Post-Extraction Compression of Pathology Foundation-Model Embeddings: Storage, Fidelity, and Downstream Utility

Anonymous authors
Paper under double-blind review

Abstract

Whole-slide learning and retrieval often reuse large stores of patch embeddings extracted by a frozen pathology foundation model. We evaluate compression of 1,294,852 CONCH v1.5 embeddings from 1,701 TCGA glioma slides (877 patients) at 10× magnification. We measure stored bytes, CPU decode time, reconstruction and retrieval distortion, attention stability, and three-class multiple-instance learning (MIL) under patient-disjoint evaluation. Per-channel INT4 achieves 8.95× measured compression and a five-seed mean macro-AUROC of 0.8400, compared with 0.8364 for FP32. The patient-clustered 95% interval for the difference between five-seed ensembles is $[-0.0005, 0.0164]$, above the specified -0.01 engineering margin. Adding one packed residual-sign bit per coordinate lowers mean attention total variation from 0.0305 to 0.0166 and increases top-5% attention-set Jaccard overlap from 0.921 to 0.955, while reducing compression to 6.92×. Normalization followed by Haar rotation reduces the largest recorded coordinate by 24.5×, but does not reduce maximum empirical excess kurtosis; the measured per-channel rotated decoder takes 3.57 seconds per bag, compared with 22.28 ms for INT4. We derive a bag-size-independent bound on gated-attention perturbation. These results support low-bit storage for this encoder, cohort, and MIL model while showing that attention fidelity and decode cost require separate evaluation.

1 Introduction

Whole-slide images (WSIs) contain more pixels than a slide-level model can ordinarily process at once. A common approach divides each slide into patches, encodes them with a frozen foundation model, and learns from the resulting bag of embeddings. Models such as CONCH, UNI, and Prov-GigaPath provide representations for computational-pathology tasks (Lu et al., 2024; Chen et al., 2024; Xu et al., 2024). Repeated training and retrieval over large collections of embeddings can make their storage and decoding consequential to the overall system.

Reducing the bit width of stored embeddings may alter more than reconstruction error. It can change nearest-neighbor rankings, the attention assigned to individual patches, and slide-level predictions. These outcomes need distinct measurements: a small average error alone does not bound every patch's attention change, and slide-level metrics can obscure changes in selected regions. Multiple slides from one patient also require patient-level separation and patient-clustered uncertainty estimates.

We study compression after feature extraction, leaving the pathology encoder fixed. We compare FP16, per-channel INT8 and INT4, INT4 with a packed sign correction, and two rotation-based settings. The evaluation links measured storage and CPU decoding to representation distortion, retrieval rankings, attention distributions, and five-seed MIL performance on a patient-disjoint TCGA glioma split. It also gives sufficient bounds on the effect of bounded embedding error on gated attention, pooled features, logits, and retrieval scores.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Three findings organize the study. First, per-channel INT4 meets the specified engineering tolerance for ensemble macro-AUROC at 8.95× measured compression in this setting. Second, an additional sign bit improves attention agreement with FP32 while increasing bytes and CPU decode time. Third, normalization followed by rotation reduces the largest recorded coordinate without improving the observed maximum excess kurtosis, and the per-channel rotated implementation has substantially higher decode latency. The conclusions concern this encoder, cohort, magnification, and downstream model; the tested codecs do not constitute new quantization primitives.

2 Related Work

Foundation models and MIL. Attention-based MIL aggregates an unordered bag while learning which instances support a slide label (Ilse et al., 2018). It underlies many WSI systems, including data-efficient and clinical-scale pipelines (Campanella et al., 2019; Lu et al., 2021). Transformer aggregators such as TransMIL model interactions among instances at greater computational cost (Shao et al., 2021). Our study isolates the feature-store representation and uses gated-attention MIL as a controlled downstream probe.

Compression and retrieval. Product quantization and compressed indices are standard in approximate nearest-neighbor search (Jégou et al., 2010; Johnson et al., 2019). Low-bit model quantization addresses weights or activations during inference (Frantar et al., 2023; Lin et al., 2024); our setting instead stores frozen, already-extracted patch descriptors for later consumers. QuIP and QuIP# use incoherence processing to suppress problematic coordinates before low-bit weight quantization (Tseng et al., 2023; Chee et al., 2024), motivating our rotation ablation.

Pathology-specific compression. PathVQ learns a vector-quantized representation of spatial tokens and a decoder, jointly supporting slide-level pretraining (Li et al., 2025). CICS learns compact features and selects informative WSI instances to reduce MIL compute (Zheng et al., 2026). These methods change or train parts of the representation pipeline. We instead test label-free storage codecs applied after extraction, preserving the frozen encoder and, for original-coordinate codecs, the coordinate system expected by an existing MIL model. Whole-slide retrieval has also motivated compact pathology representations (Chen et al., 2022; Kalra et al., 2020). Here, exact-index rankings serve as a measure of distortion relative to the uncompressed embeddings, without a new retrieval objective.

