Block-Sphere Vector Quantization

Heesang Ann*
Seoul National University
sang3798@snu.ac.kr

Joongkyu Lee*
Seoul National University
jklee0717@snu.ac.kr

Min-hwan Oh
Seoul National University
minoh@snu.ac.kr

Abstract

Vector quantization is a fundamental primitive for scalable machine learning systems, enabling memory-efficient storage, fast retrieval, and compressed inference. Recent rotation-based quantizers such as `EDEN`, `RabitQ`, and `TurboQuant` have introduced strong guarantees and empirical performance, but the surrounding comparisons have been difficult to interpret because they rely on different distortion criteria, probability regimes, and implementation assumptions. As our first contribution, we provide a unified theoretical comparison of these methods and show that their relative advantages are criterion-dependent rather than absolute: `TurboQuant` and `EDEN` are favorable for MSE distortion, `EDEN` is also effective for expected inner-product distortion, and `RabitQ` provides strong high-probability control. As our second contribution, we introduce *Block-Sphere Quantization* (`BlockQuant`), a rotation-based block quantization algorithm designed around the spherical geometry of randomly rotated vectors. Unlike coordinate-wise quantizers, `BlockQuant` quantizes blocks on the sphere, preserving the geometry of rotated embeddings more faithfully. We prove that this block-spherical design theoretically improves over the baselines considered in this paper for both reconstruction MSE and expected inner-product distortion. Our experiments on real embedding datasets and long-context LLM inference tasks show practical gains that are consistent with our theoretical improvements.

1 Introduction

Vector quantization addresses a central bottleneck in large-scale machine learning systems: storing, transmitting, and comparing massive collections of high-dimensional vectors. In many settings, quantization reduces storage cost and memory traffic while enabling efficient similarity computation. It is becoming increasingly important in LLM inference, where KV-cache memory grows with batch size and context length and can become a major bottleneck [12, 14].

Recent calibration-free quantizers using random rotation have attracted significant attention, showing strong performance for KV-cache quantization even in the low-bit regime. `TurboQuant` explicitly targets both reconstruction MSE and inner-product distortion, and reports strong performance in simple KV-cache quantization and nearest-neighbor tasks [26]. `RabitQ` and its extensions are also rotation-based quantizers, but use a different spherical approximation; they have recently been explored for both approximate nearest-neighbor search and KV-cache quantization [6, 7, 24, 8]. In parallel, `EDEN`, which also uses random rotation and coordinate-wise marginal distributions, has been revisited in relation to `TurboQuant` [21, 22, 4]. A major strength of these methods is that they come with theoretical guarantees. However, these guarantees are stated under different criteria: `TurboQuant` and `EDEN` are primarily analyzed through expected distortion measures, whereas `RabitQ` is analyzed through high-probability bit-complexity guarantees. This makes a direct theoretical comparison difficult. Moreover, some recently discussed variants, such as inner-product variants of `EDEN` and reconstruction variants of `RabitQ`, have mostly been compared empirically [4, 8].

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of rotation-based quantizers. The suffixes BSM and UB denote the best-scalar MSE variant and the unbiased inner-product variant, respectively; MSE and PROD follow the original `TurboQuant` notation. The entries for $b = 1, 2, 3, 4$ report high-dimensional approximate values. The large-b column reports the high-rate asymptotic upper bounds of the form $C \cdot 4^{-b}$. All values are computed in this work except †, reported in Zandieh et al. [26].

Quantizer for reconstruction	MSE distortion				
	$b=1$	$b=2$	$b=3$	$b=4$	Large b
$\texttt{EDEN}_{\text{BSM}}$	0.363	0.117	0.0345	0.0095	$2.721 \cdot 4^{-b}$
$\texttt{RabitQ}_{\text{BSM}}$	0.363	0.119	0.0374	0.0115	–
$\texttt{TurboQuant}_{\text{MSE}}$	$0.36^\dagger$	$0.117^\dagger$	$0.03^\dagger$	$0.009^\dagger$	$2.721 \cdot 4^{-b\,\dagger}$
$\texttt{BlockQuant}_{\text{BSM}}(p=2)$(ours)	0.363	0.108	0.0297	0.0078	$2.015 \cdot 4^{-b}$
$\texttt{BlockQuant}_{\text{BSM}}(p=3)$(ours)	**0.357**	**0.101**	**0.0271**	**0.0071**	$\mathbf{1.770} \cdot 4^{-b}$
Quantizer for inner product	**Inner product distortion**				
	$b=1$	$b=2$	$b=3$	$b=4$	Large b
$\texttt{EDEN}_{\text{UB}}$	$\frac{0.571}{d-1}$	$\frac{0.133}{d-1}$	$\frac{0.0358}{d-1}$	$\frac{0.0096}{d-1}$	$\frac{2.721}{d-1} \cdot 4^{-b}$
$\texttt{RabitQ}_{\text{UB}}$	$\frac{0.571}{d-1}$	$\frac{0.135}{d-1}$	$\frac{0.0389}{d-1}$	$\frac{0.0117}{d-1}$	–
$\texttt{TurboQuant}_{\text{PROD}}$	$\frac{1.57}{d}^\dagger$	$\frac{0.56}{d}^\dagger$	$\frac{0.18}{d}^\dagger$	$\frac{0.047}{d}^\dagger$	$\frac{17.09}{d} \cdot 4^{-b\,\dagger}$
$\texttt{BlockQuant}_{\text{UB}}$ $(p=2)$ (ours)	$\frac{0.571}{d-1}$	$\frac{0.120}{d-1}$	$\frac{0.0306}{d-1}$	$\frac{0.0078}{d-1}$	$\frac{2.015}{d-1} \cdot 4^{-b}$
$\texttt{BlockQuant}_{\text{UB}}$ $(p=3)$ (ours)	$\frac{\mathbf{0.553}}{d-1}$	$\frac{\mathbf{0.113}}{d-1}$	$\frac{\mathbf{0.0279}}{d-1}$	$\frac{\mathbf{0.071}}{d-1}$	$\frac{\mathbf{1.770}}{d-1} \cdot 4^{-b}$

In this paper, we close this gap by placing `EDEN`, `RabitQ`, and `TurboQuant` under a unified theoretical framework. We compare these methods under three criteria: reconstruction MSE, expected inner-product distortion, and high-probability bit complexity. This comparison reveals that each algorithm reflects a different design principle: `EDEN` and `TurboQuant` are strong for reconstruction MSE, `EDEN` is also effective for expected inner-product distortion, and `RabitQ` provides the strongest high-probability control.

Motivated by these findings, we propose Block-Sphere Quantization (`BlockQuant`), a rotation-based block quantizer designed to improve expected distortion. Whereas `EDEN` and `TurboQuant` use coordinate-wise marginal distributions after random rotation, `BlockQuant` groups coordinates into small blocks and uses the exact block marginal distribution induced by the unit sphere to construct its centroids. As a result, `BlockQuant` achieves the best reconstruction MSE and expected inner-product error among the rotation-based methods. Our lower-bound analysis further shows that increasing the block size moves this framework toward the ideal spherical quantization limit, and our empirical validation supports the theoretical findings. Our main contributions are summarized as follows:

- **A unified comparison of `EDEN`, `RabitQ`, and `TurboQuant`.** We provide a unified theoretical comparison of `EDEN`, `RabitQ`, and `TurboQuant` under three criteria: reconstruction MSE, expected inner-product distortion, and high-probability bit-complexity guarantees. Our analysis shows that the relative strengths of these methods depend on the criterion: `EDEN` yields strong guarantees for expectation-based distortion criteria, while `RabitQ` provides stronger high-probability guarantees.
- **Block-Sphere Quantization.** We propose Block-Sphere Quantization (`BlockQuant`), a rotation-based block quantizer that exploits the spherical geometry of randomly rotated unit vectors. Unlike coordinate-wise quantizers, `BlockQuant` quantizes low-dimensional blocks using the exact block marginal distribution induced by the unit sphere. We prove that `BlockQuant` improve both the reconstruction MSE and the expected inner-product distortion over the baselines considered in this paper, respectively (see Table 1).
- **A sharper lower bound.** We revisit the Shannon lower bound for the MSE distortion, and we obtain a corrected lower bound for our problem setting. We further show that the idealized version of `BlockQuant` with the maximum block size has an MSE upper bound that closely matches this lower bound, both in order and in the leading constant.

This indicates that incorporating block-spherical structure is a principled route toward near-optimal MSE distortion.

- **Empirical validation.** We show that `BlockQuant` achieves lower reconstruction and inner-product distortion than `EDEN`, `RabitQ`, and `TurboQuant` on real embedding data. We further demonstrate that `BlockQuant` yields practical gains for KV-cache quantization in long-context LLM inference, supporting our theoretical findings.

Related work. Classical vector quantization builds on Lloyd–Max optimization [16, 15] and high-rate theory [10, 25], with product quantization extending these ideas to blockwise compression [11]. Recently, quantization has gained increasing attention in efficient LLM inference, particularly for KV-cache compression [14, 27]. Among calibration-free methods, rotation-based quantizers, including `DRIVE/EDEN` [21, 22], `RabitQ` [6, 7], and `TurboQuant` [26], exploit the spherical geometry induced by random rotations, offering strong empirical performance and theoretical guarantees without dataset-trained codebooks. Building on recent comparisons [4, 8], we unify their theoretical guarantees and introduce `BlockQuant`, which extends coordinate-wise quantization by exploiting the exact block-marginal distribution of randomly rotated unit vectors. See Appendix A for a more detailed discussion of related work.

2 Preliminaries

Notations. We denote the unit sphere and the unit ball in $\mathbb{R}^d$ by $\mathbb{S}^{d-1}$ and $\mathbb{B}^d$, respectively. We denote the beta function by Beta and the gamma and digamma functions by Γ, respectively.

2.1 Problem Settings

We consider a randomized quantizer $\mathcal{Q}$, with quantization map $Q : \mathbb{R}^d \to \{0,1\}^{b\cdot d}$, which maps a d-dimensional vector to a $b \cdot d$-bit string. We denote the corresponding dequantization map by $Q^{-1} : \{0,1\}^{b\cdot d} \to \mathbb{R}^d$. The randomness of Q may arise, for example, from random rotations or random projection matrices. Since vector norms can be stored separately, we focus on the quantization of unit vectors $\mathbf{x} \in \mathbb{S}^{d-1}$.

Rotation-based quantizer. We focus on quantizers that employ a random rotation, including `EDEN`, `RabitQ`, and `TurboQuant`. Let $\mathbf{x} \in \mathbb{S}^{d-1}$ be a unit vector and let R be a Haar-distributed orthogonal matrix. Then the rotated vector $R\mathbf{x}$ is uniformly distributed on $\mathbb{S}^{d-1}$. Rotation-based quantizers exploit this fact by first applying the random rotation and then encoding the rotated vector using a fixed codebook $\mathcal{C}$. More formally, let P_{code} denote the nearest-codeword map associated with $\mathcal{C}$, and let P_{decode} denote the corresponding decoder. A rotation-based quantizer encodes $\mathbf{x}$ through $P_{\mathrm{code}}(R\mathbf{x})$. The raw reconstruction in the original coordinate system is $\bar{\mathbf{x}} = R^\top P_{\mathrm{decode}}\big(P_{\mathrm{code}}(R\mathbf{x})\big)$ where $R^\top = R^{-1}$. The final dequantized output need not be exactly $\bar{\mathbf{x}}$; depending on the target objective, it may apply an additional rescaling or correction to $\bar{\mathbf{x}}$.

All `EDEN`, `RabitQ`, and `TurboQuant` fit within this rotation-based framework, differing in their choices of codebooks and dequantization rules. For `EDEN` and `RabitQ`, we use the suffixes BSM and UB to denote the best-scalar-MSE and unbiased variants, respectively [4, 8]. For `TurboQuant`, we use the suffixes MSE and PROD to denote the raw reconstruction and unbiased inner-product estimation variants, respectively, following Zandieh et al. [26]. The differences among these quantizers are summarized in Table 2, with detailed descriptions provided in Appendix B.

Expected distortion metrics. For a randomized quantizer $\mathcal{Q}$, the worst-case MSE of the reconstructed vector is defined as

$$\mathcal{D}_{\mathrm{MSE}}(\mathcal{Q}) := \max_{\mathbf{x}\in\mathbb{S}^{d-1}} \mathbb{E}_Q \left[\left\| \mathbf{x} - Q^{-1}(Q(\mathbf{x})) \right\|_2^2 \right] .$$

While MSE distortion captures reconstruction quality, inner products are the fundamental quantities in similarity search and retrieval. Thus, we also define the worst-case mean squared error of inner-product estimation as

$$\mathcal{D}_{\mathrm{IP}}(\mathcal{Q}) := \max_{\mathbf{x}\in\mathbb{S}^{d-1}} \max_{\mathbf{y}\in\mathbb{S}^{d-1}} \mathbb{E}_Q \left[\left\{ \langle \mathbf{y}, Q^{-1}(Q(\mathbf{x})) \rangle - \langle \mathbf{y}, \mathbf{x} \rangle \right\}^2 \right] .$$

Table 2: Comparison of quantization and dequantization schemes. $\bar{\mathbf{x}}$ is the raw reconstruction, $\rho = \langle \mathbf{x}, \bar{\mathbf{x}} \rangle$, and $\hat{\mathbf{x}}$ is the final dequantized vector. † $\texttt{TurboQuant}_{\text{PROD}}$ directly estimates $\langle \mathbf{y}, \mathbf{x} \rangle$ as $\langle \mathbf{y}, \bar{\mathbf{x}}' \rangle + \text{IP}_{\text{QJL}}(\mathbf{y}, \mathbf{x} - \bar{\mathbf{x}}')$, where $\bar{\mathbf{x}}'$ is the raw reconstruction obtained using $(b-1)$-bit $\texttt{TurboQuant}_{\text{MSE}}$, and IP_{QJL} denotes the inner-product estimator based on the QJL algorithm [27].

Method	Codebook	Mode	Dequantization rule
EDEN	Scalar Lloyd–Max centroids optimized for $N(0,1)$, scaled by $1/\sqrt{d}$	$\texttt{EDEN}_{\text{BSM}}$	$\hat{\mathbf{x}} = \rho\bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|_2^2$
		$\texttt{EDEN}_{\text{UB}}$	$\hat{\mathbf{x}} = \bar{\mathbf{x}}/\rho$
RabitQ	Spherical codewords obtained by normalizing the uniform-grid codewords	$\texttt{RabitQ}_{\text{BSM}}$	$\hat{\mathbf{x}} = \rho\bar{\mathbf{x}}$
		$\texttt{RabitQ}_{\text{UB}}$	$\hat{\mathbf{x}} = \bar{\mathbf{x}}/\rho$
TurboQuant	Scalar Lloyd–Max centroids optimized for the exact coordinate marginal of $\text{Unif}(S^{d-1})$	$\texttt{TurboQuant}_{\text{MSE}}$	$\hat{\mathbf{x}} = \bar{\mathbf{x}}$
		$\texttt{TurboQuant}_{\text{PROD}}$	Inner product recovery rule†
BlockQuant	2^{bp} centroids jointly optimized for the block marginal $f_{p,d}$ induced by $\text{Unif}(S^{d-1})$	$\texttt{BlockQuant}_{\text{MSE}}$	$\hat{\mathbf{x}} = \bar{\mathbf{x}}$
		$\texttt{BlockQuant}_{\text{BSM}}$	$\hat{\mathbf{x}} = \rho\bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|_2^2$
		$\texttt{BlockQuant}_{\text{UB}}$	$\hat{\mathbf{x}} = \bar{\mathbf{x}}/\rho$

Our first goal is to compare the expected distortion metrics $\mathcal{D}_{\text{MSE}}$ and $\mathcal{D}_{\text{IP}}$ for EDEN, RabitQ, and TurboQuant within a common framework. For rotation-based quantizers, the expectation $\mathbb{E}_Q$ appearing in these metrics is taken with respect to the random rotation matrix. Beyond these expectation-based criteria, we also compare their high-probability guarantees for quantization error [8]. Motivated by this comparison, we then design efficient quantizers that improve both $\mathcal{D}_{\text{MSE}}$ and $\mathcal{D}_{\text{IP}}$.

3 New Theoretical Guarantees for Comparing Existing Quantizers

In this section, we compare the theoretical guarantees of EDEN, RabitQ, and TurboQuant under a unified framework. Existing results, summarized in Appendix B.1, focus on different criteria. Specifically, Vargaftik et al. [22] provides an $\mathcal{D}_{\text{MSE}}$ bound for $\texttt{EDEN}_{\text{UB}}$ but does not analyze $\mathcal{D}_{\text{IP}}$ or high-probability behavior. RabitQ establishes a high-probability bit-complexity guarantee [7] but does not provide expected-distortion bounds for $\mathcal{D}_{\text{MSE}}$ or $\mathcal{D}_{\text{IP}}$. TurboQuant provides bounds for both $\mathcal{D}_{\text{MSE}}$ and $\mathcal{D}_{\text{IP}}$ but lacks high-probability analysis. We address these gaps to enable a unified theoretical comparison. We present the expected-distortion analysis in the main text, while the high-probability bit-complexity comparison is provided in Appendix F.

Overall, in high dimension, $\texttt{EDEN}_{\text{BSM}}$ and $\texttt{TurboQuant}_{\text{MSE}}$ show the better $\mathcal{D}_{\text{MSE}}$ approximates than RabitQ, whereas $\texttt{EDEN}_{\text{UB}}$ performs best for $\mathcal{D}_{\text{IP}}$. Based on this observation, in Section 4, we propose quantizers that strictly improve upon existing methods under both criteria.

3.1 MSE Comparison

To compare reconstruction performance, we report either numerical approximations or upper bounds for the $\mathcal{D}_{\text{MSE}}$ of $\texttt{EDEN}_{\text{BSM}}$, $\texttt{RabitQ}_{\text{BSM}}$, and $\texttt{TurboQuant}_{\text{MSE}}$. We begin with $\texttt{EDEN}_{\text{BSM}}$: for small bit-widths $b \in \{1, 2, 3, 4\}$, we report numerical approximations, whereas for the large-b regime we use a high-rate upper bound.

Proposition 1 (MSE of $\texttt{EDEN}_{\text{BSM}}$)**.** *In high dimensions, the MSE of* $\texttt{EDEN}_{\text{BSM}}$ *is approximately* $\mathcal{D}_{MSE}(\texttt{EDEN}_{\text{BSM}}) \approx \mathbf{0.363},\ \mathbf{0.117},\ \mathbf{0.0345},\ \mathbf{0.0095}$ *for* $b = 1,\ 2,\ 3,\ 4$*, respectively. Moreover, for large* b*,* $\texttt{EDEN}_{\text{BSM}}$ *satisfies* $\mathcal{D}_{MSE}(\texttt{EDEN}_{\text{BSM}}) \leq \mathbf{2.721} \cdot \frac{1}{4^b}$.

Proposition 2 (MSE of $\texttt{RabitQ}_{\text{BSM}}$)**.** *Let* $\mathbf{z} = (z_1, \ldots, z_d)$ *be randomly rotated vector of input unit vector and* $R_j := \sqrt{d}\, z_j$ *be the rescaled coordinates of* $\mathbf{z}$ *for* $j = 1, \ldots, d$*. Define the function* $Q_b(u) = \text{sgn}(u) \min\left(\lfloor |u| \rfloor + \frac{1}{2}, 2^{b-1} - \frac{1}{2}\right)$*, where the value at* $u = 0$ *is irrelevant for continuous distributions. Then the MSE of* $\texttt{RabitQ}_{\text{BSM}}$ *is*

$$\mathcal{D}_{MSE}(\texttt{RabitQ}_{\text{BSM}}) = \mathbb{E}\left[\min_{\alpha > 0} \frac{1}{d} \sum_{j=1}^{d} \left(R_j - \alpha Q_b(R_j/\alpha)\right)^2\right].$$

Moreover, for large d, if we apply the gaussian approximations, the MSE bound is approximately $\mathcal{D}_{MSE}(\texttt{RabitQ}_{\text{BSM}}) \approx \mathbf{0.363},\ \mathbf{0.119},\ \mathbf{0.037},\ \mathbf{0.0115}$ for $b = 1,\ 2,\ 3,\ 4$.

Remark 1 (MSE of $\texttt{TurboQuant}_{\text{MSE}}$)**.** *For* TurboQuant*, since the reported numerical precision of Zandieh et al. [26] differs from the one used in our comparison, we recompute the approximate MSE values for $b = 1,\ 2,\ 3,\ 4$, under our notation : For large d, the MSE of* TurboQuant *is approximately $\mathcal{D}_{MSE}(\texttt{TurboQuant}_{\text{MSE}}) \approx \mathbf{0.363},\ \mathbf{0.117},\ \mathbf{0.0345},\ \mathbf{0.0095}$ for $b = 1,\ 2,\ 3,\ 4$, respectively.*

Discussion on MSE guarantees. From Propositions 1 and 2, together with Remark 1, in high dimensions, both $\texttt{EDEN}_{\text{BSM}}$ and $\texttt{TurboQuant}_{\text{MSE}}$ achieve smaller approximate MSE values than RabitQ across $b = 1,\ 2,\ 3,\ 4$. The approximated MSE values of $\texttt{RabitQ}_{\text{BSM}}$ are also comparable, but become slightly larger than those of $\texttt{TurboQuant}_{\text{MSE}}$ and $\texttt{EDEN}_{\text{BSM}}$ as b increases. Moreover, in the high-rate regime, $\texttt{EDEN}_{\text{BSM}}$ admits an MSE upper bound with the same leading constant as that of $\texttt{TurboQuant}_{\text{MSE}}$. Overall, these results suggest that the Lloyd–Max centroids used in TurboQuant and EDEN, which are optimized for coordinate-wise MSE distortion, are effective in reducing $\mathcal{D}_{\text{MSE}}$. The proofs of Proposition 1, Proposition 2 and explanation of Remark 1 are deferred to Appendix D.1, D.2, and D.3, respectively.

3.2 Inner Product Distortion Comparison

In this section, we derive expected inner product estimation error for $\texttt{EDEN}_{\text{UB}}$ and $\texttt{RabitQ}_{\text{UB}}$, then compare the bound with that of $\texttt{TurboQuant}_{\text{PROD}}$. Different from $\texttt{TurboQuant}_{\text{PROD}}$ which uses additional QJL algorithm for unbiased estimation, $\texttt{EDEN}_{\text{UB}}$ and $\texttt{RabitQ}_{\text{UB}}$ apply scalar multiplication to raw reconstruction, i.e. multiplying $\frac{1}{\langle \mathbf{x}, \bar{\mathbf{x}} \rangle}$ to $\bar{\mathbf{x}}$. In this paper, we call this type of quantizer by (rotation-based) ratio quantizer for inner product. The below theorem shows the general bound on inner product distortion of ratio quantizers.

