{Tr}\boldsymbol{H}}{d}\Big)^2 + \frac{2\,\mathrm{od}(\boldsymbol{H})}{d^2}\right] = \frac{(\operatorname{Tr}\boldsymbol{H})^2}{d} + \frac{2\,\mathrm{od}(\boldsymbol{H})}{d}. \tag{32}$$

Therefore

$$\mathbb{E}\big[\mathrm{od}(\widetilde{\boldsymbol{H}})\big] = \|\boldsymbol{H}\|_F^2 - \frac{(\operatorname{Tr}\boldsymbol{H})^2}{d} - \frac{2\,\mathrm{od}(\boldsymbol{H})}{d}, \tag{33}$$

and dividing by $\|\boldsymbol{H}\|_F^2 > 0$ gives the equality in equation 30. For the inequality, note $\mathrm{od}(\boldsymbol{H}) = \|\boldsymbol{H}\|_F^2 - \sum_j H_{jj}^2 \leq \|\boldsymbol{H}\|_F^2$, so $2\,\mathrm{od}(\boldsymbol{H})/(d\|\boldsymbol{H}\|_F^2) \leq 2/d$. □

Lemma 2 (Hanson–Wright inequality; Rudelson & Vershynin, 2013)**.** *Let* $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_d)$ *have independent, mean-zero, sub-Gaussian entries with* $\max_j \|\varepsilon_j\|_{\psi_2} \leq K$*, and let* $\boldsymbol{A} \in \mathbb{R}^{d\times d}$*. There is an absolute constant* $c_0 > 0$ *such that, for every* $t \geq 0$*,*

$$\mathbb{P}\Big(\big|\boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon} - \mathbb{E}[\boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon}]\big| > t\Big) \;\leq\; 2\exp\!\left(-c_0\min\!\left\{\frac{t^2}{K^4\|\boldsymbol{A}\|_F^2},\; \frac{t}{K^2\|\boldsymbol{A}\|_2}\right\}\right). \tag{34}$$

Theorem 4 (Concentration; Theorem 2 restated)**.** *Let* $\rho(\boldsymbol{H}) = \operatorname{Tr}\boldsymbol{H}/\|\boldsymbol{H}\|_2$*. There is an absolute constant* $c > 0$ *such that, for every* $\epsilon \in (0,1)$*,*

$$\mathbb{P}_{\boldsymbol{D}}\Big(\max_{i\leq d}\big|\widetilde{H}_{ii} - \operatorname{Tr}(\boldsymbol{H})/d\big| > \epsilon\,\operatorname{Tr}(\boldsymbol{H})/d\Big) \;\leq\; 2d\exp\big(-c\min\{\epsilon^2 r(\boldsymbol{H}),\; \epsilon\,\rho(\boldsymbol{H})\}\big). \tag{35}$$

Proof. Fix i. Since $|W_{ij}| = d^{-1/2}$, write $W_{ij} = d^{-1/2}\sigma_{ij}$ with $\sigma_{ij} \in \{-1, +1\}$ and set $\boldsymbol{S}_i \triangleq \mathrm{diag}(\sigma_{i1}, \ldots, \sigma_{id})$, which satisfies $\boldsymbol{S}_i^\top = \boldsymbol{S}_i$ and $\boldsymbol{S}_i^2 = \boldsymbol{I}_d$, hence is orthogonal. From equation 25,

$$\widetilde{H}_{ii} = \sum_{j,k} W_{ij} W_{ik} H_{jk}\, \varepsilon_j \varepsilon_k = \boldsymbol{\varepsilon}^\top \boldsymbol{A}^{(i)} \boldsymbol{\varepsilon}, \qquad \boldsymbol{A}^{(i)} \triangleq \frac{1}{d}\, \boldsymbol{S}_i \boldsymbol{H} \boldsymbol{S}_i, \tag{36}$$

because $\big(\boldsymbol{S}_i \boldsymbol{H} \boldsymbol{S}_i\big)_{jk} = \sigma_{ij} H_{jk} \sigma_{ik}$.

Norms of $\boldsymbol{A}^{(i)}$. Since $\boldsymbol{S}_i$ is orthogonal, conjugation by it preserves both the Frobenius and the spectral norm, so

$$\big\|\boldsymbol{A}^{(i)}\big\|_F = \frac{\|\boldsymbol{H}\|_F}{d}, \qquad \big\|\boldsymbol{A}^{(i)}\big\|_2 = \frac{\|\boldsymbol{H}\|_2}{d}. \tag{37}$$

Mean. For Rademacher entries $\mathbb{E}[\varepsilon_j \varepsilon_k] = \delta_{jk}$, so $\mathbb{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{A}^{(i)} \boldsymbol{\varepsilon}] = \mathrm{Tr}\, \boldsymbol{A}^{(i)} = \frac{1}{d}\,\mathrm{Tr}(\boldsymbol{S}_i \boldsymbol{H} \boldsymbol{S}_i) = \frac{1}{d}\,\mathrm{Tr}(\boldsymbol{H})$ by cyclicity of the trace and $\boldsymbol{S}_i^2 = \boldsymbol{I}_d$; this agrees with Theorem 1.

Tail bound. Rademacher variables are sub-Gaussian with $\|\varepsilon_j\|_{\psi_2} = K_0$ an absolute constant. Apply Lemma 2 with $t = \epsilon\ \mathrm{Tr}(\boldsymbol{H})/d \geq 0$:

$$\frac{t^2}{K_0^4 \|\boldsymbol{A}^{(i)}\|_F^2} = \frac{\epsilon^2 (\mathrm{Tr}\, \boldsymbol{H})^2/d^2}{K_0^4 \|\boldsymbol{H}\|_F^2/d^2} = \frac{\epsilon^2\, r(\boldsymbol{H})}{K_0^4}, \qquad \frac{t}{K_0^2 \|\boldsymbol{A}^{(i)}\|_2} = \frac{\epsilon\ \mathrm{Tr}(\boldsymbol{H})/d}{K_0^2 \|\boldsymbol{H}\|_2/d} = \frac{\epsilon\, \rho(\boldsymbol{H})}{K_0^2}, \tag{38}$$

so that, with $c \triangleq c_0 \min\{K_0^{-4}, K_0^{-2}\}$,

$$\mathbb{P}\Big(\big|\widetilde{H}_{ii} - \mathrm{Tr}(\boldsymbol{H})/d\big| > \epsilon\ \mathrm{Tr}(\boldsymbol{H})/d\Big) \leq 2\exp\big(-c \min\{\epsilon^2 r(\boldsymbol{H}),\, \epsilon\, \rho(\boldsymbol{H})\}\big). \tag{39}$$

Union bound. The event in equation 35 is the union over $i = 1, \ldots, d$ of the events just bounded, and the bound is uniform in i; summing gives the factor d. □

