APPENDIX

We now present details of the proofs of the main claims, along with additional experimental observations.

A PROOFS

A.1 Proof of Lemma 1

PROOF. Since Π is orthogonal, $\Pi^T\Pi = I$. We will prove that the lemma holds for the following three distance metrics:

1) Inner product distance with $d(\mathbf{x}, \mathbf{y}) = \langle\Pi\mathbf{x}, \Pi\mathbf{y}\rangle$. We have $\langle\Pi\mathbf{x}, \Pi\mathbf{y}\rangle = (\Pi\mathbf{x})^T(\Pi\mathbf{y}) = \mathbf{x}^T\Pi^T\Pi\mathbf{y} = \mathbf{x}^T\mathbf{y} = \langle\mathbf{x}, \mathbf{y}\rangle$.

2) Cosine distance with $d(\mathbf{x}, \mathbf{y}) = 1 - \frac{\langle\mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}$. We first prove that $\|\mathbf{x}\| = \|\Pi\mathbf{x}\|$ with $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \mathbf{x}^T\Pi^T\Pi\mathbf{x} = \|\Pi\mathbf{x}\|^2$. With $\langle\mathbf{x}, \mathbf{y}\rangle = \langle\Pi\mathbf{x}, \Pi\mathbf{y}\rangle$ proved in 1), it holds for cosine distance.

3) Euclidean distance with $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\langle\mathbf{x}, \mathbf{y}\rangle$. Since $\|\mathbf{x}\|^2 = \|\Pi\mathbf{x}\|^2$ is proved in 2) and $\langle\mathbf{x}, \mathbf{y}\rangle = \langle\Pi\mathbf{x}, \Pi\mathbf{y}\rangle$ is proved in 1), so it holds for Euclidean distance. □

A.2 Proof of Lemma 2

PROOF. Let $\mathbf{x} \in \mathbb{R}^d$ be fixed and let $r = \|\mathbf{x}\|$ denote its norm (length). Since Π is orthogonal and independent of $\mathbf{x}$, the rotated vector $\mathbf{y} = \Pi\mathbf{x}$ is uniformly distributed on the $(d-1)$-dimensional sphere

$$\mathbb{S}^{d-1}(r) = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\| = r\}.$$

By symmetry of the sphere, all dimensions $y_1, \ldots, y_d$ have the same distribution and satisfy $\mathbb{E}[y_i] = 0$, where $\mathbb{E}[\cdot]$ denotes expectation over the randomness of Π. We always have

$$\sum_{i=1}^{d} y_i^2 = \|\mathbf{y}\|^2 = r^2.$$

Taking expectation and using that the y_i^2 are identically distributed, we obtain

$$d\,\mathbb{E}[y_i^2] = \mathbb{E}\Big[\sum_{i=1}^{d} y_i^2\Big] = r^2,$$

so $\mathrm{Var}(y_i) = \mathbb{E}[y_i^2] = r^2/d$, where $\mathrm{Var}(\cdot)$ denotes variance.

Let $\mathbf{u}$ be uniform on the unit sphere $\mathbb{S}^{d-1}(1)$ and let $Z = u_1$. To analyze the high-dimensional behavior, we can represent $\mathbf{u}$ as $\mathbf{g}/\|\mathbf{g}\|$, where $\mathbf{g} = (g_1, \ldots, g_d)$ and each $g_i \sim \mathcal{N}(0, 1)$ independently. Then $Z = u_1 = g_1/\|\mathbf{g}\|$. Since

$$\|\mathbf{g}\|^2 = \sum_{i=1}^{d} g_i^2$$

is the sum of d i.i.d. random variables with $\mathbb{E}[g_i^2] = 1$, the law of large numbers implies $\|\mathbf{g}\|^2/d \to 1$ and hence $\|\mathbf{g}\|/\sqrt{d} \to 1$ in probability. Therefore,

$$\sqrt{d}\,Z = \frac{g_1}{\|\mathbf{g}\|/\sqrt{d}} \Rightarrow \mathcal{N}(0, 1)$$

by Slutsky's theorem [1].