Theorem 1 (General bound on rotation-based ratio quantizer)**.** *Let $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ and $\eta := \langle \mathbf{x}, \mathbf{y} \rangle$. Define the ratio-based inner-product estimator as $\widehat{\eta}_{\text{ratio}} := \left\langle \frac{\bar{\mathbf{x}}}{\langle \bar{\mathbf{x}}, \mathbf{x} \rangle}, \mathbf{y} \right\rangle$. Then, its expected squared error can be expressed as*

$$\mathbb{E}[(\widehat{\eta}_{\text{ratio}} - \eta)^2] = \frac{1 - \eta^2}{d - 1} \mathbb{E}\left[\frac{\|\bar{\mathbf{x}}\|_2^2 - \langle \bar{\mathbf{x}}, \mathbf{x} \rangle^2}{\langle \bar{\mathbf{x}}, \mathbf{x} \rangle^2} \right].$$

Then, we can get the inner product distortion guarantee of EDEN and RabitQ.

Corollary 1 (Expected inner product error of $\texttt{EDEN}_{\text{UB}}$)**.** *In high dimension, the inner-product error of $\texttt{EDEN}_{\text{UB}}$ is approximately $\mathcal{D}_{IP}(\texttt{EDEN}_{\text{UB}}) \leq \frac{\mathbf{0.571}}{d-1},\ \frac{\mathbf{0.133}}{d-1},\ \frac{\mathbf{0.0358}}{d-1},\ \frac{\mathbf{0.0096}}{d-1}$, for $b = 1,\ 2,\ 3,\ 4$, respectively. Moreover, for large b, $\texttt{EDEN}_{\text{UB}}$ satisfies*

$$\mathcal{D}_{IP}(\texttt{EDEN}_{\text{UB}}) \leq \frac{\mathbf{2.721}}{d - 1} 4^{-b}(1 + o(1)).$$

Corollary 2 (Expected inner product error of $\texttt{RabitQ}_{\text{UB}}$)**.** *In high dimension, the inner product error of $\texttt{RabitQ}_{\text{UB}}$ is approximately $\mathcal{D}_{IP}(\texttt{RabitQ}_{\text{UB}}) \approx \frac{\mathbf{0.571}}{d-1},\ \frac{\mathbf{0.135}}{d-1},\ \frac{\mathbf{0.0389}}{d-1},\ \frac{\mathbf{0.0117}}{d-1}$ for $b = 1,\ 2,\ 3,\ 4$, respectively.*

Discussion on inner product error guarantees. For inner-product estimation, $\texttt{EDEN}_{\text{UB}}$ gives the strongest guarantee among the methods compared here, closely followed by that of $\texttt{RabitQ}_{\text{UB}}$. For $b = 2, 3, 4$, its expected inner-product distortion is roughly four times smaller than that of $\texttt{TurboQuant}_{\text{PROD}}$, whose distortion is approximately $\mathcal{D}_{\text{IP}}(\texttt{TurboQuant}_{\text{PROD}}) \approx \frac{\mathbf{1.57}}{d},\ \frac{\mathbf{0.56}}{d},\ \frac{\mathbf{0.18}}{d},\ \frac{\mathbf{0.047}}{d}$ respectively, as stated in Proposition B.4. This comparison suggests that the rescaling step used in $\texttt{EDEN}_{\text{UB}}$ and $\texttt{RabitQ}_{\text{UB}}$ are more effective for reducing expected inner product distortion than the QJL-based correction used in $\texttt{TurboQuant}_{\text{PROD}}$. The reason is that $\texttt{TurboQuant}_{\text{PROD}}$ sacrifices one bit for the QJL-based bias correction, rather than using all bits for reconstruction. This bit allocation appears to incur an approximately constant-factor loss, which is visible as the roughly fourfold gap. The proofs of Theorem 1 and Corollaries 1, 2 are deferred to Appendix E.1, E.2, and E.3, respectively.

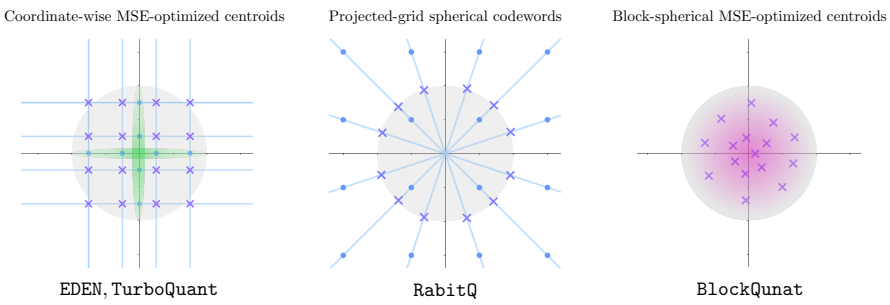

Figure 1: Conceptual comparison of codebooks used by rotation-based quantizers for $b = 2$ in a two-coordinate projection when $d > 2$. The shaded disk indicates the feasible region of the two displayed coordinates of a rotated unit vector. Left: `EDEN` and `TurboQuant` use a Cartesian-product codebook formed by coordinate-wise MSE-optimized scalar centroids. Middle: `RabitQ` uses spherical codewords obtained by projecting a $2^b \times 2^b$ uniform grid onto $\mathbb{S}^{d-1}$. Right: `BlockQuant` with block size $p = 2$ optimizes $2^{bp} = 16$ centroids directly for the two-dimensional block distribution.

4 Proposed Method

Motivated by the comparison in Section 3, we propose a new quantization scheme that improves both $\mathcal{D}_{\text{MSE}}$ and $\mathcal{D}_{\text{IP}}$. Our method follows the spirit of product quantization by grouping multiple coordinates into blocks, and extends the coordinate-wise approaches used in `EDEN` and `TurboQuant` to a block-spherical quantization scheme. This allows the quantizer to incorporate the spherical structure of the input vectors more directly, leading to improved distortion guarantees.

4.1 Block-Sphere Quantization (`BlockQuant`)

The coordinate-wise MSE-optimized centroids used in `EDEN` and `TurboQuant` lead to low MSE distortion, but they do not fully exploit the spherical structure of unit input vectors. Since quantization is performed independently across coordinates, many possible combinations of scalar centroids can lie far from the sphere. This suggests that part of the 2^{bd} code space is spent on codewords that are not well aligned with the geometry of the input domain. If the quantizer incorporates the spherical structure more directly, the bit budget can be used more efficiently, potentially leading to improved MSE and inner-product performance.

We propose *Block-Sphere Quantization* (`BlockQuant`), which addresses this limitation by constructing centroids that capture richer spherical information at the block level (Figure 1). After rotating input vectors, instead of quantizing each coordinate separately, we group the coordinates of rotated vectors $\mathbf{z}$ into blocks. Specifically, for a block size p that divides d, we decompose $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)$ where $m = d/p$ and $\mathbf{z}_j \in \mathbb{B}^p$ for $j \in [m]$. Each block $\mathbf{z}_j$ is then assigned to a centroid in a p-dimensional codebook. Each block $\mathbf{z}_j$ is then assigned to a centroid in a p-dimensional codebook. In this way, the quantizer reduces to the coordinate-wise schemes when $p = 1$, while larger block sizes allow the codebook to capture more of the spherical structure of the rotated unit vector. The following lemma characterizes the marginal distribution of each block $\mathbf{z}_j$ in $\mathbb{B}^p$.

Lemma 1 (Block marginal distribution of a uniform spherical vector)**.** *Suppose $d = mp$ and $\mathbf{x} \sim \text{Unif}(\mathbb{S}^{d-1})$. Divide vector $\mathbf{x}$ into blocks $\mathbf{x} = [\mathbf{z}_1, \ldots, \mathbf{z}_m]$ where $\mathbf{z}_j \in \mathbb{B}^p$ for $j \in [m]$. Then, each block $\mathbf{z}_j$ has density on $\mathbb{B}^p$: $f_{p,d}(\mathbf{z}_j) = \frac{\Gamma(d/2)}{\pi^{p/2}\Gamma((d-p)/2)}(1 - \|\mathbf{z}_j\|_2^2)^{\frac{d-p-2}{2}}$. Equivalently, $\mathbf{z}_j$ can decomposed as $\mathbf{z}_j = r_j\boldsymbol{\theta}_j$ for $r_j \in [0, 1]$, $\boldsymbol{\theta}_j \in \mathbb{S}^{p-1}$, where $r_j \perp \boldsymbol{\theta}_j$ and each component follows the distributions: $r_j^2 \sim \text{Beta}\left(\frac{p}{2}, \frac{d-p}{2}\right)$ and $\boldsymbol{\theta}_j \sim \text{Unif}(S^{p-1})$.*

The proof of Lemma 1 is provided in Appendix G. Then, we can formulate the block quantization as the K-means optimization problem. Let $\{\mathbf{o}_i\}_{i=1}^{2^{bp}} \subset B_p$ denote the codebook centroids. We optimize these centroids by minimizing the expected squared distance to the nearest centroid under the block marginal distribution $f_{p,d}$:

$$(\text{Distortion cost}) = \min_{\mathbf{o}_1,\ldots,\mathbf{o}_{2^{bp}} \in B_p} \int_{B_p} \min_{i \in [2^{bp}]} \|\mathbf{z} - \mathbf{o}_i\|_2^2 \, f_{p,d}(\mathbf{z}) \, \mathrm{d}\mathbf{z}. \tag{1}$$

Algorithm 1 Block-Sphere Quantization (`BlockQuant`)

Require: dimension d, block size p, bit-width b.

// Global Parameters for Setting up `BlockQuant`.

1: Generate a random rotation matrix $R \in \mathbb{R}^{d\times d}$ from Haar distribution.

2: Construct codebook by finding centroids $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{2^{bp}} \in \mathbb{B}^p$ that minimize Equation 1.

3: **procedure** QUANT($\mathbf{x}$)

4: $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m) \leftarrow R\mathbf{x}$, where $m = d/p$.

5: $\mathrm{idx}_j \leftarrow \arg\min_{i\in[2^{bp}]} \|\mathbf{z}_j - \mathbf{o}_i\|$ for every $j \in [m]$.

6: Save the alignment $\rho = \langle \mathbf{z}, \bar{\mathbf{z}}\rangle$, where $\bar{\mathbf{z}}$ aggregated vector of corresponding codewords.

7: **return** $\mathrm{idx} = [\mathrm{idx}_1, \ldots, \mathrm{idx}_m] \in \{0,1\}^{b\cdot d}$. ▷ idx's are bp-bit integers

8: **procedure** DEQUANT(idx)

9: $\bar{\mathbf{z}} = [\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_m]$ where $\bar{\mathbf{z}}_j \leftarrow c_{\mathrm{idx}_j}$ for every $j \in [m]$.

10: $\bar{\mathbf{x}} \leftarrow R^\top \bar{\mathbf{z}}$.

11: Set the rescaling parameter

$$S = \begin{cases} \dfrac{\rho}{\|\bar{\mathbf{x}}\|_2^2}, & \text{for minimizing MSE (i.e., } \texttt{BlockQuant}_{\mathrm{BSM}}\text{)}, \\ \dfrac{1}{\rho}, & \text{for unbiased reconstruction (i.e., } \texttt{BlockQuant}_{\mathrm{UB}}\text{)}, \\ 1, & \text{for raw reconstruction (i.e., } \texttt{BlockQuant}_{\mathrm{MSE}}\text{)}. \end{cases}$$

12: **return** $S \cdot \widetilde{\mathbf{x}}$.

Notably, as in `TurboQuant`, the codebook is constructed only once before quantization and is reused throughout the online quantization procedure.

For the dequantization step in Line 11 of Algorithm 1, we use objective-dependent rescaling rules for S, as in `EDEN`. Let $\bar{\mathbf{x}}$ denote the raw reconstruction before the final rescaling, and let $\rho := \langle \mathbf{x}, \bar{\mathbf{x}}\rangle$. This alignment is computed in the rotated domain in Line 6 of Algorithm 1, since $\langle \mathbf{x}, \bar{\mathbf{x}}\rangle = \langle \mathbf{x}, R^\top \bar{\mathbf{z}}\rangle = \langle R\mathbf{x}, \bar{\mathbf{z}}\rangle = \langle \mathbf{z}, \bar{\mathbf{z}}\rangle$. For $\texttt{BlockQuant}_{\mathrm{BSM}}$, the MSE-optimized variant with best scalar rescaling, we set $S = \langle \mathbf{x}, \bar{\mathbf{x}}\rangle / \|\bar{\mathbf{x}}\|_2^2 = \rho/\|\bar{\mathbf{x}}\|_2^2$. For $\texttt{BlockQuant}_{\mathrm{UB}}$, the unbiased-reconstruction variant, we set $S = 1/\rho$. When fast quantization and dequantization are preferred, we also consider the raw-reconstruction variant, denoted by $\texttt{BlockQuant}_{\mathrm{MSE}}$ following the notation of $\texttt{TurboQuant}_{\mathrm{MSE}}$. In this case, we set $S = 1$, so the value of ρ in Line 6 does not need to be stored.

4.2 Analysis of `BlockQuant`

The following theorem gives MSE bounds for the proposed algorithm with practical block sizes $p = 2$ and $p = 3$, as well as for the idealized case $p = d$, where the centroids lie on $\mathbb{S}^{d-1}$. In this case, $f_{d,d}$ in the cost function (Equation 1) is defined by the probability density function of Unif($\mathbb{S}^{d-1}$).

Theorem 2 (MSE bounds for $\texttt{BlockQuant}_{\mathrm{MSE}}$ and $\texttt{BlockQuant}_{\mathrm{BSM}}$)**.** *In high dimension, for* $\mathcal{Q} \in \{\texttt{BlockQuant}_{\mathrm{MSE}}, \texttt{BlockQuant}_{\mathrm{BSM}}\}$, $\mathcal{D}_{MSE}\left(\mathcal{Q}_{(p=2)}\right) \approx \mathbf{0.363},\ \mathbf{0.108},\ \mathbf{0.0297},\ \mathbf{0.0078}$ *and* $\mathcal{D}_{MSE}\left(\mathcal{Q}_{(p=3)}\right) \approx \mathbf{0.357},\ \mathbf{0.101},\ \mathbf{0.0271},\ \mathbf{0.0071}$*, for* $b = 1,\ 2,\ 3,\ 4$*, respectively. Moreover, for large* b*,*

$$\mathcal{D}_{MSE}\left(\mathcal{Q}_{(p=2)}\right) \le \mathbf{2.015} \cdot \frac{1}{4^b}(1 + o(1)),$$
$$\mathcal{D}_{MSE}\left(\mathcal{Q}_{(p=3)}\right) \le \mathbf{1.770} \cdot \frac{1}{4^b}(1 + o(1)),$$
$$\mathcal{D}_{MSE}\left(\mathcal{Q}_{(p=d)}\right) \le C_d \cdot \left(\frac{1}{4}\right)^{\frac{bd}{d-1}} \cdot (1 + o(1)),$$

where $C_d := \Gamma\left(1 + \frac{2}{d-1}\right)\left[2\sqrt{\pi}\,\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}\right]^{\frac{2}{d-1}}$.

Discussion of Theorem 2. The MSE guarantee of `BlockQuant` with $p = 2$ or 3 is better than, that of `EDEN`, `RabitQ`, and `TurboQuant` in both small and large bit regime (1, Remark 1). It is important to note that the these guarantees do not require access to a globally optimal codebook. The general high-rate bound in Corollary H.1 shows that the leading constants can be achieved by using explicit lattices: the optimal hexagonal lattice A_2 for $p = 2$, and $G_3^\star \le G(A_3^\star)$ for $p = 3$, where $A_3^\star$ is the body-centered cubic lattice, dual to the face-centered cubic lattice A_3. Moreover, for $p = d$, we have $C_d \approx 1.055,\ 1.008,\ 1.001$ for $d = 100,\ 1000,\ 10000$, respectively. Thus, the resulting upper bound closely matches the lower bound in Section 5, both in order and in the leading constant. This suggests that increasing the block size in `BlockQuant` moves the quantizer toward the optimal quantization limit. The proof of Theorem 2 is deferred to Appendix H.1.

Combining the ratio-estimator bound in Theorem 1 with the MSE bound in Theorem 2, we obtain the bound on the inner-product error of Algorithm 1.

Corollary 3 (Inner product error bound of Algorithm 1)**.** *In high dimension, the* $\texttt{BlockQuant}_{\mathrm{UB}}$ *with block size p satisfies* $\mathcal{D}_{IP}\left(\texttt{BlockQuant}_{\mathrm{UB}(p=2)}\right) \approx \frac{\mathbf{0.571}}{d-1},\ \frac{\mathbf{0.120}}{d-1},\ \frac{\mathbf{0.0306}}{d-1},\ \frac{\mathbf{0.0078}}{d-1}$ *for* $p = 2$ *and* $b = 1,\ 2,\ 3,\ 4$*, and* $\mathcal{D}_{IP}\left(\texttt{BlockQuant}_{\mathrm{UB}(p=3)}\right) \approx \frac{\mathbf{0.553}}{d-1},\ \frac{\mathbf{0.113}}{d-1},\ \frac{\mathbf{0.0279}}{d-1},\ \frac{\mathbf{0.0071}}{d-1}$*, for* $p = 3$ *and* $b = 1,\ 2,\ 3,\ 4$*, respectively. Moreover, for large b,*

$$\mathcal{D}_{IP}\left(\texttt{BlockQuant}_{\mathrm{UB}(p=2)}\right) \le \frac{\mathbf{2.015}}{d-1} 4^{-b}(1 + o(1)),$$
$$\mathcal{D}_{IP}\left(\texttt{BlockQuant}_{\mathrm{UB}(p=3)}\right) \le \frac{\mathbf{1.770}}{d-1} 4^{-b}(1 + o(1)).$$

Discussion of Corollary 3. Since the MSE distortion of $\texttt{EDEN}_{\mathrm{UB}}$ is approximately $\frac{\mathbf{0.571}}{d-1}$, $\frac{\mathbf{0.133}}{d-1}$, $\frac{\mathbf{0.0358}}{d-1}$, and $\frac{\mathbf{0.0096}}{d-1}$ for $b = 1,\ 2,\ 3,\ 4$, respectively (Corollary 1), the expected inner-product error guarantee for $\texttt{BlockQuant}_{\mathrm{UB}}$ with block size ≥ 2 is stronger than the corresponding bounds for other existing rotation-based quantizers. The proof of Corollary 3 is deferred to Appendix H.2.

5 Sharper Lower Bound on Distortion

We derive a lower bound for our quantization problem using the Shannon lower bound. A similar quantization lower bound based on Shannon's argument was considered in Zandieh et al. [26]; however, the entropy term used there is not directly applicable to a unit vector. (Since $\mathbb{S}^{d-1}$ has zero Lebesgue measure in $\mathbb{R}^d$, the ambient differential entropy of $\mathbf{x}$ is not finite) We correct this by applying the Shannon lower bound to the first $d-1$ coordinates of $\mathbf{x}$, whose distribution is absolutely continuous on $\mathbb{B}^{d-1}$. Consequently, unlike the previous expression, the exponent involves $\frac{bd}{(d-1)}$ rather than b. This correction reflects the fact that, although the ambient dimension is d, the unit sphere $\mathbb{S}^{d-1}$ has intrinsic dimension $d-1$.

Theorem 3 (Shannon distortion lower bound)**.** *Suppose* $\mathbf{x} \sim$ *Unif*$(\mathbb{S}^{d-1})$*. Then, for any* $b \ge 0$ *and any fixed* bd*-bit quantization map Q, the MSE is lower bounded as*

$$\mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - Q^{-1}(Q(\mathbf{x}))\|_2^2] \ge c_d \left(\frac{1}{4}\right)^{\frac{bd}{d-1}},$$

where $c_d := \frac{d-1}{2\pi e}\left(\frac{\pi^{d/2}}{\Gamma(d/2)}\right)^{2/(d-1)} \exp\left(\frac{\psi(1/2)-\psi(d/2)}{d-1}\right)$ *and* $\psi(t) := \frac{d}{dt}\log\Gamma(t) = \Gamma'(t)/\Gamma(t)$.

Disscusion on Theorem 3. The resulting theorem can be interpreted as an $\mathcal{D}_{\mathrm{MSE}}$ lower bound for rotation-based quantizers. Indeed, after a random rotation, the rotated input vector can be viewed as $\mathbf{z} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and nearest-centroid quantization with a fixed codebook can be regarded as a fixed quantization map on the sphere. Since the distance between two points on sphere is preserved under rotation, the left hand side is equal to $\mathcal{D}_{\mathrm{MSE}}$. Since $c_d \approx 0.936,\ 0.991,\ 0.999$ for $d = 100,\ 1000,\ 10000$, respectively, the lower bound in Theorem 3 demonstrates the tightness of the upper bound in Theorem 2. This suggests that, as the block size p increases, `BlockQuant` moves closer to the optimal quantization limit. The proof of Theorem3 is provided in Appendix I.

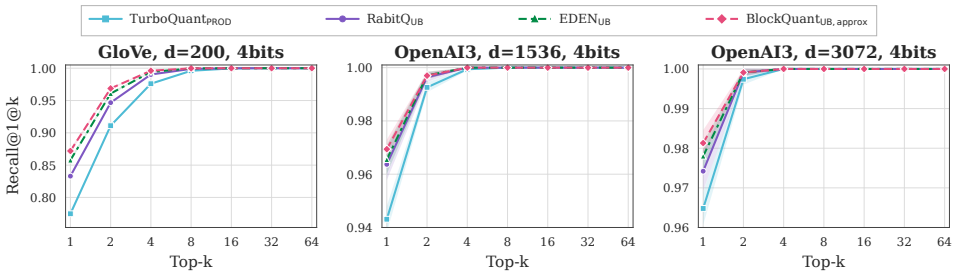

Figure 2: Recall comparison at 4 bits across different datasets.

6 Experiments

To examine whether the distortion improvements predicted by our theory translate into practical gains, we evaluate `BlockQuant` on real embedding and LLM KV-cache quantization tasks. Specifically, we consider three settings: quantization accuracy on real embeddings (Appendix L.2), nearest-neighbor search using quantized inner-product estimates (Section 6.1), and KV-cache quantization for long-context LLM inference (Section 6.2). Unless stated otherwise, we use block size $p = 3$. We largely follow the experimental environments and settings of Zandieh et al. [27]. For `BlockQuant`, we use a simple lookup-table approximation for nearest centroid assignment (Appendix L.1).

6.1 Nearest-Neighbor Search.

We evaluate retrieval quality using Recall@1@k. For each query $\mathbf{q}$, let $g(\mathbf{q})$ denote the exact top-1 neighbor computed using full-precision inner products, and let $\mathcal{A}_k(\mathbf{q})$ denote the set of top-k candidates returned by a method using quantized inner-product estimates. We define

$$\text{Recall@1@}k = \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{q} \in \mathcal{Q}} \mathbf{1}\{g(\mathbf{q}) \in \mathcal{A}_k(\mathbf{q})\},$$

where $\mathcal{Q}$ is the query set. Thus, Recall@1@k measures whether quantization preserves the exact nearest neighbor within the top-k retrieved candidates.

Figure 2 compares the methods on GloVe ($d = 200$) and OpenAI3/DBpedia ($d = 1536$ and 3072) under 4-bit compression; the corresponding 2-bit results are provided in Figure L.3. Across both datasets, the approximate version of $\texttt{BlockQuant}_{\text{UB}}$ achieves strong recall, with the largest gains in the low-k regime. This regime is especially sensitive to quantization error, since small perturbations in inner-product estimates can change the ordering of the top-ranked candidates.