Remark 3 (The benign case). A vanishing equation 24 is not sufficient for a rotation to be harmless: if $\boldsymbol{H}$ is diagonal and not a multiple of the identity then $\mathrm{od}(\boldsymbol{H}) = 0$ and $\widetilde{H}_{ii} = \mathrm{Tr}(\boldsymbol{H})/d$ deterministically, yet $\boldsymbol{R}\boldsymbol{H}\boldsymbol{R}^\top$ is no longer diagonal and carries off-diagonal energy $\|\boldsymbol{H}\|_F^2\big(1 - r(\boldsymbol{H})/d\big)$ created by the rotation. What is harmless is an isotropic block: if $\boldsymbol{H}$ restricted to a block equals $\lambda \boldsymbol{I}_n$, then $\boldsymbol{O}(\lambda \boldsymbol{I}_n)\boldsymbol{O}^\top = \lambda\, \boldsymbol{O}\boldsymbol{O}^\top = \lambda \boldsymbol{I}_n$ for every orthogonal $\boldsymbol{O} \in \mathbb{R}^{n \times n}$, so the block is invariant and $\mathcal{L}$ is unchanged. This is the hypothesis exploited in Section 4.3 and formalized in Proposition 3.

Remark 4 (Reading the two thresholds). The right-hand side of equation 35 is below 1 only if $c \min\{\epsilon^2 r(\boldsymbol{H}), \epsilon \rho(\boldsymbol{H})\} > \log(2d)$, i.e. only if

$$r(\boldsymbol{H}) > \frac{\log(2d)}{c\, \epsilon^2} \qquad \text{and} \qquad \rho(\boldsymbol{H}) > \frac{\log(2d)}{c\, \epsilon}. \tag{40}$$

By Lemma 1 both quantities are at most d and reach d only for a flat spectrum, so for any fixed accuracy ϵ the guarantee degrades as the spectrum becomes anisotropic, and is void when $r(\boldsymbol{H}) = O(1)$—the sharply decaying spectra of Section 3.1.

B PROOFS FOR SECTION 4

B.1 REVERSE WATER-FILLING

Proposition 5 (Proposition 2 restated)**.** *Let $w_1, \dots, w_d > 0$ and $R > 0$, and let $D(\boldsymbol{b}) = \sum_{k=1}^{d} w_k 2^{-2b_k}$. The problem*

$$\min_{\boldsymbol{b}\in\mathbb{R}^d} D(\boldsymbol{b}) \qquad \textit{subject to} \qquad b_k \geq 0 \ \ \forall k, \qquad \tfrac{1}{d}\textstyle\sum_k b_k = R, \tag{41}$$

has a unique minimizer

$$b_k^\star = \Big(\tfrac{1}{2}\log_2 \frac{w_k}{\theta}\Big)_+, \qquad D(\boldsymbol{b}^\star) = \sum_{k=1}^{d} \min(\theta, w_k), \tag{42}$$

where $\theta > 0$ is the unique level at which $\sum_k b_k^\star = Rd$.

Proof. Step 1 (existence and uniqueness). The feasible set $\mathcal{F} = \{\boldsymbol{b} \in \mathbb{R}^d : b_k \geq 0, \sum_k b_k = Rd\}$ is a scaled simplex: nonempty (it contains $b_k \equiv R$), convex, and compact. Each summand $b \mapsto w_k 2^{-2b}$ has second derivative $4(\ln 2)^2 w_k 2^{-2b} > 0$, so D is continuous and strictly convex on $\mathbb{R}^d$. A strictly convex continuous function on a nonempty compact convex set attains its minimum at exactly one point.

Step 2 (KKT conditions are necessary and sufficient). D is differentiable and the constraints are affine, so no constraint qualification beyond affinity is required; by convexity the KKT conditions are also sufficient. With multipliers $\mu \in \mathbb{R}$ for the budget and $\eta_k \geq 0$ for $b_k \geq 0$, the Lagrangian

$$L(\boldsymbol{b}, \mu, \boldsymbol{\eta}) = \sum_k w_k 2^{-2b_k} + \mu\Big(\sum_k b_k - Rd\Big) - \sum_k \eta_k b_k \tag{43}$$

gives the stationarity, dual feasibility and complementary slackness conditions

$$-2\ln 2\ w_k 2^{-2b_k} + \mu - \eta_k = 0, \qquad \eta_k \geq 0, \qquad \eta_k b_k = 0, \qquad \forall k. \tag{44}$$

Since $Rd > 0$, at least one b_k is positive; for such a k complementary slackness gives $\eta_k = 0$ and hence $\mu = 2\ln 2\, w_k 2^{-2b_k} > 0$. Define

$$\theta \ \triangleq\ \frac{\mu}{2\ln 2} \ > \ 0. \tag{45}$$

Case $b_k > 0$. Then $\eta_k = 0$ and equation 44 reads

$$w_k 2^{-2b_k} = \theta \qquad \Longleftrightarrow \qquad b_k = \tfrac{1}{2}\log_2 \frac{w_k}{\theta}, \tag{46}$$

and $b_k > 0$ holds if and only if $w_k > \theta$.

Case $b_k = 0$. Then equation 44 gives $\eta_k = \mu - 2\ln 2\, w_k \geq 0$, i.e. $w_k \leq \theta$.

The two cases are exhaustive and mutually exclusive, so $b_k^\star = \big(\tfrac{1}{2}\log_2(w_k/\theta)\big)_+$, the first half of equation 42.

Step 3 (existence and uniqueness of θ). Define

$$B(\theta) \ \triangleq\ \sum_{k=1}^{d} \Big(\tfrac{1}{2}\log_2 \frac{w_k}{\theta}\Big)_+, \qquad \theta \in \big(0, \max_k w_k\big]. \tag{47}$$

Each summand is continuous and non-increasing in θ, hence so is B. For $\theta < \max_k w_k$, the summand indexed by any $k \in \arg\max_k w_k$ is strictly positive and strictly decreasing, so B is strictly decreasing on $(0, \max_k w_k]$. Moreover $B(\max_k w_k) = 0$, and $B(\theta) \geq \tfrac{1}{2}\log_2(\max_k w_k/\theta) \to +\infty$ as $\theta \to 0^+$. By continuity and strict monotonicity there is exactly one $\theta \in (0, \max_k w_k)$ with $B(\theta) = Rd$, for every $Rd > 0$.

Step 4 (distortion at the optimum). Partition $\mathcal{I}_+ = \{k : w_k > \theta\}$ and $\mathcal{I}_0 = \{k : w_k \leq \theta\}$. For $k \in \mathcal{I}_+$, equation 46 gives $w_k 2^{-2b_k^\star} = \theta$; for $k \in \mathcal{I}_0$, $b_k^\star = 0$ gives $w_k 2^0 = w_k$. Summing,

$$D(\boldsymbol{b}^\star) = \sum_{k\in\mathcal{I}_+} \theta + \sum_{k\in\mathcal{I}_0} w_k = \sum_{k=1}^{d} \min(\theta, w_k). \tag{48}$$

□

Remark 5 (Exactness of the law at zero rate). The law $w_k 2^{-2b_k}$ is a high-resolution approximation, but at $b_k = 0$ it is exact. An unencoded coordinate is reconstructed by its mean, and $\boldsymbol{z}_i$ is centered, so $\widehat{z}_{ik} = 0$ and the squared error is $\mathbb{E}[z_{ik}^2] = \sigma_k^2$; its contribution to equation 8 is $\lambda_k \sigma_k^2 = w_k$, which equals $w_k 2^{-2 \cdot 0}$. The pruning branch of equation 42 therefore rests on an identity, not on an approximation.