A uniform point on $\mathbb{S}^{d-1}(r)$ can be written as $r\mathbf{u}$, so the coordinate y_i has the same distribution as rZ. Therefore

$$y_i \sim \mathcal{N}\left(0, \frac{r^2}{d}\right) \quad \text{for large } d.$$

When $\mathbf{x}$ is drawn from the whole dataset, its norm $r = \|\mathbf{x}\|$ may vary. By definition,

$$\sigma^2 = \mathbb{E}_{\mathbf{x}}[\|\mathbf{x}\|^2]/d$$

is the average value of the coordinate variance r^2/d over data vectors $\mathbf{x}$. Hence the marginal variance of y_i is σ^2, and in high dimensions the marginal distribution of y_i is well modeled by $\mathcal{N}(0, \sigma^2)$. Since all coordinates are exchangeable and become asymptotically uncorrelated under the random rotation, they can be treated as approximately independent. □

A.3 Proof of Lemma 3

PROOF. The 2-Wasserstein distance between distributions with means μ_P and μ_Q satisfy $W_2(P, Q) \geq \|\mu_P - \mu_Q\|_2$ by the triangle inequality for optimal transport. Since the optimal centroid under squared error loss is the mean, we have $\mathbf{c}_P = \mu_P$ and $\mathbf{c}_Q = \mu_Q$, giving $\|\mathbf{c}_P - \mathbf{c}_Q\|_2 \leq W_2(P, Q) \leq \varepsilon$. □

A.4 Proof of Theorem 4

PROOF. Given a vector $\mathbf{x}$ with its transformed vector $\mathbf{y} = \Pi\mathbf{x}$ partitioned into M vectors $[\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(M)}]$, as analyzed in section 3.2, we have the real distance and approximate distance as

$$d(\mathbf{q}, \mathbf{x}) = \sum_{m=1}^{M} d(\mathbf{q}_{\mathrm{JL}}^m, \mathbf{y}^{(m)}) \tag{1}$$

,

$$d_{\mathrm{approx}}(\mathbf{q}, \mathbf{x}) = \sum_{m=1}^{M} d(\mathbf{q}_{\mathrm{JL}}^m, \hat{\mathbf{y}}^{(m)}). \tag{2}$$

where $\mathbf{q}_{\mathrm{JL}}^m$ is the m-th transformed sub-vector of $\mathbf{q}$ and $\hat{\mathbf{y}}^{(m)}$ is the $\mathbf{q}_{\mathrm{JL}}^m$'s closest codeword in C^m (the codebook in the m-subspace). According to the *triangle inequality*, we have $d(\mathbf{q}, \mathbf{x}) - d_{\mathrm{approx}}(\mathbf{q}, \mathbf{x}) \leq \sum_{m=1}^{M} d(\mathbf{y}^{(m)}, \hat{\mathbf{y}}^{(m)})$. Then under the Euclidean distance metric, we have:

$$(d(\mathbf{y}^{(m)}, \hat{\mathbf{y}}^{(m)}))^2 = \|\mathbf{y}^{(m)} - \hat{\mathbf{y}}^{(m)}\|^2 = \sum_{j=1}^{D_s} (y_j^{(m)} - \hat{y}_j^{(m)})^2.$$

Since C^m is generated through the Cartesian product of one-dimensional (1D) Lloyd-Max quantizers and the codewords are positioned within their regions based on the probability distribution, the data boundaries ± 1 represent the worst-case points.

