TACT-KV: Tri-Axis Cosine Transform for Compressing Volumetric KV Caches in Medical VLMs

Chongyu Qu[1] **Ritchie Zhao**[3] **Yufan He**[3] **Dong Yang**[3] **Zhengyi Lu**[1] **Junchao Zhu**[1]
Tianyuan Yao[1] **Juming Xiong**[1] **Junlin Guo**[1] **Yanfan Zhu**[1] **Yuechen Yang**[1]
Daguang Xu[3] **Bennett A. Landman**[1,2] **Yucheng Tang**[3] **Yuankai Huo**[1,*]

[1]Vanderbilt University
[2]Vanderbilt University Medical Center
[3]NVIDIA

https://github.com/Chongyu1117/TACT-KV

Abstract

Vision-language models (VLMs) applied to 3D medical scans face a structural memory wall. A single CT or MRI volume produces tens of thousands of vision tokens, and the resulting key-value (KV) cache scales linearly with both batch and context. Existing KV compression methods operate on the cache's feature or token statistics, or apply 1D and 2D transforms, leaving the three-axis spatial structure in the cache unexploited. We propose **TACT-KV** (Tri-Axis Cosine Transform), a calibration-free KV compression method that exploits strong spatial correlations in visual KV states along all three anatomical axes of 3D scans. This correlation concentrates the KV cache energy under a frequency-domain transform. TACT-KV applies a separable 3D discrete cosine transform to the KV cache, allocates a per-layer bit budget across frequency bins via reverse water-filling (more bits to high-energy components), and quantizes each bin with a fixed scalar codebook. We derive a closed-form relation: the distortion reduction from adaptive bit allocation is dominated by the spectral concentration of the transformed cache. Across six medical VLMs and two 3D CT benchmarks, TACT-KV achieves the lowest 1-bit prediction-flip rate on eleven of twelve model and dataset pairs in our main comparison. The same advantage reproduces on four general-purpose VLMs evaluated on video tasks, indicating that the benefit of exploiting three-axis structure generalizes beyond medical imaging. At equal high-bandwidth memory (HBM), TACT-KV scales the decoding batch size on a single GPU by up to $8\times$.

1 Introduction

Vision-language models (VLMs) are increasingly used for 3D medical image analysis [31], with applications spanning computed tomography (CT) report generation, magnetic resonance imaging (MRI) anomaly detection, and multimodal diagnostic assistance [33, 20, 27]. Unlike 2D natural images, a 3D medical scan contains hundreds of slices that capture fine-grained anatomical detail such as small lesions, subtle nodules, and vascular structures driving clinical diagnosis [16, 5, 2]. The scan then produces tens of thousands of vision tokens after the vision encoder, and the resulting key-value (KV) cache scales linearly with both batch size and sequence length. Clinical sites typically operate on single-card or small-cluster GPUs [25], where this cache quickly dominates GPU memory and becomes the bottleneck during inference. Compressing the visual KV cache without sacrificing diagnostic fidelity is therefore a gating problem for clinical deployment.

*Corresponding author: Yuankai Huo (yuankai.huo@vanderbilt.edu).

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

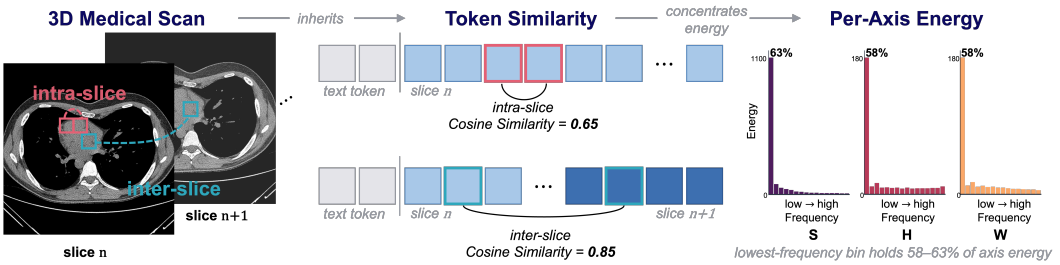

Figure 1: **Motivation.** Tokens at the same in-plane spatial position across adjacent slices in the visual KV cache exhibit high pairwise similarity, and within each slice neighboring tokens are also strongly correlated. The cache energy then concentrates in a small set of low-frequency components along every axis, a three-axis structure that previous work leaves unexploited.

KV cache compression directly addresses this bottleneck, and existing methods split into two categories by their target. (1) LLM-targeted methods, including KIVI [29], KVQuant [19], and TurboQuant [44], quantize the cache along its 1D token or channel axis. (2) VLM-targeted methods, including AKVQ-VL [35] and VL-Cache [38], exploit attention-derived saliency to allocate precision or evict cache entries. A separate line of work instead transforms the cache into the frequency domain before compression. FreqKV [21] applies a 1D discrete cosine transform (DCT) along the token sequence, FlashCache [43] extracts a low-frequency baseline and handles outliers via IDCT and uses the reconstruction residual to identify outlier KV pairs for retention, KVTC [34] performs feature-dimension PCA with adaptive bit allocation, and Fourier-VLM [41] applies a 2D DCT to vision encoder features before they enter the cache. None of these methods exploit the spatial continuity that 3D scans carry along all three axes.

A closer analog is video KV cache compression, which exploits redundancy between adjacent 2D frames [32, 40, 36]. Medical VLMs commonly treat a 3D scan as a sequence of 2D slices [33, 20], but the slice axis differs from the temporal axis. The temporal axis is motion-induced, with attention concentrated on a few salient frames, while the slice axis is anatomically continuous, with diagnostic signal spread across many slices rather than localized in a sparse subset. Directly applying video-domain compression to medical scans is therefore questionable.

3D medical scans carry a distinctive structural property. Nearby slices are highly similar, just like neighboring pixels within a single image. We observe that this structure also appears in the visual KV cache of the language model decoder: *tokens at the same spatial location across adjacent slices have high similarity (Fig. 1), and the cache shows strong autocorrelation along all three spatial axes.* This spatial correlation concentrates the cache energy in a small set of low-frequency components, which a frequency-domain transform can compress.

Motivated by this observation, we propose **TACT-KV** (Tri-Axis Cosine Transform for KV Cache). TACT-KV reshapes the visual KV cache into a 3D tensor with dimensions (slice, height, width), then applies a separable 3D DCT across these three spatial axes. After transforming the cache into the frequency domain, TACT-KV allocates bits to different frequency components using a reverse water-filling strategy and compresses them with standard scalar quantization. The method requires no calibration data and introduces no additional learnable parameters.

We further prove that the gain of water-filling over uniform bit allocation grows as the cache energy concentrates in fewer frequency components, and verify this empirically. We evaluate TACT-KV on six medical VLMs using two 3D CT benchmarks, RSNA-ICH [11] and CT-RATE [17]. Even at 1-bit quantization, TACT-KV performs close to the uncompressed cache, while existing LLM and VLM compression methods often fail. We also test TACT-KV on Video-MME [13, 20] using four general-purpose VLMs and observe similar improvements, suggesting that the method generalizes beyond medical imaging. At the same GPU memory budget, TACT-KV can increase the maximum decoding batch size by up to 8×. Our contributions are:

- We propose TACT-KV, to our knowledge the first method to treat the visual KV cache as a volumetric tensor and exploit the three-axis spatial correlation from 3D scans.

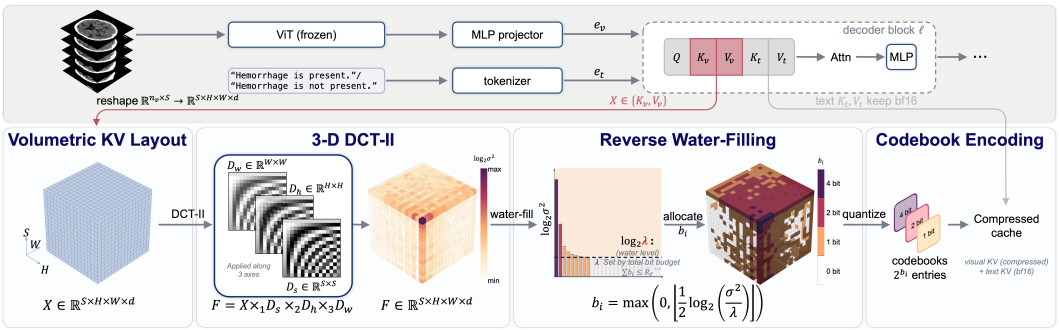

Figure 2: **TACT-KV pipeline.** A medical VLM encodes a volumetric CT scan and a clinical prompt into visual and text embeddings $\mathbf{e}_v, \mathbf{e}_t$. Inside each decoder block, visual keys/values K_v, V_v enter the four-stage pipeline below, while text K_t, V_t bypass at bf16. **Volumetric KV Layout:** the visual cache is reshaped into a (S, H, W) grid inherited from the vision encoder. **3-D DCT-II:** a separable transform along the three axes concentrates energy at the DC corner (magma colormap, $\log_2 \sigma^2$). **Reverse Water-Filling:** a per-layer bit budget is allocated above water level λ, with bins below dropped (bar color = bit width 4/2/1/0). **Codebook Encoding:** surviving bins are gain-shape encoded against the corresponding Lloyd–Max codebook (1/2/4-bit tiles shown). All spectral statistics are estimated on the fly, requiring no calibration data.

- We prove a closed-form expression for the distortion gain of water-filling over uniform bit allocation in terms of the spectral concentration of the cache, which we verify empirically.
- We evaluate TACT-KV across six medical VLMs (4B-32B), two 3D CT benchmarks, and four general-purpose VLMs on Video-MME. TACT-KV achieves the lowest 1-bit flip rate on eleven of twelve medical VLM and dataset pairs in our main comparison, while delivering up to $8\times$ larger decoding batch at equal HBM.

2 Method

We introduce TACT-KV, a calibration-free method for compressing the visual portion of the KV cache held by a decoder-only vision–language model during generation. For a backbone with L transformer layers and n_h attention heads per layer, a single sequence at layer ℓ produces a key tensor $K^{(\ell)}$ and a value tensor $V^{(\ell)}$ with per-head feature dimension d. In a vision–language model, tokens originate from two sources, a vision encoder and a text tokenizer. When the input is a volumetric medical scan or a video, visual tokens dominate the count and inherit a known three-dimensional grid $S \times H \times W$ from the vision encoder, where S is the slice or temporal axis token count and H, W are the spatial pooling dimensions. We exploit this structural prior by allocating the per-layer bit budget for the visual KV cache in a frequency-domain representation of this grid, retaining the information required by downstream tasks at low bit precision. The pipeline introduces no learned parameters and runs once per layer per sequence in a single forward pass. Figure 2 illustrates the four stages.

2.1 Volumetric KV Layout

Let $K_v \in \mathbb{R}^{B \times n_h \times n_v \times d}$ denote the visual portion of the layer-ℓ key cache and $K_t \in \mathbb{R}^{B \times n_h \times n_t \times d}$ denote the text portion, with B the batch dimension, $n_v = S \cdot H \cdot W$ the visual token count, and n_t the text token count. The value tensors V_v, V_t are defined analogously. The grid parameters (S, H, W) are emitted by the vision encoder for each input and are read directly by our pipeline. We do not estimate any per-sample structural quantity. We reshape the visual key tensor into the volumetric grid:

$$X = \text{reshape}(K_v) \in \mathbb{R}^{B \times n_h \times S \times H \times W \times d}, \tag{1}$$

and reshape the visual value tensor identically. The remaining pipeline (3-D DCT, water-filling allocation, codebook encoding) runs independently on K_v and V_v with separate per-bin variance estimates and bit budgets. The batch and head dimensions (B, n_h) act as outer indices for the remainder of the pipeline and are omitted from subsequent equations. The text portion K_t, V_t is held at bf16 (bfloat16) [22] throughout the pipeline.

2.2 3-D DCT-II

The discrete cosine transform of type II is an orthonormal energy-compacting transform [1]. For a real sequence of length M, its $M \times M$ matrix D has entries indexed by output frequency k and input sample m,

$$D[k, m] = \alpha_k \cos\Big[\frac{\pi}{M}\big(m + \tfrac{1}{2}\big)\, k\Big], \qquad 0 \le k, m \le M - 1, \tag{2}$$

with normalization constants $\alpha_0 = \sqrt{1/M}$ and $\alpha_{k\ge 1} = \sqrt{2/M}$ chosen so that $DD^\top = I$. For a first-order Markov source, the DCT-II asymptotically approaches the Karhunen–Loève transform, and variants of it underpin the block transform in JPEG and successor image and video codecs [39].

We extend the DCT-II to the reshaped tensor $X \in \mathbb{R}^{S\times H\times W\times d}$. Let $D_s \in \mathbb{R}^{S\times S}$, $D_h \in \mathbb{R}^{H\times H}$, and $D_w \in \mathbb{R}^{W\times W}$ be the DCT-II matrices on the corresponding axis lengths, each constructed by (2). The frequency-domain tensor $F \in \mathbb{R}^{S\times H\times W\times d}$ is defined by the mode product [23]:

$$F = X \times_1 D_s \times_2 D_h \times_3 D_w, \tag{3}$$

where $\times_n$ denotes matrix multiplication along the n-th tensor axis. The separable three-axis structure reduces the per-layer cost from $\mathcal{O}(n_v^2 \cdot d)$ for a non-separable transform to $\mathcal{O}\big((S + H + W) \cdot n_v \cdot d\big)$, and the parameter count from $(SHW)^2$ to $S^2 + H^2 + W^2$. The feature axis d is left untransformed and is handled by the scalar quantizer (SQ) in Section 2.4. Because D_s, D_h, D_w are each orthonormal, (3) is a lossless invertible change of variable that preserves the ℓ_2 norm. Compression is introduced in the next subsection by allocating the per-layer bit budget across frequency bins.

2.3 Reverse Water-Filling Bit Allocation

Let the per-bin variance of the frequency tensor in (3), averaged over the batch, head, and feature dimensions, be

$$\sigma_i^2 = \mathbb{E}_{B,\, n_h,\, d}\big[F_i^2\big], \qquad i = 1, \ldots, N, \quad N = SHW. \tag{4}$$

Given a per-layer bit budget R_ℓ, our goal is to allocate non-negative integer bits b_i to each frequency bin so as to minimize the total reconstruction MSE. Under the high-rate scalar-quantization distortion model for a Gaussian source [15], the per-bin distortion is $\mathcal{D}_i(b_i) = c \cdot \sigma_i^2 \cdot 2^{-2b_i}$ with c a constant determined by the quantizer geometry [7], and the problem reduces to

$$\min_{b_i \in \mathbb{Z}_{\ge 0}} \sum_i \sigma_i^2 \cdot 2^{-2b_i}, \quad \text{s.t.} \quad \sum_i b_i \le R_\ell. \tag{5}$$

A Lagrangian relaxation with a non-negativity constraint yields the reverse water-filling solution

$$b_i = \max\big(0,\ \big\lfloor \tfrac{1}{2} \log_2(\sigma_i^2/\lambda) \big\rfloor\big), \tag{6}$$

where the water level λ is determined by a bisection search on the budget equation. The $\max(0, \cdot)$ in (6) enforces non-negativity, the floor produces integer bits, and a greedy ± 1 correction enforces $\sum_i b_i = R_\ell$ exactly. In our implementation b_i is capped at 8.

The following theorem expresses the distortion improvement of reverse water-filling over uniform allocation as a closed-form function of the spectral shape. Let $\mathcal{D}_{\text{uniform}}$ and $\mathcal{D}_{\text{water-fill}}$ denote the total distortion summed over all bins under the two allocations.

Theorem 1. *Let $\{\sigma_i^2\}_{i=1}^N$ be the per-bin variances in* (4)*, and define the spectral flatness*

$$\gamma = \frac{\big(\prod_i \sigma_i^2\big)^{1/N}}{(1/N)\, \sum_i \sigma_i^2} \in (0, 1]. \tag{7}$$

Under matched bit budget R_ℓ and the high-rate Gaussian distortion model,

$$\mathcal{D}_{\text{uniform}} \,/\, \mathcal{D}_{\text{water-fill}} = \frac{(1/N) \sum_i \sigma_i^2}{\big(\prod_i \sigma_i^2\big)^{1/N}} = \frac{1}{\gamma}. \tag{8}$$

Sketch. Substituting $b_i = R_\ell/N$ into the distortion model gives $\mathcal{D}_{\text{uniform}} = c \cdot 2^{-2R_\ell/N} \cdot \sum_i \sigma_i^2$. Substituting the continuous solution of (6) into the per-bin distortion gives $\mathcal{D}_i = c\lambda$, independent of i, so $\mathcal{D}_{\text{water-fill}} = cN\lambda$. The budget equation $\sum_i b_i = R_\ell$ then yields $\lambda = \big(\prod_i \sigma_i^2\big)^{1/N} \cdot 2^{-2R_\ell/N}$. Dividing the two expressions gives (8). The full proof appears in Appendix A. □

Theorem 1 expresses $1/\gamma$ as the closed-form distortion improvement of reverse water-filling over uniform allocation.[2] As $\gamma \to 1$, the spectrum is flat and water-filling reduces to uniform allocation. As $\gamma \to 0$, energy concentrates in a small subset of bins and the closed-form $1/\gamma$ grows without bound. In practice the integer water-filling in (6) drops the inactive bins and the realized gain is finite. Figure 3 compares this closed-form prediction to the empirical noise-variance ratio on RSNA-ICH. Three of six cells fall within the 30% closure band, one cell sits at 40% (directional agreement), and two cells (Hulu-Med 4B and 7B) lie above the identity line in the low-rate regime where the high-rate Gaussian distortion model underlying Theorem 1 relaxes. When noise or motion weakens spatial correlation and flattens the cache spectrum, γ approaches one and the predicted allocation gain decreases.