3 Compression and Attention Stability

3.1 Problem and codecs

A slide is a bag $\mathcal{B} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$. A codec $(\mathcal{E}, \mathcal{D})$ stores the bag in a measured file and reconstructs $\hat{\boldsymbol{x}}_i$. We report $\mathrm{CR} = B_{\mathrm{FP32}}/B_{\mathrm{codec}}$, the ratio of measured FP32 to codec file sizes for matched bags, including codec metadata. Nominal bits per coordinate do not account for this overhead. Write $\boldsymbol{e}_i = \hat{\boldsymbol{x}}_i - \boldsymbol{x}_i$. We separate average distortion from the worst-case quantity needed for a certificate:

$$
\begin{aligned}
D_{\mathrm{RMSE}} &= \left(\frac{1}{Nd} \sum_{i=1}^{N} \|\boldsymbol{e}_i\|_2^2 \right)^{1/2}, \\
\epsilon &= \max_i \|\boldsymbol{e}_i\|_2, \\
D_{\cos} &= 1 - \frac{1}{N} \sum_i \frac{\boldsymbol{x}_i^\top \hat{\boldsymbol{x}}_i}{\|\boldsymbol{x}_i\|_2 \|\hat{\boldsymbol{x}}_i\|_2}.
\end{aligned}
\tag{1}
$$

The first and third terms describe typical reconstruction; ϵ controls downstream worst-case perturbations. We assess stored size, CPU decode time, reconstruction, and downstream outcomes rather than rely on nominal bit width alone.

Per-channel integer quantization. For bit width b and $M = 2^{b-1} - 1$, a symmetric scale is

$$s_j = \frac{\max_i |x_{ij}|}{M}, \qquad q_{ij} = \mathrm{clip}(\lfloor x_{ij}/s_j \rceil, -M, M), \qquad \hat{x}_{ij} = s_j q_{ij}. \tag{2}$$

Scales are computed over the embeddings being encoded, including the unlabeled test embeddings, so Equation 2 does not clip stored values in this experiment. Nearest rounding therefore gives the deterministic bounds

$$|e_{ij}| \le \frac{s_j}{2}, \qquad \|\boldsymbol{e}_i\|_2 \le \frac{1}{2}\|\boldsymbol{s}\|_2, \qquad D^2_{\mathrm{RMSE}} \le \frac{1}{4d}\|\boldsymbol{s}\|_2^2. \tag{3}$$

For a future embedding outside the calibration range, the error bound must also account for clipping; Section 6 discusses this restriction. INT4 codes are packed two per byte. We compare FP16, INT8, and INT4 against FP32.

Residual-sign correction. Given the INT4 reconstruction $\tilde{\boldsymbol{x}}_i$, let $\boldsymbol{r}_i = \boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i$ and let $s_{ij} = +1$ for $r_{ij} \ge 0$ and $s_{ij} = -1$ otherwise. This zero-residual convention ensures $\|\boldsymbol{s}_i\|_2^2 = d$. One additional packed bit per coordinate stores $\boldsymbol{s}_i$, while one FP32 scalar per patch stores

$$\gamma_i^* = \arg\min_{\gamma} \|\boldsymbol{r}_i - \gamma \boldsymbol{s}_i\|_2^2 = \frac{1}{d}\sum_{j=1}^{d} |r_{ij}|. \tag{4}$$

The reconstruction is $\hat{\boldsymbol{x}}_i = \tilde{\boldsymbol{x}}_i + \gamma_i^* \boldsymbol{s}_i$. This additional representation tests whether one extra residual bit improves attention agreement enough to justify its storage cost. The resulting residual error and exact squared-error reduction are

$$\|\boldsymbol{r}_i - \gamma_i^* \boldsymbol{s}_i\|_2^2 = \sum_{j=1}^{d} \big(|r_{ij}| - \gamma_i^*\big)^2, \qquad \|\boldsymbol{r}_i\|_2^2 - \|\boldsymbol{r}_i - \gamma_i^* \boldsymbol{s}_i\|_2^2 = d(\gamma_i^*)^2. \tag{5}$$

Thus the sign correction cannot increase patchwise squared error. Its nominal payload is $\lceil d/8 \rceil + 4$ bytes per patch—one packed sign bit per coordinate and one FP32 gain—in addition to the packed INT4 coarse code and amortized scale metadata.