6.2 KV-cache quantization

Experimental settings. We evaluate whether the improved distortion of `BlockQuant` translates into end-to-end LLM performance under KV-cache quantization. We apply each quantizer to the KV cache of Llama-3.1-8B-Instruct while keeping the model weights unchanged.

In the attention computation, we follow the outlier-aware configuration used in the KV-cache quantization setup [26]. For each attention head with head dimension $d_h = 128$, the 32 key channels with the largest L2 norm are treated as outlier channels and quantized at a higher bit-width, while the remaining 96 channels are quantized at a lower bit-width. In 3.5-bit setting, the outlier channels use 4-bit quantization and the non-outlier channels use 3-bit quantization, together with two additional float16 scaling values for the two subvectors. This gives an effective key-cache bit-width of $\frac{32\times4+96\times3+2\times16}{128} = 3.5$. For the value cache, we quantize the full head dimension uniformly using a 2-bit approximation. All other components, including query states, model weights, MLP layers, embeddings, and output projections, remain unquantized.

We compare our algorithm with `RabitQ`, `TurboQuant`, and `EDEN`, and report results on the Needle-In-A-Haystack [13] benchmark and LongBench-E [3]. Since the randomness induced by sampled rotation matrices has a particularly noticeable effect on LLM inference, we repeat each experiment over five random seeds and report the mean performance with standard deviations.

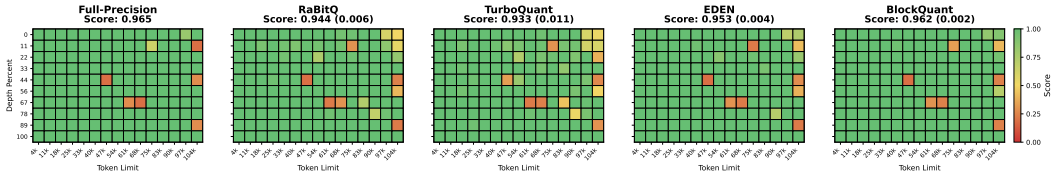

Figure 3: Evaluation of Llama-3.1-8B-Instruct on the "Needle-In-A-Haystack" benchmark over five random seeds. Results are reported as mean, with standard deviations shown in parentheses.

Table 3: Evaluation of Llama 3.1 8B Instruct on the "LongBench-E" benchmark over five random seeds. Results are reported as mean (standard deviation), with 95% bootstrap confidence intervals shown in brackets below.

Method	SingleQA	MultiQA	Summ	Few-shot	Synthetic	Code	Average
RabitQ	19.12 (0.54) [18.65, 19.59]	15.88 (0.26) [15.69, 16.12]	29.40 (0.11) [29.30, 29.50]	68.23 (0.10) [68.15, 68.32]	55.75 (0.35) [55.44, 56.05]	60.35 (0.13) [60.22, 60.46]	43.52 (0.15) [43.38, 43.65]
TurboQuant	18.60 (0.39) [18.23, 18.90]	15.79 (0.15) [15.66, 15.93]	29.11 (0.11) [29.02, 29.20]	68.19 (0.10) [68.10, 68.27]	55.31 (0.81) [54.55, 56.00]	59.72 (0.49) [59.25, 60.17]	43.20 (0.20) [43.00, 43.35]
EDEN	19.02 (0.27) [18.82, 19.29]	16.30 (0.26) [16.05, 16.50]	29.42 (0.17) [29.29, 29.58]	68.46 (0.12) [68.33, 68.55]	56.29 (0.22) [56.14, 56.51]	61.41 (0.21) [61.23, 61.59]	43.87 (0.07) [43.80, 43.93]
BlockQuant	**19.40** (0.16) [19.26, 19.53]	**16.27** (0.56) [15.83, 16.84]	**29.56** (0.04) [29.52, 29.59]	**68.47** (0.08) [68.39, 68.54]	**56.13** (0.70) [55.47, 56.63]	**61.47** (0.26) [61.24, 61.69]	**43.93** (0.15) [43.80, 44.05]
Full Cache (16-bit)	19.53	16.54	30.28	68.41	56.41	61.60	44.15

Needle-In-A-Haystack. Figure 3 shows the Needle-In-A-Haystack results across different context lengths and needle depths. The full-precision cache obtains a score of 0.965. Among the quantized methods, `BlockQuant` achieves the highest score, 0.962, with a standard deviation of 0.002, nearly matching the full-precision cache. In comparison, `EDEN`, `RabitQ`, and `TurboQuant` obtain scores of 0.953, 0.944, and 0.933, respectively. This indicates that `BlockQuant` better preserves the attention-relevant information in the KV cache, especially in long-context settings where small quantization errors can accumulate across many tokens.

LongBench-E. Table 3 reports the LongBench-E results.[2] `BlockQuant` achieves performance comparable to the strongest baseline, `EDEN` (43.93 ± 0.15 vs. 43.87 ± 0.07), while outperforming `RabitQ` (43.52) and `TurboQuant` (43.20). The score is also close to the full-cache result of 44.15, leaving only a small gap.

Overall, `BlockQuant` significantly outperforms all baselines on Needle-In-A-Haystack, and matches the strongest baseline on LongBench-E. These results show that the block-spherical construction is effective not only for standalone embedding distortion and nearest-neighbor retrieval, but also for downstream LLM inference.

7 Conclusion

In this work, we provide a unified comparison of `EDEN`, `RabitQ`, and `TurboQuant` across reconstruction MSE, expected inner-product distortion, and high-probability bit complexity, clarifying their criterion-dependent strengths. We introduce `BlockQuant`, which exploits the exact block marginals of randomly rotated unit vectors, and prove improved distortion constants for both expected error measures over coordinate-wise rotation-based quantizers. Experiments on embedding retrieval and KV-cache quantization demonstrate lower distortion, improved recall, and competitive LLM benchmark performance, supporting our theoretical findings. A promising direction for future work is to tailor `BlockQuant` to the structure of KV caches, moving beyond direct cache quantization toward more effective LLM-specific designs.

[2]The original `TurboQuant` evaluation [27] applies additional prediction post-processing, including truncation to the *first* generated line or token before scoring. To avoid potential evaluation artifacts caused by this truncation, we instead use the official LongBench-E `result.json` outputs.

Acknowledgments

This work was supported by grants from the National Research Foundation of Korea (NRF) and the Institute of Information & Communications Technology Planning & Evaluation (IITP), both funded by the Korean government (MSIT) (Nos. RS-2022-NR071853, RS-2023-00222663, RS-2025-25463302, RS-2026-25507282, RS-2026-25613008, and RS-2025-25420849). JL was additionally supported by the 2025 Google PhD Fellowship, with support from Google.org.

References

[1] Noga Alon and Bo'az Klartag. Optimal compression of approximate inner products and dimension reduction. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–650. IEEE, 2017.

[2] 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.

[3] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137, 2024.

[4] Ran Ben-Basat, Yaniv Ben-Itzhak, Gal Mendelson, Michael Mitzenmacher, Amit Portnoy, and Shay Vargaftik. A note on TurboQuant and the earlier DRIVE/EDEN line of work. *arXiv preprint arXiv:2604.18555*, 2026.

[5] Sepehr Eghbali and Ladan Tahvildari. Deep spherical quantization for image search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11682–11691, 2019.

[6] 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.

[7] 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.

[8] Jianyang Gao, Yutong Gou, Yuexuan Xu, Jifan Shi, Yongyi Yang, Shuolin Li, Raymond Chi-Wing Wong, and Cheng Long. Revisiting rabitq and turboquant: A symmetric comparison of methods, theory, and experiments. *arXiv preprint arXiv:2604.19528*, 2026.

[9] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2946–2953, 2013. doi: 10.1109/CVPR.2013.379.

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

[11] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011. doi: 10.1109/TPAMI.2010.57.

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

[13] Gregory Kamradt. Needle in a haystack-pressure testing llms, 2023. *URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack*, 2024.

[14] Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024.

[15] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

[16] Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1): 7–12, 1960. doi: 10.1109/TIT.1960.1057548.

[17] Mohammad Norouzi and David J. Fleet. Cartesian k-means. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3017–3024, 2013. doi: 10.1109/CVPR.2013.388.

[18] PF Panter and Wu Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proceedings of the IRE*, 39(1):44–48, 2006.

[19] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

[20] Tycho F. A. van der Ouderaa, Maximilian L. Croci, Agrin Hilmkil, and James Hensman. Pyramid vector quantization for LLMs. *arXiv preprint arXiv:2410.16926*, 2024. URL `https://arxiv.org/abs/2410.16926`.

[21] Shay Vargaftik, Ran Ben-Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Drive: One-bit distributed mean estimation. *Advances in Neural Information Processing Systems*, 34:362–377, 2021.

[22] Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben Itzhak, and Michael Mitzenmacher. Eden: Communication-efficient and robust distributed mean estimation for federated learning. In *International Conference on Machine Learning*, pages 21984–22014. PMLR, 2022.

[23] Jianfeng Wang, Jingdong Wang, Jingkuan Song, Xin-Shun Xu, Heng Tao Shen, and Shipeng Li. Optimized cartesian k-means. *IEEE Transactions on Knowledge and Data Engineering*, 27(1): 180–192, 2015. doi: 10.1109/TKDE.2014.2324592.

[24] Yongyi Yang, Jianyang Gao, and Wei Hu. Raana: A fast, flexible, and data-efficient post-training quantization algorithm. *arXiv preprint arXiv:2504.03717*, 2025.

[25] Paul L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–148, 1982. doi: 10.1109/TIT.1982.1056490.

[26] Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. Turboquant: Online vector quantization with near-optimal distortion rate. In *International Conference on Learning Representations (ICLR)*, 2025.

[27] Amir Zandieh, Majid Daliri, and Insu Han. Qjl: 1-bit quantized jl transform for kv cache quantization with zero overhead. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25805–25813, 2025.

[28] Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=yGnsH3gQ6U`.

Appendix

Table of Contents

A Related Work **13**

B Existing Rotation-Based Quantizers **14**
B.1 Existing Guarantees on `EDEN`, `RabitQ`, and `TurboQuant` 15

C High-Resolution Formula **16**

D MSE Analysis **16**
D.1 Proof of Proposition 1 16
D.2 Proof of Proposition 2 18
D.3 Approximated MSE of `TurboQuant` 18

E Inner Product Distortion Analysis **19**
E.1 Proof of Theorem 1 19
E.2 Proof of Corollary 1 19
E.3 Proof of Corollary 2 20

F Bit Complexity Analysis **21**
F.1 Bit Complexity of `EDEN` 21
F.2 Bit Complexity of $\texttt{TurboQuant}_{\text{PROD}}$ 23

G Block Marginal Distribution of a Uniform Spherical Vector (Proof of Lemma 1) **24**

H Expected Distortion Analysis of `BlockQuant` **25**
H.1 Proof of Theorem 2 25
H.2 Proof of Corollary 3 28

I Lower Bound Analysis (Proof of Theorem 3) **30**

J High Probability Bit Complexity Comparison **30**

K Auxiliary Lemmas **31**

L Additional Experimental Details and Results **31**
L.1 Time Complextiy and Approximate Nearest-Centroid Assignment for `BlockQuant` 31
L.2 Quantization Accuracy. 33
L.3 Nearest-Neighbor Search. 34
L.4 Quantization Efficiency 35
L.5 Computational Resources 35
L.6 Comparison with QuaRot for KV-Cache Quantization 35

M Limitations **36**

A Related Work

Classical vector quantization and high-rate theory. Vector quantization has a long history in source coding and signal compression. Classical scalar and vector quantizer design is built on the Lloyd–Max optimality conditions and the Linde–Buzo–Gray algorithm, which iteratively alternates nearest-codeword assignment and centroid updates [16, 15]. High-rate quantization theory further

characterizes the leading-order distortion of optimal vector quantizers through the Zador–Gersho formula [10, 25].

Product and block quantization for nearest-neighbor search. Product quantization (PQ) and its variants are among the most widely used vector quantization methods for approximate nearest-neighbor search. PQ decomposes a high-dimensional vector into low-dimensional subspaces and quantizes each subvector using a separate sub-codebook [11, 9, 17, 23]. Block-Sphere Quantization (BlockQuant) is related to this line of work in that it also quantizes low-dimensional blocks. The key difference is that our codebook is not learned from a dataset. BlockQuant constructs a universal block codebook for this spherical marginal distribution of randomly rotated input vectors in the same framework as EDEN, RabitQ, and TurboQuant.

Spherical quantization. Several quantization methods exploit hyperspherical geometry by normalizing vectors and quantizing their directions. For example, deep spherical quantization uses unit-sphere embeddings for supervised image retrieval, while binary spherical quantization applies spherical normalization and binary codes for visual tokenization [5, 28]. Pyramid vector quantization and related lattice-based methods also use structured codebooks to represent directions on spherical or pyramidal domains, and have recently been revisited for neural and LLM compression [20].

Although these works demonstrate the usefulness of spherical geometry, the role of the sphere is different from ours. Existing spherical quantizers typically quantize a full normalized embedding, feature vector, or model-weight direction, often using a learned or structured spherical codebook. In contrast, BlockQuant does not construct a codebook on the full sphere. After a random rotation of $x \in \mathbb{S}^{d-1}$, BlockQuant partitions the rotated vector into low-dimensional blocks $z_j \in \mathbb{B}^p$ and quantizes the Euclidean-ball marginal of each block. This marginal distribution is induced exactly by the unit-sphere geometry and is known in closed form. Thus, BlockQuant uses spherical geometry through the exact block marginals of randomly rotated unit vectors, rather than through a global spherical code or a learned hyperspherical embedding.

Rotation-based quantizers and theoretical guarantees. Several recent quantizers use randomized rotations to make coordinated istributions more regular. DRIVE and EDEN apply randomized rotations and scalar quantization for distributed mean estimation and federated learning [21, 22]. RabitQcombines randomized quantization with a ratio-based estimator and provides high-probability error guarantees for approximate nearest-neighbor search[6, 7]. TurboQuant uses the exact coordinate marginal distribution induced by random rotation to design scalar quantizers for MSE, and then combines a reconstruction quantizer with a QJL residual correction to obtain unbiased inner-product estimation[26, 27]. Recent notes have compared the EDEN/DRIVE, RabitQ, and TurboQuant lines of work, highlighting the need to evaluate these methods under a common set of criteria[4, 8]. However, existing comparisons remain fragmented: they often focus on empirical performance or algorithmic structure, while the theoretical guarantees are stated under different objectives and are therefore difficult to compare directly. Our work follows this comparison-driven perspective by separating three criteria—reconstruction MSE, expected inner-product distortion, and high-probability bit complexity—and analyzing EDEN, RabitQ,and TurboQuant under a unified framework. The insights from this comparison then motivate BlockQuant, which exploits block-spherical structure to improve expected distortion.

B Existing Rotation-Based Quantizers

EDEN [21, 22]. For a unit input vector $\mathbf{x}$, EDEN first applies a random rotation R and rescales the rotated vector by a factor η_{q}, so that a dimension-independent scalar codebook can be used. Each coordinate of the rescaled vector is then quantized to its nearest centroid. During dequantization, each code is replaced by the corresponding centroid, the inverse rotation is applied, and the result is rescaled by a scalar η_{dq}. The choice of η_{dq} depends on the target objective.

A key component of EDEN is the Lloyd–Max scalar codebook $\mathcal{C}_{\mathrm{EDEN}}$, constructed for the standard normal distribution [22]. This choice is motivated by the distribution of randomly rotated vectors: if $\mathbf{x} \in \mathbb{S}^{d-1}$ and R is a random rotation, then each coordinate of $R\mathbf{x}$ tends to follow $N(0, \frac{1}{d})$. Thus, choosing $\eta_{\mathrm{q}} = \sqrt{d}$ makes the standard-normal Lloyd–Max codebook well matched to the rotated coordinates.

The same quantization rule can be expressed equivalently by scaling the codebook rather than the rotated vector $R\mathbf{x}$. If we define $\mathcal{C}_{\text{EDEN}}^{(\text{scaled})} = \frac{1}{\sqrt{d}}\mathcal{C}_{\text{EDEN}}$, then the output of EDEN can be written as $\eta R^\top \left(P_{\text{decode}}\left(P_{\text{code}}(R\mathbf{x})\right)\right) = \eta\bar{\mathbf{x}}$ for some $\eta \in \mathbb{R}$, where P_{code} and P_{decode} denote the encoding and decoding maps with respect to $\mathcal{C}_{\text{EDEN}}^{(\text{scaled})}$. Then, $\texttt{EDEN}_{\text{BSM}}$ uses $\eta = \frac{\langle \mathbf{x}, \bar{\mathbf{x}}\rangle}{\|\bar{\mathbf{x}}\|_2^2}$, which is the best scalar for minimizing the expected squared reconstruction error, and $\texttt{EDEN}_{\text{UB}}$ uses $\eta = \frac{1}{\langle \mathbf{x}, \bar{\mathbf{x}}\rangle}$ for unbiased inner-product estimation. These two variants correspond to `EDEN-biased` and `EDEN-unbiased` in Ben-Basat et al. [4], respectively. For presentation clarity and consistency across methods, we rename them $\texttt{EDEN}_{\text{BSM}}$ and $\texttt{EDEN}_{\text{UB}}$, where the suffixes indicate the reconstruction-oriented and unbiased inner-product variants, respectively.

RabitQ [7, 8]. RabitQ is a rotation-based quantizer whose practical encoder quantizes the rotated vector using a uniform grid. Although RabitQ is implemented through this grid quantization procedure, it can be equivalently viewed as using a fixed spherical codebook $\mathcal{C}_{\text{RabitQ}}$ in the rotated domain [7]. The elements of $\mathcal{C}_{\text{RabitQ}}$ are obtained by projecting the grid codewords onto $\mathbb{S}^{d-1}$. Thus, after applying a random rotation R, RabitQ selects a spherical codeword for $R\mathbf{x}$, and the raw reconstruction $\bar{\mathbf{x}}$ is obtained by applying the inverse rotation to the decoded codeword.

RabitQ applies a scalar correction to the raw reconstruction for reconstruction or inner-product estimation, like EDEN. For notational consistency with EDEN and TurboQuant, we absorb the scalar correction into the dequantized output. Specifically, throughout this paper, we regard $\frac{\bar{\mathbf{x}}}{\langle \mathbf{x}, \bar{\mathbf{x}}\rangle}$ as the dequantized output of $\texttt{RabitQ}_{\text{UB}}$ for unbiased inner-product estimation. This convention is equivalent to the estimator in Gao et al. [7], where the correction factor is treated as part of the estimation procedure rather than as part of the dequantization output. For reconstruction, the best scalar multiple of $\bar{\mathbf{x}}$ for approximating $\mathbf{x}$ is $\frac{\langle \mathbf{x}, \bar{\mathbf{x}}\rangle}{\|\bar{\mathbf{x}}\|_2^2}\bar{\mathbf{x}} = \langle \mathbf{x}, \bar{\mathbf{x}}\rangle\bar{\mathbf{x}}$, since RabitQ uses a spherical codebook. We therefore regard $\langle \mathbf{x}, \bar{\mathbf{x}}\rangle\bar{\mathbf{x}}$ as the dequantized output of $\texttt{RabitQ}_{\text{BSM}}$. Here, $\texttt{RabitQ}_{\text{UB}}$ corresponds to the original RabitQ in Gao et al. [7], while $\texttt{RabitQ}_{\text{BSM}}$ corresponds to $\texttt{RabitQ}_{\text{MSE}}$ in Gao et al. [8]. Following the same naming convention as above, we use the suffixes to distinguish the unbiased inner-product variant from the reconstruction-oriented variant.

TurboQuant [26]. TurboQuant is a randomized rotation-based quantization method based on coordinate-wise scalar quantization, similar to EDEN. Unlike EDEN, which designs its scalar codebook using a Gaussian approximation to the marginal distribution of randomly rotated coordinates, TurboQuant constructs a Lloyd–Max codebook for the exact coordinate marginal of a uniformly random point on $\mathbb{S}^{d-1}$. In this sense, $\texttt{TurboQuant}_{\text{MSE}}$ can be viewed as an EDEN-type coordinate-wise Lloyd–Max quantizer that replaces the high-dimensional Gaussian approximation with the exact spherical marginal and uses the scaling convention $\eta_{\text{q}} = 1$ [4].

TurboQuant also provides a variant for inner product estimation, denoted by $\texttt{TurboQuant}_{\text{PROD}}$. A b-bit $\texttt{TurboQuant}_{\text{PROD}}$ first applies a $(b-1)$-bit $\texttt{TurboQuant}_{\text{MSE}}$ quantizer to obtain a raw reconstruction $\bar{\mathbf{x}}$, and then allocates the remaining one bit to a QJL-based correction for the residual $\mathbf{x} - \bar{\mathbf{x}}$ [27]. The final estimator combines the inner product with the raw reconstruction and an unbiased one-bit estimate of the residual contribution. This residual correction makes the resulting inner-product estimator unbiased.

B.1 Existing Guarantees on EDEN, RabitQ, and TurboQuant

The following proposition is performance guarantees of EDEN.

Proposition B.1 (MSE bound of $\texttt{EDEN}_{\text{UB}}$, Theorem 2.3 of Vargaftik et al. [22])**.** *For all unit input $\mathbf{x} \in \mathbb{R}^d$, $\texttt{EDEN}_{\text{UB}}$ satisfies:*

$$\mathcal{D}_{MSE}(\texttt{EDEN}_{\text{UB}}) \leq \frac{1}{\mathbb{E}\left[\left(Q(z)\right)^2\right]} - 1 + O\left(\sqrt{\frac{\log d}{d}}\right),$$

where $z \sim \mathcal{N}(0,1)$ and Q is Lloyd-max quantizer for standard normal distribution.

The following proposition is performance guarantees of RabitQ.

Proposition B.2 (Bit complexity of RabitQ, Theorem 3.2 of Gao et al. [7])**.** *For $\epsilon,\ \delta > 0$, assume that $\frac{1}{\epsilon^2}\log\frac{1}{\delta} > d$ holds. Then, to ensure that the inner product error of the estimator is bounded by ϵ with the probability of at least $1-\delta$,* RabitQ *requires $b = \Theta\big(\log\big(\frac{1}{d\epsilon^2}\log\frac{1}{\delta}\big)\big)$.*

The following propositions are performance guarantees of TurboQuant.