By symmetry of the near-Gaussian distribution, we consider the effective quantization range $[c_1, c_{D_s}]$ based on the cluster radius of the outermost centroid. The error in each dimension is bounded by the maximum distance among cluster centroids within this range: $\varepsilon_{\max} = -c_1$. The probability that $y_j^{(m)}$ falls in this range is: $P_\varepsilon = P(c_1 \leq y_j^{(m)} \leq c_{D_s}) = P(c_1 \leq y_j^{(m)} \leq -c_1) = \Phi\left(\frac{-c_1}{\sigma}\right) - \Phi\left(\frac{c_1}{\sigma}\right)$ Thus, $d(\mathbf{q}, \mathbf{x}) - d_{\mathrm{approx}}(\mathbf{q}, \mathbf{x}) \leq \sqrt{MD_s}\cdot(-c_1)$ with probability P_ε. □

A.5 Proof of Lemma 5

PROOF. In the primary level, each dimension of codewords $\mathbf{c} \in C^m$ is independent since C^m is generated through the Cartesian Product of Lloyd-Max (i.e., one-dimensional) codewords in 2. So the quantization residual inherits this independence. □

A.6 Proof of Theorem 6

Proof. We verify the decomposition for each metric.

Inner product distance. Here $d(\mathbf{a}, \mathbf{b}) = -\langle \mathbf{a}, \mathbf{b} \rangle$. By linearity of the inner product,

$$d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}}) = d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}}) + d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{r}}). \tag{3}$$

Squared Euclidean distance. Here $d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|^2$. We expand the two components:

$$d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}}) = \|\mathbf{q}_{\mathrm{JL}}\|^2 + \|\hat{\mathbf{y}}\|^2 - 2\langle \mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} \rangle, \tag{4}$$

$$d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{r}}) = \|\mathbf{q}_{\mathrm{JL}}\|^2 + \|\hat{\mathbf{r}}\|^2 - 2\langle \mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{r}} \rangle. \tag{5}$$

Summing (4)–(5) gives

$$d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}}) + d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{r}}) = 2\|\mathbf{q}_{\mathrm{JL}}\|^2 + \|\hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{r}}\|^2 - 2\langle \mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}} \rangle. \tag{6}$$

Next, expand the distance to the reconstruction:

$$\begin{aligned} d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}}) &= \|\mathbf{q}_{\mathrm{JL}} - (\hat{\mathbf{y}} + \hat{\mathbf{r}})\|^2 \\ &= \|\mathbf{q}_{\mathrm{JL}}\|^2 + \|\hat{\mathbf{y}} + \hat{\mathbf{r}}\|^2 - 2\langle \mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}} \rangle, \end{aligned} \tag{7}$$

with

$$\|\hat{\mathbf{y}} + \hat{\mathbf{r}}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{r}}\|^2 + 2\langle \hat{\mathbf{y}}, \hat{\mathbf{r}} \rangle. \tag{8}$$

Substituting (8) into (7), and comparing with (6), we obtain the exact identity:

$$d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}}) = d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}}) + d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{r}}) - \|\mathbf{q}_{\mathrm{JL}}\|^2 + 2\langle \hat{\mathbf{y}}, \hat{\mathbf{r}} \rangle. \tag{9}$$

Cosine distance. Here

$$d(\mathbf{a}, \mathbf{b}) = 1 - \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \, \|\mathbf{b}\|}.$$

For the reconstruction,

$$d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}}) = 1 - \frac{\langle \mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}} \rangle}{\|\mathbf{q}_{\mathrm{JL}}\| \, \|\hat{\mathbf{y}} + \hat{\mathbf{r}}\|}. \tag{10}$$

Linearity of the inner product gives

$$\langle \mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}} \rangle = \langle \mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} \rangle + \langle \mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{r}} \rangle.$$

Substituting this into (10) yields

$$d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}} + \hat{\mathbf{r}}) = d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{y}}) + d(\mathbf{q}_{\mathrm{JL}}, \hat{\mathbf{r}}), \tag{11}$$

since both numerator and denominator admit the same additive structure.