Rotation variants. We sample a Haar orthogonal matrix $\boldsymbol{R}$ by QR decomposition of a seeded Gaussian matrix (Mezzadri, 2007). Each embedding is normalized, rotated, and quantized as $\boldsymbol{R}\boldsymbol{x}_i/\|\boldsymbol{x}_i\|_2$; reconstruction also requires its stored norm and an inverse rotation. We compare fixed-scale and per-channel scaling; the measured per-channel rotation arm additionally includes residual signs. Orthogonality preserves exact inner products before quantization; it does not guarantee that a finite, fixed dataset has lower empirical kurtosis after one sampled rotation. For $\boldsymbol{y} = \boldsymbol{R}\boldsymbol{x}/\|\boldsymbol{x}\|_2$ and its quantized reconstruction $\hat{\boldsymbol{y}}$, orthogonality gives

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2 \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2, \qquad \mathbb{E}_{\boldsymbol{R}}\big[(\boldsymbol{R}\boldsymbol{x})_j^2\big] = \frac{\|\boldsymbol{x}\|_2^2}{d}. \tag{6}$$

The expectation concerns one fixed vector averaged over random rotations. It gives no deterministic guarantee about marginal kurtosis across the observed dataset for one sampled rotation.

3.2 A bag-size-independent gated-attention bound

At evaluation time dropout is inactive. The implemented MIL model first projects a stored embedding,

$$\boldsymbol{u}_i = \phi(\boldsymbol{x}_i) = \mathrm{ReLU}(\boldsymbol{P}\boldsymbol{x}_i + \boldsymbol{b}_P), \qquad L_\phi = \|\boldsymbol{P}\|_2, \tag{7}$$

then assigns score and attention

$$g(\boldsymbol{u}) = \boldsymbol{w}^\top [\tanh(\boldsymbol{V}\boldsymbol{u} + \boldsymbol{b}_V) \odot \sigma(\boldsymbol{U}\boldsymbol{u} + \boldsymbol{b}_U)] + b_w, \qquad a_i = \frac{\exp(g(\boldsymbol{u}_i))}{\sum_k \exp(g(\boldsymbol{u}_k))}. \tag{8}$$

Let $\|\cdot\|_2$ denote the Euclidean or spectral norm according to context. With $\boldsymbol{D}_{\tanh}$ and $\boldsymbol{D}_\sigma$ denoting diagonal Jacobians,

$$\begin{aligned}\nabla g(\boldsymbol{u}) &= \boldsymbol{V}^\top \boldsymbol{D}_{\tanh}(\boldsymbol{V}\boldsymbol{u}+\boldsymbol{b}_V)[\boldsymbol{w}\odot\sigma(\boldsymbol{U}\boldsymbol{u}+\boldsymbol{b}_U)] \\ &\quad + \boldsymbol{U}^\top \boldsymbol{D}_\sigma(\boldsymbol{U}\boldsymbol{u}+\boldsymbol{b}_U)[\boldsymbol{w}\odot\tanh(\boldsymbol{V}\boldsymbol{u}+\boldsymbol{b}_V)], \qquad \|\nabla g(\boldsymbol{u})\|_2 \le L_g, \\ L_g &= \|\boldsymbol{w}\|_2\left(\|\boldsymbol{V}\|_2+\tfrac{1}{4}\|\boldsymbol{U}\|_2\right), \qquad L_{\text{gate}} = L_\phi L_g.\end{aligned} \tag{9}$$

For arbitrary logits $\boldsymbol{z}$ and perturbation $\boldsymbol{d}$, the corresponding softmax distributions satisfy

$$\mathrm{TV}(\mathrm{softmax}(\boldsymbol{z}),\mathrm{softmax}(\boldsymbol{z}+\boldsymbol{d})) \le \tanh\left(\frac{\mathrm{osc}(\boldsymbol{d})}{4}\right), \quad \mathrm{osc}(\boldsymbol{d}) = \max_i d_i - \min_i d_i. \tag{10}$$

Theorem 1 (Gated-attention stability). If $\max_i \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|_2 \le \epsilon$, then

$$\mathrm{TV}(\boldsymbol{a},\hat{\boldsymbol{a}}) \le \tanh\left(\frac{L_{\text{gate}}\epsilon}{2}\right) \le \frac{L_{\text{gate}}\epsilon}{2}. \tag{11}$$

The bound is independent of the number of patches in the bag.

ReLU is 1-Lipschitz, so $\|\phi(\boldsymbol{x}_i)-\phi(\hat{\boldsymbol{x}}_i)\|_2 \le L_\phi\epsilon$. Equations 9 and 10 then give $\mathrm{osc}(\hat{\boldsymbol{z}}-\boldsymbol{z}) \le 2L_{\text{gate}}\epsilon$. Appendix A proves the softmax inequality and all norm steps.