Remark 6 (Invariance under a constant factor). Let the true distortion be $D_c(\boldsymbol{b}) = c \sum_k w_k 2^{-2b_k}$ for an unknown $c > 0$ common to all rates, $b_k = 0$ included. Repeating Step 2 with D_c replaces equation 46 by $c\, w_k 2^{-2b_k} = \theta'$ and the pruning condition by $c\, w_k \le \theta'$. Setting $\theta = \theta'/c$ returns both conditions verbatim, so $\boldsymbol{b}^\star$ is unchanged and only the attained value scales, $D_c(\boldsymbol{b}^\star) = c\, D(\boldsymbol{b}^\star)$.

Proposition 6 (Robustness of the allocation law)**.** *Let $\psi : [0, \infty) \to (0, \infty)$ be differentiable, strictly decreasing and strictly convex with $\psi(0) = 1$ and $\psi'(0)$ finite, and let $D_\psi(\boldsymbol{b}) = \sum_k w_k \psi(b_k)$. The minimizer of D_ψ over $\mathcal{F}$ is unique and there exists $\theta_\psi > 0$ such that*

$$b_k^\star = 0 \iff w_k \le \theta_\psi, \qquad \text{and} \qquad w_j \le w_k \implies b_j^\star \le b_k^\star. \tag{49}$$

Proof. Uniqueness follows as in Step 1, since ψ strictly convex and $w_k > 0$ make D_ψ strictly convex, and $\mathcal{F}$ is compact convex. The KKT conditions read $w_k \psi'(b_k) + \mu - \eta_k = 0$ with $\eta_k \ge 0$, $\eta_k b_k = 0$. As before $\mu = -w_k \psi'(b_k) > 0$ for any active k, because $\psi' < 0$.

Since ψ is strictly convex and differentiable, ψ' is strictly increasing, and since ψ is strictly decreasing, $\psi' < 0$; hence $\psi' : [0, \infty) \to [\psi'(0), 0)$ is a strictly increasing bijection onto its range, with strictly increasing inverse $(\psi')^{-1}$.

For $b_k > 0$ we get $\psi'(b_k) = -\mu / w_k$, which admits a solution in $(0, \infty)$ if and only if $-\mu/w_k > \psi'(0)$, i.e. if and only if $w_k > \mu / (-\psi'(0)) \triangleq \theta_\psi$. For $b_k = 0$ we get $\eta_k = \mu + w_k \psi'(0) \ge 0$, i.e. $w_k \le \theta_\psi$. This is the first claim. For the second, $b_k^\star = (\psi')^{-1}(-\mu/w_k)$ on the active set, and $w \mapsto -\mu/w$ is increasing for $\mu > 0$, so $b_k^\star$ is non-decreasing in w_k; pruned coordinates have $b_k^\star = 0$ and the smallest weights, so the monotonicity extends to all k. □

Remark 7. Taking $\psi(b) = 2^{-2b}$ recovers Proposition 2. Proposition 6 shows that the two structural features of the allocation—pruning by a threshold on w_k, and rate ordered as w_k—follow from monotonicity and convexity of the distortion-rate law alone, not from its exponential form. Together with Remark 5 and Remark 6, this is the precise sense in which the high-rate approximation cannot reorder coordinates: it may change how many are kept and how much rate each receives, but the kept ones are always those of largest w_k.

Remark 8 (Degenerate weights). If $\lambda_k = 0$ for some k then $w_k = 0$, the hypothesis $w_k > 0$ fails, and that coordinate contributes $\lambda_k (z_{ik} - \widehat{z}_{ik})^2 = 0$ to equation 8 at any rate; it is pruned outright and the proposition applies verbatim to the coordinates with $w_k > 0$.

Proposition 7 (Feasibility of the tiered allocation)**.** *Let $\mathcal{B} = \{0\} \cup \mathcal{B}^+$ with $\mathcal{B}^+ \subset \mathbb{Z}_{\ge 1}$ finite, and $\lfloor r \rfloor_{\mathcal{B}} = \max\{b \in \mathcal{B} : b \le \max(r, 0)\}$. Define*

$$B_{\mathcal{B}}(\theta) \triangleq \sum_{k=1}^{d} \Big\lfloor \tfrac{1}{2} \log_2 \frac{w_k}{\theta} \Big\rfloor_{\mathcal{B}}. \tag{50}$$

Then $B_{\mathcal{B}}$ is non-increasing on $(0, \infty)$, $B_{\mathcal{B}}(\max_k w_k) = 0$, and the bisection of Algorithm 1 (lines 4–6) maintains the invariant $B_{\mathcal{B}}(\theta_{\max}) \le Rd$ at every iteration. Consequently the allocation $b_k^\star = \lfloor \frac{1}{2} \log_2(w_k/\theta_{\max}) \rfloor_{\mathcal{B}}$ of line 7 satisfies $\frac{1}{d} \sum_k b_k^\star \le R$.

Proof. Well-posedness. $\lfloor r \rfloor_{\mathcal{B}}$ is well defined because $0 \in \mathcal{B}$ and $0 \le \max(r, 0)$, so the set is nonempty, and finite because $\mathcal{B}$ is finite.

Monotonicity. For fixed w_k, the map $\theta \mapsto \frac{1}{2} \log_2(w_k/\theta)$ is strictly decreasing, and $r \mapsto \lfloor r \rfloor_{\mathcal{B}}$ is non-decreasing; the composition is non-increasing, and so is the finite sum $B_{\mathcal{B}}$.

Initialization. At $\theta = \max_k w_k$ we have $\frac{1}{2} \log_2(w_k/\theta) \le 0$ for every k, hence $\max(\cdot, 0) = 0$ and $\lfloor \cdot \rfloor_{\mathcal{B}} = 0$, so $B_{\mathcal{B}}(\max_k w_k) = 0 \le Rd$: the invariant holds before the loop.