2.4 Codebook Encoding

The integer bit allocation b_i from (6) routes each frequency bin to a scalar quantizer of matching width. We use a gain-shape scalar quantizer as the backend. For each bin with $b_i \geq 1$, we extract the ℓ_2 norm g, normalize the residual onto the unit sphere, apply a fixed orthogonal rotation shared across all bins [6], and look up each rotated coordinate against a precomputed Lloyd–Max boundary table [30] at width b_i. The total per-layer cache occupancy is $\sum_{i:\, b_i \geq 1}(b_i \cdot d + 32)$ bits plus a small index overhead, where the fp32 scalar g contributes the constant 32 for each non-empty bin. We compare gain-shape SQ against symmetric uniform and NormalFloat backends in Section 3. Decoding inverts the steps above. We read the indices, dequantize, undo the rotation, multiply by g, and apply the transpose mode product to recover $\hat{X} = \hat{F} \times_1 D_s^\top \times_2 D_h^\top \times_3 D_w^\top$. Both the inverse DCT and the codebook dequantization are deterministic and remain transparent to the attention implementation. A fused implementation of the inverse path is described in Appendix H.

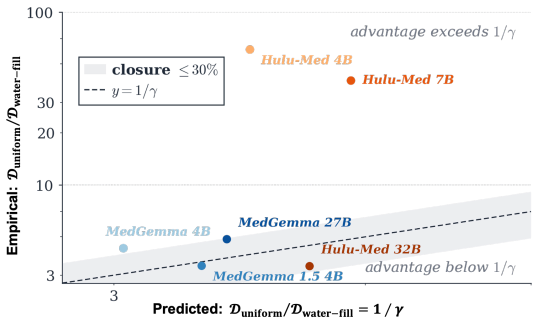

Figure 3: **Theorem 1 closure on RSNA-ICH.** Empirical $\mathcal{D}_{\text{uniform}}/\mathcal{D}_{\text{water-fill}}$ versus closed-form $1/\gamma$ at 1-bit average rate ($N = 3000$ studies, six VLMs). Shaded band marks $\leq 30\%$ deviation from $y = 1/\gamma$.

3 Experiments and Results

3.1 Experimental Setup

Datasets and VLMs. We evaluate on three benchmarks across ten VLMs in two domains. For **medical CT**, we use RSNA-ICH [11] on a stratified N=3,000 subset of the 2019 training split under the standard six-label hemorrhage protocol, and CT-RATE [17] on the full validation split (N=3,039 volumes) under the official 18-abnormality multi-label protocol. We evaluate six medical VLMs on these benchmarks: MedGemma 1.5/4B/27B [33] and Hulu-Med 4B/7B/32B [20]. For **cross-domain transfer**, we use Video-MME [13] in the no-subtitle setting on all 2,700 four-way multiple-choice questions (MCQs), with four general-purpose VLMs: Qwen2.5-VL-7B [4], Qwen3-VL-8B [3], LLaVA-OneVision-7B [26], and VideoLLaMA3-7B [45]. The lineup spans three VL variants, three vision encoders, and four LLM versions (Qwen2, Qwen2.5, Qwen3, Gemma-3).

Implementation. All experiments run on NVIDIA RTX PRO 6000 (96 GiB) GPUs. For the medical benchmarks we use rank classification on assertion-style completions [14]; for Video-MME we use the four-letter MCQ prompt with greedy decoding. All methods compress only the visual portion of the cache and use the same inputs and prompts, with non-visual tokens (chat template, boundary markers) held at bf16. This isolates the contribution of our method and is consistent with prior

[2] The result applies to the continuous relaxation that drops the floor in (6), assumes the high-rate Gaussian distortion model, and implicitly assumes that every bin receives positive rate so that $\mathcal{D}_{\text{water-fill}} = cN\lambda$. The integer floor and the ± 1 correction contribute a finite-rate penalty to the water-filling distortion, and both the high-rate model and the all-active assumption lose validity as the average bit rate per coordinate approaches zero.

Table 1: **Macro prediction-flip rate (%) relative to bf16 on RSNA-ICH and CT-RATE.** Entries include Wald 95% CI half-widths. **Bold** and underline mark the lowest and second-lowest measured values in each column. RSNA-ICH uses a stratified subset of N=3,000 studies under the six-label challenge protocol [11]. Each CT-RATE entry aggregates 54,702 predictions from N=3,039 official validation volumes under the 18-abnormality protocol [17]. Bit budgets denote nominal quantizer rates and exclude metadata. For KVTC, the budget is the average number of bits per PCA coefficient before entropy coding.

	RSNA-ICH (N=3,000)				**CT-RATE** (N=3,039)			
Method	1-bit	2-bit	3-bit	4-bit	1-bit	2-bit	3-bit	4-bit
MedGemma 1.5 4B								
TACT-KV (ours)	**3.78**$_{\pm 0.28}$	**1.91**$_{\pm 0.20}$	1.29$_{\pm 0.17}$	1.09$_{\pm 0.15}$	**4.69**$_{\pm 0.18}$	**2.33**$_{\pm 0.13}$	1.89$_{\pm 0.11}$	1.54$_{\pm 0.10}$
TurboQuant	6.27$_{\pm 0.35}$	4.47$_{\pm 0.30}$	2.16$_{\pm 0.21}$	1.41$_{\pm 0.17}$	9.14$_{\pm 0.24}$	4.61$_{\pm 0.18}$	2.46$_{\pm 0.13}$	1.87$_{\pm 0.11}$
KIVI	10.74$_{\pm 0.45}$	1.96$_{\pm 0.20}$	**1.12**$_{\pm 0.15}$	1.06$_{\pm 0.15}$	23.22$_{\pm 0.35}$	4.66$_{\pm 0.18}$	**1.59**$_{\pm 0.10}$	**1.51**$_{\pm 0.10}$
KVQuant	9.19$_{\pm 0.42}$	2.31$_{\pm 0.22}$	1.26$_{\pm 0.16}$	**1.02**$_{\pm 0.15}$	9.46$_{\pm 0.25}$	3.49$_{\pm 0.15}$	1.66$_{\pm 0.11}$	1.51$_{\pm 0.10}$
KVTC	16.55$_{\pm 0.54}$	6.93$_{\pm 0.37}$	1.53$_{\pm 0.18}$	2.26$_{\pm 0.22}$	37.75$_{\pm 0.41}$	17.91$_{\pm 0.32}$	7.04$_{\pm 0.21}$	2.86$_{\pm 0.14}$
VidKV	7.54$_{\pm 0.39}$	2.92$_{\pm 0.25}$	–	–	9.13$_{\pm 0.24}$	6.01$_{\pm 0.20}$	–	–
AKVQ-VL	28.54$_{\pm 0.66}$	2.10$_{\pm 0.21}$	–	–	20.72$_{\pm 0.34}$	4.10$_{\pm 0.17}$	–	–
MedGemma 4B								
TACT-KV (ours)	2.37$_{\pm 0.22}$	0.91$_{\pm 0.14}$	**0.64**$_{\pm 0.12}$	0.67$_{\pm 0.12}$	**2.95**$_{\pm 0.14}$	**1.98**$_{\pm 0.12}$	**1.69**$_{\pm 0.11}$	1.57$_{\pm 0.10}$
TurboQuant	4.41$_{\pm 0.30}$	1.21$_{\pm 0.16}$	0.81$_{\pm 0.13}$	0.64$_{\pm 0.12}$	7.33$_{\pm 0.22}$	3.97$_{\pm 0.16}$	2.22$_{\pm 0.12}$	2.20$_{\pm 0.12}$
KIVI	2.43$_{\pm 0.23}$	0.84$_{\pm 0.13}$	0.67$_{\pm 0.12}$	0.63$_{\pm 0.12}$	11.58$_{\pm 0.27}$	3.80$_{\pm 0.16}$	1.73$_{\pm 0.11}$	**1.56**$_{\pm 0.10}$
KVQuant	10.58$_{\pm 0.45}$	1.22$_{\pm 0.16}$	0.73$_{\pm 0.12}$	**0.61**$_{\pm 0.11}$	7.92$_{\pm 0.23}$	2.84$_{\pm 0.14}$	2.09$_{\pm 0.12}$	1.65$_{\pm 0.11}$
KVTC	7.12$_{\pm 0.38}$	24.77$_{\pm 0.63}$	3.45$_{\pm 0.27}$	1.38$_{\pm 0.17}$	16.86$_{\pm 0.31}$	13.67$_{\pm 0.29}$	5.57$_{\pm 0.19}$	2.17$_{\pm 0.12}$
VidKV	14.13$_{\pm 0.51}$	0.90$_{\pm 0.14}$	–	–	15.09$_{\pm 0.30}$	2.99$_{\pm 0.14}$	–	–
AKVQ-VL	**2.20**$_{\pm 0.21}$	**0.76**$_{\pm 0.13}$	–	–	23.28$_{\pm 0.35}$	3.58$_{\pm 0.16}$	–	–
MedGemma 27B								
TACT-KV (ours)	**0.37**$_{\pm 0.09}$	**0.26**$_{\pm 0.08}$	0.22$_{\pm 0.07}$	0.28$_{\pm 0.08}$	**2.63**$_{\pm 0.13}$	**1.33**$_{\pm 0.10}$	**1.44**$_{\pm 0.10}$	1.38$_{\pm 0.10}$
TurboQuant	4.31$_{\pm 0.30}$	0.47$_{\pm 0.10}$	0.31$_{\pm 0.08}$	0.24$_{\pm 0.07}$	9.14$_{\pm 0.24}$	3.76$_{\pm 0.16}$	1.63$_{\pm 0.11}$	1.64$_{\pm 0.11}$
KIVI	16.37$_{\pm 0.54}$	0.42$_{\pm 0.09}$	**0.21**$_{\pm 0.07}$	0.21$_{\pm 0.07}$	19.83$_{\pm 0.33}$	2.26$_{\pm 0.12}$	1.61$_{\pm 0.11}$	1.38$_{\pm 0.10}$
KVQuant	20.75$_{\pm 0.59}$	0.38$_{\pm 0.09}$	0.32$_{\pm 0.08}$	0.25$_{\pm 0.07}$	14.13$_{\pm 0.29}$	2.73$_{\pm 0.14}$	1.72$_{\pm 0.11}$	1.49$_{\pm 0.10}$
KVTC	50.57$_{\pm 0.73}$	10.56$_{\pm 0.45}$	0.27$_{\pm 0.08}$	**0.18**$_{\pm 0.06}$	28.43$_{\pm 0.38}$	6.88$_{\pm 0.21}$	2.19$_{\pm 0.12}$	**0.96**$_{\pm 0.08}$
VidKV	2.79$_{\pm 0.24}$	0.38$_{\pm 0.09}$	–	–	14.49$_{\pm 0.29}$	2.10$_{\pm 0.12}$	–	–
AKVQ-VL	15.48$_{\pm 0.53}$	0.43$_{\pm 0.10}$	–	–	12.48$_{\pm 0.28}$	1.72$_{\pm 0.11}$	–	–
Hulu-Med 7B								
TACT-KV (ours)	**6.08**$_{\pm 0.35}$	3.51$_{\pm 0.27}$	2.33$_{\pm 0.22}$	1.86$_{\pm 0.20}$	**10.44**$_{\pm 0.26}$	**4.71**$_{\pm 0.18}$	**3.43**$_{\pm 0.15}$	**2.78**$_{\pm 0.14}$
TurboQuant	36.25$_{\pm 0.70}$	28.07$_{\pm 0.66}$	40.44$_{\pm 0.72}$	52.33$_{\pm 0.73}$	52.95$_{\pm 0.42}$	35.24$_{\pm 0.40}$	43.01$_{\pm 0.41}$	36.42$_{\pm 0.40}$
KIVI	25.38$_{\pm 0.64}$	4.37$_{\pm 0.30}$	**2.25**$_{\pm 0.22}$	**1.67**$_{\pm 0.19}$	56.62$_{\pm 0.42}$	8.53$_{\pm 0.23}$	4.77$_{\pm 0.18}$	3.48$_{\pm 0.15}$
KVQuant	8.29$_{\pm 0.40}$	4.10$_{\pm 0.29}$	2.32$_{\pm 0.22}$	1.90$_{\pm 0.20}$	16.82$_{\pm 0.31}$	7.38$_{\pm 0.22}$	5.06$_{\pm 0.18}$	3.28$_{\pm 0.15}$
KVTC	60.38$_{\pm 0.71}$	13.57$_{\pm 0.50}$	3.12$_{\pm 0.25}$	1.75$_{\pm 0.19}$	48.76$_{\pm 0.42}$	43.15$_{\pm 0.42}$	6.38$_{\pm 0.20}$	3.31$_{\pm 0.15}$
VidKV	22.18$_{\pm 0.61}$	4.02$_{\pm 0.29}$	–	–	45.51$_{\pm 0.42}$	8.19$_{\pm 0.23}$	–	–
AKVQ-VL	32.76$_{\pm 0.69}$	**3.26**$_{\pm 0.26}$	–	–	56.24$_{\pm 0.42}$	6.33$_{\pm 0.20}$	–	–
Hulu-Med 4B								
TACT-KV (ours)	**8.53**$_{\pm 0.41}$	3.91$_{\pm 0.28}$	2.43$_{\pm 0.23}$	2.05$_{\pm 0.21}$	**7.22**$_{\pm 0.22}$	**3.70**$_{\pm 0.16}$	2.78$_{\pm 0.14}$	**2.33**$_{\pm 0.13}$
TurboQuant	76.86$_{\pm 0.62}$	78.61$_{\pm 0.60}$	19.47$_{\pm 0.58}$	4.30$_{\pm 0.30}$	39.06$_{\pm 0.41}$	37.06$_{\pm 0.40}$	27.99$_{\pm 0.38}$	5.69$_{\pm 0.19}$
KIVI	77.01$_{\pm 0.61}$	4.65$_{\pm 0.31}$	**2.37**$_{\pm 0.22}$	**2.00**$_{\pm 0.20}$	37.03$_{\pm 0.40}$	4.77$_{\pm 0.18}$	**2.66**$_{\pm 0.13}$	2.37$_{\pm 0.13}$
KVQuant	28.85$_{\pm 0.66}$	6.51$_{\pm 0.36}$	2.80$_{\pm 0.24}$	2.16$_{\pm 0.21}$	17.17$_{\pm 0.32}$	5.70$_{\pm 0.19}$	3.02$_{\pm 0.14}$	2.41$_{\pm 0.13}$
KVTC	40.44$_{\pm 0.72}$	75.53$_{\pm 0.63}$	5.82$_{\pm 0.34}$	3.28$_{\pm 0.26}$	56.24$_{\pm 0.42}$	27.44$_{\pm 0.37}$	6.70$_{\pm 0.21}$	2.94$_{\pm 0.14}$
VidKV	30.52$_{\pm 0.67}$	4.34$_{\pm 0.30}$	–	–	24.68$_{\pm 0.36}$	4.77$_{\pm 0.18}$	–	–
AKVQ-VL	75.68$_{\pm 0.63}$	**3.70**$_{\pm 0.28}$	–	–	36.13$_{\pm 0.40}$	3.71$_{\pm 0.16}$	–	–
Hulu-Med 32B								
TACT-KV (ours)	**9.13**$_{\pm 0.42}$	4.40$_{\pm 0.30}$	3.05$_{\pm 0.25}$	2.71$_{\pm 0.24}$	**10.04**$_{\pm 0.25}$	**4.33**$_{\pm 0.17}$	2.75$_{\pm 0.14}$	2.21$_{\pm 0.12}$
TurboQuant	24.91$_{\pm 0.63}$	6.66$_{\pm 0.36}$	4.02$_{\pm 0.29}$	3.25$_{\pm 0.26}$	22.24$_{\pm 0.35}$	6.75$_{\pm 0.21}$	3.64$_{\pm 0.16}$	2.84$_{\pm 0.14}$
KIVI	43.32$_{\pm 0.72}$	5.89$_{\pm 0.34}$	**2.95**$_{\pm 0.25}$	2.51$_{\pm 0.23}$	65.09$_{\pm 0.40}$	5.14$_{\pm 0.18}$	**2.71**$_{\pm 0.14}$	**2.06**$_{\pm 0.12}$
KVQuant	11.30$_{\pm 0.46}$	5.32$_{\pm 0.33}$	3.22$_{\pm 0.26}$	2.57$_{\pm 0.23}$	14.39$_{\pm 0.29}$	6.05$_{\pm 0.20}$	2.98$_{\pm 0.14}$	2.30$_{\pm 0.13}$
KVTC	48.36$_{\pm 0.73}$	23.02$_{\pm 0.62}$	3.31$_{\pm 0.26}$	**2.46**$_{\pm 0.23}$	52.08$_{\pm 0.42}$	28.90$_{\pm 0.38}$	3.51$_{\pm 0.15}$	2.30$_{\pm 0.13}$
VidKV	34.67$_{\pm 0.70}$	5.56$_{\pm 0.33}$	–	–	30.80$_{\pm 0.39}$	4.79$_{\pm 0.18}$	–	–
AKVQ-VL	35.31$_{\pm 0.70}$	**4.20**$_{\pm 0.29}$	–	–	50.47$_{\pm 0.42}$	5.46$_{\pm 0.19}$	–	–

VLM-targeted compression that treats visual and text tokens separately [35, 36, 38]. We report bit budgets $\{1, 2, 3, 4\}$, with primary focus on the 1-bit setting where compression methods diverge most sharply. The 8-bit setting is approximately lossless across all backends [29, 19] and is omitted. Full per-VLM configurations and protocols are in Appendix C.2.