Proposition 1 (Pooled-representation stability). Set $\epsilon_\phi = L_\phi\epsilon$. If $\max_i \|\boldsymbol{u}_i\|_2 \le R$, then for $\boldsymbol{h} = \sum_i a_i\boldsymbol{u}_i$ and $\hat{\boldsymbol{h}} = \sum_i \hat{a}_i\phi(\hat{\boldsymbol{x}}_i)$,

$$\|\boldsymbol{h}-\hat{\boldsymbol{h}}\|_2 \le \rho(\epsilon,R) := \epsilon_\phi + 2(R+\epsilon_\phi)\tanh(L_{\text{gate}}\epsilon/2). \tag{12}$$

From pooled features to a decision certificate. The implemented classifier is $\boldsymbol{\ell} = \boldsymbol{C}_2\,\mathrm{ReLU}(\boldsymbol{C}_1\boldsymbol{h}+\boldsymbol{b}_1)+\boldsymbol{b}_2$. Writing $L_{\text{cls}} = \|\boldsymbol{C}_2\|_{2\to\infty}\|\boldsymbol{C}_1\|_2$, where $\|\boldsymbol{C}_2\|_{2\to\infty}$ is its largest row norm,

$$\|\boldsymbol{\ell}-\hat{\boldsymbol{\ell}}\|_\infty \le L_{\text{cls}}\rho(\epsilon,R). \tag{13}$$

If the FP32 winning-class margin $m(\boldsymbol{\ell}) = \ell_{(1)}-\ell_{(2)}$ exceeds $2L_{\text{cls}}\rho(\epsilon,R)$, compression cannot change the predicted class. This is a sufficient per-slide condition; measured AUROC and attention changes remain necessary because global worst-case norms can be loose.

A retrieval-order certificate. For any two pooled vectors $\boldsymbol{x},\boldsymbol{y}$ used by retrieval, with $\|\boldsymbol{x}\|_2, \|\boldsymbol{y}\|_2 \le R$ and reconstruction errors at most ϵ,

$$\left|\hat{\boldsymbol{x}}^\top\hat{\boldsymbol{y}} - \boldsymbol{x}^\top\boldsymbol{y}\right| \le 2R\epsilon+\epsilon^2. \tag{14}$$

Consequently, a query's FP32 top-K set is unchanged if its similarity gap between ranks K and $K+1$ exceeds $2(2R\epsilon+\epsilon^2)$. The measured rank overlap in Section 5 tests behavior when this sufficient margin condition does not uniformly hold.

4 Experimental Protocol

4.1 Cohort, representation, and splits

We use TCGA-LGG and TCGA-GBM diagnostic slides for three-class histologic subtyping: astrocytoma, glioblastoma, and oligodendroglioma (Brat et al., 2015). The cohort comprises 1,701 slides from 877 patients, represented at each of 5×, 10×, and 20× magnification, with 6,398,609 patches across the three scales. The primary 10× analysis uses 1,294,852 CONCH v1.5 embeddings of dimension $d = 768$. Train, validation, and test contain 1,378, 156, and 167 slides from 701, 88, and 88 patients, respectively. Across the 1,701 slides, class counts for astrocytoma, glioblastoma, and oligodendroglioma are 783, 162, and 756. The partitions contain no overlapping patients or slides; every bag is nonempty, and a slide retains the same partition across magnifications.

4.2 MIL, inference, and statistics

The MIL model projects each embedding from 768 to 256 dimensions, applies gated attention of width 128, and uses dropout of 0.25 and a two-layer, three-class classifier. Each representation is trained for at most 30 epochs with Adam (learning rate 10^{-4}, weight decay 10^{-5}), early stopping based on validation macro-AUROC with patience 7, and seeds $\{1, 2, 3, 4, 5\}$. The AUROC column in Table 1 averages five separately trained model metrics. For paired inference, we instead average the five probability vectors per slide to form an ensemble; its macro-AUROC and the mean per-seed macro-AUROC are distinct quantities. For class c and $S = 5$ seeds, ensemble probabilities and ensemble macro-AUROC are

$$\bar{p}_{ic} = \frac{1}{S}\sum_{s=1}^{S} p_{ic}^{(s)}, \qquad \mathrm{AUC}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{AUC}(\mathbf{1}[y_i = c], \bar{p}_{ic}), \quad C = 3. \tag{15}$$