Combining (3), (9), and (11) establishes the theorem. □

A.7 Proofs of Theorem 7

Proof. Let $d = MD_s$. By definition, $r_1 = |c_1 - c_2|/2$ where c_1 and c_2 are the two leftmost adjacent Lloyd-Max centroids. In Lloyd-Max quantization, the maximum per-coordinate error occurs when a value falls midway between adjacent centroids. Therefore, for each sub-vector $\mathbf{y}^{(m)}$ with its quantized version $\hat{\mathbf{y}}^{(m)}$ from the primary codebook C^m, the residual satisfies

$$|r_j^{(m)}| = |y_j^{(m)} - \hat{y}_j^{(m)}| \le r_1$$

for all $j \in \{1, \ldots, D_s\}$ and $m \in \{1, \ldots, M\}$.

In the residual level, the scalar codebook C_R^m has $K_r = 2^{B_r}$ centroids on $[-r_1, r_1]$. As an upper bound on k-means performance, uniform quantization over range $2r_1$ yields worst-case error $\Delta_{\max} \le r_1/2^{B_r}$. Since $|r_j^{(m)}| \le r_1$, each residual has quantization error $|r_j^{(m)} - \hat{r}_j^{(m)}| \le \Delta_{\max}$.

Therefore, the total squared error is

$$\varepsilon_{\mathrm{JHQ}}^2 = \|\mathbf{r} - \hat{\mathbf{r}}\|^2 = \sum_{m=1}^{M} \sum_{j=1}^{D_s} |r_j^{(m)} - \hat{r}_j^{(m)}|^2 \le d\,\Delta_{\max}^2,$$

which gives $\varepsilon_{\mathrm{JHQ}} \le \sqrt{d}\,\Delta_{\max} \le \sqrt{d} \cdot \frac{r_1}{2^{B_r}} = \frac{\sqrt{MD_s}\,r_1}{2^{B_r}} = \frac{r_1 \varepsilon_{\mathrm{JQ}}}{(-c_1)2^{B_r}}.$ □

Remark. *This theorem bounds the reconstruction error $\|\mathbf{y} - (\hat{\mathbf{y}} + \hat{\mathbf{r}})\|$. By Theorem 6, the ADC computation in query processing has the same distance as directly computing distances to reconstructed vectors. Therefore, bounding the reconstruction error directly bounds the search quality.*

B ADDITIONAL QUANTITATIVE OBSERVATIONS

This part of the appendix complements the experimental results in the main body. All settings follow Section 5, and the additional results reported here are consistent with the trends discussed in Sections 5.2, 5.3, 5.4, and 5.5.

B.1 Additional Speed-Accuracy Trade-off

Figure 1(a) complements the speed–accuracy evaluation in Section 5.2 by reporting QPS vs. recall@10 on the remaining three datasets (OpenAI3-1536, Vogue-768, and BGE-M3-1024). The experimental setup is identical to that used for Figure 1a in the main body.

On the million-scale OpenAI3-1536 and Vogue-768 datasets, JQ provides the best speed–accuracy trade-off: its primary codes are already sufficiently accurate, so the extra residual refinement in JHQ brings limited accuracy gains but adds overhead and thus lower QPS at fixed recall. As the dataset size grows (e.g., on BGE-M3-1024), the relative benefit of JHQ increases: shorter primary codes help prune candidates more aggressively, partially offsetting the residual-level cost, in line with the trend observed for Stella-TREC24-1024 in the main body. Across all three datasets, our methods remain substantially faster than IRVQ and LSQ++ at comparable recall, and JQ is consistently more accurate than PQ and OPQ under the same throughput.

Dataset characteristics. These three benchmarks also span diverse dataset characteristics in terms of dimensionality, LID, and RC, corresponding to increasingly challenging ANN regimes. On datasets with high LID and small RC, all methods become harder to accelerate, yet our methods still maintain a clear advantage. For example, on Vogue-768, which exhibits high LID and small RC, JQ achieves 1595 QPS at 90% recall, about 2.3× faster than PQ and 1056× faster than IRVQ.

Table 1: Additional validation of distance error bounds.