Proposition B.3 (Expected distortion bound of $\texttt{TurboQuant}_{\text{MSE}}$, Theorem 1 of Zandieh et al. [26])**.** *The b-bit $\texttt{TurboQuant}_{\text{MSE}}$ achieves the following distortion rate:*

- $\mathcal{D}_{MSE}(\texttt{TurboQuant}_{\text{MSE}}) \lesssim \frac{\sqrt{3}\pi}{2}\cdot\frac{1}{4^b}$
- *For $b = 1,\ 2,\ 3,\ 4$, $\mathcal{D}_{MSE}(\texttt{TurboQuant}_{\text{MSE}}) \approx 0.36, 0.117, 0.03, 0.009$*

Proposition B.4 (Expected distortion bound of $\texttt{TurboQuant}_{\text{PROD}}$, Theorem 2 of Zandieh et al. [26])**.** *The b-bit $\texttt{TurboQuant}_{\text{PROD}}$ achieves the following distortion rate:*

- $\mathcal{D}_{IP}(\texttt{TurboQuant}_{\text{PROD}}) \lesssim \frac{\sqrt{3}\pi^2}{d}\cdot\frac{1}{4^b}$
- *For $b = 1,\ 2,\ 3,\ 4$, $\mathcal{D}_{IP}(\texttt{TurboQuant}_{\text{PROD}}) \approx \frac{1.57}{d}, \frac{0.56}{d}, \frac{0.18}{d}, \frac{0.047}{d}$.*

C High-Resolution Formula

In this section we briefly state high-resolution formula of high-rate quantization analysis [10, 25]. Let $\mathcal{M}$ be an m-dimensional smooth Riemannian manifold with volume measure $\mathrm{d}V_{\mathcal{M}}$, and let μ be a probability measure on $\mathcal{M}$ with density f with respect to $\mathrm{d}V_{\mathcal{M}}$.

Definition C.1 (Zador–Gersho constant)**.** *Let $\mathcal{T} \subset \mathcal{M}$ be a bounded measurable subset of M with positive volume, and let $\bar{\mathbf{x}}_{\mathcal{T}} := \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} \mathbf{x}\ \mathrm{d}V_{\mathcal{M}}(\mathbf{x})$ be its centroid. Its normalized second moment is*

$$G(\mathcal{T}) := \frac{1}{m\,|\mathcal{T}|^{1+2/m}}\int_{\mathcal{T}} \|u-\bar{u}_{\mathcal{T}}\|_2^2\ \mathrm{d}V_{\mathcal{M}}(\mathbf{x}).$$

Then, the Zador–Gersho constant $G^{\star}_{\mathcal{M}}$ is defined as the infimum of $G(\mathcal{T})$ over all tessellating m-dimensional cells:

$$G^{\star}_{\mathcal{M}} := \inf\{G(\mathcal{T}) : \mathcal{T} \text{ tiles } \mathcal{M} \text{ by translations}\}.$$

Definition C.2 (Source factor)**.** *The source factor $J_{\mathcal{M}}$ is defined as*

$$J_{\mathcal{M}} := \left(\int_{\mathcal{M}} f(\mathbf{x})^{\frac{m}{m+2}}\ \mathrm{d}V_{\mathcal{M}}(\mathbf{x})\right)^{\frac{m+2}{m}}.$$

Proposition C.1 (General high-rate Zador–Gersho formula)**.** *Let $\mathcal{D}^*_{MSE}(K)$ be the optimal K-point quantization distortion for $\mathcal{M}$, i.e.*

$$\mathcal{D}^*_{MSE}(K) := \inf_{\mathcal{C}\subset\mathbb{R}^d,\,|\mathcal{C}|\le K}\int_{\mathcal{M}}\min_{c\in\mathcal{C}}\|\mathbf{x}-c\|_2^2 f(\mathbf{x})\ \mathrm{d}V_{\mathcal{M}}(\mathbf{x}).$$

Then, the leading-order optimal distortion is

$$\mathcal{D}^*_{MSE}(K) = \left(G^*_{\mathcal{M}} + o(1)\right) J_{\mathcal{M}} K^{-2/m}, \qquad K\to\infty. \tag{C.1}$$

The Panter–Dite high-resolution formula [18] used for scalar fixed-rate quantization is the $m=1$ specialization of eq. C.1. For a one-dimensional density f and a scalar quantizer with K levels,

$$\mathcal{D}^*_{\text{MSE}}(K) = \left(\frac{1}{12}+o(1)\right)\left(\int f(x)^{1/3}\ \mathrm{d}x\right)^3 K^{-2}, \qquad K\to\infty.$$

D MSE Analysis

D.1 Proof of Proposition 1

D.1.1 Proof for Small $b = 1,\ 2,\ 3,\ 4$

Proof. Let $\mathbf{z} = R\mathbf{x}$, let $R_j := \sqrt{d}\,z_j$, and set $K = 2^b$. Let $Q_b : \mathbb{R}\to\{q_1,\dots,q_K\}$ be the K-level Lloyd–Max scalar quantizer for $Z\sim N(0,1)$. Thus there are thresholds $-\infty = t_0 <$

$t_1 < \cdots < t_K = \infty$ such that $Q_b(r) = q_i$ for $r \in [t_{i-1}, t_i)$, $q_i = \mathbb{E}[Z \mid Z \in [t_{i-1}, t_i)]$, and $t_i = (q_i + q_{i+1})/2$. Set $\bar{\mathbf{z}} := d^{-1/2}(Q_b(R_1), \ldots, Q_b(R_d))$ and $\bar{\mathbf{x}} := R^\top \bar{\mathbf{z}}$. Since R is orthogonal, $\rho_d := \langle \mathbf{x}, \bar{\mathbf{x}} \rangle = \langle \mathbf{z}, \bar{\mathbf{z}} \rangle = d^{-1} \sum_{j=1}^d R_j Q_b(R_j)$ and $\psi_d^2 := \|\bar{\mathbf{x}}\|_2^2 = d^{-1} \sum_{j=1}^d Q_b(R_j)^2$.

The dequantized output of $\texttt{EDEN}_{\text{BSM}}$ uses the reconstruction-optimal scalar $\alpha^\star = \rho_d/\psi_d^2$. Therefore, for each realization of the random rotation,

$$\|\mathbf{x} - \alpha^\star \bar{\mathbf{x}}\|_2^2 = \min_{\alpha \in \mathbb{R}} \|\mathbf{x} - \alpha \bar{\mathbf{x}}\|_2^2 = 1 - \frac{\rho_d^2}{\psi_d^2}. \tag{D.1}$$

For fixed b, the empirical averages in ρ_d and ψ_d^2 are approximated in high dimension by their Gaussian counterparts:

$$\rho_d = \mathbb{E}[Z Q_b(Z)] + o_d(1), \qquad \psi_d^2 = \mathbb{E}[Q_b(Z)^2] + o_d(1).$$

By the Lloyd–Max centroid condition,

$$\mathbb{E}[Z \mid Q_b(Z)] = Q_b(Z), \qquad \mathbb{E}[Z Q_b(Z)] = \mathbb{E}[Q_b(Z)^2] =: m_b.$$

Substituting these approximations into (D.1) gives

$$\mathbb{E}_R \left[\|\mathbf{x} - \alpha^\star \bar{\mathbf{x}}\|_2^2 \right] \approx 1 - \frac{m_b^2}{m_b} = 1 - m_b.$$

Equivalently, if $e_b := \mathbb{E}\left[(Z - Q_b(Z))^2\right]$, then the centroid condition also gives $e_b = 1 - m_b$, and hence the high-dimensional approximate MSE of $\texttt{EDEN}_{\text{BSM}}$ is e_b (which is also the approximate MSE of $\texttt{TurboQuant}_{\text{MSE}}$). For $b = 1$, $Q_1(Z) = \sqrt{2/\pi}\,\text{sgn}(Z)$, so $m_1 = 2/\pi$ and $e_1 = 1 - 2/\pi \approx 0.363380$. For $b = 2, 3, 4$, solving the Lloyd–Max equations gives the following scalar distortions.

b	$K = 2^b$	$m_b = \mathbb{E}[Q_b(Z)^2]$	$e_b = 1 - m_b$
1	2	0.6366197724	0.3633802276
2	4	0.8825181522	0.1174818478
3	8	0.9654522392	0.0345477608
4	16	0.9904989920	0.0095010080

Thus, in high dimensions,

$$\mathcal{D}_{\text{MSE}}(\texttt{EDEN}_{\text{BSM}}) \approx 0.363,\ 0.117,\ 0.0345,\ 0.0095$$

for $b = 1,\ 2,\ 3,\ 4$, respectively. □

D.1.2 Proof for Large Bit-Width

Proof. Since $\texttt{EDEN}_{\text{BSM}}$ chooses the best scalar multiple of the raw reconstruction $\bar{\mathbf{x}}$, we have, for every realization of the random rotation,

$$\|\mathbf{x} - \widehat{\mathbf{x}}_{\texttt{EDEN}_{\text{BSM}}}\|_2^2 = \min_{\alpha \in \mathbb{R}} \|\mathbf{x} - \alpha \bar{\mathbf{x}}\|_2^2 \le \|\mathbf{x} - \bar{\mathbf{x}}\|_2^2.$$

In the high-dimensional approximation, the scaled rotated coordinates are asymptotically standard normal. Therefore, when $d \to \infty$,

$$\mathcal{D}_{\text{MSE}}(\texttt{EDEN}_{\text{BSM}}) \le \mathbb{E}[(Z - Q_b(Z))^2],$$

where $Z \sim N(0,1)$ and Q_b is the b-bit Gaussian Lloyd–Max quantizer. By the Panter–Dite high-rate formula (Appendix C),

$$\mathcal{D}_{\text{MSE}}(\texttt{EDEN}_{\text{BSM}}) \le \mathbb{E}[(Z - Q_b(Z))^2] \le \frac{\pi\sqrt{3}}{2} 4^{-b}(1 + o_b(1)),$$

Hence, in high-demensional, $\texttt{EDEN}_{\text{BSM}}$ and $\texttt{TurboQuant}_{\text{MSE}}$ has the same leading constant.

We note that if high-rate analysis based on the exact spherical marginal density

$$f_{1,d}(s) = \frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d-1)/2)} (1 - s^2)^{(d-3)/2}, \qquad -1 \le s \le 1,$$

is applied for $\texttt{EDEN}_{\text{BSM}}$, the coefficient of 4^{-b} would be $\frac{d+3}{d+6} \frac{\pi d}{12} \frac{\Gamma(d/2)}{\Gamma((d-1)/2)} \left[\frac{\Gamma((d+3)/6)}{\Gamma((d+6)/6)}\right]^3$, which converges to $\pi\sqrt{3}/2$ as $d \to \infty$. □

D.2 Proof of Proposition 2

Proof. Let $\bar{\mathbf{x}}$ be the selected RabitQ codeword with $\|\mathbf{x}\|_2 = 1$ and set $\rho := \langle \mathbf{x}, \bar{\mathbf{x}} \rangle$. The MSE-best scalar reconstruction along $\bar{\mathbf{x}}$ is $\hat{\mathbf{x}} := \rho \bar{\mathbf{x}}$, and since $\|\mathbf{x}\|_2 = \|\bar{\mathbf{x}}\|_2 = 1$,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \|\mathbf{x} - \rho\bar{\mathbf{x}}\|_2^2 = 1 - \rho^2. \tag{D.2}$$

Let $\mathcal{G}_b := \{-(2^b - 1)/2 + u : u = 0, 1, \ldots, 2^b - 1\}^d$. The unrotated RabitQ codebook is $\mathcal{C}_{\mathrm{RabitQ}} = \{\mathbf{g}/\|\mathbf{g}\|_2 : \mathbf{g} \in \mathcal{G}_b\}$, and the algorithm selects the codeword closest to $\mathbf{z}$. Therefore the squared angular error is

$$1 - \rho^2 = \min_{t>0,\, \mathbf{g} \in \mathcal{G}_b} \|\mathbf{z} - t\mathbf{g}\|_2^2.$$

For $\alpha = t\sqrt{d}$, this objective becomes

$$\|\mathbf{z} - t\mathbf{g}\|_2^2 = \sum_{j=1}^{d} (z_j - t g_j)^2 = \frac{1}{d} \sum_{j=1}^{d} (R_j - \alpha g_j)^2.$$

For fixed α, the minimizing g_j is the nearest grid point to R_j/α, namely $g_j = Q_b(R_j/\alpha)$. Combining this with (D.2) gives

$$\mathcal{D}_{\mathrm{MSE}}(\texttt{RabitQ}_{\mathrm{BSM}}) = \mathbb{E}[1 - \rho^2] = \mathbb{E}\left[\min_{\alpha > 0} \frac{1}{d} \sum_{j=1}^{d} \left(R_j - \alpha Q_b(R_j/\alpha)\right)^2 \right]. \tag{D.3}$$

For large d, the rescaled coordinates $R_j = \sqrt{d}\, z_j$ are approximately standard normal in the marginal sense, and empirical averages concentrate. Thus, for $Z \sim N(0,1)$, the right-hand side of (D.3) is approximated by $\min_{\alpha>0} \phi(\alpha)$, where $\phi(\alpha) := \mathbb{E}[(Z - \alpha Q_b(Z/\alpha))^2]$. If ζ denotes the standard normal density, symmetry gives

$$\phi(\alpha) = 2 \sum_{k=0}^{2^{b-1}-2} \int_{k\alpha}^{(k+1)\alpha} \left(r - \alpha \left(k + \frac{1}{2} \right) \right)^2 \zeta(r)\, dr + 2 \int_{(2^{b-1}-1)\alpha}^{\infty} \left(r - \alpha \left(2^{b-1} - \frac{1}{2} \right) \right)^2 \zeta(r)\, dr. \tag{D.4}$$

For $b = 1$, $Q_1(Z/\alpha) = \frac{1}{2}\operatorname{sgn}(Z)$, so $\phi(\alpha) = 1 - \alpha\sqrt{2/\pi} + \alpha^2/4$ and the minimum is $1 - 2/\pi \approx 0.363380$ at $\alpha = 2\sqrt{2/\pi}$. For $b = 2, 3, 4$, numerical minimization of (D.4) gives the following values.

b	minimizing α	$\min_{\alpha>0} \phi(\alpha)$
1	1.5957691216	0.3633802276
2	0.9956867007	0.1188460504
3	0.5860194285	0.0374396594
4	0.3352006088	0.0115428844

These are the stated Gaussian-approximation constants. □

D.3 Approximated MSE of TurboQuant

Let $\mathbf{z} = R\mathbf{x}$ be the randomly rotated input vector and define the rescaled coordinate $R_j := \sqrt{d}\, z_j$. TurboQuant applies coordinate-wise scalar quantization, where the 2^b centroids are chosen to minimize the one-dimensional MSE distortion. In the high-dimensional regime, each R_j is well approximated by $Z \sim N(0,1)$. Hence the b-bit scalar codebook can be computed from the Lloyd–Max problem

$$\mathcal{A}_b^{\mathrm{TQ}} := \arg \min_{\substack{\mathcal{A} \subset \mathbb{R} \\ |\mathcal{A}| = 2^b}} \mathbb{E}_{Z \sim N(0,1)} \left[\min_{a \in \mathcal{A}} (Z - a)^2 \right]$$

which is identical to the original codebook of EDEN [22]. If q_b denotes the nearest-centroid map associated with $\mathcal{A}_b^{\mathrm{TQ}}$, then the dequantized vector is

$$\bar{\mathbf{x}} \approx R^\top \frac{1}{\sqrt{d}} \big(q_b(R_1), \ldots, q_b(R_d) \big),$$

and concentration of empirical averages gives

$$\mathcal{D}_{\text{MSE}}(\texttt{TurboQuant}_{\text{MSE}}) = \mathbb{E}\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2 \approx \mathbb{E}_{Z\sim N(0,1)}\left[(Z - q_b(Z))^2\right]$$

Solving this scalar Lloyd–Max problem gives

$$\mathcal{D}_{\text{MSE}}(\texttt{TurboQuant}_{\text{MSE}}) \approx 0.363380,\ 0.117482,\ 0.034548,\ 0.009501,$$

for $b = 1,\ 2,\ 3,\ 4$, respectively.

E Inner Product Distortion Analysis

E.1 Proof of Theorem 1

Proof. We first record the isotropy supplied by the random rotation. Fix $\mathbf{x}$ with $\|\mathbf{x}\|_2 = 1$, and consider any orthogonal matrix U satisfying $U\mathbf{x} = \mathbf{x}$, i.e., any rotation in the stabilizer of $\mathbf{x}$. By Haar invariance, $RU^\top \overset{d}{=} R$. Moreover, since $U^\top\mathbf{x} = \mathbf{x}$, the rotated input is unchanged, $(RU^\top)\mathbf{x} = R\mathbf{x}$, while the final inverse rotation gives $\bar{\mathbf{x}}_{RU^\top} = U\bar{\mathbf{x}}_R$. Consequently, $\rho := \langle\bar{\mathbf{x}}, \mathbf{x}\rangle$ is unchanged, since $\langle U\bar{\mathbf{x}}_R, \mathbf{x}\rangle = \langle\bar{\mathbf{x}}_R, U^\top\mathbf{x}\rangle = \langle\bar{\mathbf{x}}_R, \mathbf{x}\rangle$, and $\psi := \|\bar{\mathbf{x}}\|_2$ is also unchanged. On the other hand, the residual component $\mathbf{v} := \bar{\mathbf{x}} - \rho\mathbf{x}$, which lies in $\mathbf{x}^\perp$, is transformed as $\mathbf{v} \mapsto U\mathbf{v}$. Thus, after conditioning on the scalar quantities (ρ, ψ), the only remaining randomness in $\mathbf{v}$ is its direction inside $\mathbf{x}^\perp$. Since the stabilizer of $\mathbf{x}$ acts transitively on directions in $\mathbf{x}^\perp$, the conditional distribution of $\mathbf{v}$ is directionally isotropic in $\mathbf{x}^\perp$.

Now decompose the reconstruction as $\bar{\mathbf{x}} = \rho\mathbf{x} + \mathbf{v}$, where $\mathbf{v} \perp \mathbf{x}$ by the definition of $\rho = \langle\bar{\mathbf{x}}, \mathbf{x}\rangle$. Since $\|\mathbf{x}\|_2 = 1$, this decomposition gives $\|\mathbf{v}\|_2^2 = \|\bar{\mathbf{x}}\|_2^2 - \rho^2 = \psi^2 - \rho^2$. Let $\phi := \langle\mathbf{x}, \mathbf{y}\rangle$ and assume $\rho > 0$. For the ratio estimator, we have $\widehat{\phi}_{\text{ratio}} = \langle\bar{\mathbf{x}}, \mathbf{y}\rangle/\rho$, and hence $\widehat{\phi}_{\text{ratio}} - \phi = \langle\bar{\mathbf{x}}, \mathbf{y}\rangle/\rho - \langle\mathbf{x}, \mathbf{y}\rangle = \langle\mathbf{v}, \mathbf{y}\rangle/\rho$. Writing $\mathbf{y}_\perp := \mathbf{y} - \phi\mathbf{x}$, we have $\mathbf{y}_\perp \in \mathbf{x}^\perp$ and $\|\mathbf{y}_\perp\|_2^2 = 1 - \phi^2$. Moreover, since $\mathbf{v} \perp \mathbf{x}$, $\langle\mathbf{v}, \mathbf{y}\rangle = \langle\mathbf{v}, \mathbf{y}_\perp\rangle$. By the conditional isotropy of $\mathbf{v}$ in $\mathbf{x}^\perp$, its conditional mean in every fixed direction of $\mathbf{x}^\perp$ is zero, and thus $\mathbb{E}[\langle\mathbf{v}, \mathbf{y}_\perp\rangle \mid \rho, \psi] = 0$. Therefore $\mathbb{E}[\widehat{\phi}_{\text{ratio}} \mid \rho, \psi] = \phi$, so $\widehat{\phi}_{\text{ratio}}$ is unbiased.

The same conditional isotropy implies that the squared length $\|\mathbf{v}\|_2^2 = \psi^2 - \rho^2$ is spread uniformly over the $d-1$ dimensions of $\mathbf{x}^\perp$, giving $\mathbb{E}[\langle\mathbf{v}, \mathbf{y}_\perp\rangle^2 \mid \rho, \psi] = (\psi^2 - \rho^2)\|\mathbf{y}_\perp\|_2^2/(d-1)$. Consequently,

$$\mathbb{E}\left[(\widehat{\phi}_{\text{ratio}} - \phi)^2\right] = \frac{1 - \phi^2}{d-1}\mathbb{E}\left[\frac{\psi^2 - \rho^2}{\rho^2}\right].$$

□

E.2 Proof of Corollary 1

E.2.1 Proof for Small $b = 1,\ 2,\ 3,\ 4$

Proof. Let $\mathbf{z} = R\mathbf{x}$, $R_j := \sqrt{d}\, z_j$, and let Q_b be the 2^b-level Lloyd–Max scalar quantizer for $Z \sim N(0, 1)$. As in the proof of Proposition 1, set $\bar{\mathbf{z}} := d^{-1/2}(Q_b(R_1), \ldots, Q_b(R_d))$ and $\bar{\mathbf{x}} := R^\top\bar{\mathbf{z}}$. Then $\rho_d := \langle\mathbf{x}, \bar{\mathbf{x}}\rangle = d^{-1}\sum_{j=1}^d R_j Q_b(R_j)$ and $\psi_d^2 := \|\bar{\mathbf{x}}\|_2^2 = d^{-1}\sum_{j=1}^d Q_b(R_j)^2$.

The $\texttt{EDEN}_{\text{UB}}$ estimator is $\widehat{\eta}_{\text{ratio}} := \langle\bar{\mathbf{x}}, \mathbf{y}\rangle/\rho_d$. Applying Theorem 1 with $\rho = \rho_d$ and $\psi = \psi_d$, for $\eta = \langle\mathbf{x}, \mathbf{y}\rangle$ we get

$$\mathbb{E}\left[(\widehat{\eta}_{\text{ratio}} - \eta)^2\right] = \frac{1 - \eta^2}{d-1}\mathbb{E}\left[\frac{\psi_d^2}{\rho_d^2} - 1\right].$$

Taking the supremum over $\mathbf{y} \in \mathbb{S}^{d-1}$ gives $\mathcal{D}_{\text{IP}}(\texttt{EDEN}_{\text{UB}}) \le (d-1)^{-1}(\mathbb{E}[\psi_d^2/\rho_d^2] - 1)$ (= holds when $\eta = 0$). The concentration argument used in Proposition 1 gives $\mathbb{E}[\psi_d^2/\rho_d^2] \le m_b^{-1} + O(\sqrt{\log d/d})$, where $m_b := \mathbb{E}[Q_b(Z)^2]$. Hence

$$\mathcal{D}_{\text{IP}}(\texttt{EDEN}_{\text{UB}}) \le \frac{1}{d-1}\left(\frac{1}{m_b} - 1 + O\left(\sqrt{\frac{\log d}{d}}\right)\right).$$

For Lloyd–Max centroids, $\mathbb{E}[Z \mid Q_b(Z)] = Q_b(Z)$, so $\mathbb{E}[ZQ_b(Z)] = m_b$. Thus, with $e_b := \mathbb{E}[(Z - Q_b(Z))^2]$, we have $e_b = 1 - m_b$ and $m_b^{-1} - 1 = e_b/(1 - e_b) =: B_b^{\text{EDEN}}$. From the proof of Proposition 1, the constants are

b	e_b	$m_b = 1 - e_b$	$B_b^{\mathrm{EDEN}} = e_b/(1-e_b)$
1	0.3633802276	0.6366197724	0.5707963268
2	0.1174818478	0.8825181522	0.1331211687
3	0.0345477608	0.9654522392	0.0357840185
4	0.0095010080	0.9904989920	0.0095921430

This proves the stated small-bit inner-product bounds for $\texttt{EDEN}_{\mathrm{UB}}$. □

E.2.2 Proof for Large Bit-Width

Proof. Let $\bar{\mathbf{x}}$ be the unscaled coordinate-wise high-rate reconstruction used in the proof of Proposition 1. The $\texttt{EDEN}_{\mathrm{UB}}$ estimator is $\widehat{\eta}_{\texttt{EDEN}_{\mathrm{UB}}} := \langle \bar{\mathbf{x}}, \mathbf{y}\rangle / \langle \bar{\mathbf{x}}, \mathbf{x}\rangle$. Write $\mathbf{e} := \bar{\mathbf{x}} - \mathbf{x}$, $t := \langle \mathbf{x}, \mathbf{e}\rangle$, and $\mathbf{u} := \mathbf{e} - t\mathbf{x}$. Then $\mathbf{u} \perp \mathbf{x}$, $\bar{\mathbf{x}} = (1+t)\mathbf{x} + \mathbf{u}$, $\rho_d := \langle \bar{\mathbf{x}}, \mathbf{x}\rangle = 1 + t$, and $\psi_d^2 - \rho_d^2 = \|\mathbf{u}\|_2^2$.