Let $\mathcal{G}_p$ be all held-out slide indices for patient p, $p = 1, \ldots, P$ with $P = 88$. For bootstrap replicate b, we draw patient multiplicities

$$\begin{aligned} (m_1^{(b)}, \ldots, m_P^{(b)}) &\sim \mathrm{Multinomial}\big(P; \tfrac{1}{P}, \ldots, \tfrac{1}{P}\big), \\ \mathcal{I}^{(b)} &= \biguplus_{p=1}^{P} m_p^{(b)} \mathcal{G}_p, \\ \Delta^{(b)} &= \mathrm{AUC}_{\mathrm{codec}}(\mathcal{I}^{(b)}) - \mathrm{AUC}_{\mathrm{FP32}}(\mathcal{I}^{(b)}). \end{aligned} \tag{16}$$

Every occurrence of a sampled patient carries all of that patient's slides, preserving within-patient dependence. From 2,000 paired replicates, we report the percentile interval $[\hat{Q}_{0.025}(\Delta), \hat{Q}_{0.975}(\Delta)]$. The representation meets the specified engineering tolerance when

$$\hat{Q}_{0.025}(\Delta) > -\delta, \qquad \delta = 0.01. \tag{17}$$

Bootstrap draws lacking one of the three classes are discarded and redrawn. The 0.01 margin is an engineering tolerance for this outcome and cohort; it has no established clinical interpretation.

We evaluate both retrained models and zero-shot transfer, in which the FP32 model is evaluated directly on compressed features. Attention stability uses the seed-1 FP32 model on all 167 test slides. For attention distributions $\boldsymbol{a}, \hat{\boldsymbol{a}}$ and their top-K index sets $S_K, \hat{S}_K$, the primary distortions are

$$\mathrm{TV}(\boldsymbol{a}, \hat{\boldsymbol{a}}) = \frac{1}{2}\sum_i |a_i - \hat{a}_i|, \qquad J_K = \frac{|S_K \cap \hat{S}_K|}{|S_K \cup \hat{S}_K|}, \qquad K = \lceil 0.05N \rceil. \tag{18}$$

4.3 Retrieval and systems measurements

For every codec, mean, max, and generalized-mean pooled slide vectors are searched with an exact inner-product index over all 1,701 slides. If $\mathcal{N}_{10}^{(0)}(q)$ and $\mathcal{N}_{10}^{(c)}(q)$ are the FP32 and codec neighbor sets after removing the query itself, rank overlap is

$$O_{10}^{(c)} = \frac{1}{Q}\sum_{q=1}^{Q} \frac{|\mathcal{N}_{10}^{(0)}(q) \cap \mathcal{N}_{10}^{(c)}(q)|}{10}. \tag{19}$$

We also report cosine reconstruction. These are distortion measures, not an external clinical retrieval validation. Other slides from the same patient are not removed, a limitation discussed in Section 6.

Storage ratios and CPU decode times are measured on the same five sampled bags, with ten decode repetitions per bag; timings exclude the MIL forward pass and data transfer to the GPU. For comparison, an aggregate container spanning all three magnifications attains 11.04× with INT4 when lossless NPZ compression exploits code redundancy. This container-level ratio is separate from the matched storage ratios in Table 1.

Table 1: Primary results for 10× CONCH embeddings from TCGA glioma. AUROC averages five training seeds; Δ and its 95% interval compare five-seed ensembles with FP32 using 2,000 patient-clustered paired bootstrap replicates on 167 test slides (88 patients). CR is the measured FP32-to-codec file-size ratio, and decode is CPU time per bag on five sampled bags with ten repetitions. Total variation (TV) and top-5% attention-set Jaccard overlap (Jacc.) compare each representation with FP32 on 167 test slides. Dashes indicate unmeasured attention metrics for the fixed-scale rotation setting.

Codec	CR↑ (×)	Decode (ms/bag)	AUROC	Δ [95% CI]	TV↓	Jacc.↑
FP32	1.0	7.05	0.8364	reference	0	1.000
FP16	2.0	9.98	0.8380	+.0040 [−.0042, .0177]	0.0001	1.000
INT8-PC	4.1	17.55	0.8384	+.0074 [−.0068, .0292]	0.0017	0.996
INT4-PC	8.95	22.28	0.8400	+.0063 [−.0005, .0164]	0.0305	0.921
INT4+sign	6.92	38.18	0.8353	+.0022 [−.0019, .0080]	0.0166	0.955
Rot-INT4+sign	6.88	3566.01	0.8362	+.0002 [−.0065, .0058]	0.0189	0.946
Rot-fixed	6.4	35.87	0.8220	−.0143 [−.0382, .0089]	–	–