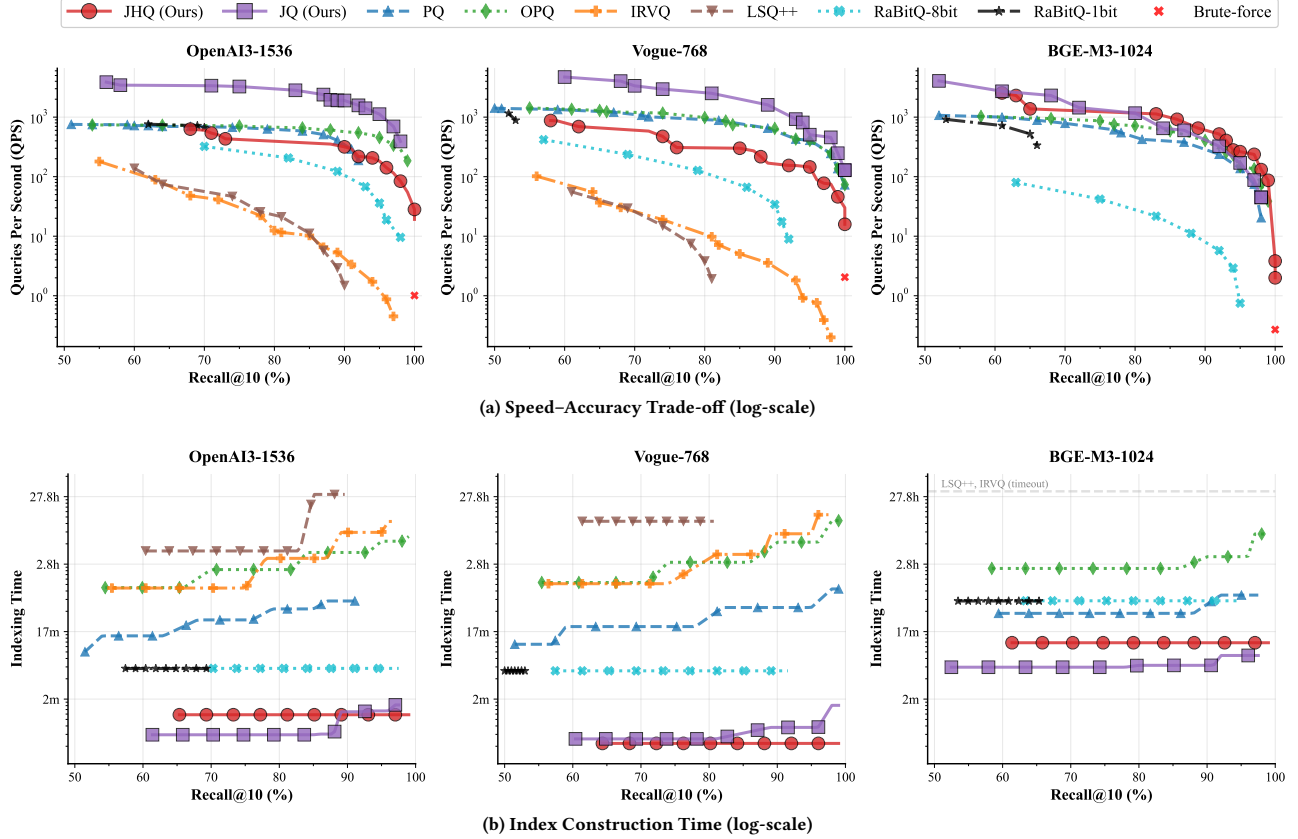| Dataset	JQ		JHQ	
	Emp	Thr	Emp	Thr
OpenAI3-1536	0.326	0.675	0.0069	0.0422
Vogue-768	0.599	0.675	0.0226	0.0422
BGE-M3-1024	0.335	0.675	0.0093	0.0422

Figure 1: Additional efficiency evaluation across three datasets: (a) QPS vs. recall@10; (b) index construction time.