Preservation. At each iteration the midpoint θ is tested. If $B_{\mathcal{B}}(\theta) > Rd$ the algorithm sets $\theta_{\min} \leftarrow \theta$ and leaves $\theta_{\max}$ unchanged, so the invariant is preserved; otherwise $B_{\mathcal{B}}(\theta) \le Rd$ and the algorithm sets $\theta_{\max} \leftarrow \theta$, for which the invariant holds by assumption of that branch. By induction $B_{\mathcal{B}}(\theta_{\max}) \le Rd$ after every iteration, and line 7 evaluates exactly $B_{\mathcal{B}}(\theta_{\max})$. □

Remark 9 (Realized rate). Proposition 7 guarantees feasibility, not tightness: since $\lfloor\cdot\rfloor_{\mathcal{B}}$ rounds down and no greedy top-up follows, the realized rate $\frac{1}{d}\sum_k b_k^\star$ may be strictly below R, by at most $\max\mathcal{B}$ bits per coordinate that is rounded. Adding the f bits of n_i (Section 4.4), the bits actually stored per vector correspond to

$$R_{\text{eff}} = \frac{1}{d}\Big(\sum_{k=1}^{d} b_k^\star + f\Big), \tag{51}$$

so R is a budget imposed on the allocation and R_{eff} is the realized rate of the index.

B.2 Tier Isometry

Lemma 3 (Invertibility on the active set)**.** *For every $k \in \mathcal{I}_+$ we have $\lambda_k > 0$, so $\boldsymbol{\Lambda}_b$ is positive definite and $\boldsymbol{\Lambda}_b^{-1/2}$ is well defined for every $b \in \mathcal{B}^+$.*

Proof. By construction $k \in \mathcal{I}_+$ means $b_k^\star \geq 1$, which by equation 42 requires $w_k > \theta > 0$. Since $w_k = \lambda_k \sigma_k^2$ and $\sigma_k^2 \geq 0$, this forces $\lambda_k > 0$. As $\boldsymbol{\Lambda}_b = \operatorname{diag}(\lambda_k)_{k\in\mathcal{I}_b}$ with $\mathcal{I}_b \subseteq \mathcal{I}_+$ for $b \in \mathcal{B}^+$, the matrix is diagonal with strictly positive entries. □

Proposition 8 (Proposition 3 restated)**.** *Let $\widehat{\boldsymbol{y}}_{i,b}$ be the reconstruction of $\boldsymbol{y}_{i,b} = \boldsymbol{R}_b\boldsymbol{u}_{i,b}$ returned by $\mathcal{Q}_b$, and let $\widetilde{\boldsymbol{z}}_i$ have blocks $\boldsymbol{\Lambda}_b^{-1/2}\boldsymbol{R}_b^\top\widehat{\boldsymbol{y}}_{i,b}$ for $b \in \mathcal{B}^+$ and zero on $\mathcal{I}_0$. For every orthogonal $\boldsymbol{R}_b$,*

$$\mathcal{L}\big(\boldsymbol{V}(\boldsymbol{z}_i - \widetilde{\boldsymbol{z}}_i)\big) = \sum_{b\in\mathcal{B}^+} \big\|\boldsymbol{y}_{i,b} - \widehat{\boldsymbol{y}}_{i,b}\big\|_2^2 \;+\; \sum_{k\in\mathcal{I}_0} \lambda_k z_{ik}^2. \tag{52}$$

Proof. Step 1 (reduction to a weighted sum). Applying equation 8 to the residual $\boldsymbol{e}_i = \boldsymbol{V}(\boldsymbol{z}_i - \widetilde{\boldsymbol{z}}_i)$ and using that $\{\mathcal{I}_b\}_{b\in\mathcal{B}}$ partitions $\{1,\dots,d\}$,

$$\mathcal{L}\big(\boldsymbol{V}(\boldsymbol{z}_i - \widetilde{\boldsymbol{z}}_i)\big) = \sum_{k=1}^{d} \lambda_k\big(z_{ik} - \widetilde{z}_{ik}\big)^2 = \sum_{b\in\mathcal{B}^+}\sum_{k\in\mathcal{I}_b} \lambda_k\big(z_{ik} - \widetilde{z}_{ik}\big)^2 + \sum_{k\in\mathcal{I}_0} \lambda_k z_{ik}^2, \tag{53}$$

the last sum because $\widetilde{z}_{ik} = 0$ for $k \in \mathcal{I}_0$ by hypothesis.

Step 2 (absorbing the weights). Fix $b \in \mathcal{B}^+$ and write $\boldsymbol{z}_{i,b}, \widetilde{\boldsymbol{z}}_{i,b}$ for the tier-b blocks. By Lemma 3, $\boldsymbol{\Lambda}_b^{1/2}$ is a diagonal matrix with entries $\lambda_k^{1/2} > 0$, so

$$\sum_{k\in\mathcal{I}_b} \lambda_k\big(z_{ik} - \widetilde{z}_{ik}\big)^2 = \sum_{k\in\mathcal{I}_b} \Big(\lambda_k^{1/2} z_{ik} - \lambda_k^{1/2}\widetilde{z}_{ik}\Big)^2 = \Big\|\boldsymbol{\Lambda}_b^{1/2}\big(\boldsymbol{z}_{i,b} - \widetilde{\boldsymbol{z}}_{i,b}\big)\Big\|_2^2. \tag{54}$$

Step 3 (identifying the two terms). By equation 10, $\boldsymbol{\Lambda}_b^{1/2}\boldsymbol{z}_{i,b} = \boldsymbol{u}_{i,b}$. By the definition of $\widetilde{\boldsymbol{z}}_i$,

$$\boldsymbol{\Lambda}_b^{1/2}\widetilde{\boldsymbol{z}}_{i,b} = \boldsymbol{\Lambda}_b^{1/2}\,\boldsymbol{\Lambda}_b^{-1/2}\boldsymbol{R}_b^\top\widehat{\boldsymbol{y}}_{i,b} = \boldsymbol{R}_b^\top\widehat{\boldsymbol{y}}_{i,b}. \tag{55}$$

Moreover $\boldsymbol{y}_{i,b} = \boldsymbol{R}_b\boldsymbol{u}_{i,b}$ and $\boldsymbol{R}_b$ orthogonal give $\boldsymbol{u}_{i,b} = \boldsymbol{R}_b^\top\boldsymbol{y}_{i,b}$. Hence

$$\boldsymbol{\Lambda}_b^{1/2}\big(\boldsymbol{z}_{i,b} - \widetilde{\boldsymbol{z}}_{i,b}\big) = \boldsymbol{R}_b^\top\boldsymbol{y}_{i,b} - \boldsymbol{R}_b^\top\widehat{\boldsymbol{y}}_{i,b} = \boldsymbol{R}_b^\top\big(\boldsymbol{y}_{i,b} - \widehat{\boldsymbol{y}}_{i,b}\big). \tag{56}$$

Step 4 (isometry). Since $\boldsymbol{R}_b^\top$ is orthogonal it preserves the Euclidean norm, so

$$\Big\|\boldsymbol{R}_b^\top\big(\boldsymbol{y}_{i,b} - \widehat{\boldsymbol{y}}_{i,b}\big)\Big\|_2^2 = \big\|\boldsymbol{y}_{i,b} - \widehat{\boldsymbol{y}}_{i,b}\big\|_2^2. \tag{57}$$

Substituting into Step 1 and summing over $b \in \mathcal{B}^+$ gives equation 52. □

Remark 10 (The order of the two operations). Scaling before rotating is not interchangeable with the reverse. Rotating first replaces the tier block $\boldsymbol{\Lambda}_b$ by $\boldsymbol{R}_b\boldsymbol{\Lambda}_b\boldsymbol{R}_b^\top$, which is diagonal only if $\boldsymbol{\Lambda}_b \propto \boldsymbol{I}_{d_b}$; the subsequent equalization, which can only rescale coordinates, would then leave off-diagonal terms and reinstate inside the tier the cross-talk removed in Section 4.1. In the order used by VENICE the block is $\boldsymbol{I}_{d_b}$ when the rotation is applied, and $\boldsymbol{R}_b\boldsymbol{I}_{d_b}\boldsymbol{R}_b^\top = \boldsymbol{I}_{d_b}$ for every orthogonal $\boldsymbol{R}_b$, which is the benign regime identified in Section 3.3.