By Theorem 1, after taking the supremum over $\mathbf{y} \in \mathbb{S}^{d-1}$,

$$\mathcal{D}_{\mathrm{IP}}(\texttt{EDEN}_{\mathrm{UB}}) = \frac{1}{d-1}\mathbb{E}\left[\frac{\psi_d^2 - \rho_d^2}{\rho_d^2}\right] = \frac{1}{d-1}\mathbb{E}\left[\frac{\|\mathbf{u}\|_2^2}{(1+t)^2}\right].$$

Thus the ratio correction removes the radial component from the numerator exactly, so the leading contribution comes only from $\|\mathbf{u}\|_2^2$. In the high-rate regime, $\mathbb{E}\|\mathbf{e}\|_2^2 = O(4^{-b})$ and $\mathbb{E}\|\mathbf{e}\|_2^4 = O(4^{-2b})$. Since $|t| \le \|\mathbf{e}\|_2$ and $\|\mathbf{u}\|_2 \le \|\mathbf{e}\|_2$, Hölder's inequality gives $\mathbb{E}\|\mathbf{u}\|_2^2|t| \le \mathbb{E}\|\mathbf{e}\|_2^3 \le (\mathbb{E}\|\mathbf{e}\|_2^4)^{3/4} = o(4^{-b})$. Hence the denominator $(1+t)^2$ only changes the expression by a lower-order term, and therefore $\mathbb{E}[\|\mathbf{u}\|_2^2/(1+t)^2] = \mathbb{E}\|\mathbf{u}\|_2^2 + o(4^{-b})$.

It remains to recall the tangential energy from the high-rate MSE proof. Let $f_{1,d}$ be the one-dimensional marginal density of a coordinate of a uniform vector on $\mathbb{S}^{d-1}$, and define $L_d := \int_{-1}^{1} f_{1,d}(s)^{1/3}\,ds$ and $M_d := \int_{-1}^{1} s^2 f_{1,d}(s)^{1/3}\,ds$. The Panter–Dite formula gives the leading coordinate-wise high-rate distortion, and the radial calculation applies the same local error expansion to the component along $\mathbf{x}$. Thus,

$$\mathbb{E}[\|\mathbf{e}\|_2^2] = \frac{d}{12}L_d^3 4^{-b}(1+o(1)), \qquad \mathbb{E}[t^2] = \frac{d}{12}L_d^2 M_d 4^{-b}(1+o(1)).$$

Since $\mathbf{e} = t\mathbf{x} + \mathbf{u}$ with $\mathbf{u} \perp \mathbf{x}$, the tangential energy is obtained by subtracting the radial energy: $\mathbb{E}[\|\mathbf{u}\|_2^2] = \mathbb{E}[\|\mathbf{e}\|_2^2] - \mathbb{E}[t^2] = \frac{d}{12}L_d^2(L_d - M_d)4^{-b}(1+o(1))$. Also, because $f_{1,d}(s)^{1/3}$ is proportional to $(1-s^2)^{(d-3)/6}$,

$$\frac{M_d}{L_d} = \frac{\int_{-1}^{1} s^2(1-s^2)^{(d-3)/6}\,ds}{\int_{-1}^{1}(1-s^2)^{(d-3)/6}\,ds} = \frac{\mathrm{B}(3/2, (d+3)/6)}{\mathrm{B}(1/2, (d+3)/6)} = \frac{3}{d+6}.$$

Thus $\mathbb{E}[\|\mathbf{u}\|_2^2] = \frac{d+3}{d+6}\frac{d}{12}L_d^3 4^{-b}(1+o(1))$. Using $L_d^3 = \pi\frac{\Gamma(d/2)}{\Gamma((d-1)/2)}\left[\frac{\Gamma((d+3)/6)}{\Gamma((d+6)/6)}\right]^3$, we obtain

$$\mathcal{D}_{\mathrm{IP}}(\texttt{EDEN}_{\mathrm{UB}}) \le \frac{1}{d-1}\frac{d+3}{d+6}\frac{\pi d}{12}\frac{\Gamma(d/2)}{\Gamma((d-1)/2)}\left[\frac{\Gamma((d+3)/6)}{\Gamma((d+6)/6)}\right]^3 4^{-b}(1+o(1)).$$

Finally, $\Gamma(x+a)/\Gamma(x+b) \sim x^{a-b}$ implies that the coefficient multiplying $4^{-b}/(d-1)$ converges to $\pi\sqrt{3}/2 \approx 2.721$. This proves the high-rate inner-product bound for $\texttt{EDEN}_{\mathrm{UB}}$. □

E.3 Proof of Corollary 2

Proof. The factor inside the expectation in Theorem 1 is nonnegative and independent of $\mathbf{y}$, so the worst case over $\mathbf{y} \in \mathbb{S}^{d-1}$ is attained when $\phi = 0$. For a fixed $\mathbf{x}$,

$$\mathcal{D}_{\mathrm{IP}}(\texttt{RabitQ}) = \frac{1}{d-1}\mathbb{E}_Q\left[\frac{\|\bar{\mathbf{x}}\|_2^2 - \langle \bar{\mathbf{x}}, \mathbf{x}\rangle^2}{\langle \bar{\mathbf{x}}, \mathbf{x}\rangle^2}\right],$$

and rotation invariance makes this quantity independent of the particular $\mathbf{x} \in \mathbb{S}^{d-1}$.

The codewords of RabitQ are normalized, so $\|\bar{\mathbf{x}}\|_2 = 1$. Let $\rho := \langle \bar{\mathbf{x}}, \mathbf{x} \rangle$ and $\Delta := 1 - \rho^2$. As in the proof of Proposition 2, $\Delta = \|\mathbf{x} - \rho\bar{\mathbf{x}}\|_2^2$ is the reconstruction error of $\texttt{RabitQ}_{\text{BSM}}$. Hence

$$\mathcal{D}_{\text{IP}}(\texttt{RabitQ}) = \frac{1}{d-1}\mathbb{E}_Q\left[\frac{\Delta}{1-\Delta}\right]. \tag{E.1}$$

Let $\mathbf{z} = R\mathbf{x}$ and $R_j := \sqrt{d}\, z_j$. For $Q_b(u) := \text{sign}(u)\min(\lfloor |u| \rfloor + 1/2, 2^{b-1} - 1/2)$, the same MSE proof gives

$$\Delta = \min_{\alpha > 0} \frac{1}{d}\sum_{j=1}^{d} \left(R_j - \alpha Q_b(R_j/\alpha)\right)^2. \tag{E.2}$$

We now use the same high-dimensional Gaussian approximation as in the proof of Proposition 2. Let $Z \sim N(0,1)$ with density ζ. For fixed b, the Gaussian limit of the empirical objective in (E.2) is

$$\begin{aligned}\phi(\alpha) &:= \mathbb{E}\left[(Z - \alpha Q_b(Z/\alpha))^2\right] \\ &= 2\sum_{k=0}^{2^{b-1}-2} \int_{k\alpha}^{(k+1)\alpha} \left(r - \alpha\left(k + \frac{1}{2}\right)\right)^2 \zeta(r)\, dr + 2\int_{(2^{b-1}-1)\alpha}^{\infty} \left(r - \alpha\left(2^{b-1} - \frac{1}{2}\right)\right)^2 \zeta(r)\, dr,\end{aligned} \tag{E.3}$$

where the summation term is empty when $b = 1$. Define $\Delta_z := \min_{\alpha>0}\phi(\alpha)$ and $\kappa_b^{\text{RQ}} := \Delta_z/(1-\Delta_z)$. By concentration of the empirical averages, for fixed b and $d \to \infty$, $\Delta = \Delta_z + o_{\mathbb{P}}(1)$. Since $\Delta_z < 1$ for the bit-widths considered here and the denominator stays bounded away from zero with high probability, $\mathbb{E}_Q[\Delta/(1-\Delta)] = \kappa_b^{\text{RQ}} + o(1)$. Equation (E.1) therefore gives

$$\mathcal{D}_{\text{IP}}(\texttt{RabitQ}) = \frac{\kappa_b^{\text{RQ}} + o(1)}{d-1}.$$

The minimization in (E.3), computed in Proposition 2, gives

b	Optimal value of α	Δ_z	$\kappa_b^{\text{RQ}} = \Delta_z/(1-\Delta_z)$
1	1.595769	0.363380	0.570796
2	0.995687	0.118846	0.134875
3	0.586019	0.037440	0.038896
4	0.335201	0.011543	0.011678

Thus, for $b = 1, 2, 3, 4$, the leading constants are $0.570796/(d-1)$, $0.134875/(d-1)$, $0.038896/(d-1)$, and $0.011678/(d-1)$, respectively. □

F Bit Complexity Analysis

We prove the bit-complexity guarantees for $\texttt{EDEN}_{\text{UB}}$ and $\texttt{TurboQuant}_{\text{PROD}}$ separately in Sections F.1 and F.2. Each subsection contains the corresponding formal statement, namely Theorems F.1 and F.2, the high-probability residual bound, and the proof.

F.1 Bit Complexity of EDEN

Theorem F.1 (Bit complexity of $\texttt{EDEN}_{\text{UB}}$ (formal))**.** *Let $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ and let $0 < \epsilon, \delta < 1$. There exist universal constants $C, c > 0$ such that, if $\log(4/\delta) \le cd$, then for every bit-width $b \ge 1$,*

$$\mathbb{P}_Q\left[\left|\text{IP}_{\text{EDEN}}^{(b)}(\mathbf{x},\mathbf{y}) - \langle \mathbf{x}, \mathbf{y}\rangle\right| > C\left(2^{-b}\sqrt{\frac{\log(4/\delta)}{d}} + \frac{\log(4/\delta)}{d}\right)\right] \le \delta.$$

Consequently, whenever $\epsilon \ge C\log(4/\delta)/d$, to ensure

$$\mathbb{P}_Q\left[\left|\text{IP}_{\text{EDEN}}^{(b)}(\mathbf{x},\mathbf{y}) - \langle \mathbf{x}, \mathbf{y}\rangle\right| > \epsilon\right] \le \delta,$$

it is sufficient to take

$$b \ge \max\left\{1, \left\lceil \frac{1}{2}\log_2\left(\frac{C\log(4/\delta)}{d\epsilon^2}\right)\right\rceil\right\}$$

bits per dimension.

F.1.1 High probability residual bound

Lemma F.1 (High probability residual bound EDEN)**.** *Assume that the unscaled b-bit* EDEN *reconstruction satisfies*

$$\max_{\mathbf{x}\in\mathbb{S}^{d-1}} \mathbb{E}_Q\left[\|\mathbf{x}-\bar{\mathbf{x}}\|_2^2\right] \le C_{\mathrm{ed}} 4^{-b}.$$

Let $\mathbf{x},\mathbf{y}\in\mathbb{S}^{d-1}$ be fixed, and let $\bar{\mathbf{x}}$ be the unscaled dequantized codeword produced by the b-bit EDEN *quantizer. Then, for every $0<\delta<1$,*

$$\mathbb{P}_Q\left[\|\mathbf{x}-\bar{\mathbf{x}}\|_2 > \sqrt{C_{\mathrm{ed}}}\,2^{-b} + C_{\mathrm{levy}}\sqrt{\frac{\log(4/\delta)}{d}}\right] \le \frac{\delta}{2}, \tag{F.1}$$

where $C_{\mathrm{levy}}>0$ is the universal constant from Lévy's concentration.

Proof. Fix b,d and $\mathbf{x}\in\mathbb{S}^{d-1}$. Let R be the random rotation, write $\mathbf{z}=R\mathbf{x}$, and let $\mathcal{C}_{\mathrm{EDEN}}^{(b)}$ be the fixed rotated-coordinate product codebook after the $1/\sqrt{d}$ scaling. If $\bar{\mathbf{z}}\in\arg\min_{\mathbf{o}_i\in\mathcal{C}_{\mathrm{EDEN}}^{(b)}}\|\mathbf{z}-\mathbf{o}_i\|_2$ and $\bar{\mathbf{x}}=R^\top\bar{\mathbf{z}}$, then $\|\mathbf{x}-\bar{\mathbf{x}}\|_2=\|\mathbf{z}-\bar{\mathbf{z}}\|_2$ by orthogonality of R. Define $f(\mathbf{z}):=\mathrm{dist}(\mathbf{z},\mathcal{C}_{\mathrm{EDEN}}^{(b)})=\min_{\mathbf{o}_i\in\mathcal{C}_{\mathrm{EDEN}}^{(b)}}\|\mathbf{z}-\mathbf{o}_i\|_2$, so $f(\mathbf{z})=\|\mathbf{x}-\bar{\mathbf{x}}\|_2$.

The distance-to-a-set map is 1-Lipschitz: for any $\mathbf{z},\mathbf{z}'\in\mathbb{S}^{d-1}$, $f(\mathbf{z})-f(\mathbf{z}')\le\|\mathbf{z}-\mathbf{o}_{i^\star(\mathbf{z}')}\|_2-\|\mathbf{z}'-\mathbf{o}_{i^\star(\mathbf{z}')}\|_2\le\|\mathbf{z}-\mathbf{z}'\|_2$, and reversing $\mathbf{z},\mathbf{z}'$ gives $|f(\mathbf{z})-f(\mathbf{z}')|\le\|\mathbf{z}-\mathbf{z}'\|_2$. Set $L=\log(4/\delta)$. By Lévy's concentration inequality, after increasing C_{levy} if necessary,

$$\mathbb{P}_Q\left[f(\mathbf{z})>\mathbb{E}_Q f(\mathbf{z})+C_{\mathrm{levy}}\sqrt{\frac{L}{d}}\right]\le\frac{\delta}{2}.$$

Moreover, Jensen's inequality and the assumed MSE bound give $\mathbb{E}_Q f(\mathbf{z})\le(\mathbb{E}_Q f(\mathbf{z})^2)^{1/2}=(\mathbb{E}_Q\|\mathbf{x}-\bar{\mathbf{x}}\|_2^2)^{1/2}\le\sqrt{C_{\mathrm{ed}}}\,2^{-b}$. Combining the two estimates proves (F.1). □

F.1.2 Proof of Theorem F.1

Fix $b\ge1$. Let $\bar{\mathbf{x}}$ be the unscaled b-bit EDEN codeword, set $\rho=\langle\mathbf{x},\bar{\mathbf{x}}\rangle$, and define $\widehat{\mathbf{x}}_{\mathrm{EDEN}}^{(b)}=\bar{\mathbf{x}}/\rho$. The $\mathtt{EDEN}_{\mathrm{UB}}$ estimator is $\mathrm{IP}_{\mathrm{EDEN}}^{(b)}(\mathbf{x},\mathbf{y}):=\langle\widehat{\mathbf{x}}_{\mathrm{EDEN}}^{(b)},\mathbf{y}\rangle=\langle\bar{\mathbf{x}},\mathbf{y}\rangle/\rho$. Let $\eta=\langle\mathbf{x},\mathbf{y}\rangle$, $\mathbf{v}=\bar{\mathbf{x}}-\rho\mathbf{x}$, and $\mathbf{y}_\perp=\mathbf{y}-\eta\mathbf{x}$. Then $\mathbf{v}\perp\mathbf{x}$, $\mathbf{y}_\perp\perp\mathbf{x}$, and

$$\mathrm{IP}_{\mathrm{EDEN}}^{(b)}(\mathbf{x},\mathbf{y})-\langle\mathbf{x},\mathbf{y}\rangle=\frac{\langle\mathbf{v},\mathbf{y}\rangle}{\rho}=\frac{\langle\mathbf{v},\mathbf{y}_\perp\rangle}{\rho}.$$

We next record the isotropy supplied by the Haar rotation. Write the EDEN output as $\bar{\mathbf{x}}_R=R^\top Q_0(R\mathbf{x})$, where Q_0 is the fixed rotated-coordinate scalar quantize–dequantize map. If $U\mathbf{x}=\mathbf{x}$, then $RU^\top\overset{d}{=}R$ and $\bar{\mathbf{x}}_{RU^\top}=(RU^\top)^\top Q_0(RU^\top\mathbf{x})=U\bar{\mathbf{x}}_R$. Hence ρ and $\|\mathbf{v}\|_2$ are invariant, while $\mathbf{v}$ is transformed into $U\mathbf{v}$. Therefore, conditional on ρ and $\|\mathbf{v}\|_2$, the direction of $\mathbf{v}$ is rotationally invariant in $\mathbf{x}^\perp$.

Set $L=\log(4/\delta)$ and $\alpha_b:=\sqrt{C_{\mathrm{ed}}}\,2^{-b}+C_{\mathrm{levy}}\sqrt{L/d}$. By Lemma F.1, with probability at least $1-\delta/2$, $\|\mathbf{x}-\bar{\mathbf{x}}\|_2\le\alpha_b$. The EDEN MSE bound from the previous section gives a universal constant $C_{\mathrm{ed}}<4$ for the unscaled reconstruction. Hence, by choosing the universal constant c small enough, $L\le cd$ implies $\alpha_b\le\alpha_0$ for some fixed $\alpha_0<1$ and all $b\ge1$. On this residual event, $\|\mathbf{v}\|_2\le\|\bar{\mathbf{x}}-\mathbf{x}\|_2\le\alpha_b$, and since $\|\bar{\mathbf{x}}-\mathbf{x}\|_2^2=\|\bar{\mathbf{x}}\|_2^2+1-2\rho$ with $\|\bar{\mathbf{x}}\|_2^2\ge0$, we also have $\rho\ge(1-\alpha_b^2)/2\ge(1-\alpha_0^2)/2=:c_\rho>0$.

Conditional on ρ and $\|\mathbf{v}\|_2$, write $\mathbf{v}=\|\mathbf{v}\|_2\boldsymbol{\theta}$ with $\boldsymbol{\theta}$ uniform on the unit sphere in $\mathbf{x}^\perp$. For fixed $\mathbf{y}_\perp$, Lévy's concentration on this $(d-2)$-dimensional sphere gives universal constants $C_{\mathrm{sph}},c_{\mathrm{sph}}>0$ such that, whenever $L\le c_{\mathrm{sph}}d$,

$$\mathbb{P}_Q\left[|\langle\mathbf{v},\mathbf{y}_\perp\rangle|>C_{\mathrm{sph}}\|\mathbf{v}\|_2\|\mathbf{y}_\perp\|_2\sqrt{\frac{L}{d}}\;\middle|\;\rho,\|\mathbf{v}\|_2\right]\le\frac{\delta}{2}.$$

Since $\|\mathbf{y}_\perp\|_2 \le 1$, a union bound yields, with probability at least $1-\delta$,

$$\left|\mathrm{IP}^{(b)}_{\mathrm{EDEN}}(\mathbf{x},\mathbf{y}) - \langle \mathbf{x},\mathbf{y}\rangle\right| \le \frac{C_{\mathrm{sph}}}{c_\rho}\alpha_b\sqrt{\frac{L}{d}} \le C\left(2^{-b}\sqrt{\frac{L}{d}} + \frac{L}{d}\right).$$

This proves the high-probability bound after taking $c \le c_{\mathrm{sph}}$ and replacing L by $\log(4/\delta)$.

It remains to choose b for target accuracy ϵ. If $\epsilon \ge C_0 L/d$ with $C_0 \ge 2C$, then $CL/d \le \epsilon/2$. If

$$b \ge \max\left\{1, \left\lceil \frac{1}{2}\log_2\left(\frac{C_1 L}{d\epsilon^2}\right)\right\rceil\right\},$$

then, after increasing C_1 if necessary, $C2^{-b}\sqrt{L/d} \le \epsilon/2$: if $C_1 L/(d\epsilon^2) > 1$, this follows from the lower bound on b, while if $C_1 L/(d\epsilon^2) \le 1$, it follows from $\sqrt{L/d} \le \epsilon/\sqrt{C_1}$ and $2^{-b} \le 1$. Thus the error is at most ϵ with probability at least $1-\delta$, and enlarging the theorem constant C to dominate C_0, C_1 completes the proof.

F.2 Bit Complexity of $\texttt{TurboQuant}_{\mathrm{PROD}}$

Theorem F.2 (Bit complexity of $\texttt{TurboQuant}_{\mathrm{PROD}}$ (formal))**.** *Let $\mathbf{x},\mathbf{y} \in \mathbb{S}^{d-1}$ and let $0 < \epsilon, \delta < 1$. There exist universal constants $C, c > 0$ such that, if $\log(4/\delta) \le cd$, then for every bit-width $b \ge 1$,*

$$\mathbb{P}_Q\left[\left|\mathrm{IP}^{(b)}_{\mathrm{TQ}}(\mathbf{x},\mathbf{y}) - \langle \mathbf{x},\mathbf{y}\rangle\right| > C\left(2^{-(b-1)}\sqrt{\frac{\log(4/\delta)}{d}} + \frac{\log(4/\delta)}{d}\right)\right] \le \delta.$$

Consequently, whenever $\epsilon \ge C\log(4/\delta)/d$, to ensure

$$\mathbb{P}_Q\left[\left|\mathrm{IP}^{(b)}_{\mathrm{TQ}}(\mathbf{x},\mathbf{y}) - \langle \mathbf{x},\mathbf{y}\rangle\right| > \epsilon\right] \le \delta,$$

it is sufficient to take

$$b \ge 1 + \max\left\{0, \left\lceil \frac{1}{2}\log_2\left(\frac{C\log(4/\delta)}{d\epsilon^2}\right)\right\rceil\right\}$$

bits per dimension.

F.2.1 High probability residual bound

We first introduce a high-probability bound on the residual norm of $\texttt{TurboQuant}_{\mathrm{MSE}}$, obtained by applying Lévy's concentration inequality.