Additional Distance Error. Table 1 complements Table **??** by reporting empirical maximum distance errors for JQ and JHQ on the same three datasets. We use the same parameter settings ($D_s = 8$, $B = 8$, $B_r = 4$), which yield the same theoretical bounds 6.75×10^{-1} for JQ and 4.22×10^{-2} for JHQ on all datasets. As shown in Table 1, these empirical errors are all below the corresponding theoretical bounds for both JQ and JHQ, further validating Theorems 4 and 7.

B.2 Additional Index Build Efficiency

Figure 1(b) reports additional index construction time on the remaining three datasets (OpenAI3-1536, Vogue-768, and BGE-M3-1024). Our methods JQ and JHQ achieve the fastest index construction across all these datasets. It is mainly because that the JL transformation's dimensional independence enables quick Lloyd–Max centroid computation in $O(K)$ time per subspace, while other methods like PQ require expensive k-means clustering with $O(InK)$ and OPQ further needs $O(d^3)$ time to train the rotation matrix; LSQ++ is even more costly since learning its codebook is NP-hard. JHQ takes slightly longer than JQ because the residual level introduces additional overhead, but this extra cost is small compared with the large gap to the baselines, even on the largest BGE-M3-1024 dataset.

B.3 Additional Ablation Study

We extend the ablation study in Section 5.4 to the remaining four datasets (OpenAI3-1536, Vogue-768, ArXiv-Abstracts-768, and BGE-M3-1024), as shown in Figure 2.

Impact of JL Transformation. Figure 2(a)–(d) compare JQ with and without the JL transformation. On all four datasets, JQ with JL achieves higher QPS once recall exceeds about 70% and reaches a much higher maximum recall, whereas the variant without JL is only slightly faster at very low recall and saturates at a much lower recall level because it skips the rotation. This is consistent with the observations in the main body and supports using JL to map features to a near-Gaussian space, which enables efficient codeword computation.

Impact of Hierarchical Design. Figure 2(e)–(h) compare JQ with the hierarchical JHQ index. On the three million-scale datasets (OpenAI3-1536, ArXiv-Abstracts-768, and Vogue-768), JQ is generally faster at fixed recall, since its primary codes are already accurate and the residual refinement in JHQ brings limited accuracy gains but adds computation. On the ten-million-scale BGE-M3-1024 dataset, the gap between JQ and JHQ becomes smaller and JHQ starts to provide a better speed–accuracy trade-off at high recall, which is consistent with the behavior observed on Stella-TREC24-1024 and our analysis in Section 5.2.

Scalar vs. Vector Quantization. Figure 3 provides the speed–accuracy curves for the scalar vs. vector quantization ablation on
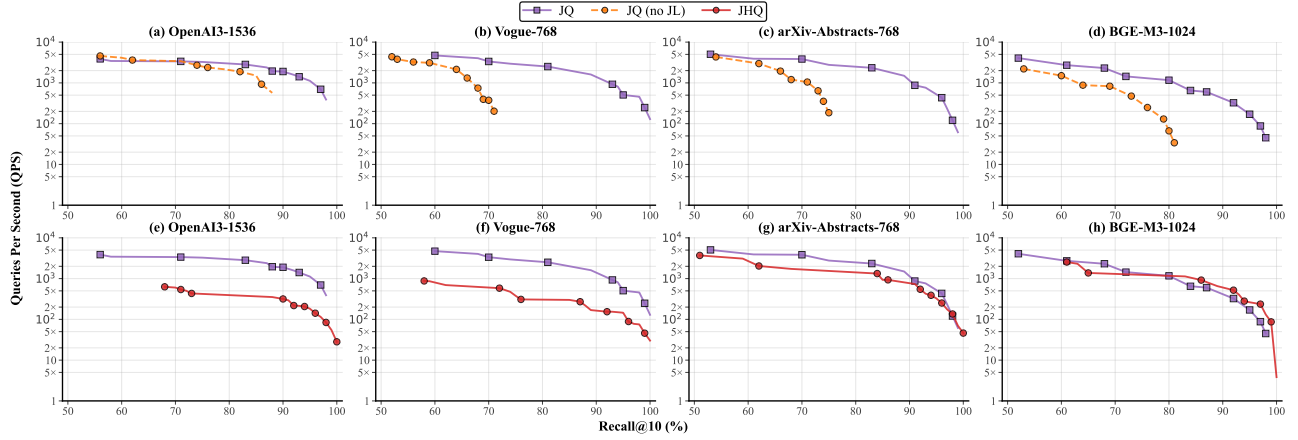
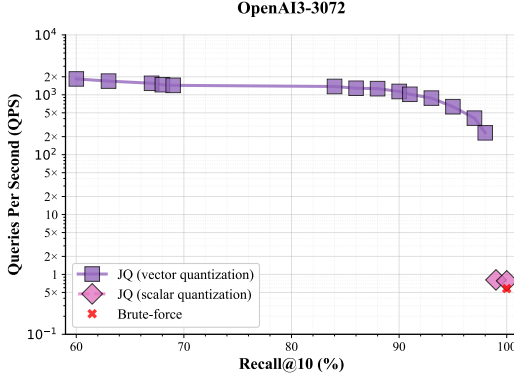Figure 2: Additional ablation study. (a)–(d) JL transformation; (e)–(h) hierarchical design.