Remark 11 (Tier rotations for an arbitrary tier size). Proposition 3 requires only that $\boldsymbol{R}_b$ be orthogonal, whereas a Walsh–Hadamard matrix of order n exists only for restricted n and $d_b = |\mathcal{I}_b|$ is arbitrary. Writing the binary expansion $d_b = \sum_{j=1}^{J} 2^{p_j}$ with distinct exponents p_j, VENICE uses

$$\boldsymbol{R}_b = \mathrm{blockdiag}\big(\boldsymbol{W}_{2^{p_1}} \boldsymbol{D}_{b,1}, \ldots, \boldsymbol{W}_{2^{p_J}} \boldsymbol{D}_{b,J}\big)\, \boldsymbol{\Pi}_b, \tag{58}$$

where $\boldsymbol{\Pi}_b$ is a fixed random permutation of the tier coordinates and the $\boldsymbol{D}_{b,j}$ are independent Rademacher sign matrices. Each diagonal block is orthogonal and $\boldsymbol{\Pi}_b$ is a permutation matrix, hence $\boldsymbol{R}_b$ is orthogonal for every d_b and no padding is needed. Theorem 1 and Theorem 2 then apply within each block, with d replaced by the block size 2^{p_j}, the largest of which satisfies $2^{p_J} > d_b/2$; the permutation prevents the block partition from aligning with the coordinate order.

B.3 NORM RESTORATION

Lemma 4 (Contraction of squared-error decoders). *Let $\boldsymbol{z}$ be a random vector in $\mathbb{R}^n$ with $\mathbb{E}\|\boldsymbol{z}\|_2^2 < \infty$, let $c(\cdot)$ be any measurable encoder, and let the decoder be optimal in squared error, $\widehat{\boldsymbol{z}} = \mathbb{E}\big[\boldsymbol{z} \mid c(\boldsymbol{z})\big]$. Then*

$$\mathbb{E}\big\|\widehat{\boldsymbol{z}}\big\|_2^2 = \mathbb{E}\big\|\boldsymbol{z}\big\|_2^2 - \mathbb{E}\big\|\boldsymbol{z} - \widehat{\boldsymbol{z}}\big\|_2^2 \;\le\; \mathbb{E}\big\|\boldsymbol{z}\big\|_2^2, \tag{59}$$

with equality if and only if $\boldsymbol{z} = \widehat{\boldsymbol{z}}$ almost surely. Moreover, for every code value c, writing $\widehat{\boldsymbol{z}}_c = \mathbb{E}[\boldsymbol{z} \mid c(\boldsymbol{z}) = c]$,

$$\mathbb{E}\big[\|\boldsymbol{z}\|_2^2 \,\big|\, c(\boldsymbol{z}) = c\big] - \big\|\widehat{\boldsymbol{z}}_c\big\|_2^2 = \mathrm{Tr}\,\mathrm{Cov}\,\big[\boldsymbol{z} \,\big|\, c(\boldsymbol{z}) = c\big]. \tag{60}$$

Proof. Set $\boldsymbol{r} \triangleq \boldsymbol{z} - \widehat{\boldsymbol{z}}$, so $\boldsymbol{z} = \widehat{\boldsymbol{z}} + \boldsymbol{r}$. The vector $\widehat{\boldsymbol{z}}$ is $\sigma(c(\boldsymbol{z}))$-measurable by construction, and

$$\mathbb{E}\big[\boldsymbol{r} \mid c(\boldsymbol{z})\big] = \mathbb{E}\big[\boldsymbol{z} \mid c(\boldsymbol{z})\big] - \widehat{\boldsymbol{z}} = \boldsymbol{0}. \tag{61}$$

By the tower property, pulling out the $\sigma(c(\boldsymbol{z}))$-measurable factor,

$$\mathbb{E}\big\langle \widehat{\boldsymbol{z}}, \boldsymbol{r}\big\rangle = \mathbb{E}\Big[\mathbb{E}\big[\langle \widehat{\boldsymbol{z}}, \boldsymbol{r}\rangle \mid c(\boldsymbol{z})\big]\Big] = \mathbb{E}\Big[\big\langle \widehat{\boldsymbol{z}}, \mathbb{E}[\boldsymbol{r} \mid c(\boldsymbol{z})]\big\rangle\Big] = 0, \tag{62}$$

which is the orthogonality principle. Expanding,

$$\mathbb{E}\|\boldsymbol{z}\|_2^2 = \mathbb{E}\|\widehat{\boldsymbol{z}}\|_2^2 + 2\,\mathbb{E}\langle \widehat{\boldsymbol{z}}, \boldsymbol{r}\rangle + \mathbb{E}\|\boldsymbol{r}\|_2^2 = \mathbb{E}\|\widehat{\boldsymbol{z}}\|_2^2 + \mathbb{E}\|\boldsymbol{z} - \widehat{\boldsymbol{z}}\|_2^2, \tag{63}$$

which gives equation 59; the inequality is $\mathbb{E}\|\boldsymbol{z} - \widehat{\boldsymbol{z}}\|_2^2 \ge 0$, an equality if and only if $\boldsymbol{z} = \widehat{\boldsymbol{z}}$ a.s. Conditioning on $\{c(\boldsymbol{z}) = c\}$, where $\widehat{\boldsymbol{z}}_c$ is deterministic, the same expansion gives

$$\mathbb{E}\big[\|\boldsymbol{z}\|_2^2 \mid c\big] = \|\widehat{\boldsymbol{z}}_c\|_2^2 + \mathbb{E}\big[\|\boldsymbol{z} - \widehat{\boldsymbol{z}}_c\|_2^2 \mid c\big], \tag{64}$$

and $\mathbb{E}[\|\boldsymbol{z} - \widehat{\boldsymbol{z}}_c\|_2^2 \mid c] = \mathrm{Tr}\,\mathrm{Cov}[\boldsymbol{z} \mid c]$ by definition of the conditional covariance, since $\widehat{\boldsymbol{z}}_c$ is the conditional mean. □

Remark 12 (Why the correction must be per datapoint). Equation 59 states that the contraction exists; equation 60 states that it is not uniform. The length deficit of a code is its within-cell dispersion $\mathrm{Tr}\,\mathrm{Cov}[\boldsymbol{z} \mid c]$, a quantity that differs from cell to cell, so the ratio $\|\widehat{\boldsymbol{z}}_i\|_2/\|\boldsymbol{z}_i\|_2$ varies across datapoints. This makes precise the claim of Section 3.2 that the shrinkage is heterogeneous: a common factor leaves MIPS orderings unchanged, whereas a factor that varies with i does not. Neither statement requires the encoder to be trained, optimal, or of any particular form; only the decoder is assumed to be the conditional mean, which is the necessary condition for squared-error optimality.