Lemma F.2 (High probability residual bound $\texttt{TurboQuant}_{\mathrm{MSE}}$)**.** *Assume that the b-bit $\texttt{TurboQuant}_{\mathrm{MSE}}$ reconstruction satisfies*

$$\max_{\mathbf{x}\in\mathbb{S}^{d-1}} \mathbb{E}_Q\left[\|\mathbf{x}-\bar{\mathbf{x}}\|_2^2\right] \le C_{\mathrm{mse}} 4^{-b}.$$

Let $\mathbf{x} \in \mathbb{S}^{d-1}$ be fixed, and let $\bar{\mathbf{x}}$ be the corresponding dequantized codeword. Then, for every $0 < \delta < 1$,

$$\mathbb{P}_Q\left[\|\mathbf{x}-\bar{\mathbf{x}}\|_2 > \sqrt{C_{\mathrm{mse}}}\,2^{-b} + C_{\mathrm{levy}}\sqrt{\frac{\log(2/\delta)}{d}}\right] \le \delta, \tag{F.2}$$

where $C_{\mathrm{levy}} > 0$ is a universal constant from Lévy's concentration.

Proof. Fix b, d and $\mathbf{x} \in \mathbb{S}^{d-1}$. Let R be the random rotation, write $\mathbf{z} = R\mathbf{x}$, and let $\mathcal{C}^{(b)}_{\mathrm{TQ}}$ be the fixed rotated-coordinate $\texttt{TurboQuant}_{\mathrm{MSE}}$ codebook. If $\bar{\mathbf{z}} \in \arg\min_{\mathbf{o}_i \in \mathcal{C}^{(b)}_{\mathrm{TQ}}} \|\mathbf{z}-\mathbf{o}_i\|_2$ and $\bar{\mathbf{x}} = R^\top \bar{\mathbf{z}}$, then $\|\mathbf{x}-\bar{\mathbf{x}}\|_2 = \|\mathbf{z}-\bar{\mathbf{z}}\|_2$. Define $f(\mathbf{z}) := \mathrm{dist}(\mathbf{z}, \mathcal{C}^{(b)}_{\mathrm{TQ}}) = \min_{\mathbf{o}_i \in \mathcal{C}^{(b)}_{\mathrm{TQ}}} \|\mathbf{z}-\mathbf{o}_i\|_2$, so $f(\mathbf{z}) = \|\mathbf{x}-\bar{\mathbf{x}}\|_2$.

As above, f is 1-Lipschitz because $f(\mathbf{z}) - f(\mathbf{z}') \le \|\mathbf{z}-\mathbf{o}_{i^\star(\mathbf{z}')}\|_2 - \|\mathbf{z}'-\mathbf{o}_{i^\star(\mathbf{z}')}\|_2 \le \|\mathbf{z}-\mathbf{z}'\|_2$, and the reverse inequality follows by swapping $\mathbf{z}, \mathbf{z}'$. Hence Lévy's concentration gives $\mathbb{P}_Q[f(\mathbf{z}) > \mathbb{E}_Q f(\mathbf{z}) + t] \le 2\exp(-cdt^2)$ for a universal $c > 0$. Taking $t = C_{\mathrm{levy}}\sqrt{\log(2/\delta)/d}$ makes this probability at most δ. Jensen's inequality and the assumed MSE guarantee give $\mathbb{E}_Q f(\mathbf{z}) \le (\mathbb{E}_Q f(\mathbf{z})^2)^{1/2} = (\mathbb{E}_Q\|\mathbf{x}-\bar{\mathbf{x}}\|_2^2)^{1/2} \le \sqrt{C_{\mathrm{mse}}}\,2^{-b}$. Combining the two estimates proves (F.2). □

F.2.2 Proof of Theorem F.2

Fix a total bit-width $b \geq 1$. Let $\bar{\mathbf{x}}$ be the reconstruction produced by the $(b-1)$-bit $\texttt{TurboQuant}_{\text{MSE}}$ stage and set $\mathbf{r} = \mathbf{x} - \bar{\mathbf{x}}$. The $\texttt{TurboQuant}_{\text{PROD}}$ estimator is $\text{IP}^{(b)}_{\text{TQ}}(\mathbf{x}, \mathbf{y}) := \langle \mathbf{y}, \bar{\mathbf{x}} \rangle + \text{IP}_{\text{QJL}}(\mathbf{y}, \mathbf{r})$. Since $\langle \mathbf{y}, \mathbf{x} \rangle = \langle \mathbf{y}, \bar{\mathbf{x}} \rangle + \langle \mathbf{y}, \mathbf{r} \rangle$,

$$\text{IP}^{(b)}_{\text{TQ}}(\mathbf{x}, \mathbf{y}) - \langle \mathbf{x}, \mathbf{y} \rangle = \text{IP}_{\text{QJL}}(\mathbf{y}, \mathbf{r}) - \langle \mathbf{y}, \mathbf{r} \rangle.$$

Set $L = \log(4/\delta)$. Applying Lemma F.2 to the $(b-1)$-bit residual with failure probability $\delta/2$ gives, with probability at least $1 - \delta/2$,

$$\|\mathbf{r}\|_2 \leq \sqrt{C_{\text{mse}}}\, 2^{-(b-1)} + C_{\text{levy}} \sqrt{\frac{L}{d}}.$$

Conditional on this residual, Lemma K.4 with $\eta = \delta/2$ gives, since $\|\mathbf{y}\|_2 = 1$ and $L \leq c_{\text{qjl}} d$, that with conditional probability at least $1 - \delta/2$,

$$|\text{IP}_{\text{QJL}}(\mathbf{y}, \mathbf{r}) - \langle \mathbf{y}, \mathbf{r} \rangle| \leq C_{\text{qjl}} \|\mathbf{r}\|_2 \sqrt{\frac{L}{d}}.$$

A union bound gives both events with probability at least $1 - \delta$, and on their intersection

$$\left| \text{IP}^{(b)}_{\text{TQ}}(\mathbf{x}, \mathbf{y}) - \langle \mathbf{x}, \mathbf{y} \rangle \right| \leq C \left(2^{-(b-1)} \sqrt{\frac{L}{d}} + \frac{L}{d} \right),$$

for a universal constant $C > 0$. This proves the first claim after replacing L by $\log(4/\delta)$ and taking $c \leq c_{\text{qjl}}$.

For the bit-width claim, if $\epsilon \geq C_0 L/d$ and $C_0 \geq 2C$, then $CL/d \leq \epsilon/2$. If

$$b \geq 1 + \max \left\{ 0, \left\lceil \frac{1}{2} \log_2 \left(\frac{C_1 L}{d\epsilon^2} \right) \right\rceil \right\},$$

then $2^{-(b-1)} \leq (C_1 L/(d\epsilon^2))^{-1/2}$ when $C_1 L/(d\epsilon^2) > 1$, while the same inequality is trivial when $C_1 L/(d\epsilon^2) \leq 1$. Hence $C 2^{-(b-1)} \sqrt{L/d} \leq C\epsilon/\sqrt{C_1} \leq \epsilon/2$ for $C_1 \geq 4C^2$. Combining the two terms gives the desired probability bound, and enlarging the theorem constant C to dominate C_0, C_1 completes the proof.

G Block Marginal Distribution of a Uniform Spherical Vector (Proof of Lemma 1)

Proof. By symmetry, it suffices to prove the claim for the first block $\mathbf{z}_1$. Let $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_m) \sim N(0, I_d)$, where $\mathbf{g}_j \in \mathbb{R}^p$ are independent standard Gaussian blocks. The standard Gaussian representation of the uniform distribution on the sphere gives $\mathbf{x} \overset{d}{=} \frac{\mathbf{g}}{\|\mathbf{g}\|_2}$, and hence,

$$\mathbf{z}_1 \overset{d}{=} \frac{\mathbf{g}_1}{\sqrt{\|\mathbf{g}_1\|_2^2 + \sum_{k=2}^m \|\mathbf{g}_k\|_2^2}}.$$

Define $U = \|\mathbf{g}_1\|_2^2$ and $V = \sum_{k=2}^m \|\mathbf{g}_k\|_2^2$. Then $U \sim \chi^2_p$ and $V \sim \chi^2_{d-p}$, and U and V are independent. Moreover, the Gaussian direction $\boldsymbol{\theta}_1 := \mathbf{g}_1/\|\mathbf{g}_1\|_2$ is uniform on $\mathbb{S}^{p-1}$ and is independent of U and V. Therefore, we have $\mathbf{z}_1 \overset{d}{=} \sqrt{\frac{U}{U+V}}\, \boldsymbol{\theta}_1$. It follows from the standard beta–chi-square relationship that

$$r_1^2 := \frac{U}{U+V} \sim \text{Beta}\left(\frac{p}{2}, \frac{d-p}{2} \right),$$

and r_1 is independent of $\boldsymbol{\theta}_1$. This proves the polar decomposition.

It remains to derive the density with respect to Lebesgue measure on $\mathbb{B}^p$. Let $a = p/2$ and $b = (d-p)/2$. Since $r_1^2 \sim \text{Beta}(a, b)$, the density of r_1 on $[0, 1]$ is

$$f_{r_1}(r) = \frac{2}{B(a,b)} r^{p-1} (1 - r^2)^{\frac{d-p-2}{2}}, \qquad 0 \leq r \leq 1.$$

Since $\boldsymbol{\theta}_1$ is uniform on $\mathbb{S}^{p-1}$ and independent of r_1, the density $f_{p,d}$ of $\mathbf{z}_1 = r_1\boldsymbol{\theta}_1$ must be radial. Using the polar-coordinate identity $d\mathbf{z} = r^{p-1}\, dr\, d\sigma(\boldsymbol{\theta})$, where $d\sigma$ denotes surface-area measure on $\mathbb{S}^{p-1}$, we get

$$f_{p,d}(r\boldsymbol{\theta}) = \frac{f_{r_1}(r)}{|\mathbb{S}^{p-1}|r^{p-1}} = \frac{2}{B(a,b)|\mathbb{S}^{p-1}|}(1-r^2)^{\frac{d-p-2}{2}}.$$

Finally, using $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $|\mathbb{S}^{p-1}| = 2\pi^{p/2}/\Gamma(p/2)$, we obtain

$$f_{p,d}(\mathbf{z}_1) = \frac{\Gamma(d/2)}{\pi^{p/2}\Gamma((d-p)/2)}\left(1-\|\mathbf{z}_1\|_2^2\right)^{\frac{d-p-2}{2}}, \qquad \mathbf{z}_1 \in \mathbb{B}^p.$$

□

H Expected Distortion Analysis of `BlockQuant`

H.1 Proof of Theorem 2

H.1.1 Proof for Small $b = 1,\ 2,\ 3,\ 4$

Proof. Let p be the block size and let $\mathcal{C}_{\mathrm{BQ}}^{(p)}$ be the codebook minimizing the distortion cost in Equation 1. For d-dimensional b-bit compression,

$$\mathcal{D}_{\mathrm{MSE}}(\texttt{BlockQuant}) = \frac{d}{p}\int_{\mathbb{B}^p} \min_{\mathbf{o}\in\mathcal{C}_{\mathrm{BQ}}^{(p)}} \|\mathbf{u}-\mathbf{o}\|_2^2 f_{p,d}(\mathbf{u})\, d\mathbf{u}\,.$$

Define $\mathbf{R}_j = \sqrt{d}\,\mathbf{z}_j$ and $\mathbf{a}_i = \sqrt{d}\,\mathbf{o}_i$. Then $\|\mathbf{z}_j - \mathbf{o}_i\|_2^2 = d^{-1}\|\mathbf{R}_j - \mathbf{a}_i\|_2^2$, and the density of $\mathbf{R}_j$ is

$$h_{p,d}(\mathbf{r}) = \frac{1}{d^{p/2}} f_{p,d}\left(\frac{\mathbf{r}}{\sqrt{d}}\right) = \frac{\Gamma(d/2)}{d^{p/2}\pi^{p/2}\Gamma((d-p)/2)}\left(1-\frac{\|\mathbf{r}\|_2^2}{d}\right)^{(d-p-2)/2} \mathbb{1}\{\|\mathbf{r}\|_2 \le \sqrt{d}\}\,.$$

After this rescaling,

$$\mathcal{D}_{\mathrm{MSE}}(\texttt{BlockQuant}) = \frac{1}{p}\inf_{\mathcal{A}\subset\sqrt{d}\mathbb{B}^p,\,|\mathcal{A}|\le 2^{bp}} \int_{\sqrt{d}\mathbb{B}^p} \min_{\mathbf{a}\in\mathcal{A}} \|\mathbf{r}-\mathbf{a}\|_2^2 h_{p,d}(\mathbf{r})\, d\mathbf{r}.$$

For fixed p, $h_{p,d}$ converges pointwise to the standard p-dimensional Gaussian density $\zeta_p(\mathbf{r}) = (2\pi)^{-p/2}\exp(-\|\mathbf{r}\|_2^2/2)$. Hence the finite-rate MSE is approximated by $\mathcal{D}_{\mathrm{MSE}}(\texttt{BlockQuant}) = p^{-1}\phi_p^\star + o_d(1)$, where

$$\phi_p^\star := \inf_{\mathcal{A}\subset\mathbb{R}^p,\,|\mathcal{A}|\le 2^{bp}} \int_{\mathbb{R}^p} \min_{\mathbf{a}\in\mathcal{A}} \|\mathbf{r}-\mathbf{a}\|_2^2 \zeta_p(\mathbf{r})\, d\mathbf{r}\,.$$

For a candidate rescaled codebook $\mathcal{A} = \{\mathbf{a}_1,\dots,\mathbf{a}_{2^{bp}}\}$, let $V_i(\mathcal{A}) := \{\mathbf{r} : \|\mathbf{r}-\mathbf{a}_i\|_2 \le \|\mathbf{r}-\mathbf{a}_{i'}\|_2 \text{ for all } i'\}$. Then the Gaussian objective is

$$\phi_p(\mathcal{A}) = \sum_{i=1}^{2^{bp}} \int_{V_i(\mathcal{A})} \|\mathbf{r}-\mathbf{a}_i\|_2^2 \zeta_p(\mathbf{r})\, d\mathbf{r}\,.$$

If the cells are fixed, minimizing the i-th term gives $0 = 2\int_{V_i}(\mathbf{a}_i - \mathbf{r})\zeta_p(\mathbf{r})\, d\mathbf{r}$, so each centroid must be the Gaussian conditional mean of its own Voronoi cell, i.e., $\mathbf{a}_i = \int_{V_i}\mathbf{r}\zeta_p(\mathbf{r})\, d\mathbf{r} / \int_{V_i}\zeta_p(\mathbf{r})\, d\mathbf{r}$. Applying Lloyd optimization for $N(0, I_p)$ gives the following near-minimum values of ϕ_p.

p	b	ϕ_p	ϕ_p/p
2	1	0.726760	0.363380
2	2	0.214970	0.107485
2	3	0.059433	0.029716
2	4	0.015516	0.007758
3	1	1.068772	0.356257
3	2	0.303994	0.101331
3	3	0.081462	0.027154
3	4	0.021173	0.007058

We next show that the same Gaussian-approximation value applies to $\texttt{BlockQuant}_{\text{BSM}}$. Let $Q_{\mathcal{A}}(\mathbf{r})$ be the nearest-centroid map for the Gaussian codebook and define

$$A_p := \frac{1}{p}\int_{\mathbb{R}^p} \langle \mathbf{r}, Q_{\mathcal{A}}(\mathbf{r})\rangle \zeta_p(\mathbf{r})\, d\mathbf{r}, \qquad M_p := \frac{1}{p}\int_{\mathbb{R}^p} \|Q_{\mathcal{A}}(\mathbf{r})\|_2^2 \zeta_p(\mathbf{r})\, d\mathbf{r}\,.$$

The centroid condition implies $A_p = M_p$, since on each Voronoi cell V_i,

$$\int_{V_i} \langle \mathbf{r}, \mathbf{a}_i\rangle \zeta_p(\mathbf{r})\, d\mathbf{r} = \left\langle \int_{V_i} \mathbf{r}\zeta_p(\mathbf{r})\, d\mathbf{r}, \mathbf{a}_i \right\rangle = \|\mathbf{a}_i\|_2^2 \int_{V_i} \zeta_p(\mathbf{r})\, d\mathbf{r}$$

For the raw reconstruction $\bar{\mathbf{x}}$, the high-dimensional Gaussian approximation and the law of large numbers give $\langle \mathbf{x}, \bar{\mathbf{x}}\rangle \approx A_p$ and $\|\bar{\mathbf{x}}\|_2^2 \approx M_p$. On the other hand, we have that

$$\frac{\phi_p}{p} = \frac{1}{p}\int \|\mathbf{r} - Q_{\mathcal{A}}(\mathbf{r})\|_2^2 \zeta_p(\mathbf{r})\, d\mathbf{r} = 1 - 2A_p + M_p = 1 - M_p.$$

Therefore the best-scalar reconstruction satisfies

$$\mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{BSM}}) \approx 1 - \frac{A_p^2}{M_p} = 1 - M_p = \frac{\phi_p}{p}.$$

Thus, $\texttt{BlockQuant}_{\text{MSE}}$ and $\texttt{BlockQuant}_{\text{BSM}}$ have the same finite-rate Gaussian-approximation value. Specifically, for $\mathcal{Q} \in \{\texttt{BlockQuant}_{\text{MSE}}, \texttt{BlockQuant}_{\text{BSM}}\}$ and $b = 1,\ 2,\ 3,\ 4$,

$$\mathcal{D}_{\text{MSE}}(\mathcal{Q}_{(p=2)}) \approx 0.363380,\ 0.107485,\ 0.029716,\ 0.007758$$
$$\mathcal{D}_{\text{MSE}}(\mathcal{Q}_{(p=3)}) \approx 0.356257,\ 0.101331,\ 0.027154,\ 0.007058.$$

□

H.1.2 Proof for Large Bit-Width with $p \neq d$

Let $\bar{\mathbf{x}}$ be the raw reconstruction used by $\texttt{BlockQuant}_{\text{MSE}}$. Then, best-scalar variant $\texttt{BlockQuant}_{\text{BSM}}$ returns $\frac{\langle \mathbf{x}, \bar{\mathbf{x}}\rangle}{\|\bar{\mathbf{x}}\|_2^2}\bar{\mathbf{x}}$, and for every rotation matrix R, we have

$$\left\| \mathbf{x} - \frac{\langle \mathbf{x}, \bar{\mathbf{x}}\rangle}{\|\bar{\mathbf{x}}\|_2^2}\bar{\mathbf{x}} \right\|_2^2 = \min_{\alpha\in\mathbb{R}} \|\mathbf{x} - \alpha\bar{\mathbf{x}}\|_2^2 \le \|\mathbf{x} - \bar{\mathbf{x}}\|_2^2.$$

That means, for every block size p, it holds that

$$\mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{BSM}(p)}) \le \mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{MSE}(p)}),$$

so any high-rate upper bound for $\texttt{BlockQuant}_{\text{MSE}}$ also applies to $\texttt{BlockQuant}_{\text{BSM}}$. Thus, in this section (Section H.1.2) and the next section (Section H.1.3), we focus on bounding MSE of $\texttt{BlockQuant}_{\text{MSE}}$.

For a large number of centroids, we use the high-rate Zador–Gersho formula to derive the MSE bound of $\texttt{BlockQuant}_{\text{MSE}}$. The following corollary is a restatement of Proposition C.1 in our setting.

Corollary H.1 (MSE bound of Algorithm 1 for many centroids)**.** *Let d, m, and p be integers with $d = mp$. If Algorithm 1 is run with b-bit compression and m blocks, then for any $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{S}^{d-1}$,*

$$\mathcal{D}_{\textit{MSE}}(\texttt{BlockQuant}) \lesssim d\, G_p^\star J_{p,d}\, 4^{-b},$$

where $G_p^\star := G_{\mathbb{B}^p}^\star$ and $J_{p,d} := \left(\int_{\mathbb{B}_p} f_{p,d}(\mathbf{z})^{p/(p+2)}\, d\mathbf{z}\right)^{(p+2)/p}$.

We now prove the large-bit bound for $p \neq d$.

Proof. Recall that each block $\mathbf{z}_j$ has density $f_{p,d}(\mathbf{z}) = \frac{\Gamma(d/2)}{\pi^{p/2}\Gamma((d-p)/2)}(1 - \|\mathbf{z}\|_2^2)^{(d-p-2)/2}$ on $\mathbb{B}^p$. Set $\beta_{p,d} := p(d-p-2)/(2(p+2))$. Then

$$\int_{\mathbb{B}_p} f_{p,d}(\mathbf{z})^{p/(p+2)}\, d\mathbf{z} = \left(\frac{\Gamma(d/2)}{\pi^{p/2}\Gamma((d-p)/2)}\right)^{p/(p+2)} \int_{\mathbb{B}_p} (1 - \|\mathbf{z}\|^2)^{\beta_{p,d}}\, d\mathbf{z}.$$

Using polar coordinates and the change of variables $u = r^2$,

$$\int_{\mathbb{B}_p} (1 - \|\mathbf{z}\|^2)^{\beta_{p,d}}\, d\mathbf{z} = \frac{\pi^{p/2}}{\Gamma(p/2)} \int_0^1 u^{p/2-1}(1-u)^{\beta_{p,d}}\, du = \pi^{p/2} \frac{\Gamma(\beta_{p,d}+1)}{\Gamma(\beta_{p,d}+1+p/2)}.$$

Substitution gives

$$J_{p,d} = \pi \frac{\Gamma(d/2)}{\Gamma((d-p)/2)} \left(\frac{\Gamma(\beta_{p,d}+1)}{\Gamma(\beta_{p,d}+1+p/2)} \right)^{(p+2)/p}.$$

Case $p = 2$. Since $G_2^\star = 5/(36\sqrt{3})$ and $J_{2,d} = 8\pi(d-2)/d^2$,

$$\mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{BSM}}\,(p=2)) \le \mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{MSE}}\,(p=2)) \le \frac{10\pi}{9\sqrt{3}} \left(1 - \frac{2}{d}\right) 4^{-b} \approx 2.015 \cdot 4^{-b}.$$

Case $p = 3$. Since $G_3^\star \le G(A_3^\star) = 19/(192\, 2^{1/3}) \approx 0.0785432812$ and $J_{3,d} = \pi \frac{\Gamma(d/2)}{\Gamma((d-3)/2)} \left[\frac{\Gamma((3d-5)/10)}{\Gamma((3d+10)/10)}\right]^{5/3}$,

$$\begin{aligned}
\mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{BSM}}\,(p=3)) &\le \mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{MSE}}\,(p=3)) \\
&\le 0.0785432812\, \pi d \frac{\Gamma(d/2)}{\Gamma((d-3)/2)} \left[\frac{\Gamma((3d-5)/10)}{\Gamma((3d+10)/10)}\right]^{5/3} 4^{-b} \\
&\approx 1.770 \cdot 4^{-b}.
\end{aligned}$$

□

Remark H.1. *When $p = 1$ ($\texttt{TurboQuant}_{\text{MSE}}$), since $G_1^\star = 1/12$ and $J_{1,d} \approx 6\sqrt{3}\pi/d$, the MSE distortion satisfies $\mathcal{D}_{MSE}(\texttt{TurboQuant}_{\text{MSE}}) \le (\sqrt{3}\pi/2)4^{-b} \approx 2.721 \cdot 4^{-b}$.*

H.1.3 Proof for $p = d$

Proof. When $p = d$, Algorithm 1 uses a single block. The block source is therefore not the full-dimensional density $f_{p,d}$ on $\mathbb{B}^p$ used above; for any fixed $\mathbf{x} \in \mathbb{S}^{d-1}$, the rotated vector $\mathbf{z} := R\mathbf{x}$ is uniform on $\mathbb{S}^{d-1}$, so the intrinsic source dimension is $n = d - 1$. Let $A_{d-1} := \mathcal{H}^{d-1}(\mathbb{S}^{d-1}) = 2\pi^{d/2}/\Gamma(d/2)$. With respect to surface measure $d\sigma$, the density of $\mathbf{z}$ is $f(\mathbf{z}) = A_{d-1}^{-1}$.