Metrics. The primary metric is the **macro prediction-flip rate** relative to the bf16 reference, the fraction of per-study per-label binary predictions whose label changes under quantization, averaged over labels. Flip rate measures consistency with the bf16 decision boundary, a desirable property when a deployed system swaps in a compressed cache. We complement flip rate with macro AUROC and 95% bootstrap CIs. For Video-MME we report per-question accuracy on the four-way MCQ.

3.2 Main Results: Medical CT

Table 1 reports macro prediction-flip rate on RSNA-ICH and CT-RATE for all six medical VLMs against six representative baselines.

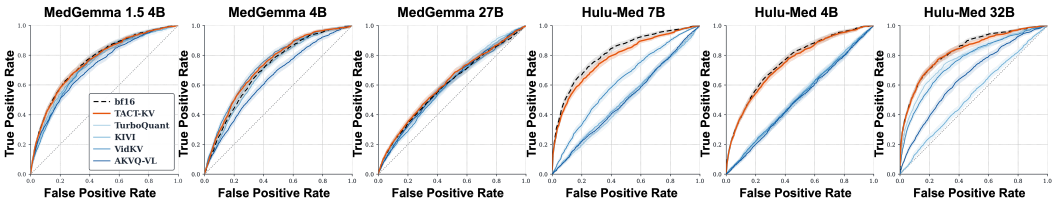

Figure 4: **Macro-averaged ROC curves at 1-bit on RSNA-ICH** (N=3,000)**.** One panel per medical VLM. The black dashed line marks the bf16 reference; translucent bands are 95% CIs from 1,000 bootstrap resamples.

TACT-KV achieves the lowest 1-bit flip rate on eleven of twelve VLM–dataset pairs. Across the eleven leading pairs, the next-best method has an average flip rate 2.60× that of TACT-KV. Several baselines incur large prediction changes at 1-bit. KIVI's per-channel grouped-uniform primitive collapses on Hulu-Med 4B RSNA-ICH (77% flip) and Hulu-Med 32B CT-RATE (65%). TurboQuant exceeds 35% flip on two of six VLMs. VidKV and AKVQ-VL reach worst-case 1-bit flip rates of 45.5% and 75.7% respectively. We attribute these failures to fixed quantization grids that, at 1-bit, lack the dynamic range to absorb the heavy-tailed visual KV distribution [9, 42, 28, 12]. TACT-KV avoids this regime by allocating bits adaptively in the spectral domain, where the cache energy is concentrated.

TACT-KV stays within the bf16 confidence band on every VLM at 1-bit. Figure 4 shows macro-averaged ROC curves at 1-bit on RSNA-ICH. TACT-KV tracks the bf16 reference within its 95% bootstrap CI on all six VLMs, while baseline curves drift outside the band on the Hulu-Med variants. The contrast is sharper on Hulu-Med than MedGemma because the Hulu-Med backbone uses full attention at every layer [20], exposing all layers to the 1-bit codebook, whereas the Gemma-3 backbone interleaves five sliding-window layers per global layer [37] and we leave the sliding-window layers at bf16 following the production convention of vLLM [24].

3.3 Theorem 1 Closure

We extend the closure check of Theorem 1 from the six RSNA-ICH points in Figure 3 to all twelve VLM–dataset pairs at 1-bit average rate. We isolate the allocation step by holding the 3D DCT and Gain-shape SQ (§2.4) fixed and swapping reverse water-filling for uniform allocation. The empirical ratio $\hat{r}$ is computed from the per-study per-label logit-margin variance under the two allocations and serves as a first-order proxy for the cache distortion ratio $\mathcal{D}_{\text{uniform}}/\mathcal{D}_{\text{water-fill}}$ predicted by Theorem 1. Table 2 shows that **eight of twelve pairs fall within** 50% **of the closed-form** $1/\gamma_K$. The remaining four are Hulu-Med 4B and 7B on both datasets, where the 1-bit setting falls outside the high-rate regime. Results at 2 and 4 bits are provided in Appendix F.

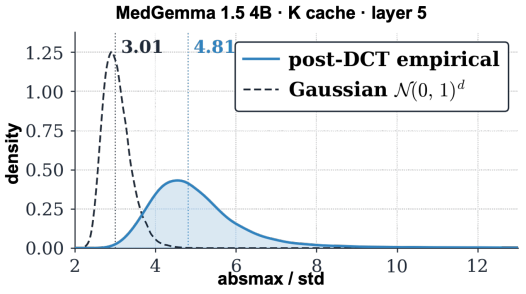

Figure 5: **Post-DCT K-cache coefficient distribution is heavy-tailed.**

3.4 Ablation Studies

Quantizer attribution. We substitute Gain-shape SQ in the 3-axis DCT + reverse water-filling pipeline with two alternatives, a symmetric uniform quantizer with per-vector absmax scaling and NormalFloat [10]. Table 3 shows that gain-shape SQ achieves the lowest flip rate across all settings, with the largest gap at 1-bit. We attribute this gap to the heavy-tailed post-DCT coefficient distribution (Figure 5, median absmax-to-std 4.81 versus the Gaussian reference 3.01 at d=256). Naive absmax scaling assigns codebook levels to the spectrum tail rather than the bulk, while gain-shape SQ's decomposition factors out per-vector magnitude before quantization. Per-quantizer reconstruction MSE is in Appendix G.

Table 2: **Theorem 1 closure on twelve VLM-dataset pairs at 1-bit average rate.** For each pair we report the K-cache spectral flatness γ_K, the closed-form prediction $1/\gamma_K$, the empirical ratio $\hat{r}$, and the closure $|\hat{r} - 1/\gamma_K|/(1/\gamma_K)$. **Bold** marks closure within 30% of $1/\gamma_K$, underline within 50%. The four unmarked entries (Hulu-Med 4B and 7B) lie outside the high-rate regime in which Theorem 1 holds.

	RSNA-ICH				CT-RATE			
VLM	γ_K	$1/\gamma_K$	$\hat{r}$	Closure	γ_K	$1/\gamma_K$	$\hat{r}$	Closure
MedGemma 1.5 4B	0.279	$3.58\times$	$3.41\times$	**4.9%**	0.171	$5.85\times$	$4.21\times$	**28.0%**
MedGemma 4B	0.327	$3.06\times$	$4.30\times$	40.6%	0.200	$5.00\times$	$7.49\times$	49.8%
MedGemma 27B	0.265	$3.77\times$	$4.85\times$	**28.6%**	0.180	$5.56\times$	$6.70\times$	**20.6%**
Hulu-Med 7B	0.206	$4.85\times$	$40.1\times$	727%	0.132	$7.58\times$	$38.0\times$	401%
Hulu-Med 4B	0.253	$3.95\times$	$60.9\times$	1440%	0.162	$6.17\times$	$100.3\times$	1524%
Hulu-Med 32B	0.224	$4.46\times$	$3.38\times$	**24.2%**	0.152	$6.58\times$	$3.72\times$	43.5%

Table 3: **Quantizer attribution: macro flip rate (%) on RSNA-ICH under three scalar quantizers inside the 3-axis DCT + reverse water-filling pipeline.** The DCT and water-filling stages are held fixed and only the per-bin scalar quantizer differs. Gain-shape SQ is the configuration used in TACT-KV. **Bold** marks the lowest entry in a column, underline marks the second-lowest. Each entry aggregates $18{,}000$ binary predictions.

	MedGemma 1.5 4B				Hulu-Med 7B			
Quantizer	1-bit	2-bit	3-bit	4-bit	1-bit	2-bit	3-bit	4-bit
Gain-shape SQ	**3.78**	**1.91**	**1.29**	**1.09**	**6.08**	**3.51**	**2.33**	**1.86**
Uniform sym	24.04	15.56	8.45	1.41	36.25	25.88	4.94	2.12
NormalFloat [10]	23.26	12.40	5.34	1.29	37.05	24.41	11.74	6.60

Transform and bit-allocation ablations. With gain-shape SQ fixed, Table 4 isolates the roles of the transform and bit allocation. The reduced-axis variants follow the axis choices of FreqKV [21] and Fourier-VLM [41]. Flip rate increases as fewer spatial axes are transformed. Removing the DCT is particularly costly on Hulu-Med 7B, where flip rate rises from 6.08% to 42.20%. Uniform allocation also degrades both models, showing that the transform and adaptive allocation each contribute to prediction fidelity.

Table 4: **Transform and allocation ablations.** RSNA-ICH flip rate (%) at 1 bit.

Configuration	MedGemma 1.5 4B	Hulu-Med 7B
3-axis DCT + water-filling (ours)	**3.78**	**6.08**
2-axis DCT + water-filling	5.00	6.21
1-axis DCT + water-filling	6.00	9.65
w/o DCT + water-filling	6.61	42.20
3-axis DCT + uniform allocation	9.97	25.76

3.5 Cross-Domain Generalization: Video-MME

We evaluate on Video-MME (no-subtitle) across four general-purpose VLMs. Table 5 reports per-question accuracy on the four-way MCQ with Wald 95% CI half-widths. TACT-KV is the highest-accuracy method on every VLM at 1-bit, and the worst-case drop from bf16 is 4.77 pp on Qwen3-VL-8B. Other baselines collapse at 1-bit (KIVI below 13% on every VLM). The advantage of three-axis structural compression therefore extends beyond medical imaging to video tasks.

3.6 Deployment Profiling

We profile the iso-HBM decoding behavior of TACT-KV on a single 96 GiB NVIDIA RTX PRO 6000 GPU at $S = 60$ axial slices. To avoid materializing the full reconstructed cache at decode, we implement a tile-fused inverse path that integrates index dequantization and the inverse three-axis DCT into the FlashAttention [8] kernel. The per-step intermediate footprint is then bounded by the FlashAttention block size and is independent of B (kernel details in Appendix H).

Cache savings support larger batches and higher throughput. We increase B until the timed decode loop runs out of memory and report the largest batch that fits for each method in Table 6. Across the three scales, the per-request KV cache size shrinks by 6.9–$7.6\times$, and the largest decoding batch a single GPU sustains grows by $4.00\times$, $6.25\times$, and $8.00\times$ respectively. The larger batches translate into 1.17–$2.02\times$ higher aggregate decoding throughput. The batch ratio rises monotonically

Table 5: **Video-MME no-subtitle accuracy (%) of the compressed cache across four general-purpose VLMs.** Each entry is per-question accuracy on 2,700 four-way MCQs followed by the Wald 95% CI half-width. **Bold** marks the highest accuracy in a column, underline marks the second-highest, ranked over methods with measured values.

Method	1-bit	2-bit	3-bit	4-bit	1-bit	2-bit	3-bit	4-bit
	Qwen2.5-VL-7B (bf16: $60.00_{\pm 1.85}$***)***				***Qwen3-VL-8B (bf16:*** $72.33_{\pm 1.69}$***)***			
TACT-KV (ours)	**57.15**$_{\pm 1.87}$	59.52$_{\pm 1.85}$	60.00$_{\pm 1.85}$	59.93$_{\pm 1.85}$	**67.56**$_{\pm 1.77}$	71.00$_{\pm 1.71}$	71.69$_{\pm 1.70}$	71.93$_{\pm 1.69}$
TurboQuant	23.04$_{\pm 1.59}$	25.04$_{\pm 1.63}$	21.33$_{\pm 1.55}$	44.85$_{\pm 1.88}$	14.33$_{\pm 1.32}$	54.26$_{\pm 1.88}$	70.93$_{\pm 1.71}$	71.67$_{\pm 1.70}$
KIVI	12.04$_{\pm 1.23}$	59.70$_{\pm 1.85}$	**60.37**$_{\pm 1.84}$	**59.96**$_{\pm 1.85}$	3.44$_{\pm 0.69}$	71.30$_{\pm 1.71}$	**71.89**$_{\pm 1.70}$	**72.22**$_{\pm 1.69}$
VidKV	19.70$_{\pm 1.50}$	59.33$_{\pm 1.85}$	–	–	26.63$_{\pm 1.67}$	**71.67**$_{\pm 1.70}$	–	–
AKVQ-VL	39.44$_{\pm 1.84}$	**59.89**$_{\pm 1.85}$	–	–	7.37$_{\pm 0.99}$	71.37$_{\pm 1.71}$	–	–
	LLaVA-OneVision-7B (bf16: $59.26_{\pm 1.85}$***)***				***VideoLLaMA3-7B (bf16:*** $65.19_{\pm 1.80}$***)***			
TACT-KV (ours)	**55.63**$_{\pm 1.87}$	58.70$_{\pm 1.86}$	**58.78**$_{\pm 1.86}$	58.93$_{\pm 1.86}$	**62.00**$_{\pm 1.83}$	64.89$_{\pm 1.80}$	**65.07**$_{\pm 1.80}$	**64.93**$_{\pm 1.80}$
TurboQuant	39.56$_{\pm 1.84}$	47.67$_{\pm 1.88}$	40.22$_{\pm 1.85}$	38.04$_{\pm 1.83}$	40.81$_{\pm 1.85}$	56.44$_{\pm 1.87}$	27.41$_{\pm 1.68}$	52.78$_{\pm 1.88}$
KIVI	7.63$_{\pm 1.00}$	58.30$_{\pm 1.86}$	58.59$_{\pm 1.86}$	**59.48**$_{\pm 1.85}$	7.44$_{\pm 0.99}$	64.67$_{\pm 1.80}$	65.04$_{\pm 1.80}$	64.70$_{\pm 1.80}$
VidKV	40.00$_{\pm 1.85}$	**58.85**$_{\pm 1.86}$	–	–	40.56$_{\pm 1.85}$	**65.04**$_{\pm 1.80}$	–	–
AKVQ-VL	18.96$_{\pm 1.48}$	58.56$_{\pm 1.86}$	–	–	12.44$_{\pm 1.25}$	64.19$_{\pm 1.81}$	–	–

Table 6: **Memory and wall-clock performance on one 96 GiB RTX PRO 6000 GPU at $S = 60$ slices.** Factors in parentheses give batch expansion, cache reduction, and throughput gain relative to bf16. Cache size is measured at $B = 1$ after prefill. Decode step time and throughput are averaged over five CT-RATE volumes at each method's maximum batch. Headroom is the device-reported capacity (102.6 GB) minus model weights.

Method	Max decoding B	Per-request cache (MB)	Decode step (ms)	Throughput (tok/s)
Hulu-Med 4B (Weights 9.0 GB, HBM headroom 93.6 GB)				
bf16	32	2333	165	194
TACT-KV	128$_{(4.00\times)}$	309$_{(7.55\times)}$	467	274$_{(1.42\times)}$
Hulu-Med 7B (Weights 16.1 GB, HBM headroom 86.5 GB)				
bf16	64	915	136	472
TACT-KV	400$_{(6.25\times)}$	132$_{(6.93\times)}$	723	553$_{(1.17\times)}$
Hulu-Med 32B (Weights 66.4 GB, HBM headroom 36.2 GB)				
bf16	8	4140	121	66
TACT-KV	64$_{(8.00\times)}$	543$_{(7.62\times)}$	481	133$_{(2.02\times)}$

with the weight-to-headroom fraction. When weights occupy 66.4 GB of 102.6 GB and the cache becomes the dominant non-weight term (Hulu-Med 32B), the cache savings translate most directly into a batch-capacity multiplier. At smaller scales, the larger non-cache headroom dilutes the gain.

TACT-KV also reduces latency at the same batch size. For each model, we use the largest batch supported by bf16 and retain the same five CT-RATE volumes. Table 7 shows that TACT-KV reduces step latency on all three models, with speedups of 1.10–1.45×. Both methods generate the same number of tokens per step at fixed B, so the latency reduction gives the same factor of improvement in aggregate throughput.

Table 7: **Decode latency at matched batch sizes.** Each model uses bf16's maximum B. Parentheses show speedup.