Proposition 9 (Effect of norm restoration). *Let $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ be symmetric positive definite, write $\|\boldsymbol{p}\|_{\boldsymbol{A}}^2 = \boldsymbol{p}^\top \boldsymbol{A}\boldsymbol{p}$ and $\langle \boldsymbol{p}, \boldsymbol{r}\rangle_{\boldsymbol{A}} = \boldsymbol{p}^\top \boldsymbol{A}\boldsymbol{r}$, and let $\boldsymbol{z}, \boldsymbol{v} \in \mathbb{R}^m \setminus \{\boldsymbol{0}\}$. For $t \ge 0$ set $\widehat{\boldsymbol{z}}(t) = t\,\boldsymbol{v}/\|\boldsymbol{v}\|_2$, so that $\widehat{\boldsymbol{z}}(\|\boldsymbol{v}\|_2) = \boldsymbol{v}$. Then, with*

$$s \triangleq \frac{\|\boldsymbol{v}\|_{\boldsymbol{A}}^2}{\|\boldsymbol{v}\|_2^2} > 0, \qquad t^\star \triangleq \frac{\langle \boldsymbol{z}, \boldsymbol{v}\rangle_{\boldsymbol{A}}\, \|\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_{\boldsymbol{A}}^2}, \tag{65}$$

the map $t \mapsto \|\boldsymbol{z} - \widehat{\boldsymbol{z}}(t)\|_{\boldsymbol{A}}^2$ is strictly convex and satisfies

$$\big\|\boldsymbol{z} - \widehat{\boldsymbol{z}}(t)\big\|_{\boldsymbol{A}}^2 = s\,(t - t^\star)^2 + \Big(\|\boldsymbol{z}\|_{\boldsymbol{A}}^2 - s\,(t^\star)^2\Big), \tag{66}$$

so it is uniquely minimized at $t = t^\star$. Consequently, replacing the decoded $\boldsymbol{v}$ by the restored $\widehat{\boldsymbol{z}}(n)$ changes the error by exactly

$$\big\|\boldsymbol{z} - \widehat{\boldsymbol{z}}(n)\big\|_{\boldsymbol{A}}^2 - \big\|\boldsymbol{z} - \boldsymbol{v}\big\|_{\boldsymbol{A}}^2 = s\Big[(n - t^\star)^2 - \big(\|\boldsymbol{v}\|_2 - t^\star\big)^2\Big], \tag{67}$$

which is negative if and only if $|n - t^\star| < \big|\|\boldsymbol{v}\|_2 - t^\star\big|$.

Proof. Since $\|\widehat{\boldsymbol{z}}(t)\|_{\boldsymbol{A}}^2 = t^2\|\boldsymbol{v}\|_{\boldsymbol{A}}^2/\|\boldsymbol{v}\|_2^2 = st^2$ and $\langle \boldsymbol{z}, \widehat{\boldsymbol{z}}(t)\rangle_{\boldsymbol{A}} = t\langle \boldsymbol{z}, \boldsymbol{v}\rangle_{\boldsymbol{A}}/\|\boldsymbol{v}\|_2 = s\,t\,t^\star$ by equation 65,

$$g(t) \triangleq \big\|\boldsymbol{z} - \widehat{\boldsymbol{z}}(t)\big\|_{\boldsymbol{A}}^2 = \|\boldsymbol{z}\|_{\boldsymbol{A}}^2 - 2s\,t\,t^\star + s\,t^2, \tag{68}$$

a quadratic with $g''(t) = 2s > 0$. Completing the square gives equation 66, hence strict convexity and the unique minimizer $t^\star$. Evaluating equation 66 at $t = n$ and at $t = \|\boldsymbol{v}\|_2$, where $\widehat{\boldsymbol{z}}(\|\boldsymbol{v}\|_2) = \boldsymbol{v}$, and subtracting, the constant term cancels and equation 67 follows. The sign condition is immediate since $s > 0$. □

Remark 13 (Norm restoration is not an isometry). Proposition 3 accounts for $\widetilde{\boldsymbol{z}}_i$, the reconstruction *before* Section 4.4; the deployed reconstruction is $\widehat{\boldsymbol{z}}_i$ of equation 12. Applying Proposition 9 on the active block with $\boldsymbol{A} = \boldsymbol{\Lambda}_{\mathcal{I}_+}$, $\boldsymbol{z} = \boldsymbol{z}_{i,\mathcal{I}_+}$, $\boldsymbol{v} = \widetilde{\boldsymbol{z}}_{i,\mathcal{I}_+}$ and $n = n_i$, the first term of equation 52 changes by exactly equation 67: the step lowers the retrieval loss if and only if the restored length is closer than the decoded one to $t^\star$, the length that minimizes the $\boldsymbol{\Lambda}$-weighted error along the decoded direction. It is therefore not an isometry, and not error-optimal either, since $n_i = \|\boldsymbol{z}_{i,\mathcal{I}_+}\|_2 \neq t^\star$ in general. This is deliberate: equation 1 is a surrogate for what retrieval actually scores, the ranking induced by $\langle \boldsymbol{q}, \widehat{\boldsymbol{x}}_i\rangle$, and by equation 60 the length deficit varies across datapoints and therefore perturbs that ranking, whereas a uniform factor would not.

B.4 Additional Remarks

Remark 14 (Diagonalizing $\boldsymbol{H}$ versus diagonalizing $\boldsymbol{\Sigma}$). With $\boldsymbol{H} = \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{\Sigma} = \boldsymbol{H} - \boldsymbol{\mu}\boldsymbol{\mu}^\top$, the eigenbasis $\boldsymbol{V}$ of $\boldsymbol{H}$ gives $\boldsymbol{V}^\top\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{\Lambda} - \boldsymbol{\nu}\boldsymbol{\nu}^\top$, whose off-diagonal entries are $-\nu_j\nu_k$; this vanishes for all $j \neq k$ if and only if $\boldsymbol{\nu}$ has at most one nonzero entry, i.e. if and only if $\boldsymbol{\mu}$ is an eigenvector of $\boldsymbol{H}$. Diagonalizing $\boldsymbol{H}$ therefore leaves a rank-one correlation in the data, and the two matrices admit a common eigenbasis only in that degenerate case. VENICE diagonalizes $\boldsymbol{H}$ because $\boldsymbol{H}$ is the matrix that equation 1 charges: equation 8 is an identity for every residual, whatever the residual correlation of the data, whereas the eigenbasis of $\boldsymbol{\Sigma}$ would decorrelate the data while leaving $\mathrm{od}(\boldsymbol{V}_{\boldsymbol{\Sigma}}^\top\boldsymbol{H}\boldsymbol{V}_{\boldsymbol{\Sigma}}) \neq 0$, hence a cross-talk term no separable objective can charge.