In the $p = d$ case, the total number of codewords is $K = 2^{bd}$. For $\mathcal{C} \subset \mathbb{R}^d$ with $|\mathcal{C}| \le K$, define $D_{\text{sph}}(\mathcal{C}) := \int_{\mathbb{S}^{d-1}} \min_{\mathbf{o}\in\mathcal{C}} \|\mathbf{z} - \mathbf{o}\|_2^2 f(\mathbf{z})\, d\sigma(\mathbf{z})$ and $D_{\text{sph}}^\star(K) := \inf_{|\mathcal{C}|\le K} D_{\text{sph}}(\mathcal{C})$. This is the single-block analogue of the objective above.

Formally, the intrinsic n-dimensional Zador–Gersho formula gives

$$D_{\text{sph}}^\star(K) \le n G_n^\star \left(\int_{\mathbb{S}^{d-1}} f(\mathbf{z})^{n/(n+2)}\, d\sigma(\mathbf{z}) \right)^{(n+2)/n} K^{-2/n}(1 + o(1)).$$

Here curvature contributes only lower-order error, because on a cell of diameter r, squared Euclidean and tangent-plane distances differ by $O(r^4)$. Since f is constant, $\int_{\mathbb{S}^{d-1}} f(\mathbf{z})^{n/(n+2)} d\sigma(\mathbf{z}) = A_{d-1}^{2/(n+2)}$, so

$$D_{\text{sph}}^\star(K) \le (d-1) G_{d-1}^\star \left(\frac{2\pi^{d/2}}{\Gamma(d/2)} \right)^{2/(d-1)} K^{-2/(d-1)}(1 + o(1)).$$

Since $G_{d-1}^\star$ is not available in closed form, we use an explicit random-coding comparison.

Let $V_n := \pi^{n/2}/\Gamma(1 + n/2)$ be the unit-ball volume in $\mathbb{R}^n$, and draw $\mathcal{C}_{\text{rand}} = \{\mathbf{o}_1, \dots, \mathbf{o}_K\}$ with $\mathbf{o}_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1})$. Fix $\mathbf{z} \in \mathbb{S}^{d-1}$ and set $T_i := \|\mathbf{z} - \mathbf{o}_i\|_2^2$, $T_{(1)} := \min_i T_i$. For small t, the cap $\{\mathbf{o} : \|\mathbf{z} - \mathbf{o}\|_2^2 \le t\}$ has surface area $V_n t^{n/2}(1 + o(1))$, hence $\mathbb{P}(T_i \le t) = (V_n/A_{d-1}) t^{n/2}(1 + o(1))$. With $t = K^{-2/n} s$,

$$\mathbb{P}(T_{(1)} > K^{-2/n} s) = \left(1 - \mathbb{P}(T_i \le K^{-2/n} s)\right)^K \to \exp\left(-\frac{V_n}{A_{d-1}} s^{n/2} \right).$$

Using $\mathbb{E}T_{(1)} = \int_0^\infty \mathbb{P}(T_{(1)} > t)\, dt$ and the change of variables $u = (V_n/A_{d-1})s^{n/2}$ yields

$$\mathbb{E}_{\mathcal{C}_{\text{rand}}} T_{(1)} = \Gamma\left(1 + \frac{2}{n}\right)\left(\frac{A_{d-1}}{V_n}\right)^{2/n} K^{-2/n}(1 + o(1)).$$

The ideal codebook cannot be worse than the random comparison. Substituting $n = d - 1$ and $V_{d-1} = \pi^{(d-1)/2}/\Gamma((d+1)/2)$ gives

$$D^\star_{\text{sph}}(K) \le C_d K^{-2/(d-1)}(1 + o(1)), \qquad C_d := \Gamma\left(1 + \frac{2}{d-1}\right)\left[2\sqrt{\pi}\,\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}\right]^{2/(d-1)}.$$

Under the b-bit-per-coordinate convention, $K = 2^{bd}$ and $K^{-2/(d-1)} = (1/4)^{bd/(d-1)}$.

It remains to translate the spherical source bound back to randomized MSE. For fixed $\mathbf{x} \in \mathbb{S}^{d-1}$, let $Q_{\mathcal{C}}(\mathbf{z}) := \arg\min_{\mathbf{o}\in\mathcal{C}} \|\mathbf{z} - \mathbf{o}\|_2^2$ and $\bar{\mathbf{x}} := R^\top Q_{\mathcal{C}}(R\mathbf{x})$. Since rotations preserve distance and $R\mathbf{x} \sim \text{Unif}(\mathbb{S}^{d-1})$,

$$\mathbb{E}_R\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2 = \int_{\mathbb{S}^{d-1}} \min_{\mathbf{o}\in\mathcal{C}} \|\mathbf{z} - \mathbf{o}\|_2^2 f(\mathbf{z})\, d\sigma(\mathbf{z}).$$

The right-hand side is independent of $\mathbf{x}$. Choosing ideal spherical centroids gives

$$\mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{BSM}(p=d)}) = \max_{\mathbf{x}\in\mathbb{S}^{d-1}} \mathbb{E}_R\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2 \le C_d\left(\frac{1}{4}\right)^{bd/(d-1)}(1 + o(1)).$$

Since $\texttt{BlockQuant}_{\text{BSM}}$ is the best scalar multiple of the same raw reconstruction $\bar{\mathbf{x}}$,

$$\mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{BSM}(p=d)}) \le \mathcal{D}_{\text{MSE}}(\texttt{BlockQuant}_{\text{MSE}(p=d)}) \le C_d\left(\frac{1}{4}\right)^{bd/(d-1)}(1 + o(1)).$$

The values $C_{100} \approx 1.055$, $C_{1000} \approx 1.008$, and $C_{10000} \approx 1.001$ follow by evaluating the log-gamma expression, and $C_d \to 1$ by the gamma-ratio asymptotic. □

H.2 Proof of Corollary 3

H.2.1 Proof for Small $b = 1,\ 2,\ 3,\ 4$

Proof. Let $\bar{\mathbf{x}}_{p,b}$ be the unscaled reconstruction produced by b-bit $\texttt{BlockQuant}_{\text{BSM}(p)}$, and set $\rho_{p,b} := \langle \bar{\mathbf{x}}_{p,b}, \mathbf{x}\rangle$ and $\psi_{p,b} := \|\bar{\mathbf{x}}_{p,b}\|_2$. The corresponding ratio estimator is $\widehat{\eta}_{\text{ratio}} := \langle \bar{\mathbf{x}}_{p,b}, \mathbf{y}\rangle / \langle \bar{\mathbf{x}}_{p,b}, \mathbf{x}\rangle$. Since $\texttt{BlockQuant}_{\text{BSM}(p)}$ is obtained by a Haar random rotation, a fixed block quantizer, and the inverse rotation, Theorem 1 applies. Thus, with $\eta = \langle \mathbf{x}, \mathbf{y}\rangle$,

$$\mathbb{E}[(\widehat{\eta}_{\text{ratio}} - \eta)^2] = \frac{1 - \eta^2}{d-1}\mathbb{E}\left[\frac{\psi_{p,b}^2 - \rho_{p,b}^2}{\rho_{p,b}^2}\right].$$

Taking the supremum over $\mathbf{y} \in \mathbb{S}^{d-1}$ gives $\mathcal{D}_{\text{IP}}(\texttt{BlockQuant}_{\text{UB}(p)}) = (d-1)^{-1}\mathbb{E}[(\psi_{p,b}^2 - \rho_{p,b}^2)/\rho_{p,b}^2]$.

It remains to evaluate the deterministic equivalent of the nonlinear factor. As in the small-bit proof of Theorem 2, write $\mathbf{R}_j = \sqrt{d}\,\mathbf{z}_j$ and approximate its law by $N(0, I_p)$. Let $Q_{p,b}$ be the nearest-neighbor Gaussian block quantizer with 2^{bp} centroids and define $\delta_{p,b} := p^{-1}\mathbb{E}\|\mathbf{R} - Q_{p,b}(\mathbf{R})\|_2^2$, where $\mathbf{R} \sim N(0, I_p)$. The centroid condition gives $\mathbb{E}[\mathbf{R} \mid Q_{p,b}(\mathbf{R})] = Q_{p,b}(\mathbf{R})$, hence $\mathbb{E}\langle \mathbf{R}, Q_{p,b}(\mathbf{R})\rangle = \mathbb{E}\|Q_{p,b}(\mathbf{R})\|_2^2$. If $m_{p,b} := p^{-1}\mathbb{E}\|Q_{p,b}(\mathbf{R})\|_2^2$, then $\delta_{p,b} = 1 - m_{p,b}$.

After rescaling back to the unit sphere, the empirical block averages satisfy $\rho_{p,b} = m_{p,b} + o_d(1)$ and $\psi_{p,b}^2 = m_{p,b} + o_d(1)$ under the high-dimensional Gaussian approximation. Therefore

$$\mathbb{E}\left[\frac{\psi_{p,b}^2 - \rho_{p,b}^2}{\rho_{p,b}^2}\right] \approx \frac{m_{p,b} - m_{p,b}^2}{m_{p,b}^2} = \frac{\delta_{p,b}}{1 - \delta_{p,b}}.$$

Substituting the small-bit MSE constants from Theorem 2 gives the following leading inner-product constants.

p	b	$\delta_{p,b}$	$\delta_{p,b}/(1-\delta_{p,b})$
2	1	0.363380	0.570796
2	2	0.107485	0.120429
2	3	0.029716	0.030626
2	4	0.007758	0.007819
3	1	0.356257	0.553415
3	2	0.101331	0.112757
3	3	0.027154	0.027912
3	4	0.007058	0.007108

Since $\mathcal{D}_{\text{IP}}(\texttt{BlockQuant}_{\text{UB}(p)}) \approx \frac{1}{d-1}\frac{\delta_{p,b}}{1-\delta_{p,b}}$, dividing these coefficients by $d-1$ proves the claimed constants for $b = 1,\ 2,\ 3,\ 4$. □

H.2.2 Proof for Large Bit-Width

Proof. Let $\bar{\mathbf{x}}_{p,b}$ be the unscaled high-rate $\texttt{BlockQuant}_{\text{BSM}(p)}$ reconstruction of $\mathbf{x}$, and set $\rho_{p,b} := \langle \bar{\mathbf{x}}_{p,b}, \mathbf{x}\rangle$ and $\psi_{p,b} := \|\bar{\mathbf{x}}_{p,b}\|_2$. Write $\mathbf{e}_{p,b} := \bar{\mathbf{x}}_{p,b} - \mathbf{x}$, $t_{p,b} := \langle \mathbf{x}, \mathbf{e}_{p,b}\rangle$, and $\mathbf{u}_{p,b} := \mathbf{e}_{p,b} - t_{p,b}\mathbf{x}$. Then $\mathbf{u}_{p,b} \perp \mathbf{x}$, $\bar{\mathbf{x}}_{p,b} = (1+t_{p,b})\mathbf{x} + \mathbf{u}_{p,b}$, $\rho_{p,b} = 1 + t_{p,b}$, and $\psi_{p,b}^2 - \rho_{p,b}^2 = \|\mathbf{u}_{p,b}\|_2^2$.

By Theorem 1, taking the worst case over $\mathbf{y} \in \mathbb{S}^{d-1}$ gives

$$\mathcal{D}_{\text{IP}}(\texttt{BlockQuant}_{\text{UB}(p)}) = \frac{1}{d-1}\mathbb{E}\left[\frac{\psi_{p,b}^2 - \rho_{p,b}^2}{\rho_{p,b}^2}\right] = \frac{1}{d-1}\mathbb{E}\left[\frac{\|\mathbf{u}_{p,b}\|_2^2}{(1+t_{p,b})^2}\right].$$

Thus only the tangential error enters the numerator, while the random denominator remains inside the expectation.

Under the same high-rate regularity used in Corollary H.1, $\mathbb{E}[\|\mathbf{e}_{p,b}\|_2^2] = O(4^{-b})$ and $\mathbb{E}[\|\mathbf{e}_{p,b}\|_2^4] = O(4^{-2b})$. Since $|t_{p,b}| \le \|\mathbf{e}_{p,b}\|_2$ and $\|\mathbf{u}_{p,b}\|_2 \le \|\mathbf{e}_{p,b}\|_2$, the denominator does not change the leading order:

$$\left|\mathbb{E}\left[\frac{\|\mathbf{u}_{p,b}\|_2^2}{(1+t_{p,b})^2}\right] - \mathbb{E}[\|\mathbf{u}_{p,b}\|_2^2]\right| \le C\cdot\mathbb{E}[\|\mathbf{u}_{p,b}\|_2^2|t_{p,b}|] + o(4^{-b}) \le C\cdot\mathbb{E}\|\mathbf{e}_{p,b}\|_2^3 + o(4^{-b}) = o(4^{-b}),$$

where $\mathbb{E}[\|\mathbf{e}_{p,b}\|_2^3] \le (\mathbb{E}[\|\mathbf{e}_{p,b}\|_2^4])^{3/4} = O(2^{-3b})$. Hence

$$\mathbb{E}\left[\frac{\|\mathbf{u}_{p,b}\|_2^2}{(1+t_{p,b})^2}\right] = \mathbb{E}[\|\mathbf{u}_{p,b}\|_2^2] + o(4^{-b}) \le \mathbb{E}[\|\mathbf{e}_{p,b}\|_2^2] + o(4^{-b}).$$

This uses only the total high-rate MSE constant. By Theorem 2, equivalently Corollary H.1, $\mathbb{E}\|\mathbf{e}_{p,b}\|_2^2 \le dG_p^\star J_{p,d} 4^{-b}(1+o(1))$. Therefore

$$\mathcal{D}_{\text{IP}}(\texttt{BlockQuant}_{\text{UB}(p)}) \le \frac{dG_p^\star J_{p,d}}{d-1}\cdot 4^{-b}(1+o(1)).$$

For $p=2$, $dG_2^\star J_{2,d} = \frac{10\pi}{9\sqrt{3}}(1-2/d)$, so

$$\mathcal{D}_{\text{IP}}(\texttt{BlockQuant}_{\text{UB}(p=2)}) \le \frac{1}{d-1}\frac{10\pi}{9\sqrt{3}}\left(1-\frac{2}{d}\right)4^{-b}(1+o(1)) \le \frac{\mathbf{2.015}}{d-1}4^{-b}(1+o(1)).$$

For $p=3$,

$$dG_3^\star J_{3,d} \le 0.0785432812\,\pi d\frac{\Gamma(d/2)}{\Gamma((d-3)/2)}\left[\frac{\Gamma((3d-5)/10)}{\Gamma((3d+10)/10)}\right]^{5/3},$$

which gives

$$\begin{aligned}\mathcal{D}_{\text{IP}}(\texttt{BlockQuant}_{\text{UB}(p=3)}) &\le \frac{0.0785432812\,\pi d}{d-1}\frac{\Gamma(d/2)}{\Gamma((d-3)/2)}\left[\frac{\Gamma((3d-5)/10)}{\Gamma((3d+10)/10)}\right]^{5/3}4^{-b}(1+o(1))\\ &\approx \frac{\mathbf{1.770}}{d-1}4^{-b}(1+o(1)).\end{aligned}$$

This proves the high-rate bounds in the corollary. □

I Lower Bound Analysis (Proof of Theorem 3)

Proof. Write $\mathbf{x} = (\mathbf{u}, X_d)$, where $\mathbf{u} \in \mathbb{R}^{d-1}$ denotes the first $d-1$ coordinates of $\mathbf{x}$. Also write $\widehat{\mathbf{x}} = Q^{-1}(Q(\mathbf{x}))$ and let $\widehat{\mathbf{u}}$ be the first $d-1$ coordinates of $\widehat{\mathbf{x}}$. Since coordinate projection cannot increase Euclidean distance, $\|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2 \geq \|\mathbf{u} - \widehat{\mathbf{u}}\|_2^2$. Moreover, $\widehat{\mathbf{u}}$ is determined by the bd-bit message $Q(\mathbf{x})$, so by data processing $I(\mathbf{u}; \widehat{\mathbf{u}}) \leq H(Q(\mathbf{x})) \leq bd$ bits. Hence

$$\mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - Q^{-1}(Q(\mathbf{x}))\|_2^2] \geq D_{\mathbf{u}}(bd), \qquad D_{\mathbf{u}}(B) := \inf_{I(\mathbf{u};\widetilde{\mathbf{u}}) \leq B} \mathbb{E}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_2^2,$$

where mutual information is measured in bits. It remains to lower bound the distortion-rate function of the projected spherical source $\mathbf{u}$.

We first record the law of $\mathbf{u}$. The sphere is the union, up to the equator of surface measure zero, of the two graphs $u \mapsto (u, \pm\sqrt{1 - \|u\|_2^2})$ over the open unit ball in $\mathbb{R}^{d-1}$. The surface element of either graph is $(1 - \|u\|_2^2)^{-1/2} du$. Dividing the contribution of the two sheets by $|\mathbb{S}^{d-1}| = 2\pi^{d/2}/\Gamma(d/2)$ gives the density

$$f_{\mathbf{u}}(u) = \frac{\Gamma(d/2)}{\pi^{d/2}} (1 - \|u\|_2^2)^{-1/2} \mathbb{1}\{\|u\|_2 < 1\}.$$

Therefore, with h_e denoting differential entropy in nats,

$$h_e(\mathbf{u}) = \log\left(\frac{\pi^{d/2}}{\Gamma(d/2)}\right) + \frac{1}{2}\mathbb{E}\log(1 - \|\mathbf{u}\|_2^2).$$

The radial variable satisfies $\|\mathbf{u}\|_2^2 \sim \mathrm{Beta}((d-1)/2, 1/2)$, and the beta identity $\mathbb{E}\log(1-U) = \psi(\beta) - \psi(\alpha+\beta)$ for $U \sim \mathrm{Beta}(\alpha, \beta)$ gives

$$h_e(\mathbf{u}) = \log\left(\frac{\pi^{d/2}}{\Gamma(d/2)}\right) + \frac{1}{2}\{\psi(1/2) - \psi(d/2)\}.$$

Now set $n = d - 1$. For any reconstruction $\widetilde{\mathbf{u}}$ satisfying $I(\mathbf{u}; \widetilde{\mathbf{u}}) \leq B$, let $D = \mathbb{E}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_2^2$ and $\mathbf{e} = \mathbf{u} - \widetilde{\mathbf{u}}$. Since conditioning cannot increase differential entropy and translation does not change it, $I(\mathbf{u}; \widetilde{\mathbf{u}}) = h_2(\mathbf{u}) - h_2(\mathbf{u} \mid \widetilde{\mathbf{u}}) \geq h_2(\mathbf{u}) - h_2(\mathbf{e})$, where $h_2 = h_e/\log 2$. Among all n-dimensional errors with second moment at most D, the isotropic Gaussian has the largest entropy, so $h_2(\mathbf{e}) \leq \frac{n}{2}\log_2(2\pi e D/n)$. Thus every B-bit reconstruction must satisfy

$$D \geq \frac{n}{2\pi e} 2^{\frac{2}{n}(h_2(\mathbf{u}) - B)} = \frac{n}{2\pi e} \exp\left(\frac{2h_e(\mathbf{u})}{n}\right) 2^{-2B/n}.$$

Applying this with $B = bd$ and substituting the entropy formula above yields

$$D_{\mathbf{u}}(bd) \geq \frac{d-1}{2\pi e} \left(\frac{\pi^{d/2}}{\Gamma(d/2)}\right)^{2/(d-1)} \cdot \exp\left(\frac{\psi(1/2) - \psi(d/2)}{d-1}\right) 2^{-2bd/(d-1)}.$$

Since $2^{-2bd/(d-1)} = (\frac{1}{4})^{bd/(d-1)}$, this is exactly the claimed lower bound. The numerical values of c_d follow by direct evaluation of the displayed formula. □

J High Probability Bit Complexity Comparison

Unlike the other two methods, `RabitQ` analyzes the high-probability behavior of the inner-product error and derives the corresponding bit-complexity guarantee. As pointed out by Gao et al. [8], applying Chebyshev's inequality only to the variance bound for $\texttt{TurboQuant}_{\text{PROD}}$ in Proposition B.4 is not sufficient to recover the optimal bit complexity of Alon and Klartag [1] stated in Lemma K.2. We show, however, that a sharper high-probability guarantee for $\texttt{TurboQuant}_{\text{PROD}}$ can be obtained by combining the high-probability guarantee of $\texttt{TurboQuant}_{\text{MSE}}$ with that of `QJL`. In particular, this refined analysis shows that $\texttt{TurboQuant}_{\text{PROD}}$ attains the optimal bit complexity in a certain low-accuracy regime. The similar bit-complexity guarantee also holds for $\texttt{EDEN}_{\text{UB}}$.

Theorem J.1 (Bit complexity of $\texttt{EDEN}_{\text{UB}}$ and $\texttt{TurboQuant}_{\text{PROD}}$ (informal))**.** *Let d be the input dimension and let $\epsilon, \delta \in (0,1)$. Suppose that $\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right) \lesssim d$. Then, for both $\texttt{EDEN}_{\text{UB}}$ and $\texttt{TurboQuant}_{\text{PROD}}$, it suffices to use $b = \Theta\left(\log\left(\frac{1}{d\epsilon^2}\log\frac{1}{\delta}\right)\right)$ bits per dimension to ensure that the inner-product estimation error is at most ϵ with failure probability at most δ.*

Discussion on bit complexity. The above theorem shows that $\texttt{EDEN}_{\text{UB}}$ and $\texttt{TurboQuant}_{\text{PROD}}$ achieves the optimal bit complexity of Alon and Klartag [1] when $\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right) \lesssim d$. This guarantee is weaker than that of RabitQ in Lemma B.2, since RabitQ achieves the same optimal bit complexity in the high-accuracy regime $d \leq \frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)$, which is the regime considered by Alon and Klartag [1]. This limitation comes from the MSE-oriented construction of EDEN and TurboQuant: although it effectively reduces expected reconstruction error, it does not directly provide sharp high-probability control of the residual distribution. The proof is deferred to Appendix F.