		Decode step (ms)	
Model	B	bf16	TACT-KV
Hulu-Med 4B	32	165	**114**$_{(1.45\times)}$
Hulu-Med 7B	64	136	**119**$_{(1.14\times)}$
Hulu-Med 32B	8	121	**110**$_{(1.10\times)}$

4 Discussion and Conclusion

We presented TACT-KV, a transform-coding pipeline for compressing the visual KV cache of vision–language models on volumetric token grids. TACT-KV reshapes the cache into a (S, H, W) tensor, applies a separable three-axis DCT-II for spatial decorrelation, allocates a per-layer bit budget by reverse water-filling, and quantizes each frequency bin with a scalar codebook. The pipeline is calibration-free and adds no learned parameters. The volumetric prior we exploit is absent from the text and natural-image settings that prior KV compressors target, and we prove that the distortion gain of reverse water-filling over uniform allocation has a closed form, $1/\gamma$, in the spectral flatness of the transformed cache. Across six medical VLMs on two 3D CT benchmarks, TACT-KV achieves the lowest 1-bit flip rate on eleven of twelve model and dataset pairs in our main comparison. The method also transfers to four general-purpose VLMs on Video-MME, and its gain in maximum decoding batch size at equal HBM increases with model scale, reaching 8× on Hulu-Med 32B. The margin

between TACT-KV and 1D quantizers on these benchmarks also measures how much volumetric structure scalar codebooks alone cannot recover, and that margin widens as the input becomes more spatially coherent. Integrating TACT-KV with attention-aware eviction on the text portion of the cache is a natural next step and orthogonal to the bit-allocation savings reported here.

Limitation. The closed-form prediction of Theorem 1 holds in the high-rate regime in which every frequency bin receives non-trivial bit allocation, and the lowest-rate operating points on the smaller Hulu-Med variants fall outside this regime (Section 3.3). Our pipeline also compresses the visual portion of the cache only, leaving text tokens at bf16, so the savings reported here are tied to inputs in which visual tokens dominate. The volumetric prior is validated on chest CT and head CT only, and whether it transfers to other 3D modalities such as MRI, PET, and 3D ultrasound is not measured. Although the same pipeline transfers to Video-MME without modification, the temporal axis differs in origin from the slice axis, and the gains we observe there are empirical rather than predicted by the volumetric assumption that motivates the design.

Broader Impact. Lower memory requirements can broaden access to volumetric medical VLMs at clinical sites with limited GPU capacity [25]. Prediction changes introduced by compression could lead to missed findings or unnecessary follow-up in clinical use. Evaluation on local data should therefore examine per-label prediction changes and diagnostic accuracy, especially for rare findings that aggregate metrics may obscure.

Acknowledgments

This research was supported by the National Institutes of Health (NIH) under grant R01DK135597, with $20,000 representing 66% of the total project costs. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This material is based upon work supported by the U.S. National Science Foundation under Award Number 2434229. A total of $5,000, representing 17% of the total project costs, was financed with this funding. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation. Non-governmental support was provided by the Vanderbilt Internal Grants. A total of $5,000, representing 17% of the total project costs, was financed with non-governmental funds.

The authors declare no additional competing interests.

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[2] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[5] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pages rs–3, 2024.

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

[7] Thomas M Cover, Joy A Thomas, et al. Entropy, relative entropy and mutual information. *Elements of information theory*, 2(1):12–13, 1991.

[8] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

[9] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.

[10] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[11] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020.

[12] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

[13] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.

[14] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Zenodo*, 2021.

[15] Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

[16] Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. *arXiv preprint arXiv:2510.20639*, 2025.

[17] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024.

[18] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19, 2026.

[19] Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun S Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303, 2024.

[20] Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

[21] Jushi Kai, Boyi Zeng, Yixuan Wang, Haoli Bai, Ziwei He, Bo Jiang, and Zhouhan Lin. Freqkv: Frequency domain key-value compression for efficient context window extension. *arXiv preprint arXiv:2505.00570*, 2025.

[22] Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019.

[23] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[24] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[25] Qizhen Lan, Aaron Choi, Jun Ma, Bo Wang, Zhaogming Zhao, Xiaoqian Jiang, and Yu-Chun Hsu. From performance to practice: Knowledge-distilled segmentator for on-premises clinical workflows. *arXiv preprint arXiv:2601.09191*, 2026.

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[27] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[28] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024.

[29] 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.

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

[31] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

[32] Vishnu Sai, Dheeraj Sai, Girish Varma, Priyesh Shukla, et al. Dual-signal adaptive kv-cache optimization for long-form video understanding in vision-language models. *arXiv preprint arXiv:2602.14236*, 2026.

[33] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[34] Konrad Staniszewski and Adrian Łańcucki. Kv cache transform coding for compact storage in llm inference. *arXiv preprint arXiv:2511.01815*, 2025.

[35] Zunhai Su, Wang Shen, Linge Li, Zhe Chen, Hanyu Wei, Huangqi Yu, and Kehong Yuan. Akvq-vl: Attention-aware kv cache adaptive 2-bit quantization for vision-language models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[36] Keda Tao, Haoxuan You, Yang Sui, Can Qin, and Huan Wang. Plug-and-play 1. x-bit kv cache quantization for video large language models. *arXiv preprint arXiv:2503.16257*, 2025.

[37] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[38] Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. Vl-cache: Sparsity and modality-aware kv cache compression for vision-language model inference acceleration. *arXiv preprint arXiv:2410.23317*, 2024.

[39] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

[40] Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4065–4078, 2024.

[41] Huanyu Wang, Jushi Kai, Haoli Bai, Lu Hou, Bo Jiang, Ziwei He, and Zhouhan Lin. Fourier-vlm: Compressing vision tokens in the frequency domain for large vision-language models. *arXiv preprint arXiv:2508.06038*, 2025.

[42] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR, 2023.

[43] Yaoxin Yang, Peng Ye, Xudong Tan, Chongjun Tu, Maosen Zhao, Jia Hao, and Tao Chen. Revisiting multimodal kv cache compression: A frequency-domain-guided outlier-kv-aware approach. *arXiv preprint arXiv:2511.16786*, 2025.

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

[45] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Appendix

Appendix contents

A Proof of Theorem 1 **15**

B Per-Axis Spectral Flatness **17**

C Reproducibility Statement **18**
C.1 Dataset Details 18
C.2 Per-VLM Implementation Details 19

D Per-Subtype 1-bit Results **21**
D.1 RSNA-ICH 21
D.2 CT-RATE 24
D.3 Video-MME 30

E Additional Baseline Comparisons **31**
E.1 Implementation Details 31

F Closure at Higher Bit Rates **32**

G Reconstruction Error of the Inner Scalar Quantizer **32**

H Tile-Fused Inverse-DCT Attention Kernel **33**

I Detailed Related Work **34**

J Limitations **35**

A Proof of Theorem 1

We give the full proof of Theorem 1 from Section 2.3. The proof proceeds under the continuous-bit relaxation of (6) and the high-rate Gaussian distortion model used to derive the per-bin distortion $\mathcal{D}_i(b_i) = c\sigma_i^2 \cdot 2^{-2b_i}$. The two regime caveats noted in the footnote of Section 2.3—integer rounding of b_i and breakdown of the high-rate model at very low average rate—are discussed in Remarks A.2 and A.3 at the end of this appendix.

Preliminaries

Fix a layer index ℓ and a per-bin variance vector $(\sigma_1^2, \ldots, \sigma_N^2)$ with $\sigma_i^2 > 0$, defined as in (4) of Section 2.3. The high-rate Gaussian scalar-quantization model assigns to each bin the per-bin distortion

$$\mathcal{D}_i(b_i) = c \cdot \sigma_i^2 \cdot 2^{-2b_i}, \qquad i = 1, \ldots, N, \tag{9}$$

where $c > 0$ is the quantizer-geometry constant (independent of i). Let $\mathcal{D}(\mathbf{b}) = \sum_{i=1}^N \mathcal{D}_i(b_i)$ denote the total reconstruction MSE under bit allocation $\mathbf{b} = (b_1, \ldots, b_N)$. We work under the continuous relaxation $b_i \in \mathbb{R}_{\geq 0}$ and the matched budget constraint $\sum_{i=1}^N b_i = R_\ell$.

For convenience we restate the theorem before proving it.

Theorem 1 (restated from Section 2.3)**.** *Let $\{\sigma_i^2\}_{i=1}^N$ be per-bin variances, and define the spectral flatness*

$$\gamma = \frac{\left(\prod_{i=1}^N \sigma_i^2\right)^{1/N}}{(1/N)\sum_{i=1}^N \sigma_i^2} \in (0, 1].$$

Under matched bit budget R_ℓ and the high-rate Gaussian distortion model (9)*, the total distortion under uniform allocation and under reverse water-filling allocation satisfies*

$$\frac{\mathcal{D}_{\text{uniform}}}{\mathcal{D}_{\text{water-fill}}} = \frac{1}{\gamma}.$$

Proof. The proof has five steps. Step 1 evaluates the uniform-allocation distortion. Steps 2 and 3 derive the water-filling solution from the KKT conditions of the constrained optimization. Step 4 specializes to the high-rate regime in which all bins are active. Step 5 forms the ratio.

Step 1: Uniform-allocation distortion. Setting $b_i = R_\ell/N$ for every i in (9) gives

$$\mathcal{D}_{\text{uniform}} \;=\; \sum_{i=1}^N c \cdot \sigma_i^2 \cdot 2^{-2R_\ell/N} \;=\; c \cdot 2^{-2R_\ell/N} \cdot \sum_{i=1}^N \sigma_i^2. \tag{10}$$

Step 2: Convexity and KKT conditions for water-filling. The water-filling problem is the constrained optimization

$$\min_{\mathbf{b} \in \mathbb{R}_{\geq 0}^N} \; \sum_{i=1}^N c \cdot \sigma_i^2 \cdot 2^{-2b_i} \quad \text{subject to} \quad \sum_{i=1}^N b_i \;=\; R_\ell. \tag{11}$$

The objective is a sum of strictly convex functions of the form $b \mapsto c\sigma_i^2 \cdot 2^{-2b}$ (the second derivative is $4(\ln 2)^2 \cdot c\sigma_i^2 \cdot 2^{-2b} > 0$). The equality constraint is affine and the inequality constraints $b_i \geq 0$ are convex, so (11) is a convex program. The KKT conditions are therefore both necessary and sufficient for global optimality.

Introduce a multiplier $\mu \in \mathbb{R}$ for the budget equality and multipliers $\nu_i \geq 0$ for the constraints $-b_i \leq 0$. The Lagrangian is

$$L(\mathbf{b}, \mu, \boldsymbol{\nu}) \;=\; \sum_{i=1}^N c\sigma_i^2 \cdot 2^{-2b_i} \;+\; \mu\left(\sum_{i=1}^N b_i - R_\ell\right) \;-\; \sum_{i=1}^N \nu_i b_i.$$

Stationarity in b_i yields

$$\frac{\partial L}{\partial b_i} \;=\; -2\ln 2 \cdot c\sigma_i^2 \cdot 2^{-2b_i} \;+\; \mu \;-\; \nu_i \;=\; 0 \tag{12}$$

for every i. Define the water level

$$\lambda \;:=\; \frac{\mu}{2\,c\,\ln 2}, \tag{13}$$

which is positive because $c > 0$ and $\mu > 0$ (any candidate optimum has $\mu > 0$ since the objective strictly decreases in each b_i). Equation (12) rearranges to $c\sigma_i^2 \cdot 2^{-2b_i} = (\mu - \nu_i)/(2\ln 2)$, equivalently

$$\sigma_i^2 \cdot 2^{-2b_i} \;=\; \lambda - \frac{\nu_i}{2\,c\,\ln 2}. \tag{14}$$

Step 3: Active set and per-bin distortion. Complementary slackness gives $\nu_i \cdot b_i = 0$ for each i. We split bins into two cases.

Active bins. If $b_i > 0$, then $\nu_i = 0$ and (14) reduces to $\sigma_i^2 \cdot 2^{-2b_i} = \lambda$, equivalent to

$$b_i \;=\; \tfrac{1}{2}\log_2(\sigma_i^2/\lambda). \tag{15}$$

For (15) to give $b_i > 0$ we need $\sigma_i^2 > \lambda$. The corresponding per-bin distortion under (9) is

$$\mathcal{D}_i \;=\; c \cdot \sigma_i^2 \cdot 2^{-2b_i} \;=\; c\,\lambda, \tag{16}$$

which is the same constant for every active bin.

Inactive bins. If $b_i = 0$, then (12) gives $\nu_i = \mu - 2\,c\sigma_i^2\,\ln 2 = 2\,c\,\ln 2\,(\lambda - \sigma_i^2)$, and the constraint $\nu_i \geq 0$ requires $\sigma_i^2 \leq \lambda$.

Defining the active set

$$A \;:=\; \{i : \sigma_i^2 > \lambda\}, \qquad K := |A|,$$

we have $b_i > 0$ if and only if $i \in A$, and the per-active-bin distortion is the constant $c\lambda$.

Step 4: High-rate regime. The high-rate Gaussian distortion model presumes that every bin receives non-trivial bit allocation, i.e., that R_ℓ is large enough that the optimal water level λ falls strictly below the smallest variance:

$$\lambda \;<\; \min_i \sigma_i^2. \tag{17}$$

In this regime $A = \{1, \ldots, N\}$ and $K = N$, so (15) and (16) hold for every i. Substituting the per-bin allocation (15) into the budget constraint $\sum_{i=1}^N b_i = R_\ell$ yields

$$\sum_{i=1}^N \tfrac{1}{2}\log_2(\sigma_i^2/\lambda) \;=\; R_\ell \;\iff\; \tfrac{1}{2}\left(\sum_{i=1}^N \log_2 \sigma_i^2 \;-\; N\log_2\lambda\right) \;=\; R_\ell,$$

which solves to

$$\log_2\lambda \;=\; \frac{1}{N}\sum_{i=1}^N \log_2\sigma_i^2 \;-\; \frac{2R_\ell}{N}, \qquad \text{equivalently} \qquad \lambda \;=\; \left(\prod_{i=1}^N \sigma_i^2\right)^{1/N} \cdot 2^{-2R_\ell/N}. \tag{18}$$

The total water-filling distortion is therefore

$$\mathcal{D}_{\text{water-fill}} \;=\; \sum_{i=1}^N c\,\lambda \;=\; c\,N\,\lambda \;=\; c\,N\,\left(\prod_{i=1}^N \sigma_i^2\right)^{1/N} \cdot 2^{-2R_\ell/N}. \tag{19}$$

Step 5: Ratio. Dividing (10) by (19),

$$\frac{\mathcal{D}_{\text{uniform}}}{\mathcal{D}_{\text{water-fill}}} \;=\; \frac{c \cdot 2^{-2R_\ell/N} \cdot \sum_{i=1}^N \sigma_i^2}{c\,N\,\left(\prod_{i=1}^N \sigma_i^2\right)^{1/N} \cdot 2^{-2R_\ell/N}} \;=\; \frac{(1/N)\sum_{i=1}^N \sigma_i^2}{\left(\prod_{i=1}^N \sigma_i^2\right)^{1/N}} \;=\; \frac{1}{\gamma},$$

where the last equality uses the definition of γ in (7). Both occurrences of c and the common factor $2^{-2R_\ell/N}$ cancel, so the ratio is independent of the bit budget and of the quantizer-geometry constant. □

Remarks

Remark A.1 (Range of γ and tightness). The arithmetic-geometric mean inequality applied to the positive reals $\sigma_1^2, \ldots, \sigma_N^2$ gives

$$\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 \;\geq\; \left(\prod_{i=1}^{N}\sigma_i^2\right)^{1/N},$$

with equality if and only if $\sigma_1^2 = \cdots = \sigma_N^2$. Combined with $\sigma_i^2 > 0$ for all i, this shows $\gamma \in (0, 1]$, and consequently $1/\gamma \geq 1$. The water-filling advantage equals one (uniform is already optimal) precisely when the cache spectrum is flat, and grows without bound as the spectrum concentrates on a small number of bins.

Remark A.2 (Integer rounding penalty). The implementation rounds b_i via the floor in (6) and applies a greedy ± 1 correction to enforce $\sum_i b_i = R_\ell$ exactly. This restricts the feasible set of bit allocations to $\mathbb{Z}^N_{\geq 0}$, so the integer-allocated water-fill distortion satisfies

$$\mathcal{D}^{\text{int}}_{\text{water-fill}} \;\geq\; \mathcal{D}^{\text{cont}}_{\text{water-fill}},$$

where the right-hand side is the continuous-relaxation value used in the proof above. The empirical ratio $\mathcal{D}_{\text{uniform}}/\mathcal{D}^{\text{int}}_{\text{water-fill}}$ is therefore tightened toward 1 relative to the closed-form $1/\gamma$, and is visible as a small finite-rate gap on the points inside the closure band of Figure 3.

Remark A.3 (Breakdown of the high-rate model at low average rate). The high-rate Gaussian distortion model (9) is an asymptotic approximation for large b_i and Gaussian-distributed source coordinates. When the average bit rate R_ℓ/N becomes small, two effects pull the empirical ratio away from the closed-form $1/\gamma$. First, condition (17) can fail and bins with $\sigma_i^2 \leq \lambda$ receive zero bits, in which case the active-set argument in Step 3 yields a partition-dependent expression for $\mathcal{D}_{\text{water-fill}}$ rather than the clean $cN\lambda$ form of Step 4. Second, on heavy-tailed empirical post-DCT distributions (most pronounced for the Hulu-Med 4B/7B K-cache in Section 3.3), the actual scalar-quantization distortion at $b_i = 1$ deviates from the high-rate prediction $c\sigma_i^2 \cdot 2^{-2}$ in a direction that disproportionately penalizes uniform allocation, since uniform spreads bits across low-energy bins where the high-rate model is loosest. The two effects can combine, and the proof above does not control either, which is why we report the four out-of-band entries in Section 3.3 for completeness rather than as confirmations of the closed form.