C Experimental Details

C.1 Benchmark Datasets

Summary specifications and domain characteristics for the five datasets used in Section 5 are listed below:

- **GloVe** (d=100, 1M vectors) (Pennington et al., 2014), comprising pre-trained dense word representations.
- **BigANN** (d=128, 100M vectors) (Jegou et al., 2011), containing unnormalized SIFT visual descriptors.
- **GIST-960** (d=960, 1M vectors) (Jegou et al., 2011), containing holistic GIST visual scene descriptors.
- **OpenAI** (d=1536, 2M vectors) (Neelakantan et al., 2022; Simhadri et al., 2022), consisting of dense sentence-level text embeddings.
- **Wiki** (d=3072, 2M vectors) (Wikimedia Foundation, 2026; Gemini Team, Google, 2026), capturing Gemini document representations of Wikipedia articles.

C.2 Base Quantization Codecs

Below we summarize the encoding mechanism and design characteristics of the six base codecs wrapped by Venice:

- **EDEN** (Vargaftik et al., 2022) combines Hadamard preconditioning with normal-calibrated Lloyd–Max scalar quantization and analytical bias correction.
- **RaBitQ** (Extended-RaBitQ) (Gao & Long, 2024; Gao et al., 2025) applies a randomized Haar rotation followed by scale-invariant sign-binarization (and its multi-bit grid extension), backed by concentration bounds.
- **TurboQuant** (Zandieh et al., 2026) employs a randomized QR rotation paired with Beta-marginal Lloyd–Max quantization and unbiased QJL estimation.
- **RSLM** (Lenhardt et al., 2026) uses a linear-time Block-FWHT coupled with vector-adaptive norm correction and Gauss-calibrated Lloyd–Max quantization.
- **PQ** (Jegou et al., 2011) performs a Cartesian subspace decomposition into 2-dimensional blocks encoded via data-dependent k-means codebooks.
- **CAQ** (Li et al., 2025) refines codes via linear-time coordinate descent, performing directional cosine maximization and scale-invariant code adjustment.

D Additional Experimental Results

This appendix reports supplementary results that complement those presented in Section 5.2: the two tighter rate regimes at Recall@10@10, the full sweep at Recall@1@1, and the leave-one-out ablation across all three rates.

D.1 Recall@10@10 at 2 and 1 Bits per Dimension

Figure 3 extends Figure 1 to 2 and 1 bits per dimension. Every codec improves on every dataset at both rates, GloVe included: the inert regime of Section 5.2 is confined to 4 bits, and once the budget is tight enough to make the allocation non-uniform the basis starts to help on GloVe as well.

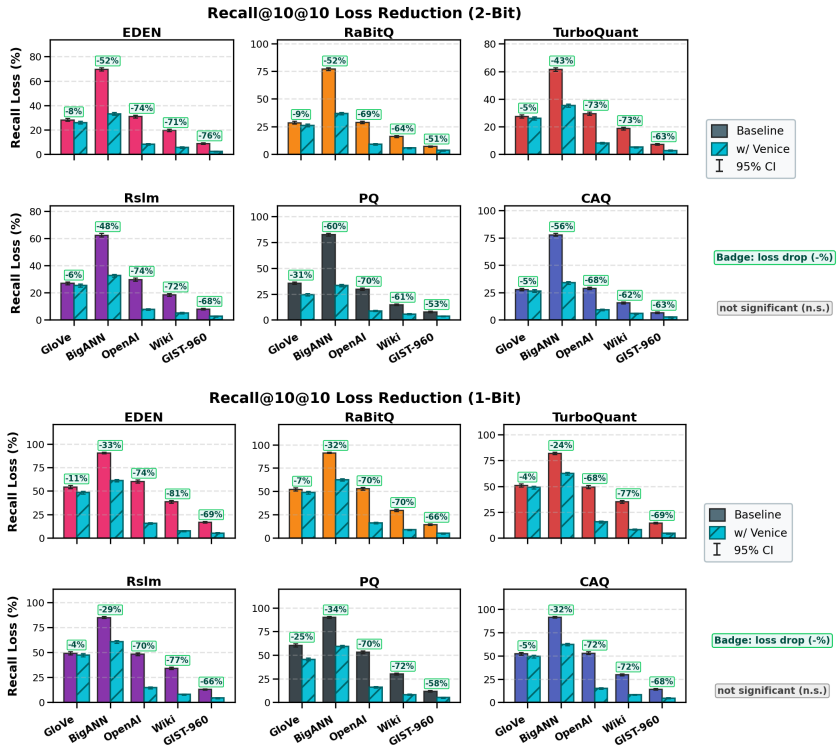

Figure 3: Recall loss ($1 - \text{Recall@10@10}$) at 2 bits (top) and 1 bit (bottom) per dimension, with and without Venice. Badges as in Figure 1.

D.2 Recall@1@1 at 4, 2, and 1 Bits per Dimension

Recall@1@1 requires the exact nearest neighbour to be returned in first position and is the stricter of the two metrics. Figure 4 repeats the full sweep under it. The whole picture is matching the conclusions made for Recall@10@10 in the main text.