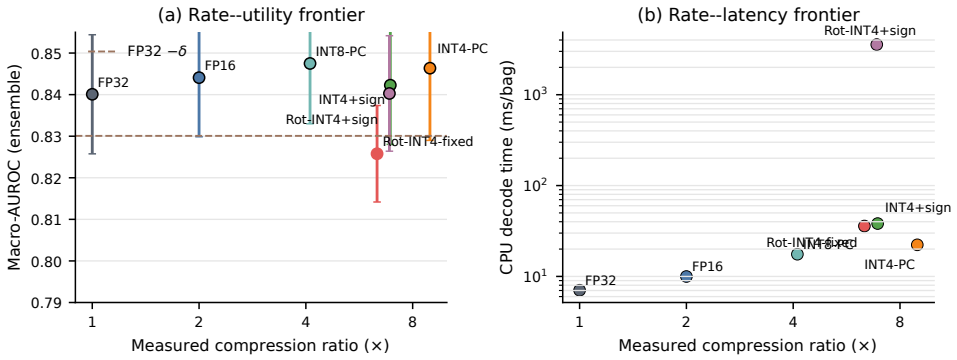

Figure 1: Measured storage ratio and downstream performance for the 10× TCGA glioma analysis. Panel (a) shows mean test macro-AUROC with standard deviations across five seeds; outlined markers have patient-clustered ensemble interval lower bounds above the −0.01 engineering tolerance. Panel (b) shows CPU decode time per bag measured on five sampled bags with ten repetitions.

5 Results

5.1 Compression and slide-level performance

Table 1 and Figure 1 report storage, latency, and slide-level performance. Per-channel INT4 has 8.95× measured compression and mean macro-AUROC 0.8400, compared with 0.8364 for FP32. The ensemble difference is +0.0063, with a patient-clustered 95% interval of $[-0.0005, 0.0164]$; its lower bound exceeds the specified −0.01 tolerance. The difference between the displayed mean AUROCs is 0.0036, because these means are computed over separate seeds rather than from the ensembled probabilities. Without retraining, evaluation of the FP32 ensemble on INT4 features changes ensemble AUROC by +0.00044 in this test cohort.

FP16 and INT8 also meet the specified tolerance, at lower storage ratios than INT4. The fixed-scale rotation setting averages 0.8220 AUROC, and its ensemble interval crosses the −0.01 boundary. The per-channel rotated setting averages 0.8362 and meets the tolerance, but CPU decoding takes 3.57 seconds per bag, versus 22.28 ms for INT4. Among these measured settings, INT4 provides the largest storage reduction of the settings that meet the tolerance and decode within 40 ms per bag.

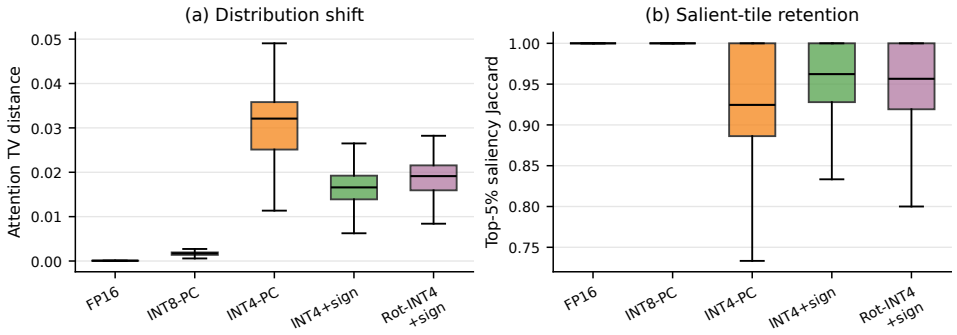

Figure 2: Attention agreement with the FP32 seed-1 MIL model on 167 held-out TCGA glioma slides. Distributions are compared by total variation and top-5% patch-set Jaccard overlap. Boxes show medians and interquartile ranges; whiskers extend to 1.5× the interquartile range.

5.2 Residual signs improve attention agreement

Relative to INT4, the INT4+sign setting reduces mean attention TV by 45.6%, from 0.0305 to 0.0166, and raises mean top-5% Jaccard overlap from 0.9214 to 0.9547 (Figure 2). Median and worst-decile TV are 0.0166/0.0214 with the sign correction and 0.0321/0.0393 with INT4. Compression falls from 8.95× to 6.92×, and CPU decode time rises from 22.28 to 38.18 ms per bag. These measurements quantify attention agreement with the FP32 model; they do not establish that either attention map identifies diagnostically relevant tissue.

The rotated residual setting yields TV 0.0189 and Jaccard 0.9464, compared with 0.0166 and 0.9547 for the original-coordinate sign correction. Orthogonal rotation preserves pre-quantization inner products, but these attention measurements depend on the reconstructed coordinates supplied to the gated network.