B Per-Axis Spectral Flatness

This appendix reports the per-axis spectral flatness of the visual KV cache that motivates the 3-axis design.

Protocol

For each VLM-dataset pair we sample 20 studies, run prefill at the per-VLM official slice budget, and read out the per-bin variance tensor $\sigma^2 \in \mathbb{R}^{S\times H\times W}_{\geq 0}$ defined in (4). For each axis $\alpha \in \{s, h, w\}$ with corresponding size $N_\alpha \in \{S, H, W\}$ we project σ^2 onto its 1-D marginal by summing over the other two axes,

$$m_s[i] = \sum_{j=1}^{H}\sum_{k=1}^{W}\sigma^2_{ijk}, \qquad m_h[j] = \sum_{i=1}^{S}\sum_{k=1}^{W}\sigma^2_{ijk}, \qquad m_w[k] = \sum_{i=1}^{S}\sum_{j=1}^{H}\sigma^2_{ijk},$$

and apply the flatness functional of (7) to the marginal,

$$\gamma^\alpha_K = \frac{\left(\prod_{i=1}^{N_\alpha} m_\alpha[i]\right)^{1/N_\alpha}}{(1/N_\alpha)\sum_{i=1}^{N_\alpha} m_\alpha[i]} \in (0, 1].$$

We report per-axis γ^α_K averaged across all transformer layers and all sampled studies for each pair.

Table 8: **Per-axis spectral flatness γ_K^α of the visual K cache.** The joint γ_K column reproduces the value reported in Table 2. Per-axis γ_K^α is computed on the marginal variance projection along axis α, averaged over transformer layers and over 20 sampled studies per pair.

	RSNA-ICH				CT-RATE			
VLM	γ_K	γ_K^s	γ_K^h	γ_K^w	γ_K	γ_K^s	γ_K^h	γ_K^w
MedGemma 1.5 4B	0.279	0.389	0.458	0.412	0.171	0.215	0.473	0.447
MedGemma 4B	0.327	0.448	0.534	0.468	0.200	0.249	0.534	0.480
MedGemma 27B	0.265	0.443	0.440	0.379	0.180	0.258	0.432	0.396
Hulu-Med 7B	0.206	0.416	0.354	0.273	0.132	0.279	0.345	0.271
Hulu-Med 4B	0.253	0.444	0.392	0.336	0.162	0.282	0.367	0.320
Hulu-Med 32B	0.224	0.500	0.415	0.300	0.152	0.350	0.403	0.296

Results

Table 8 reports γ_K^α on the K cache across the six medical VLMs and two CT benchmarks.

Every axis is spectrally concentrated. The largest entry of γ_K^α across the 36 axis-VLM-dataset entries of Table 8 is 0.534, on MedGemma 4B at the H axis on both datasets, so $\gamma_K^\alpha < 0.55$ everywhere. The marginal-wise water-filling advantage from Theorem 1 therefore satisfies $1/\gamma_K^\alpha > 1.87\times$ on every axis. Each dropped axis forfeits at least this margin under the marginal-wise application of the theorem, in agreement with the monotone $3 \to 2 \to 1$-axis degradation in Table 4. The pattern of axis-wise minima is split between the slice axis on six pairs and the width axis on the other six, with the H axis never the most concentrated, so any single dropped axis includes a strongly concentrated direction on at least half of the pairs.

C Reproducibility Statement

C.1 Dataset Details

This appendix documents the official source, license, and sampling protocol of each benchmark used in Section 3, in the order they appear in Section 3.1.

RSNA-ICH. The 2019 RSNA Intracranial Hemorrhage Detection Challenge releases a public head-CT dataset under the RSNA Brain CT Hemorrhage Challenge Terms of Use and Attribution [11]. Access is granted for non-commercial academic research and education, redistribution of any portion of the data is prohibited, and any use must cite the dataset as "Brain CT Hemorrhage Dataset, Copyright RSNA, 2019" alongside the Flanders et al. reference. Each study carries six binary labels, the global `any` flag and the five hemorrhage subtypes `epidural`, `intraparenchymal`, `intraventricular`, `subarachnoid`, and `subdural`. After de-duplication and removal of studies with corrupted axial DICOM windows, $N = 21{,}744$ usable training-split studies remain. We evaluate on a stratified subset of $N = 3{,}000$ studies obtained by stratifying the full pool on the (`any`, `epidural`) cross-product, the rare-subtype `epidural` drives the smallest stratum and is locked in to remove sample-to-sample variance across methods, then proportionally allocating $3{,}000$ slots and drawing without replacement at fixed seed 20260423. Table 9 reports the per-label prevalence in the resulting subset alongside the full training pool. Every label prevalence in the subset matches the training pool to within 0.23 percentage points, so the subset preserves the underlying class distribution rather than selecting a sweet spot. The sampling script and the resulting manifest of study UIDs are released with the supplementary code.

CT-RATE. CT-RATE [18] is a public chest-CT benchmark released under CC BY-NC-SA 4.0, providing volumes paired with reconstructed reports and 18 binary multi-label abnormality annotations. The license restricts use to non-commercial academic research and prohibits redistribution of the dataset or any portion thereof. The official validation split contains $N = 3{,}039$ volumes from $1{,}304$ held-out patients. We evaluate on the full validation split with no further sampling. Each volume is processed under the per-VLM official preprocessing recipe (slice count and HU windowing inherited from the published evaluation pipeline of each VLM family), and predictions are made on each of the 18 abnormalities independently for every volume, yielding $18 \times 3{,}039 = 54{,}702$ binary predictions per entry of the main results table.

Table 9: **Per-label prevalence on the RSNA-ICH evaluation subset versus the full training pool.** The subset is obtained by proportional stratified sampling on the (`any`, `epidural`) cross-product at seed 20260423. Differences in prevalence are reported in absolute percentage points.

	Training pool (N=21,744)		**Subset** (N=3,000)		
Label	positives	rate (%)	positives	rate (%)	Δ (pp)
`any`	8,882	40.85	1,226	40.87	+0.02
`epidural`	354	1.63	49	1.63	0.00
`intraparenchymal`	5,321	24.47	729	24.30	−0.17
`intraventricular`	3,692	16.98	506	16.87	−0.11
`subarachnoid`	3,932	18.08	536	17.87	−0.21
`subdural`	3,812	17.53	519	17.30	−0.23

Video-MME. Video-MME [13] is a multimodal benchmark covering 900 videos and 2,700 human-annotated four-way multiple-choice questions across short, medium, and long durations. The official terms restrict use to academic research and prohibit commercial use, distribution, publication, copying, dissemination, or modification of the data without prior approval from the authors. We evaluate on all 2,700 questions in the no-subtitle setting using the official Video-MME no-subtitle prompt verbatim, and extract the answer letter via the rule-based regex of the `lmms-eval` reference implementation. Each VLM uses its own video sampling recipe drawn from the public reference for that model, with the per-VLM frame schedule, pixel budget, and processor settings documented in Appendix C.2.

C.2 Per-VLM Implementation Details

This appendix records the per-VLM configuration of every evaluation in Section 3, in the order each topic is referenced from the main text.

Hardware, software stack, and compute resources. All experiments run on a single 96 GiB NVIDIA RTX PRO 6000 Blackwell Server Edition GPU. Models are loaded in `bfloat16` via the HuggingFace `transformers` library with `attn_implementation="flash_attention_2"`. The KV-cache compression backends are PyTorch reimplementations integrated through the standard `transformers.cache_utils.DynamicCache` API and shipped with the supplementary code release. A single ⟨VLM, dataset, method, bit⟩ entry in Table 1 or Table 5 corresponds to one full eval pass on this GPU, and the wall-clock per entry scales primarily with VLM size, slice budget, and the number of full-attention layers covered by the compression backend. For TACT-KV at 1-bit, the RSNA-ICH pass (N=3,000 studies, 60-slice budget) takes approximately 1.5 h on MedGemma 1.5 4B and MedGemma 4B (five full-attention layers compressed), 2.5–3 h on Hulu-Med 4B and Hulu-Med 7B (all transformer layers compressed), 3 h on MedGemma 27B, and 8 h on Hulu-Med 32B; the corresponding CT-RATE pass (N=3,039, 85-slice budget) takes approximately 3.5 h on the 4B-scale variants, 3 h on Hulu-Med 7B, and 8 h on MedGemma 27B and Hulu-Med 32B. The Video-MME pass (N=2,700 four-way MCQs) finishes in under 3 h on Qwen2.5-VL-7B, LLaVA-OneVision-7B, and VideoLLaMA3-7B under their published sparse-sampling configurations, and in approximately a day on Qwen3-VL-8B under the dense cookbook configuration described below. Per-entry wall-clock for the four baseline backends sits within the same range, with the no-transform KIVI and TurboQuant somewhat faster than TACT-KV and the saliency-based AKVQ-VL somewhat slower at 1-bit. Aggregated over the cells reported in the main tables, the ablations of Section 3.4, and the Theorem 1 closure replicas of Section 3.3, the measured runs together account for on the order of 1,500 GPU-hours on the GPU class above; preliminary kernel iterations, calibration sweeps, and runs discarded for protocol changes contribute a comparable additional budget, for a project total in the low thousands of GPU-hours.

Medical CT shared protocol. For both RSNA-ICH and CT-RATE we use rank classification on assertion-style completion pairs. For each volume we build a single shared visual prefix that interleaves vision tokens with the text markers `SLICE 1`, `SLICE 2`, …, prefill it once, snapshot the resulting cache, and for each abnormality forward only the completion tokens against the restored cache. The label is decided by comparing the mean per-token log-probability of the positive completion against the negative completion. RSNA-ICH uses the pair (`X hemorrhage is present.`, `X hemorrhage is not present.`) for $X \in \{$`intracranial`, `epidural`, `intraparenchymal`, `intraventricular`, `subarachnoid`, `subdural`$\}$. CT-RATE uses (`X is present.`, `X is not`

present.) for each of the 18 abnormalities. The shared-prefix cache pattern is reused identically across compression backends so that the evaluated quantization step is the only factor that changes between methods.

Slice budget and HU windowing. Each volume is subsampled to a fixed slice budget along the axial axis using the MedGemma evenly-spaced indexing rule, $\text{idx}(i) = \lfloor i \cdot (n_{\text{native}} - 1)/n_{\text{budget}} \rceil$ for $i = 1, \ldots, n_{\text{budget}}$. The default budget is 60 slices for RSNA-ICH and 85 slices for CT-RATE, matching the published MedGemma evaluation recipe. Each subsampled slice is mapped to a three-channel image by triple HU windowing with the MedGemma standard windows (R, G, B) $= ((-1024, 1024), (-135, 215), (0, 80))$. The same preprocessing is applied for all six medical VLMs to keep the visual input identical across methods and bit budgets.

Video-MME shared protocol. For Video-MME we use the official no-subtitle prompt verbatim, `"Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option.\n{question}\n{options}\nThe best answer is:"`, with options as the four letter-prefixed strings shipped by the dataset. We generate up to 8 tokens by greedy decoding (`do_sample=False`) and extract the four-way letter via the rule-based regex of the `lmms-eval` reference implementation. The same prompt is shared by all four general-purpose VLMs.

MedGemma family (1.5 4B, 4B, 27B). The MedGemma family is loaded from the official HuggingFace checkpoints `medgemma-1.5-4b-it`, `medgemma-4b-it`, and `medgemma-27b-it`. All three models inherit the Gemma-3 decoder, which interleaves five sliding-window layers per global full-attention layer. MedGemma 1.5 4B and MedGemma 4B each have 5 full-attention layers and 29 sliding-window layers (34 total). MedGemma 27B has 10 full-attention and 52 sliding-window layers (62 total). All three models set the sliding-window length to 1,024 tokens, which already caps the per-layer cache on sliding-window layers via the standard dynamic sliding-window cache layer of the HuggingFace transformers library. We therefore leave sliding-window layers at bf16 and apply the compression backend only to full-attention layers, following the production convention of vLLM. The bypass is implemented uniformly across compression backends in the supplementary code release.

Hulu-Med family (4B, 7B, 32B). The Hulu-Med family is loaded from the official HuggingFace checkpoints `Hulu-Med-4B`, `Hulu-Med-7B`, and `Hulu-Med-32B`. The family ships a custom processor and chat template that emit an image-token marker for each image block, so we load the weights through the standard causal-language-model loader with remote code enabled. Hulu-Med 4B uses a Qwen3 backbone with 36 transformer layers, Hulu-Med 7B uses a Qwen2 backbone with 28 layers, and Hulu-Med 32B uses a Qwen2 backbone with 64 layers. The text-config of all three models marks every layer as full-attention and leaves the sliding-window field unset, so the compression backend applies to every transformer layer with no bypass. The image processor caps the total visual-token budget at approximately 16,000 tokens and redistributes the per-slice token count when the input exceeds this budget, so the effective context length plateaus around 14–17 k tokens at slice budgets at or above 60.

Qwen2.5-VL-7B-Instruct. Loaded from the official HuggingFace checkpoint `Qwen2.5-VL-7B-Instruct`. Video frames are sampled by the official Qwen-VL utility under the model-card sparse configuration, with a maximum pixel count of 360×420, fps $= 1.0$, and a frame cap of 32. The dense cookbook configuration shipped with the Qwen3-VL repository drops Video-MME accuracy on the short-bucket by approximately 10 percentage points on this model, so we follow the model card.

Qwen3-VL-8B-Instruct. Loaded from the official HuggingFace checkpoint `Qwen3-VL-8B-Instruct`. Video frames are sampled by the Qwen-VL utility under the dense cookbook configuration of the Qwen3-VL repository, with a total pixel budget of 20 M, a minimum pixel budget of $64 \cdot 32 \cdot 32$, sample fps $= 2$, a frame cap of 2,048, image patch size 16, and per-video metadata returned by the processor. The generic `lmms-eval` sparse configuration under-performs the cookbook by approximately 20 percentage points on this model, so we follow the cookbook.

LLaVA-OneVision-7B. Loaded from the official HuggingFace checkpoint `llava-onevision-qwen2-7b-ov-hf`. Video frames are sampled at 32 uniformly spaced indices over the full video duration and passed to the processor as a stacked array of shape $(32, h, w, 3)$.

VideoLLaMA3-7B. Loaded from the official HuggingFace checkpoint `VideoLLaMA3-7B` through the standard causal-language-model loader with remote code enabled. Video frames are sampled at fps $= 1$ with a cap of 64 frames per video, reduced from the `lmms-eval` default of 180 to fit the per-entry compute budget. The 64-frame setting reproduces the published Table 8 accuracy of the model release within ± 2 percentage points. We disable the model's data-dependent visual-token compression because it drops visual tokens whose inter-frame pixel difference falls below a threshold and produces a variable visual-token count per video, which is incompatible with the fixed (S, H, W) reshape used by our 3-axis transform. With compression disabled, the visual-token count is deterministic at $T \times (H/m) \times (W/m)$ for video length T and projector merge factor m.

KV-cache compression configuration. All compression backends operate on the visual portion of the KV cache only and hold the non-visual portion (chat template, image-boundary markers, slice labels, the prompt suffix) at bf16. The visual layout (S, H, W) is built once per prefill from the input-id stream and the image-token index in the model config, and is handed to the backend before the prefill cache is filled. Bit budgets are reported at $\{1, 2, 3, 4\}$ on the visual portion. The 8-bit setting is approximately lossless across all backends and is omitted from the main tables. The TACT-KV, TurboQuant, KIVI, VidKV, and AKVQ-VL backends are each implemented as a subclass of the standard dynamic-cache class in the HuggingFace transformers library and shipped in the supplementary code release. Gain-shape SQ is not a stand-alone baseline but is the per-bin scalar codebook used inside TACT-KV, and is also the codebook varied in the quantizer ablation of Section 3.4.

D Per-Subtype 1-bit Results

This appendix expands the macro prediction-flip rate and AUROC reported in Table 1 into the underlying per-subtype results at 1-bit on the three benchmarks evaluated in Section 3. Each medical-CT VLM is reported in its own table with two vertically stacked panels. The top panel reports the macro prediction-flip rate ($\%$) against the bf16 reference and the bottom panel reports AUROC, with the bf16 column included in the bottom panel as the uncompressed reference. The bf16 column is omitted from the top panel because the flip rate of bf16 against itself is zero by definition. Half-widths follow the Wald 95% binomial CI for flip rate (with the Wilson upper bound used for $p = 0$ entries) and the percentile bootstrap 95% CI ($1{,}000$ resamples) for AUROC.

D.1 RSNA-ICH

Tables 10–15 report per-subtype 1-bit results on RSNA-ICH for the six medical VLMs. The six subtypes follow the official challenge protocol [11]: the global `any` hemorrhage flag and the five anatomical subtypes `epidural`, `intraparenchymal`, `intraventricular`, `subarachnoid`, and `subdural`. The macro row at the bottom of each panel is the unweighted mean over the six subtypes and reproduces the value reported in the corresponding row of Table 1.