K Auxiliary Lemmas

Lemma K.1 (Shannon's lower bound on distortion, Lemma 2 in Zandieh et al. [26])**.** *Let $\mathbf{x} \in \mathbb{R}^d$ be a random vector with finite differential entropy $h(\mathbf{x})$. Then, for any $b \geq 0$, and any quantization map Q, the following Shannon Lower Bound holds:*

$$\mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - Q^{-1}(Q(\mathbf{x}))\|_2^2] \geq \frac{d}{2\pi e} \cdot 2^{\frac{2}{d}(h(\mathbf{x})-bd)}.$$

Lemma K.2 (High-probability error guarantees, Alon and Klartag [1], Gao et al. [7])**.** *Let d be the dimension, and let $\epsilon, \delta \in (0,1)$. Suppose that an error bound ϵ with failure probability at most δ is required, and assume that $\frac{1}{\epsilon^2}\log\frac{1}{\delta} > d$. Then the minimum number of bits required to achieve such a guarantee is $\Theta\big(d\log\left(\frac{1}{d\epsilon^2}\log\frac{1}{\delta}\right)\big)$.*

Lemma K.3 (High-probability angular error decay of RabitQ)**.** *Let $\mathbf{x}$, and $\bar{\mathbf{x}}$ be a unit d-dimensional vector and its quantized vector by* RabitQ*, respectively. Then, for any $L > 0$, we have*

$$\mathbb{P}\left\{\sqrt{1-\langle \mathbf{x}, \bar{\mathbf{x}}\rangle^2} > \frac{L}{2^b} + \frac{c_1}{\sqrt{\delta}} \cdot \exp\left(-\frac{c_0}{2}L^2\right)\right\} < \delta$$

where c_0 and c_1 are absolute constants.

Lemma K.4 (High probability inner product distortion bound of QJL, Lemma 3.5 of Zandieh et al. [27])**.** *Let $S \in \mathbb{R}^{d\times d}$ have i.i.d. standard Gaussian rows. For fixed $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$, define the inner product estimator IP_{QJL} of* QJL*, i.e. $\text{IP}_{\text{QJL}}(\mathbf{y}, \mathbf{x}) := \frac{\sqrt{\pi/2}}{d}||\mathbf{x}||\langle S\mathbf{y}, \text{sign}(S\mathbf{x})\rangle$. Then, there exist universal constants $c_{\text{qjl}}, C_{\text{qjl}} > 0$ such that, for all $0 < \delta < 1$ with $\log(2/\delta) \leq c_{\text{qjl}}d$,*

$$\mathbb{P}_S\left[|\text{IP}_{\text{QJL}}(\mathbf{y}, \mathbf{x}) - \langle \mathbf{y}, \mathbf{x}\rangle| > C_{\text{qjl}}\sqrt{\frac{\log(2/\delta)}{d}} \;\middle|\; \mathbf{x}\right] \leq \delta$$

L Additional Experimental Details and Results

L.1 Time Complextiy and Approximate Nearest-Centroid Assignment for BlockQuant

For block size p, each vector contains d/p blocks, quantized using a shared codebook of $K = 2^{bp}$ centroids.Exact nearest-centroid assignment requires $O(pK)$ time per block, $O(dK)$ per vector, and $O(Nd2^{bp})$ for N vectors, while shared codebook storage requires $O(p2^{bp})$ scalar entries. Thus, both the assignment cost and codebook storage grow exponentially with p. In contrast, the centroid indices occupy exactly Nbd bits, independently of p.

To reduce the assignment cost, we use a lookup-table approximation. The exact encoding step of BlockQuant assigns each rotated block $z_j \in \mathbb{R}^p$ to its nearest codebook centroid:

$$\text{idx}_j = \arg\min_{i\in[K]} \|z_j - c_i\|_2^2, \qquad K = 2^{bp},$$

where b is the bit-width per coordinate and p is the block size. Since the number of blocks is $m = d/p$, the exact assignment costs $O(mKp)$ distance evaluations per vector. This becomes expensive for larger b and p; for example, when $p = 3$ and $b = 4$, each block has $K = 2^{12} = 4096$ candidate centroids.

To reduce this cost, we use a lookup-table approximation that replaces the full nearest-centroid search by a small candidate search. The approximation changes only the encoding step; the codebook construction, dequantization, and ratio rescaling for inner-product estimation remain unchanged.

Cartesian LUT construction. We partition the block domain into a Cartesian grid. Let each coordinate axis be divided into L bins, producing L^p grid cells. For a cell indexed by $u \in [L]^p$, let $G_u \subset \mathbb{R}^p$ denote the cell and let g_u be its center. For each grid center g_u, we precompute the k closest codebook centroids:

$$\mathcal{C}_{\mathrm{top}k}(u) := \arg\mathrm{top}k_{i\in[K]}\big(-\|g_u - c_i\|_2^2\big).$$

Equivalently, $\mathcal{C}_{\mathrm{top}k}(u)$ stores the indices of the k smallest values among

$$\{\|g_u - c_i\|_2^2 : i \in [K]\}.$$

This table is built once after the codebook is constructed and is reused for all input vectors.

Approximate assignment. At quantization time, for each block z_j, we first find the grid cell $G_{u(z_j)}$ containing z_j. Instead of comparing z_j with all K centroids, we compare it only with the precomputed candidate set for that cell:

$$\widetilde{\mathrm{idx}}_j = \arg\min_{i\in\mathcal{C}_{\mathrm{top}k}(u(z_j))} \|z_j - c_i\|_2^2.$$

Thus the exact search space $[K]$ is replaced by the much smaller candidate set $\mathcal{C}_{\mathrm{top}k}(u(z_j))$. The online assignment cost is reduced from $O(mKp)$ to $O(mkp)$, plus the negligible cost of locating the grid cell.

After the approximate indices are obtained, dequantization proceeds in the same way as exact `BlockQuant`:

$$\widetilde{z}' = (c_{\widetilde{\mathrm{idx}}_1}, \ldots, c_{\widetilde{\mathrm{idx}}_m}), \qquad \widetilde{x} = \Pi^\top \widetilde{z}'.$$

For inner-product estimation, we use the same ratio correction as in $\texttt{BlockQuant}_{\mathrm{UB}}$:

$$\widehat{x}_{\mathrm{approx}} = \frac{1}{\widetilde{\rho}}\widetilde{x}, \qquad \widetilde{\rho} = \langle z, \widetilde{z}'\rangle.$$

Approximation error. The approximation is exact whenever the true nearest centroid belongs to the stored candidate set:

$$\mathrm{idx}_j \in \mathcal{C}_{\mathrm{top}k}(u(z_j)).$$

Even when this does not hold, the loss is controlled by the grid resolution. Let

$$D_i(z) := \|z - c_i\|_2^2$$

and let r_L be the maximum distance between a point in a grid cell and its center:

$$r_L := \max_{u\in[L]^p} \max_{z\in G_u} \|z - g_u\|_2.$$

If all blocks, grid centers, and centroids lie in a bounded set with norm at most R, then for any $z \in G_u$,

$$|D_i(z) - D_i(g_u)| = \big|\|z - c_i\|_2^2 - \|g_u - c_i\|_2^2\big| \le 4Rr_L.$$

Therefore, if $\widetilde{i}(z)$ denotes the approximate index and $i^\star(z)$ denotes the exact nearest-centroid index, then

$$D_{\widetilde{i}(z)}(z) \le D_{i^\star(z)}(z) + 8Rr_L.$$

Thus the additional per-block squared-distance error vanishes as the grid is refined. Increasing L decreases the discretization error, while increasing k increases the probability that the exact nearest centroid is included in the candidate set.

Complexity. The LUT requires one-time preprocessing cost $O(L^pKp)$ to compute distances from all grid centers to all centroids, and memory $O(L^pk)$ to store the candidate indices. Since we use small block sizes, in particular $p = 3$, this preprocessing is modest. The online assignment cost is $O(mkp)$, which is substantially smaller than the exact cost $O(mKp)$ when $k \ll K$. In our experiments, this approximate assignment is used for `BlockQuant` unless otherwise specified.

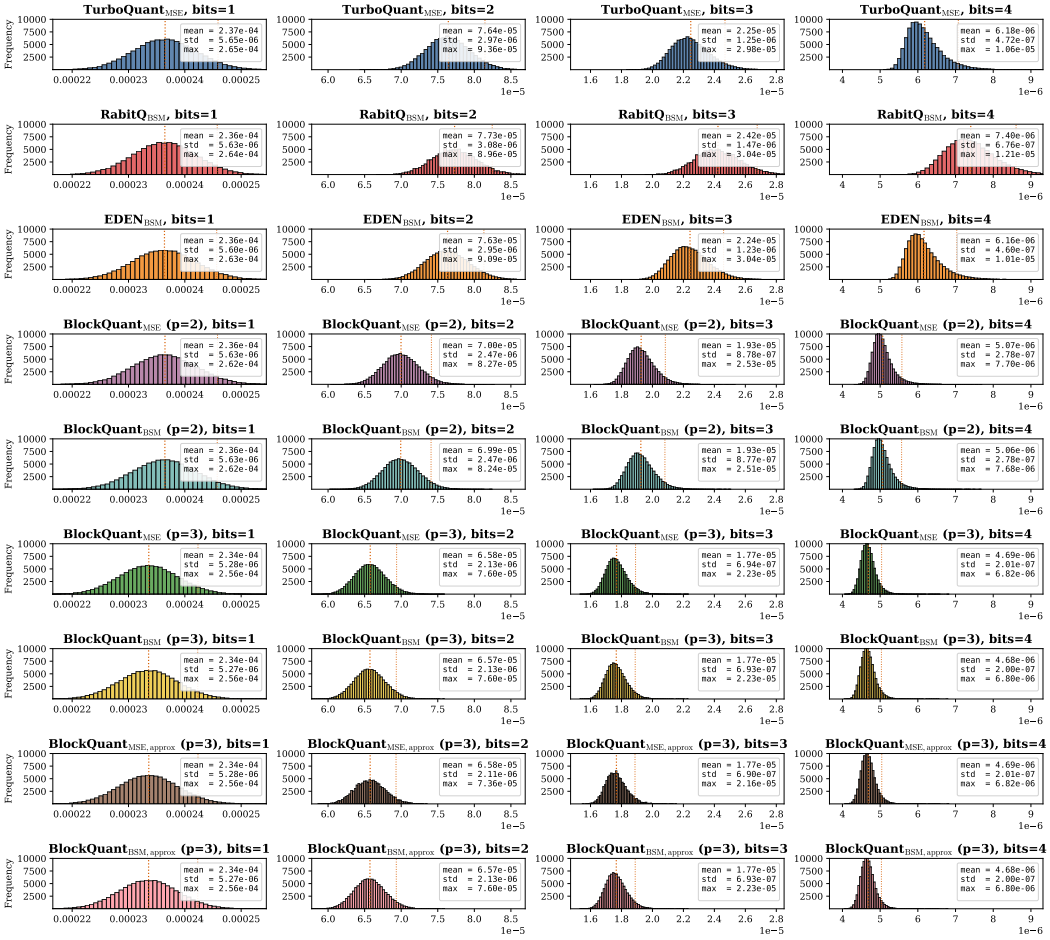

Figure L.1: Distribution of MSE.

L.2 Quantization Accuracy.

Reconstruction MSE. We first evaluate reconstruction accuracy on DBpedia Entities [19] using $1,536$-dimensional embeddings. We sample $100,000$ database vectors and normalize them to unit norm, matching the setting of our theoretical analysis. For each database vector $\mathbf{x}_i$, we measure the squared reconstruction error $e_i = \|\widehat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2$, where $\widehat{\mathbf{x}}_i$ denotes the reconstructed vector.

Figure L.1 shows the distribution of the squared reconstruction error. Both $\texttt{BlockQuant}_{\text{MSE}}$ and $\texttt{BlockQuant}_{\text{BSM}}$ with block size $p = 3$ achieve the smallest distortion among the compared methods, supporting the advantage of block-spherical centroids over coordinate-wise codebooks. We also observe no noticeable degradation from the approximate nearest-centroid search, suggesting that the approximation preserves the expected-distortion advantage of `BlockQuant` in practice. These results further indicate that $\texttt{BlockQuant}_{\text{BSM}}$ can serve as an effective reconstruction method in practical settings.

Inner product error. We next evaluate inner-product estimation accuracy on DBpedia Entities [19] using $1,536$-dimensional embeddings, with $100,000$ database vectors and $1,000$ query vectors. All vectors are normalized, and only the database vectors are quantized. For each pair $(\mathbf{x}_i, \mathbf{y}_j)$, we measure the inner-product estimation error $e_{ij} = \langle \widehat{\mathbf{x}}_i, \mathbf{y}_j \rangle - \langle \mathbf{x}_i, \mathbf{y}_j \rangle$, where $\widehat{\mathbf{x}}_i$ is the dequantized output of each algorithm.

Figure L.2 shows that the ratio-based estimators are centered near zero, confirming their empirical unbiasedness. Moreover, $\texttt{BlockQuant}_{\text{UB}}$ yields a more concentrated error distribution than the

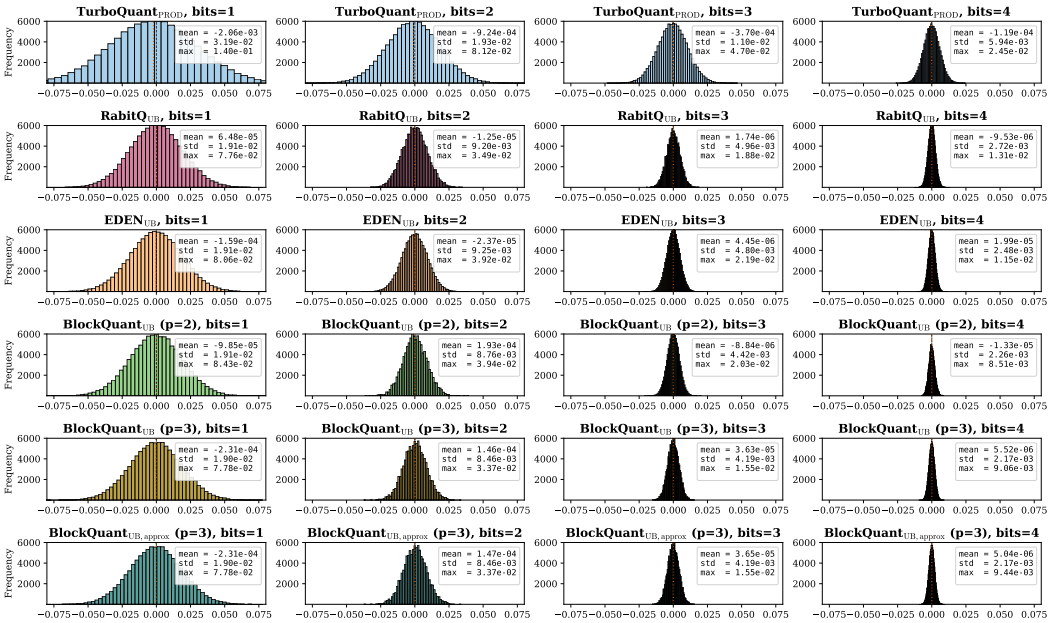

Figure L.2: Distribution of inner product error

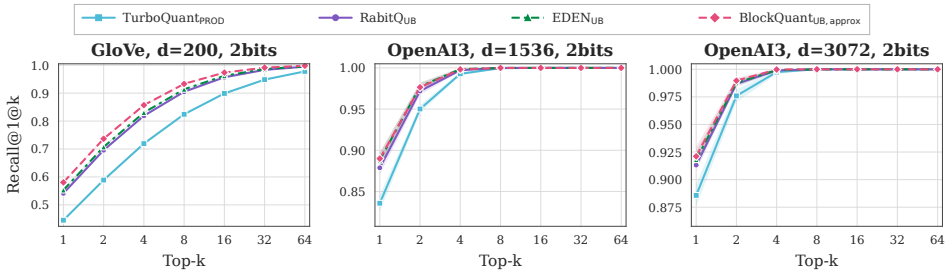

Figure L.3: Recall comparison at 2 bits across different datasets.

coordinate-wise baselines, especially at moderate bit-widths. This is consistent with Corollary 3, which predicts a smaller expected inner-product distortion for the block-spherical construction.

L.3 Nearest-Neighbor Search.

We evaluate retrieval quality using Recall@1@k, which measures whether quantization preserves the exact nearest neighbor within the top-k retrieved candidates, rather than only measuring average inner-product estimation error.

Figure 2 compares the methods on GloVe ($d = 200$) and OpenAI3/DBpedia ($d = 1536$ and 3072) under 4-bit compression; the corresponding 2-bit results are provided in Figure L.3. Across both datasets, the approximate version of $\texttt{BlockQuant}_{\text{UB}}$ achieves strong recall, with the largest gains in the low-k regime. This regime is especially sensitive to quantization error, since small perturbations in inner-product estimates can change the ordering of the top-ranked candidates.

These results show that the smaller expected inner-product distortion of $\texttt{BlockQuant}_{\text{UB}}$ translates into improved nearest-neighbor retrieval under the same bit budget. In particular, the block-spherical construction improves not only pointwise estimation accuracy but also the ranking quality that is central to approximate nearest-neighbor search.

L.4 Quantization Efficiency

We compare GPU-based quantization runtime across bit-widths in Tables L.1 and L.2. Among baselines, EDEN variants are consistently the fastest and nearly constant across bit-widths, reflecting their coordinate-wise structure. RABITQ incurs moderate overhead, while TURBOQUANT becomes slower at higher bit-widths due to additional correction steps.

Table L.1: Runtime comparison for IP-based methods (seconds).

Method	Bitwidth			
	1-bit	2-bit	3-bit	4-bit
$\texttt{TurboQuant}_{\text{PROD}}$	**0.0225**	0.0658	0.0627	0.0661
$\texttt{RabitQ}_{\text{UB}}$	0.0237	0.0391	0.0402	0.0392
$\texttt{EDEN}_{\text{UB}}$	**0.0220**	**0.0222**	**0.0221**	**0.0223**
$\texttt{BlockQuant}_{\text{UB}}(p=2)$	0.0238	0.0259	0.0372	0.0805
$\texttt{BlockQuant}_{\text{UB}}(p=3)$	0.0236	0.0318	0.1009	0.6573
$\texttt{BlockQuant}_{\text{UB,approx}}(p=3)$	0.0251	**0.0252**	**0.0254**	**0.0270**

Table L.2: Runtime comparison for non-IP methods (seconds).

Method	Bitwidth			
	1-bit	2-bit	3-bit	4-bit
$\texttt{TurboQuant}_{\text{MSE}}$	**0.0219**	**0.0220**	**0.0218**	**0.0222**
$\texttt{RabitQ}_{\text{BSM}}$	0.0247	0.0392	0.0394	0.0402
$\texttt{EDEN}_{\text{BSM}}$	0.0220	**0.0222**	**0.0221**	**0.0223**
$\texttt{BlockQuant}_{\text{MSE}}(p=2)$	0.0220	0.0242	0.0345	0.0778
$\texttt{BlockQuant}_{\text{MSE}}(p=3)$	**0.0218**	0.0301	0.0961	0.6531
$\texttt{BlockQuant}_{\text{MSE,approx}}(p=3)$	0.0236	0.0237	0.0237	0.0265
$\texttt{BlockQuant}_{\text{MSE}}(p=3)$	0.0249	0.0284	0.0294	0.0350
$\texttt{BlockQuant}_{\text{MSE,approx}}(p=3)$	0.0273	0.0273	0.0277	0.0340

For `BlockQuant`, the exact assignment cost grows rapidly with both block size and bit-width, especially for $p = 3$, where the runtime reaches 0.6573 seconds at 4 bits. In contrast, the approximate version removes this dependence on codebook size and remains nearly constant across bit-widths (e.g., 0.0251–0.0270 seconds for $\texttt{BlockQuant}_{\text{UB,approx}}$). A similar trend holds for MSE and UR variants in Table L.2.

Overall, the LUT-based approximation reduces the complexity from full codebook search to a small candidate search, bringing `BlockQuant` to a runtime comparable with the fastest baselines while preserving its accuracy advantages.

L.5 Computational Resources

Except for the KV-cache quantization experiments, all experiments are conducted on a GPU server with eight NVIDIA GeForce RTX 3090 GPUs, each with 24 GiB of VRAM, together with a dual-socket CPU server containing two Intel Xeon Gold 6226R processors, for a total of 32 cores and 32 threads. The KV-cache quantization experiments are conducted on a separate server with four NVIDIA H100 SXM5 GPUs, each with 80 GiB of VRAM, and two Intel Xeon Platinum 8592+ processors, totaling 128 cores and 128 threads.

L.6 Comparison with QuaRot for KV-Cache Quantization

Experimental settings. To complement the comparisons with general-purpose rotation-based quantizers, we evaluate the KV-cache component of `QuaRot` [2], which combines Hadamard rotations with group-wise scalar quantization. We adapt this component to the KV-cache-only setting of Section 6.2, using Llama-3.1-8B-Instruct with 4-bit keys and 4-bit values for both methods, while

keeping all other model components in full precision. We follow the same benchmark and evaluation protocols as in Section 6.2. For `BlockQuant`, we report results over five independent random rotation seeds. The adapted `QuaRot` baseline uses fixed exact Hadamard transforms and is evaluated once.

Results. Table L.3 shows that `BlockQuant` achieves a Needle-In-A-Haystack score of 0.964, improving over `QuaRot` (0.955) and closely matching the full-cache result (0.965). On LongBench-E, Table L.4 shows similar average performance for `BlockQuant` (44.15), `QuaRot` (44.18), and the full cache (44.15). Thus, at 4-bit key and value precision, `BlockQuant` improves retrieval accuracy on Needle-In-A-Haystack while maintaining comparable performance on LongBench-E.

Table L.3: Evaluation of Llama-3.1-8B-Instruct on Needle-In-A-Haystack with 4-bit keys and values. For `BlockQuant`, results are reported as mean (standard deviation) over five random rotation seeds, with the 95% bootstrap confidence interval shown in brackets below. Bold denotes the best mean among quantized methods.

Method	**K/V bits**	**Average score**
`QuaRot`	4/4	0.955
`BlockQuant`	4/4	**0.964** (0.003) [0.963, 0.966]
Full Cache	16/16	0.965

Table L.4: Evaluation of Llama-3.1-8B-Instruct on LongBench-E with 4-bit keys and values. For `BlockQuant`, results are reported as mean (standard deviation) over five random rotation seeds, with 95% bootstrap confidence intervals shown in brackets below. Bold denotes the best mean among quantized methods.

Method (K/V bits)	**SingleQA**	**MultiQA**	**Summ**	**Few-shot**	**Synthetic**	**Code**	**Average**
`QuaRot` (4/4)	19.53	**16.86**	**30.26**	68.25	**56.61**	**61.54**	**44.18**
`BlockQuant` (4/4)	**19.73** (0.18) [19.57, 19.86]	16.52 (0.20) [16.35, 16.67]	30.20 (0.08) [30.15, 30.27]	**68.52** (0.11) [68.45, 68.61]	56.24 (0.36) [55.96, 56.55]	61.50 (0.09) [61.44, 61.58]	44.15 (0.09) [44.08, 44.22]
Full Cache (16/16)	19.53	16.54	30.28	68.41	56.41	61.60	44.15

M Limitations

The codebook size of `BlockQuant` grows exponentially with the block size, increasing storage and exact nearest-centroid search costs. We therefore focus on small blocks in our experiments. Although approximate search reduces the computational cost in the tested settings, efficient quantization with larger blocks remains an open direction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims made in the abstract and introduction accurately reflect the paper's contribution

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It is provided in Appendix M.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We present proof for our claims in appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experiments support our claims

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide sufficient instructions and codes in supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We stated about the experimental settings in our main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We use shaded region for statistical significance of the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide our experimental setup in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conducted in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We mention this in supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.