Figure 3: Ablation Study: scalar vs. vector quantization

OpenAI3-3072 discussed in the main body. Consistent with Section 5.4, scalar quantization preserves accuracy (recall > 99%) but yields much lower throughput: at 99% recall it is about 285× slower than JQ and nearly as slow as brute-force search. We therefore

do not repeat this ablation on the remaining datasets, since scalar quantization is already dominated by JQ with vector quantization.

B.4 Additional Sensitivity Analysis of Refinement Factor α

The refinement factor α in JHQ controls the candidate set size for residual-level refinement. Figure 4 extends the analysis in Section 5.5 to the remaining three datasets (OpenAI3-1536, Vogue-768, and BGE-M3-1024). Across all cases, $\alpha = 4.0$ provides the best or near-best speed–accuracy trade-off, while $\alpha = 2.0$ consistently yields lower QPS at high recall because the candidate set is too small and may miss true neighbors. Larger values, such as $\alpha = 8.0$ and $\alpha_{max} = N/k$, introduce more residual distance computations and thus reduce QPS, especially on the larger BGE-M3-1024 dataset. These results are consistent with the main-body observations.

REFERENCES

[1] Aad W Van der Vaart. 2000. *Asymptotic statistics*. Vol. 3. Cambridge university press.
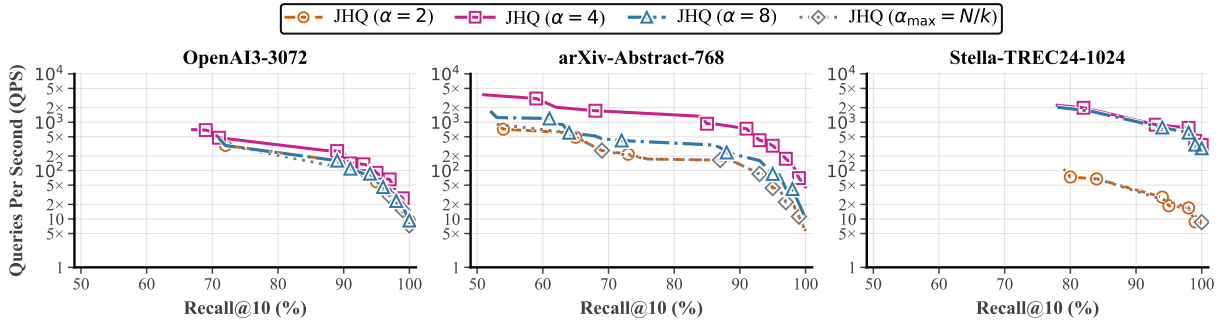
Figure 4: Additional sensitivity analysis on α.