Table 10: **Per-subtype 1-bit results on RSNA-ICH for MedGemma 1.5 4B (N=3,000).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Subtype	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
any	—	$2.33_{\pm 0.54}$	$3.73_{\pm 0.68}$	$4.27_{\pm 0.72}$	$4.07_{\pm 0.71}$	$14.61_{\pm 1.26}$
epidural	—	$0.83_{\pm 0.32}$	$2.03_{\pm 0.50}$	$5.97_{\pm 0.85}$	$1.83_{\pm 0.48}$	$31.92_{\pm 1.67}$
intraparenchymal	—	$5.60_{\pm 0.82}$	$7.77_{\pm 0.96}$	$10.57_{\pm 1.10}$	$10.11_{\pm 1.08}$	$27.42_{\pm 1.60}$
intraventricular	—	$7.77_{\pm 0.96}$	$10.27_{\pm 1.09}$	$19.94_{\pm 1.43}$	$13.88_{\pm 1.24}$	$37.56_{\pm 1.73}$
subarachnoid	—	$1.13_{\pm 0.38}$	$4.80_{\pm 0.77}$	$7.90_{\pm 0.97}$	$4.77_{\pm 0.76}$	$33.69_{\pm 1.69}$
subdural	—	$5.04_{\pm 0.78}$	$9.04_{\pm 1.03}$	$15.77_{\pm 1.30}$	$10.61_{\pm 1.10}$	$26.05_{\pm 1.57}$
Macro	—	$\mathbf{3.78}_{\pm 0.28}$	$\mathbf{6.27}_{\pm 0.35}$	$\mathbf{10.74}_{\pm 0.45}$	$\mathbf{7.54}_{\pm 0.39}$	$\mathbf{28.54}_{\pm 0.66}$
AUROC						
any	$0.804_{\pm 0.016}$	$0.802_{\pm 0.016}$	$0.788_{\pm 0.016}$	$0.763_{\pm 0.019}$	$0.778_{\pm 0.017}$	$0.741_{\pm 0.019}$
epidural	$0.693_{\pm 0.075}$	$0.675_{\pm 0.075}$	$0.662_{\pm 0.068}$	$0.673_{\pm 0.076}$	$0.667_{\pm 0.068}$	$0.698_{\pm 0.064}$
intraparenchymal	$0.839_{\pm 0.016}$	$0.833_{\pm 0.017}$	$0.812_{\pm 0.017}$	$0.821_{\pm 0.017}$	$0.784_{\pm 0.018}$	$0.791_{\pm 0.018}$
intraventricular	$0.812_{\pm 0.019}$	$0.809_{\pm 0.018}$	$0.805_{\pm 0.018}$	$0.787_{\pm 0.020}$	$0.784_{\pm 0.018}$	$0.766_{\pm 0.021}$
subarachnoid	$0.700_{\pm 0.024}$	$0.701_{\pm 0.023}$	$0.706_{\pm 0.023}$	$0.690_{\pm 0.025}$	$0.701_{\pm 0.022}$	$0.654_{\pm 0.026}$
subdural	$0.731_{\pm 0.022}$	$0.725_{\pm 0.023}$	$0.713_{\pm 0.023}$	$0.693_{\pm 0.024}$	$0.710_{\pm 0.023}$	$0.667_{\pm 0.026}$
Macro	$\mathbf{0.763}_{\pm 0.017}$	$\mathbf{0.758}_{\pm 0.018}$	$\mathbf{0.748}_{\pm 0.017}$	$\mathbf{0.738}_{\pm 0.018}$	$\mathbf{0.737}_{\pm 0.017}$	$\mathbf{0.720}_{\pm 0.017}$

Table 11: **Per-subtype 1-bit results on RSNA-ICH for MedGemma 4B (N=3,000).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Subtype	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
any	—	$1.70_{\pm 0.46}$	$4.60_{\pm 0.75}$	$1.57_{\pm 0.44}$	$17.14_{\pm 1.35}$	$0.30_{\pm 0.20}$
epidural	—	$9.87_{\pm 1.07}$	$13.80_{\pm 1.23}$	$11.77_{\pm 1.15}$	$31.19_{\pm 1.66}$	$12.34_{\pm 1.18}$
intraparenchymal	—	$1.20_{\pm 0.39}$	$4.67_{\pm 0.76}$	$0.30_{\pm 0.20}$	$14.54_{\pm 1.26}$	$0.13_{\pm 0.13}$
intraventricular	—	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.03_{\pm 0.06}$	$1.53_{\pm 0.44}$	$0.00_{\pm 0.07}$
subarachnoid	—	$1.27_{\pm 0.40}$	$3.27_{\pm 0.64}$	$0.77_{\pm 0.31}$	$17.41_{\pm 1.36}$	$0.37_{\pm 0.22}$
subdural	—	$0.17_{\pm 0.15}$	$0.13_{\pm 0.13}$	$0.17_{\pm 0.15}$	$2.94_{\pm 0.60}$	$0.03_{\pm 0.06}$
Macro	—	$\mathbf{2.37}_{\pm 0.22}$	$\mathbf{4.41}_{\pm 0.30}$	$\mathbf{2.43}_{\pm 0.22}$	$\mathbf{14.13}_{\pm 0.51}$	$\mathbf{2.20}_{\pm 0.21}$
AUROC						
any	$0.783_{\pm 0.017}$	$0.801_{\pm 0.015}$	$0.799_{\pm 0.015}$	$0.758_{\pm 0.018}$	$0.800_{\pm 0.016}$	$0.695_{\pm 0.019}$
epidural	$0.660_{\pm 0.078}$	$0.680_{\pm 0.075}$	$0.631_{\pm 0.075}$	$0.644_{\pm 0.067}$	$0.670_{\pm 0.080}$	$0.597_{\pm 0.087}$
intraparenchymal	$0.730_{\pm 0.020}$	$0.752_{\pm 0.019}$	$0.762_{\pm 0.019}$	$0.736_{\pm 0.021}$	$0.760_{\pm 0.020}$	$0.650_{\pm 0.021}$
intraventricular	$0.742_{\pm 0.022}$	$0.765_{\pm 0.021}$	$0.762_{\pm 0.020}$	$0.699_{\pm 0.024}$	$0.785_{\pm 0.020}$	$0.674_{\pm 0.025}$
subarachnoid	$0.682_{\pm 0.022}$	$0.686_{\pm 0.023}$	$0.674_{\pm 0.022}$	$0.676_{\pm 0.023}$	$0.696_{\pm 0.021}$	$0.650_{\pm 0.023}$
subdural	$0.649_{\pm 0.023}$	$0.682_{\pm 0.022}$	$0.676_{\pm 0.023}$	$0.661_{\pm 0.023}$	$0.704_{\pm 0.023}$	$0.599_{\pm 0.025}$
Macro	$\mathbf{0.707}_{\pm 0.018}$	$\mathbf{0.728}_{\pm 0.017}$	$\mathbf{0.717}_{\pm 0.017}$	$\mathbf{0.696}_{\pm 0.017}$	$\mathbf{0.736}_{\pm 0.017}$	$\mathbf{0.644}_{\pm 0.020}$

Table 12: **Per-subtype 1-bit results on RSNA-ICH for MedGemma 27B (N=3,000).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Subtype	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
any	—	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.53_{\pm 0.26}$	$0.00_{\pm 0.07}$	$6.94_{\pm 0.91}$
epidural	—	$0.53_{\pm 0.26}$	$9.04_{\pm 1.03}$	$96.00_{\pm 0.70}$	$1.77_{\pm 0.47}$	$70.01_{\pm 1.64}$
intraparenchymal	—	$0.00_{\pm 0.07}$	$0.03_{\pm 0.06}$	$0.13_{\pm 0.13}$	$0.00_{\pm 0.07}$	$8.24_{\pm 0.98}$
intraventricular	—	$0.20_{\pm 0.16}$	$8.17_{\pm 0.98}$	$0.33_{\pm 0.21}$	$9.57_{\pm 1.05}$	$2.87_{\pm 0.60}$
subarachnoid	—	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.23_{\pm 0.17}$
subdural	—	$1.47_{\pm 0.43}$	$8.64_{\pm 1.01}$	$1.23_{\pm 0.39}$	$5.37_{\pm 0.81}$	$4.57_{\pm 0.75}$
Macro	—	$\mathbf{0.37}_{\pm 0.09}$	$\mathbf{4.31}_{\pm 0.30}$	$\mathbf{16.37}_{\pm 0.54}$	$\mathbf{2.79}_{\pm 0.24}$	$\mathbf{15.48}_{\pm 0.53}$
AUROC						
any	$0.657_{\pm 0.020}$	$0.677_{\pm 0.019}$	$0.691_{\pm 0.019}$	$0.659_{\pm 0.019}$	$0.652_{\pm 0.020}$	$0.626_{\pm 0.020}$
epidural	$0.451_{\pm 0.090}$	$0.462_{\pm 0.084}$	$0.561_{\pm 0.080}$	$0.535_{\pm 0.076}$	$0.519_{\pm 0.090}$	$0.537_{\pm 0.081}$
intraparenchymal	$0.683_{\pm 0.022}$	$0.687_{\pm 0.022}$	$0.651_{\pm 0.024}$	$0.645_{\pm 0.023}$	$0.625_{\pm 0.023}$	$0.672_{\pm 0.022}$
intraventricular	$0.661_{\pm 0.026}$	$0.680_{\pm 0.024}$	$0.625_{\pm 0.027}$	$0.641_{\pm 0.028}$	$0.620_{\pm 0.028}$	$0.638_{\pm 0.027}$
subarachnoid	$0.573_{\pm 0.027}$	$0.584_{\pm 0.027}$	$0.590_{\pm 0.026}$	$0.552_{\pm 0.029}$	$0.542_{\pm 0.027}$	$0.563_{\pm 0.027}$
subdural	$0.582_{\pm 0.028}$	$0.588_{\pm 0.027}$	$0.590_{\pm 0.028}$	$0.591_{\pm 0.026}$	$0.552_{\pm 0.028}$	$0.572_{\pm 0.028}$
Macro	$\mathbf{0.601}_{\pm 0.020}$	$\mathbf{0.613}_{\pm 0.019}$	$\mathbf{0.618}_{\pm 0.020}$	$\mathbf{0.604}_{\pm 0.018}$	$\mathbf{0.585}_{\pm 0.021}$	$\mathbf{0.601}_{\pm 0.020}$

Table 13: **Per-subtype 1-bit results on RSNA-ICH for Hulu-Med 4B (N=3,000).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Subtype	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
any	—	$16.07_{\pm 1.31}$	$77.93_{\pm 1.48}$	$77.99_{\pm 1.48}$	$27.79_{\pm 1.60}$	$75.45_{\pm 1.54}$
epidural	—	$19.47_{\pm 1.42}$	$27.74_{\pm 1.60}$	$28.41_{\pm 1.61}$	$28.62_{\pm 1.62}$	$28.39_{\pm 1.61}$
intraparenchymal	—	$2.07_{\pm 0.51}$	$94.83_{\pm 0.79}$	$95.40_{\pm 0.75}$	$34.92_{\pm 1.71}$	$93.70_{\pm 0.87}$
intraventricular	—	$0.00_{\pm 0.07}$	$96.93_{\pm 0.62}$	$95.60_{\pm 0.73}$	$0.17_{\pm 0.15}$	$94.76_{\pm 0.80}$
subarachnoid	—	$0.03_{\pm 0.06}$	$96.77_{\pm 0.63}$	$96.97_{\pm 0.61}$	$44.26_{\pm 1.78}$	$95.00_{\pm 0.78}$
subdural	—	$13.50_{\pm 1.22}$	$66.99_{\pm 1.68}$	$67.72_{\pm 1.67}$	$47.36_{\pm 1.79}$	$66.78_{\pm 1.69}$
Macro	—	$\mathbf{8.53}_{\pm 0.41}$	$\mathbf{76.86}_{\pm 0.62}$	$\mathbf{77.01}_{\pm 0.61}$	$\mathbf{30.52}_{\pm 0.67}$	$\mathbf{75.68}_{\pm 0.63}$
AUROC						
any	$0.832_{\pm 0.015}$	$0.815_{\pm 0.015}$	$0.524_{\pm 0.020}$	$0.488_{\pm 0.020}$	$0.532_{\pm 0.021}$	$0.512_{\pm 0.020}$
epidural	$0.623_{\pm 0.079}$	$0.627_{\pm 0.079}$	$0.513_{\pm 0.079}$	$0.590_{\pm 0.078}$	$0.475_{\pm 0.084}$	$0.491_{\pm 0.077}$
intraparenchymal	$0.849_{\pm 0.016}$	$0.827_{\pm 0.016}$	$0.520_{\pm 0.024}$	$0.507_{\pm 0.024}$	$0.487_{\pm 0.025}$	$0.506_{\pm 0.024}$
intraventricular	$0.827_{\pm 0.018}$	$0.820_{\pm 0.020}$	$0.533_{\pm 0.027}$	$0.500_{\pm 0.027}$	$0.544_{\pm 0.028}$	$0.509_{\pm 0.028}$
subarachnoid	$0.693_{\pm 0.023}$	$0.685_{\pm 0.024}$	$0.513_{\pm 0.027}$	$0.507_{\pm 0.027}$	$0.527_{\pm 0.026}$	$0.517_{\pm 0.026}$
subdural	$0.727_{\pm 0.022}$	$0.713_{\pm 0.022}$	$0.495_{\pm 0.026}$	$0.511_{\pm 0.026}$	$0.539_{\pm 0.027}$	$0.495_{\pm 0.028}$
Macro	$\mathbf{0.759}_{\pm 0.019}$	$\mathbf{0.748}_{\pm 0.018}$	$\mathbf{0.516}_{\pm 0.020}$	$\mathbf{0.517}_{\pm 0.018}$	$\mathbf{0.517}_{\pm 0.019}$	$\mathbf{0.505}_{\pm 0.018}$

Table 14: **Per-subtype 1-bit results on RSNA-ICH for Hulu-Med 7B (N=3,000).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Subtype	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
any	—	$6.14_{\pm 0.86}$	$34.11_{\pm 1.70}$	$27.04_{\pm 1.59}$	$22.18_{\pm 1.49}$	$35.76_{\pm 1.72}$
epidural	—	$0.10_{\pm 0.11}$	$25.61_{\pm 1.56}$	$6.64_{\pm 0.89}$	$2.43_{\pm 0.55}$	$13.68_{\pm 1.23}$
intraparenchymal	—	$4.83_{\pm 0.77}$	$31.48_{\pm 1.66}$	$17.44_{\pm 1.36}$	$12.74_{\pm 1.19}$	$29.59_{\pm 1.63}$
intraventricular	—	$14.40_{\pm 1.26}$	$66.72_{\pm 1.69}$	$71.09_{\pm 1.62}$	$67.75_{\pm 1.67}$	$63.18_{\pm 1.73}$
subarachnoid	—	$9.67_{\pm 1.06}$	$33.81_{\pm 1.69}$	$21.37_{\pm 1.47}$	$18.85_{\pm 1.40}$	$31.32_{\pm 1.66}$
subdural	—	$1.33_{\pm 0.41}$	$25.78_{\pm 1.57}$	$8.67_{\pm 1.01}$	$9.14_{\pm 1.03}$	$23.05_{\pm 1.51}$
Macro	—	$\mathbf{6.08}_{\pm 0.35}$	$\mathbf{36.25}_{\pm 0.70}$	$\mathbf{25.38}_{\pm 0.64}$	$\mathbf{22.18}_{\pm 0.61}$	$\mathbf{32.76}_{\pm 0.69}$
AUROC						
any	$0.848_{\pm 0.015}$	$0.804_{\pm 0.017}$	$0.470_{\pm 0.021}$	$0.499_{\pm 0.021}$	$0.609_{\pm 0.021}$	$0.514_{\pm 0.021}$
epidural	$0.757_{\pm 0.064}$	$0.734_{\pm 0.065}$	$0.511_{\pm 0.083}$	$0.499_{\pm 0.094}$	$0.610_{\pm 0.075}$	$0.465_{\pm 0.082}$
intraparenchymal	$0.868_{\pm 0.017}$	$0.828_{\pm 0.019}$	$0.494_{\pm 0.025}$	$0.496_{\pm 0.024}$	$0.582_{\pm 0.024}$	$0.501_{\pm 0.025}$
intraventricular	$0.882_{\pm 0.017}$	$0.874_{\pm 0.018}$	$0.470_{\pm 0.028}$	$0.514_{\pm 0.030}$	$0.669_{\pm 0.024}$	$0.532_{\pm 0.028}$
subarachnoid	$0.750_{\pm 0.023}$	$0.730_{\pm 0.024}$	$0.481_{\pm 0.026}$	$0.506_{\pm 0.026}$	$0.632_{\pm 0.025}$	$0.476_{\pm 0.027}$
subdural	$0.779_{\pm 0.023}$	$0.725_{\pm 0.026}$	$0.534_{\pm 0.027}$	$0.523_{\pm 0.029}$	$0.626_{\pm 0.027}$	$0.521_{\pm 0.027}$
Macro	$\mathbf{0.814}_{\pm 0.015}$	$\mathbf{0.782}_{\pm 0.016}$	$\mathbf{0.493}_{\pm 0.020}$	$\mathbf{0.506}_{\pm 0.020}$	$\mathbf{0.622}_{\pm 0.018}$	$\mathbf{0.502}_{\pm 0.020}$