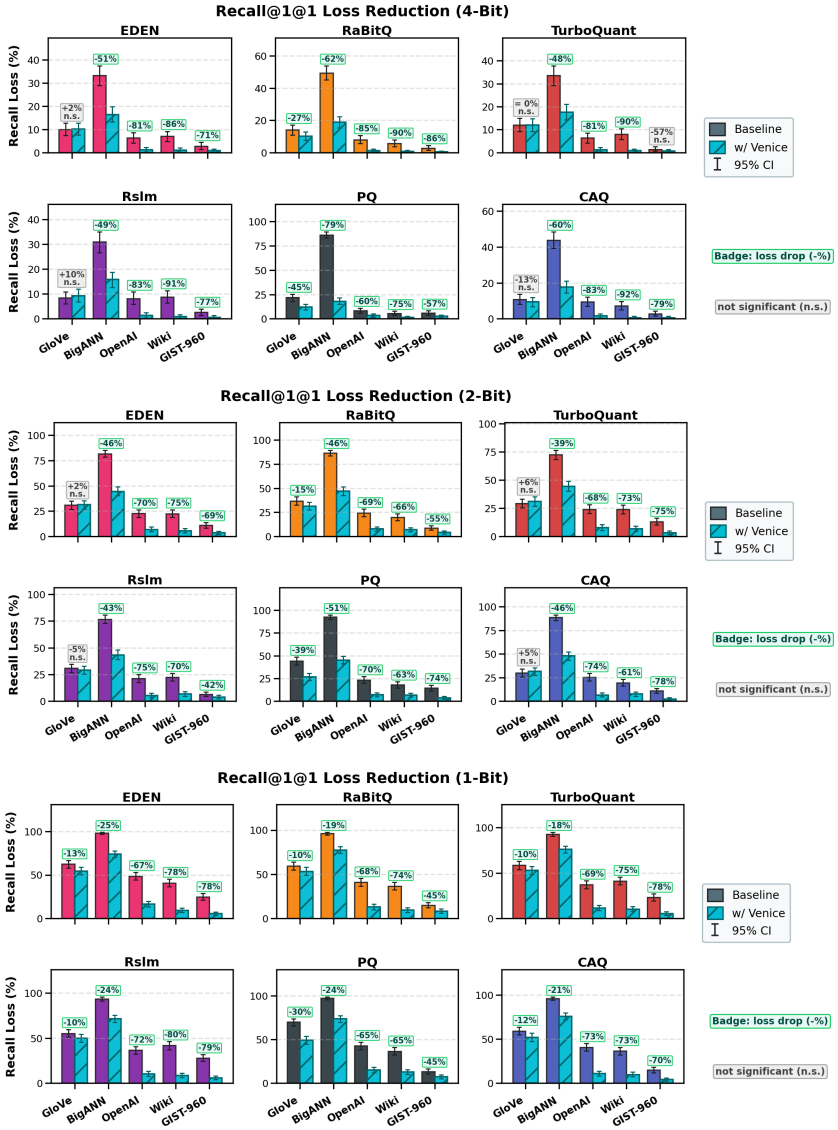

Figure 4: Recall loss ($1 - \text{Recall@1@1}$) at 4, 2 and 1 bits per dimension (top to bottom), with and without VENICE. Badges as in Figure 1.

D.3 LEAVE-ONE-OUT PILLAR ABLATION

Figures 5 and 6 removes one pillar at a time from VENICE applied to PQ and RABITQ, respectively. They show results referenced from the main text in Section 5.2. RABITQ results are representative of similar results for other data-independent codecs we evaluated—as they do both rotation and norm restoration.

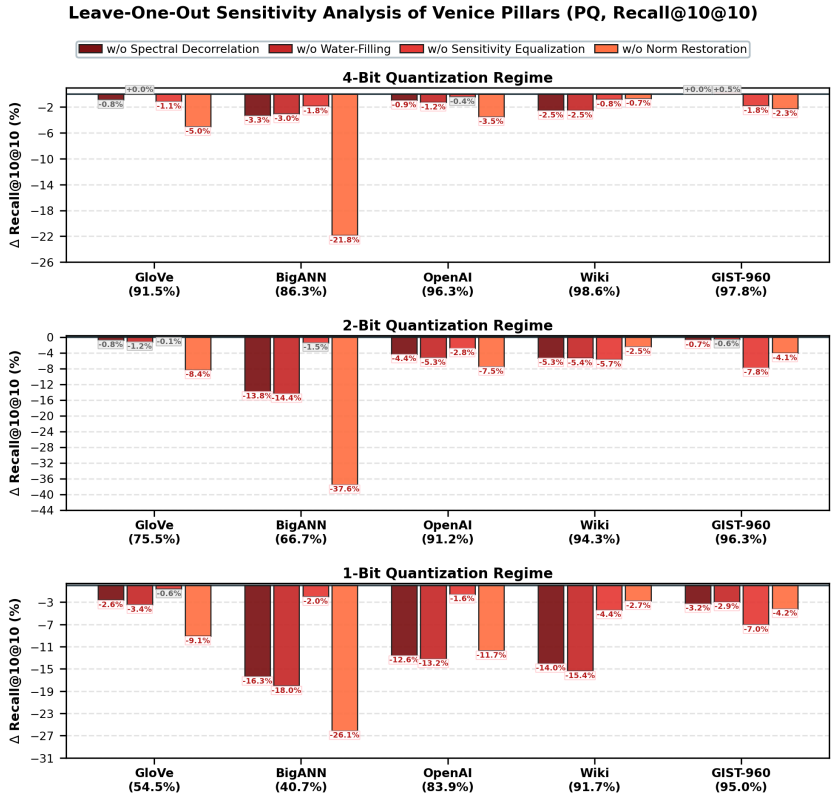

Figure 5: Leave-one-out ablation of the VENICE pillars on PQ: change in Recall@10@10 when each pillar is removed, at 4, 2 and 1 bits per dimension. In parentheses, the recall of the full pipeline; grey: not significant.

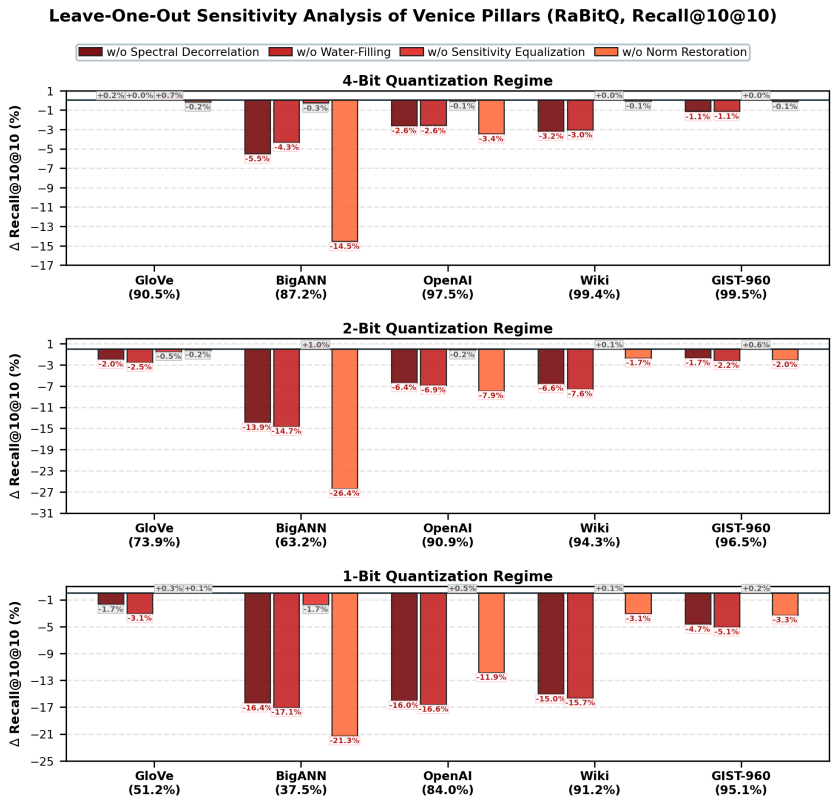

Figure 6: Leave-one-out ablation of the VENICE pillars on RABITQ: change in Recall@10@10 when each pillar is removed, at 4, 2 and 1 bits per dimension. In parentheses, the recall of the full pipeline; grey: not significant.