5.3 Embedding spectrum and rotation geometry

Among 41,783 sampled patches, the centered embedding matrix has participation ratio 13.7. The first 20 principal components account for 82.0% of variance and the first 50 for 97.9% (Figure 3a). This concentration describes the sampled embeddings; it does not isolate why the downstream model tolerates quantization.

For the sampled embeddings, the recorded maximum absolute coordinate changes from 4.895 before processing to 0.1996 after normalization and Haar rotation, a 24.5× reduction (Figure 3b). This comparison does not isolate the contribution of rotation from that of normalization. Maximum empirical coordinate excess kurtosis increases from 3.38 to 4.79. The expected coordinate energy in Equation 6 averages over rotations of a fixed vector, so it does not imply lower dataset-level marginal kurtosis for this one rotation.

5.4 Retrieval overlap depends on pooling

Figure 4 compares exact-index rankings with FP32. Under mean pooling, top-10 neighbor overlap is 0.9905 for INT4, 0.9945 for INT4+sign, and 0.9962 for rotated residuals. INT4 mean patch cosine similarity is 0.99605; both residual variants exceed 0.9988. Max pooling is less stable: INT4 overlap falls to 0.8315, whereas INT4+sign and rotated residuals reach 0.8819 and 0.8827. GeM lies between mean and max pooling. Compression conclusions therefore depend on the downstream aggregation operator.

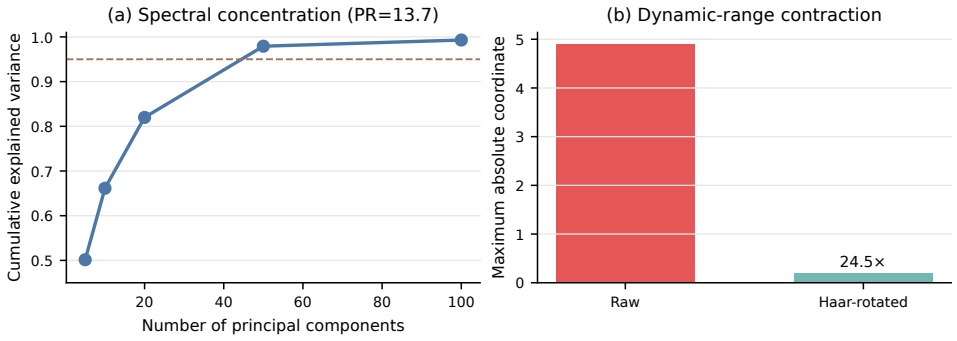

Figure 3: Geometry of 41,783 sampled, centered CONCH embeddings. (a) Principal-component variance spectrum; the first 20 components explain 82.0% of variance. (b) Maximum absolute coordinate before processing and after normalization and Haar rotation. Across coordinates, maximum empirical excess kurtosis increases from 3.38 to 4.79.

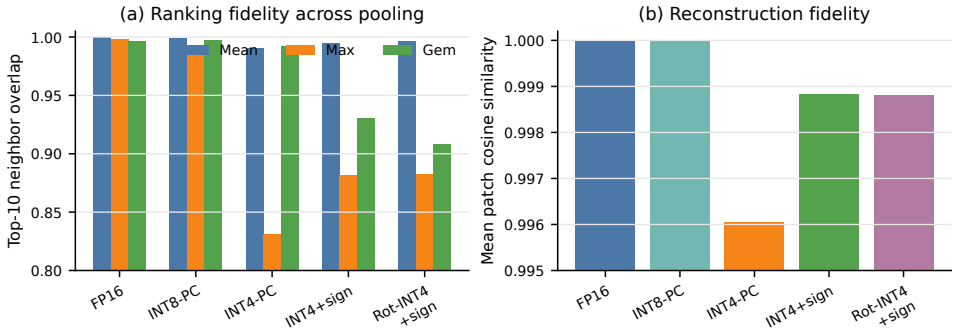

Figure 4: Representation fidelity over all 1,701 TCGA glioma slides. (a) Mean fraction of the FP32 top-10 nearest neighbors retained by each codec, using exact inner-product search after mean, max, or generalized-mean pooling. (b) Mean cosine similarity between original and reconstructed patch embeddings. The query slide is excluded from retrieval, but other slides from its patient remain.

6 Limitations

The slide-level results come from TCGA glioma, CONCH v1.5 embeddings at 10×, and one gated-attention MIL architecture. Compression effects on a different pathology encoder, magnification, institution, or aggregator remain unmeasured. The per-channel scales were fitted to the full encoded collection, including unlabeled test features. This setting applies to compression of an existing collection, whereas an inductive codec must fix its scales on training data and account for out-of-range values. The deterministic bounds in Equation 3 apply to the evaluated unclipped collection.