Table 15: **Per-subtype 1-bit results on RSNA-ICH for Hulu-Med 32B (N=3,000).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Subtype	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
any	—	$6.27_{\pm 0.87}$	$12.57_{\pm 1.19}$	$31.41_{\pm 1.66}$	$26.98_{\pm 1.59}$	$42.56_{\pm 1.77}$
epidural	—	$8.57_{\pm 1.00}$	$12.77_{\pm 1.19}$	$15.67_{\pm 1.30}$	$15.04_{\pm 1.28}$	$34.59_{\pm 1.70}$
intraparenchymal	—	$3.93_{\pm 0.70}$	$9.60_{\pm 1.05}$	$21.94_{\pm 1.48}$	$14.34_{\pm 1.25}$	$43.46_{\pm 1.77}$
intraventricular	—	$18.54_{\pm 1.39}$	$34.08_{\pm 1.70}$	$69.49_{\pm 1.65}$	$59.84_{\pm 1.75}$	$27.75_{\pm 1.60}$
subarachnoid	—	$4.80_{\pm 0.77}$	$58.75_{\pm 1.76}$	$92.66_{\pm 0.93}$	$65.08_{\pm 1.71}$	$29.72_{\pm 1.64}$
subdural	—	$12.64_{\pm 1.19}$	$21.71_{\pm 1.48}$	$28.74_{\pm 1.62}$	$26.72_{\pm 1.58}$	$33.79_{\pm 1.69}$
Macro	—	$\mathbf{9.13}_{\pm 0.42}$	$\mathbf{24.91}_{\pm 0.63}$	$\mathbf{43.32}_{\pm 0.72}$	$\mathbf{34.67}_{\pm 0.70}$	$\mathbf{35.31}_{\pm 0.70}$
AUROC						
any	$0.875_{\pm 0.013}$	$0.864_{\pm 0.014}$	$0.815_{\pm 0.016}$	$0.540_{\pm 0.021}$	$0.782_{\pm 0.017}$	$0.688_{\pm 0.018}$
epidural	$0.752_{\pm 0.057}$	$0.751_{\pm 0.069}$	$0.683_{\pm 0.073}$	$0.558_{\pm 0.079}$	$0.645_{\pm 0.083}$	$0.580_{\pm 0.079}$
intraparenchymal	$0.893_{\pm 0.015}$	$0.883_{\pm 0.016}$	$0.824_{\pm 0.019}$	$0.530_{\pm 0.024}$	$0.824_{\pm 0.017}$	$0.689_{\pm 0.022}$
intraventricular	$0.917_{\pm 0.014}$	$0.905_{\pm 0.015}$	$0.889_{\pm 0.018}$	$0.542_{\pm 0.027}$	$0.847_{\pm 0.019}$	$0.726_{\pm 0.025}$
subarachnoid	$0.774_{\pm 0.024}$	$0.781_{\pm 0.022}$	$0.740_{\pm 0.025}$	$0.522_{\pm 0.027}$	$0.745_{\pm 0.023}$	$0.648_{\pm 0.024}$
subdural	$0.785_{\pm 0.021}$	$0.772_{\pm 0.022}$	$0.721_{\pm 0.024}$	$0.515_{\pm 0.028}$	$0.713_{\pm 0.024}$	$0.618_{\pm 0.026}$
Macro	$\mathbf{0.833}_{\pm 0.014}$	$\mathbf{0.826}_{\pm 0.015}$	$\mathbf{0.779}_{\pm 0.017}$	$\mathbf{0.535}_{\pm 0.017}$	$\mathbf{0.760}_{\pm 0.019}$	$\mathbf{0.658}_{\pm 0.018}$

D.2 CT-RATE

Tables 16–21 report per-abnormality 1-bit results on the CT-RATE validation split for the six medical VLMs. The eighteen abnormalities follow the official CT-CLIP zero-shot multi-label protocol [18]. The macro row at the bottom of each panel is the unweighted mean over the eighteen abnormalities and reproduces the value reported in the corresponding row of Table 1.

Table 16: **Per-abnormality 1-bit results on CT-RATE for MedGemma 1.5 4B (N=3,039).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Abnormality	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
Medical material	—	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.59_{\pm 0.27}$	$0.00_{\pm 0.07}$	$8.00_{\pm 0.96}$
Arterial wall calcification	—	$0.00_{\pm 0.07}$	$6.55_{\pm 0.88}$	$26.82_{\pm 1.58}$	$0.07_{\pm 0.09}$	$17.14_{\pm 1.34}$
Cardiomegaly	—	$3.85_{\pm 0.68}$	$2.96_{\pm 0.60}$	$39.06_{\pm 1.73}$	$5.56_{\pm 0.81}$	$21.19_{\pm 1.45}$
Pericardial effusion	—	$6.84_{\pm 0.90}$	$16.16_{\pm 1.31}$	$53.90_{\pm 1.77}$	$6.84_{\pm 0.90}$	$16.85_{\pm 1.33}$
Coronary artery wall calcification	—	$19.97_{\pm 1.42}$	$34.02_{\pm 1.68}$	$48.37_{\pm 1.78}$	$20.11_{\pm 1.43}$	$36.72_{\pm 1.71}$
Hiatal hernia	—	$3.59_{\pm 0.66}$	$1.48_{\pm 0.43}$	$37.08_{\pm 1.72}$	$2.44_{\pm 0.55}$	$27.94_{\pm 1.60}$
Lymphadenopathy	—	$3.39_{\pm 0.64}$	$8.72_{\pm 1.00}$	$10.37_{\pm 1.08}$	$9.05_{\pm 1.02}$	$30.60_{\pm 1.64}$
Emphysema	—	$1.09_{\pm 0.37}$	$1.41_{\pm 0.42}$	$14.61_{\pm 1.26}$	$1.58_{\pm 0.44}$	$9.87_{\pm 1.06}$
Atelectasis	—	$3.69_{\pm 0.67}$	$5.46_{\pm 0.81}$	$6.52_{\pm 0.88}$	$6.42_{\pm 0.87}$	$38.63_{\pm 1.73}$
Lung nodule	—	$12.18_{\pm 1.16}$	$17.90_{\pm 1.36}$	$29.35_{\pm 1.62}$	$23.72_{\pm 1.51}$	$27.90_{\pm 1.59}$
Lung opacity	—	$4.15_{\pm 0.71}$	$11.02_{\pm 1.11}$	$18.85_{\pm 1.39}$	$15.10_{\pm 1.27}$	$18.03_{\pm 1.37}$
Pulmonary fibrotic sequela	—	$8.65_{\pm 1.00}$	$15.20_{\pm 1.28}$	$20.70_{\pm 1.44}$	$16.12_{\pm 1.31}$	$15.63_{\pm 1.29}$
Pleural effusion	—	$5.00_{\pm 0.77}$	$17.18_{\pm 1.34}$	$25.21_{\pm 1.54}$	$17.93_{\pm 1.36}$	$22.44_{\pm 1.48}$
Mosaic attenuation pattern	—	$7.60_{\pm 0.94}$	$12.70_{\pm 1.18}$	$37.61_{\pm 1.72}$	$27.18_{\pm 1.58}$	$31.49_{\pm 1.65}$
Peribronchial thickening	—	$1.91_{\pm 0.49}$	$4.84_{\pm 0.76}$	$21.32_{\pm 1.46}$	$5.50_{\pm 0.81}$	$15.73_{\pm 1.29}$
Consolidation	—	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$3.06_{\pm 0.61}$
Bronchiectasis	—	$2.04_{\pm 0.50}$	$2.11_{\pm 0.51}$	$3.13_{\pm 0.62}$	$3.52_{\pm 0.66}$	$20.30_{\pm 1.43}$
Interlobular septal thickening	—	$0.53_{\pm 0.26}$	$6.75_{\pm 0.89}$	$24.55_{\pm 1.53}$	$3.26_{\pm 0.63}$	$11.39_{\pm 1.13}$
Macro	—	$\mathbf{4.69}_{\pm 0.18}$	$\mathbf{9.14}_{\pm 0.24}$	$\mathbf{23.22}_{\pm 0.35}$	$\mathbf{9.13}_{\pm 0.24}$	$\mathbf{20.72}_{\pm 0.34}$
AUROC						
Medical material	$0.541_{\pm 0.036}$	$0.541_{\pm 0.035}$	$0.526_{\pm 0.034}$	$0.469_{\pm 0.035}$	$0.563_{\pm 0.038}$	$0.497_{\pm 0.035}$
Arterial wall calcification	$0.526_{\pm 0.023}$	$0.532_{\pm 0.024}$	$0.517_{\pm 0.025}$	$0.490_{\pm 0.024}$	$0.512_{\pm 0.024}$	$0.464_{\pm 0.021}$
Cardiomegaly	$0.573_{\pm 0.037}$	$0.580_{\pm 0.037}$	$0.538_{\pm 0.038}$	$0.563_{\pm 0.038}$	$0.549_{\pm 0.037}$	$0.549_{\pm 0.031}$
Pericardial effusion	$0.515_{\pm 0.046}$	$0.528_{\pm 0.045}$	$0.501_{\pm 0.044}$	$0.522_{\pm 0.042}$	$0.502_{\pm 0.044}$	$0.530_{\pm 0.041}$
Coronary artery wall calcification	$0.546_{\pm 0.022}$	$0.567_{\pm 0.021}$	$0.501_{\pm 0.025}$	$0.475_{\pm 0.025}$	$0.545_{\pm 0.023}$	$0.485_{\pm 0.023}$
Hiatal hernia	$0.503_{\pm 0.032}$	$0.503_{\pm 0.033}$	$0.514_{\pm 0.033}$	$0.509_{\pm 0.031}$	$0.511_{\pm 0.032}$	$0.487_{\pm 0.031}$
Lymphadenopathy	$0.563_{\pm 0.025}$	$0.572_{\pm 0.024}$	$0.547_{\pm 0.025}$	$0.542_{\pm 0.024}$	$0.543_{\pm 0.025}$	$0.516_{\pm 0.024}$
Emphysema	$0.516_{\pm 0.027}$	$0.507_{\pm 0.027}$	$0.491_{\pm 0.027}$	$0.480_{\pm 0.027}$	$0.479_{\pm 0.025}$	$0.485_{\pm 0.026}$
Atelectasis	$0.458_{\pm 0.024}$	$0.460_{\pm 0.024}$	$0.450_{\pm 0.025}$	$0.445_{\pm 0.025}$	$0.457_{\pm 0.025}$	$0.467_{\pm 0.025}$
Lung nodule	$0.458_{\pm 0.020}$	$0.452_{\pm 0.020}$	$0.447_{\pm 0.021}$	$0.458_{\pm 0.020}$	$0.448_{\pm 0.021}$	$0.455_{\pm 0.019}$
Lung opacity	$0.636_{\pm 0.022}$	$0.634_{\pm 0.021}$	$0.617_{\pm 0.023}$	$0.558_{\pm 0.021}$	$0.605_{\pm 0.022}$	$0.578_{\pm 0.021}$
Pulmonary fibrotic sequela	$0.514_{\pm 0.024}$	$0.510_{\pm 0.023}$	$0.511_{\pm 0.023}$	$0.500_{\pm 0.023}$	$0.517_{\pm 0.024}$	$0.501_{\pm 0.023}$
Pleural effusion	$0.593_{\pm 0.032}$	$0.604_{\pm 0.033}$	$0.597_{\pm 0.034}$	$0.574_{\pm 0.030}$	$0.557_{\pm 0.034}$	$0.576_{\pm 0.034}$
Mosaic attenuation pattern	$0.461_{\pm 0.034}$	$0.469_{\pm 0.034}$	$0.483_{\pm 0.035}$	$0.481_{\pm 0.035}$	$0.471_{\pm 0.034}$	$0.485_{\pm 0.035}$
Peribronchial thickening	$0.458_{\pm 0.036}$	$0.465_{\pm 0.035}$	$0.438_{\pm 0.036}$	$0.412_{\pm 0.033}$	$0.455_{\pm 0.037}$	$0.452_{\pm 0.035}$
Consolidation	$0.654_{\pm 0.028}$	$0.641_{\pm 0.028}$	$0.573_{\pm 0.026}$	$0.506_{\pm 0.026}$	$0.609_{\pm 0.027}$	$0.563_{\pm 0.026}$
Bronchiectasis	$0.578_{\pm 0.033}$	$0.576_{\pm 0.033}$	$0.519_{\pm 0.035}$	$0.559_{\pm 0.032}$	$0.502_{\pm 0.034}$	$0.557_{\pm 0.033}$
Interlobular septal thickening	$0.562_{\pm 0.044}$	$0.571_{\pm 0.044}$	$0.568_{\pm 0.044}$	$0.499_{\pm 0.042}$	$0.575_{\pm 0.045}$	$0.489_{\pm 0.040}$
Macro	$\mathbf{0.536}_{\pm 0.011}$	$\mathbf{0.540}_{\pm 0.011}$	$\mathbf{0.519}_{\pm 0.012}$	$\mathbf{0.502}_{\pm 0.010}$	$\mathbf{0.522}_{\pm 0.011}$	$\mathbf{0.508}_{\pm 0.010}$