The retrieval experiment measures agreement with FP32 rankings, not diagnostic retrieval validity. It excludes the query slide but retains other slides from the same patient; patient-excluded retrieval is needed before interpreting label agreement. The 0.01 macro-AUROC tolerance is an engineering choice rather than a clinically justified threshold. Retrospective TCGA histology labels, the single test cohort, and the absence of pathologist-reviewed localization also limit clinical interpretation of slide predictions and attention maps.

Train-calibrated rate sweeps, matched-byte comparisons with learned or product quantization, patient-excluded retrieval, additional magnifications and encoders, and external cohorts would test the scope of these findings. The current evidence supports conclusions only for the evaluated representation and cohort.

7 Conclusion

Per-channel INT4 compresses the evaluated pathology embeddings by 8.95× while the patient-clustered interval for ensemble macro-AUROC remains above the specified −0.01 engineering tolerance. A residual-sign bit improves agreement with the FP32 attention distribution at a measurable storage and decode cost. Normalization followed by rotation reduces the largest recorded coordinate, but the tested per-channel decoder is much slower and empirical maximum excess kurtosis increases. Assessing embedding storage therefore requires direct measurements of bytes, decode time, downstream utility, and attention and retrieval distortion in the intended setting.

Reproducibility Statement

The cohort counts, patient-disjoint split, training configuration, evaluation metrics, bootstrap procedure, and storage and timing protocol are specified in the main text and Appendix B. Reproducing the reported intervals and figures additionally requires the slide and patient identifiers, per-seed predictions, per-slide attention vectors, codec byte counts, and decode timing records.

Ethics Statement

This retrospective analysis uses TCGA data and pre-extracted, de-identified embeddings. The specified engineering tolerance does not establish clinical safety. Attention agreement with an FP32 model does not replace a pathologist's assessment of diagnostic regions.

LLM / AI Usage Statement

In accordance with the ICLR Author Guide, large language models (LLMs) were used as general-purpose assistive tools for plotting code, literature exploration, grammar proofreading, and LaTeX formatting and citation audit. No scientific hypotheses, mathematical derivations, or empirical findings were generated by LLMs. All analyses, codec implementations, mathematical derivations, and experimental results were executed, audited, and verified by the authors, who take full responsibility for all content.

References

Daniel J. Brat, Roel G. W. Verhaak, Kenneth D. Aldape, W. K. Alfred Yung, Sofie R. Salama, Lee A. D. Cooper, Carlos Gilberto Carlotti, Kristen A. Phillips, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. New England Journal of Medicine, 372(26):2481–2498, 2015.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, 25(8):1301–1309, 2019.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In Proceedings of the 41st International Conference on Machine Learning (ICML), volume 235 of Proceedings of Machine Learning Research, pp. 7095–7143. PMLR, 2024.

Chengkuan Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Andrew J. Schaumberg, and Faisal Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. Nature Biomedical Engineering, 6(12):1420–1434, 2022. doi: 10.1038/s41551-022-00929-8.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. Nature Medicine, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In International Conference on Learning Representations (ICLR), 2023.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Proceedings of the 35th International Conference on Machine Learning (ICML), volume 80 of Proceedings of Machine Learning Research, pp. 2127–2136. PMLR, 2018.

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, 2010.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3):535–547, 2019.

Shiv M. Kalra, H. R. Tizhoosh, C. Choi, S. Shah, P. Diamandis, C. J. V. Campbell, and L. Pantanowitz. Yottapixel retrieval of whole slide images for pathology. IEEE Transactions on Medical Imaging, 39(11):3482–3493, 2020.

Honglin Li, Zhongyi Shui, Yunlong Zhang, Chenglu Zhu, and Lin Yang. PathVQ: Reforming computational pathology foundation models for whole-slide image analysis via vector quantization. arXiv preprint arXiv:2503.06482, 2025.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for LLM compression and acceleration. In Proceedings of Machine Learning and Systems (MLSys), volume 6, 2024.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, 5(6):555–570, 2021.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. Nature Medicine, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

Francesco Mezzadri. How to generate random matrices from the classical compact groups. Notices of the American Mathematical Society, 54(5):592–604, 2007.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, and Xiangyang Ji. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In Advances in Neural Information Processing Systems, volume 34, pp. 2136–2147, 2021.

Albert Tseng, Jerry Chee, Qinghao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In Advances in Neural Information Processing Systems (NeurIPS), volume 36, pp. 60086–60117, 2023.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Bifulco, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. Nature, 630 (8015):181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Tingting Zheng, Hongxun Yao, Sicheng Zhao, and Yi Xiao. Content-aware information compression and selection for whole slide image analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, 2026.