Table 17: **Per-abnormality 1-bit results on CT-RATE for MedGemma 4B (N=3,039).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Abnormality	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
Medical material	—	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.56_{\pm 0.27}$	$0.00_{\pm 0.07}$	$1.45_{\pm 0.43}$
Arterial wall calcification	—	$0.03_{\pm 0.06}$	$0.07_{\pm 0.09}$	$0.69_{\pm 0.29}$	$0.26_{\pm 0.18}$	$17.34_{\pm 1.35}$
Cardiomegaly	—	$0.13_{\pm 0.13}$	$0.16_{\pm 0.14}$	$6.75_{\pm 0.89}$	$1.09_{\pm 0.37}$	$12.77_{\pm 1.19}$
Pericardial effusion	—	$0.03_{\pm 0.06}$	$0.00_{\pm 0.07}$	$0.23_{\pm 0.17}$	$0.20_{\pm 0.16}$	$15.60_{\pm 1.29}$
Coronary artery wall calcification	—	$0.43_{\pm 0.23}$	$0.39_{\pm 0.22}$	$13.16_{\pm 1.20}$	$0.76_{\pm 0.31}$	$25.77_{\pm 1.56}$
Hiatal hernia	—	$0.72_{\pm 0.30}$	$1.41_{\pm 0.42}$	$4.51_{\pm 0.74}$	$29.09_{\pm 1.61}$	$2.60_{\pm 0.57}$
Lymphadenopathy	—	$1.58_{\pm 0.44}$	$2.30_{\pm 0.53}$	$6.91_{\pm 0.90}$	$52.95_{\pm 1.77}$	$10.79_{\pm 1.10}$
Emphysema	—	$11.22_{\pm 1.12}$	$20.70_{\pm 1.44}$	$26.79_{\pm 1.57}$	$30.14_{\pm 1.63}$	$37.18_{\pm 1.72}$
Atelectasis	—	$9.35_{\pm 1.04}$	$25.63_{\pm 1.55}$	$27.34_{\pm 1.58}$	$11.68_{\pm 1.14}$	$46.69_{\pm 1.77}$
Lung nodule	—	$0.79_{\pm 0.31}$	$6.48_{\pm 0.88}$	$1.41_{\pm 0.42}$	$18.62_{\pm 1.38}$	$4.24_{\pm 0.72}$
Lung opacity	—	$3.98_{\pm 0.70}$	$31.36_{\pm 1.65}$	$5.56_{\pm 0.81}$	$66.17_{\pm 1.68}$	$4.87_{\pm 0.77}$
Pulmonary fibrotic sequela	—	$3.59_{\pm 0.66}$	$4.41_{\pm 0.73}$	$30.57_{\pm 1.64}$	$4.41_{\pm 0.73}$	$59.10_{\pm 1.75}$
Pleural effusion	—	$0.03_{\pm 0.06}$	$0.03_{\pm 0.06}$	$0.56_{\pm 0.27}$	$0.10_{\pm 0.11}$	$10.53_{\pm 1.09}$
Mosaic attenuation pattern	—	$8.75_{\pm 1.00}$	$17.37_{\pm 1.35}$	$36.82_{\pm 1.71}$	$20.30_{\pm 1.43}$	$43.07_{\pm 1.76}$
Peribronchial thickening	—	$0.03_{\pm 0.06}$	$0.56_{\pm 0.27}$	$0.26_{\pm 0.18}$	$6.98_{\pm 0.91}$	$8.52_{\pm 0.99}$
Consolidation	—	$8.46_{\pm 0.99}$	$15.17_{\pm 1.28}$	$21.88_{\pm 1.47}$	$19.35_{\pm 1.40}$	$42.55_{\pm 1.76}$
Bronchiectasis	—	$3.88_{\pm 0.69}$	$5.59_{\pm 0.82}$	$24.51_{\pm 1.53}$	$4.74_{\pm 0.76}$	$59.07_{\pm 1.75}$
Interlobular septal thickening	—	$0.00_{\pm 0.07}$	$0.23_{\pm 0.17}$	$0.00_{\pm 0.07}$	$4.80_{\pm 0.76}$	$16.85_{\pm 1.33}$
Macro	—	$\mathbf{2.95}_{\pm 0.14}$	$\mathbf{7.33}_{\pm 0.22}$	$\mathbf{11.58}_{\pm 0.27}$	$\mathbf{15.09}_{\pm 0.30}$	$\mathbf{23.28}_{\pm 0.35}$
AUROC						
Medical material	$0.550_{\pm 0.034}$	$0.545_{\pm 0.035}$	$0.546_{\pm 0.035}$	$0.520_{\pm 0.036}$	$0.522_{\pm 0.035}$	$0.533_{\pm 0.036}$
Arterial wall calcification	$0.459_{\pm 0.023}$	$0.459_{\pm 0.023}$	$0.461_{\pm 0.023}$	$0.456_{\pm 0.023}$	$0.435_{\pm 0.023}$	$0.499_{\pm 0.025}$
Cardiomegaly	$0.528_{\pm 0.034}$	$0.522_{\pm 0.034}$	$0.515_{\pm 0.035}$	$0.541_{\pm 0.034}$	$0.477_{\pm 0.032}$	$0.493_{\pm 0.033}$
Pericardial effusion	$0.466_{\pm 0.039}$	$0.469_{\pm 0.039}$	$0.446_{\pm 0.035}$	$0.478_{\pm 0.039}$	$0.445_{\pm 0.036}$	$0.491_{\pm 0.041}$
Coronary artery wall calcification	$0.435_{\pm 0.024}$	$0.434_{\pm 0.024}$	$0.437_{\pm 0.024}$	$0.430_{\pm 0.024}$	$0.408_{\pm 0.024}$	$0.478_{\pm 0.025}$
Hiatal hernia	$0.492_{\pm 0.031}$	$0.492_{\pm 0.030}$	$0.471_{\pm 0.029}$	$0.506_{\pm 0.030}$	$0.458_{\pm 0.027}$	$0.514_{\pm 0.030}$
Lymphadenopathy	$0.501_{\pm 0.024}$	$0.500_{\pm 0.024}$	$0.501_{\pm 0.025}$	$0.495_{\pm 0.025}$	$0.473_{\pm 0.024}$	$0.493_{\pm 0.023}$
Emphysema	$0.470_{\pm 0.027}$	$0.466_{\pm 0.027}$	$0.459_{\pm 0.027}$	$0.485_{\pm 0.026}$	$0.451_{\pm 0.026}$	$0.486_{\pm 0.027}$
Atelectasis	$0.466_{\pm 0.025}$	$0.460_{\pm 0.023}$	$0.438_{\pm 0.024}$	$0.469_{\pm 0.024}$	$0.419_{\pm 0.023}$	$0.472_{\pm 0.023}$
Lung nodule	$0.448_{\pm 0.021}$	$0.447_{\pm 0.021}$	$0.443_{\pm 0.021}$	$0.465_{\pm 0.020}$	$0.448_{\pm 0.020}$	$0.467_{\pm 0.020}$
Lung opacity	$0.554_{\pm 0.021}$	$0.551_{\pm 0.021}$	$0.565_{\pm 0.021}$	$0.559_{\pm 0.023}$	$0.519_{\pm 0.021}$	$0.548_{\pm 0.021}$
Pulmonary fibrotic sequela	$0.489_{\pm 0.023}$	$0.487_{\pm 0.022}$	$0.492_{\pm 0.024}$	$0.482_{\pm 0.022}$	$0.481_{\pm 0.022}$	$0.490_{\pm 0.024}$
Pleural effusion	$0.519_{\pm 0.035}$	$0.504_{\pm 0.034}$	$0.483_{\pm 0.033}$	$0.502_{\pm 0.031}$	$0.459_{\pm 0.032}$	$0.511_{\pm 0.034}$
Mosaic attenuation pattern	$0.502_{\pm 0.038}$	$0.491_{\pm 0.038}$	$0.485_{\pm 0.038}$	$0.487_{\pm 0.040}$	$0.472_{\pm 0.035}$	$0.511_{\pm 0.038}$
Peribronchial thickening	$0.414_{\pm 0.037}$	$0.411_{\pm 0.037}$	$0.401_{\pm 0.034}$	$0.418_{\pm 0.034}$	$0.390_{\pm 0.034}$	$0.401_{\pm 0.035}$
Consolidation	$0.541_{\pm 0.027}$	$0.538_{\pm 0.026}$	$0.552_{\pm 0.027}$	$0.529_{\pm 0.027}$	$0.502_{\pm 0.028}$	$0.533_{\pm 0.026}$
Bronchiectasis	$0.481_{\pm 0.035}$	$0.472_{\pm 0.037}$	$0.456_{\pm 0.036}$	$0.458_{\pm 0.033}$	$0.437_{\pm 0.033}$	$0.471_{\pm 0.037}$
Interlobular septal thickening	$0.485_{\pm 0.040}$	$0.479_{\pm 0.040}$	$0.474_{\pm 0.040}$	$0.492_{\pm 0.040}$	$0.456_{\pm 0.039}$	$0.470_{\pm 0.040}$
Macro	$\mathbf{0.489}_{\pm 0.011}$	$\mathbf{0.485}_{\pm 0.011}$	$\mathbf{0.479}_{\pm 0.011}$	$\mathbf{0.487}_{\pm 0.011}$	$\mathbf{0.458}_{\pm 0.010}$	$\mathbf{0.492}_{\pm 0.011}$

Table 18: **Per-abnormality 1-bit results on CT-RATE for MedGemma 27B (N=3,039).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Abnormality	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
Medical material	—	$0.03_{\pm 0.06}$	$0.07_{\pm 0.09}$	$3.39_{\pm 0.64}$	$0.33_{\pm 0.20}$	$3.59_{\pm 0.66}$
Arterial wall calcification	—	$0.43_{\pm 0.23}$	$12.47_{\pm 1.17}$	$19.45_{\pm 1.41}$	$18.33_{\pm 1.38}$	$16.42_{\pm 1.32}$
Cardiomegaly	—	$0.00_{\pm 0.07}$	$0.53_{\pm 0.26}$	$8.16_{\pm 0.97}$	$6.91_{\pm 0.90}$	$0.16_{\pm 0.14}$
Pericardial effusion	—	$10.86_{\pm 1.11}$	$22.28_{\pm 1.48}$	$24.25_{\pm 1.52}$	$24.25_{\pm 1.52}$	$21.36_{\pm 1.46}$
Coronary artery wall calcification	—	$3.13_{\pm 0.62}$	$5.20_{\pm 0.79}$	$5.20_{\pm 0.79}$	$5.20_{\pm 0.79}$	$5.20_{\pm 0.79}$
Hiatal hernia	—	$0.13_{\pm 0.13}$	$1.84_{\pm 0.48}$	$14.22_{\pm 1.24}$	$9.64_{\pm 1.05}$	$7.04_{\pm 0.91}$
Lymphadenopathy	—	$0.89_{\pm 0.33}$	$6.09_{\pm 0.85}$	$18.10_{\pm 1.37}$	$22.51_{\pm 1.48}$	$4.05_{\pm 0.70}$
Emphysema	—	$7.17_{\pm 0.92}$	$42.58_{\pm 1.76}$	$12.41_{\pm 1.17}$	$64.82_{\pm 1.70}$	$10.37_{\pm 1.08}$
Atelectasis	—	$2.07_{\pm 0.51}$	$4.80_{\pm 0.76}$	$26.85_{\pm 1.58}$	$4.64_{\pm 0.75}$	$28.59_{\pm 1.61}$
Lung nodule	—	$12.57_{\pm 1.18}$	$17.54_{\pm 1.35}$	$48.21_{\pm 1.78}$	$22.77_{\pm 1.49}$	$24.51_{\pm 1.53}$
Lung opacity	—	$0.43_{\pm 0.23}$	$2.70_{\pm 0.58}$	$3.59_{\pm 0.66}$	$0.59_{\pm 0.27}$	$15.00_{\pm 1.27}$
Pulmonary fibrotic sequela	—	$0.10_{\pm 0.11}$	$0.82_{\pm 0.32}$	$74.73_{\pm 1.55}$	$0.07_{\pm 0.09}$	$40.34_{\pm 1.74}$
Pleural effusion	—	$1.71_{\pm 0.46}$	$11.91_{\pm 1.15}$	$5.33_{\pm 0.80}$	$5.33_{\pm 0.80}$	$5.30_{\pm 0.80}$
Mosaic attenuation pattern	—	$4.18_{\pm 0.71}$	$25.90_{\pm 1.56}$	$10.66_{\pm 1.10}$	$37.38_{\pm 1.72}$	$21.45_{\pm 1.46}$
Peribronchial thickening	—	$0.53_{\pm 0.26}$	$1.12_{\pm 0.37}$	$1.45_{\pm 0.43}$	$0.53_{\pm 0.26}$	$5.79_{\pm 0.83}$
Consolidation	—	$0.03_{\pm 0.06}$	$0.00_{\pm 0.07}$	$49.46_{\pm 1.78}$	$0.00_{\pm 0.07}$	$1.22_{\pm 0.39}$
Bronchiectasis	—	$0.00_{\pm 0.07}$	$0.23_{\pm 0.17}$	$19.38_{\pm 1.41}$	$22.70_{\pm 1.49}$	$0.92_{\pm 0.34}$
Interlobular septal thickening	—	$3.16_{\pm 0.62}$	$8.49_{\pm 0.99}$	$12.11_{\pm 1.16}$	$14.77_{\pm 1.26}$	$13.36_{\pm 1.21}$
Macro	—	$\mathbf{2.63}_{\pm 0.13}$	$\mathbf{9.14}_{\pm 0.24}$	$\mathbf{19.83}_{\pm 0.33}$	$\mathbf{14.49}_{\pm 0.29}$	$\mathbf{12.48}_{\pm 0.28}$
AUROC						
Medical material	$0.527_{\pm 0.036}$	$0.536_{\pm 0.036}$	$0.508_{\pm 0.035}$	$0.505_{\pm 0.034}$	$0.539_{\pm 0.036}$	$0.524_{\pm 0.036}$
Arterial wall calcification	$0.446_{\pm 0.023}$	$0.444_{\pm 0.023}$	$0.459_{\pm 0.023}$	$0.434_{\pm 0.022}$	$0.429_{\pm 0.023}$	$0.463_{\pm 0.022}$
Cardiomegaly	$0.523_{\pm 0.035}$	$0.521_{\pm 0.036}$	$0.515_{\pm 0.035}$	$0.489_{\pm 0.032}$	$0.498_{\pm 0.036}$	$0.514_{\pm 0.035}$
Pericardial effusion	$0.511_{\pm 0.040}$	$0.489_{\pm 0.038}$	$0.466_{\pm 0.041}$	$0.473_{\pm 0.043}$	$0.459_{\pm 0.039}$	$0.462_{\pm 0.042}$
Coronary artery wall calcification	$0.440_{\pm 0.024}$	$0.449_{\pm 0.025}$	$0.487_{\pm 0.022}$	$0.411_{\pm 0.024}$	$0.503_{\pm 0.025}$	$0.469_{\pm 0.024}$
Hiatal hernia	$0.500_{\pm 0.028}$	$0.499_{\pm 0.029}$	$0.493_{\pm 0.032}$	$0.471_{\pm 0.030}$	$0.496_{\pm 0.029}$	$0.493_{\pm 0.028}$
Lymphadenopathy	$0.483_{\pm 0.024}$	$0.482_{\pm 0.024}$	$0.483_{\pm 0.022}$	$0.469_{\pm 0.023}$	$0.484_{\pm 0.023}$	$0.492_{\pm 0.023}$
Emphysema	$0.476_{\pm 0.028}$	$0.469_{\pm 0.028}$	$0.488_{\pm 0.027}$	$0.463_{\pm 0.026}$	$0.461_{\pm 0.026}$	$0.467_{\pm 0.027}$
Atelectasis	$0.467_{\pm 0.025}$	$0.456_{\pm 0.025}$	$0.454_{\pm 0.025}$	$0.463_{\pm 0.024}$	$0.437_{\pm 0.024}$	$0.469_{\pm 0.024}$
Lung nodule	$0.456_{\pm 0.020}$	$0.458_{\pm 0.021}$	$0.455_{\pm 0.021}$	$0.456_{\pm 0.021}$	$0.447_{\pm 0.021}$	$0.452_{\pm 0.021}$
Lung opacity	$0.550_{\pm 0.022}$	$0.539_{\pm 0.021}$	$0.527_{\pm 0.021}$	$0.526_{\pm 0.022}$	$0.515_{\pm 0.021}$	$0.519_{\pm 0.022}$
Pulmonary fibrotic sequela	$0.501_{\pm 0.023}$	$0.503_{\pm 0.023}$	$0.518_{\pm 0.023}$	$0.494_{\pm 0.023}$	$0.492_{\pm 0.023}$	$0.509_{\pm 0.023}$
Pleural effusion	$0.500_{\pm 0.033}$	$0.484_{\pm 0.033}$	$0.485_{\pm 0.032}$	$0.504_{\pm 0.033}$	$0.459_{\pm 0.030}$	$0.515_{\pm 0.033}$
Mosaic attenuation pattern	$0.491_{\pm 0.036}$	$0.490_{\pm 0.036}$	$0.505_{\pm 0.039}$	$0.432_{\pm 0.034}$	$0.480_{\pm 0.036}$	$0.491_{\pm 0.037}$
Peribronchial thickening	$0.404_{\pm 0.034}$	$0.410_{\pm 0.034}$	$0.393_{\pm 0.032}$	$0.416_{\pm 0.034}$	$0.378_{\pm 0.031}$	$0.423_{\pm 0.036}$
Consolidation	$0.489_{\pm 0.026}$	$0.478_{\pm 0.026}$	$0.491_{\pm 0.026}$	$0.515_{\pm 0.027}$	$0.466_{\pm 0.027}$	$0.492_{\pm 0.026}$
Bronchiectasis	$0.469_{\pm 0.035}$	$0.455_{\pm 0.035}$	$0.446_{\pm 0.032}$	$0.442_{\pm 0.034}$	$0.430_{\pm 0.032}$	$0.452_{\pm 0.035}$
Interlobular septal thickening	$0.486_{\pm 0.042}$	$0.469_{\pm 0.040}$	$0.437_{\pm 0.038}$	$0.471_{\pm 0.040}$	$0.431_{\pm 0.040}$	$0.480_{\pm 0.039}$
Macro	$\mathbf{0.484}_{\pm 0.010}$	$\mathbf{0.480}_{\pm 0.011}$	$\mathbf{0.478}_{\pm 0.010}$	$\mathbf{0.469}_{\pm 0.010}$	$\mathbf{0.467}_{\pm 0.009}$	$\mathbf{0.483}_{\pm 0.010}$

Table 19: **Per-abnormality 1-bit results on CT-RATE for Hulu-Med 4B (N=3,039).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Abnormality	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
Medical material	—	$0.00_{\pm 0.07}$	$18.46_{\pm 1.38}$	$2.47_{\pm 0.55}$	$0.00_{\pm 0.07}$	$4.21_{\pm 0.71}$
Arterial wall calcification	—	$23.69_{\pm 1.51}$	$43.24_{\pm 1.76}$	$44.16_{\pm 1.77}$	$50.61_{\pm 1.78}$	$43.30_{\pm 1.76}$
Cardiomegaly	—	$8.06_{\pm 0.97}$	$69.23_{\pm 1.64}$	$78.87_{\pm 1.45}$	$67.49_{\pm 1.67}$	$77.76_{\pm 1.48}$
Pericardial effusion	—	$0.00_{\pm 0.07}$	$92.76_{\pm 0.92}$	$98.55_{\pm 0.43}$	$2.83_{\pm 0.59}$	$90.85_{\pm 1.03}$
Coronary artery wall calcification	—	$10.99_{\pm 1.11}$	$14.35_{\pm 1.25}$	$14.94_{\pm 1.27}$	$72.89_{\pm 1.58}$	$14.45_{\pm 1.25}$
Hiatal hernia	—	$37.51_{\pm 1.72}$	$54.13_{\pm 1.77}$	$60.84_{\pm 1.74}$	$61.07_{\pm 1.73}$	$60.45_{\pm 1.74}$
Lymphadenopathy	—	$1.58_{\pm 0.44}$	$8.69_{\pm 1.00}$	$3.06_{\pm 0.61}$	$3.82_{\pm 0.68}$	$4.64_{\pm 0.75}$
Emphysema	—	$0.00_{\pm 0.07}$	$12.60_{\pm 1.18}$	$0.92_{\pm 0.34}$	$0.03_{\pm 0.06}$	$2.93_{\pm 0.60}$
Atelectasis	—	$6.28_{\pm 0.86}$	$14.87_{\pm 1.26}$	$9.31_{\pm 1.03}$	$15.76_{\pm 1.30}$	$10.17_{\pm 1.07}$
Lung nodule	—	$1.18_{\pm 0.38}$	$12.21_{\pm 1.16}$	$4.18_{\pm 0.71}$	$6.84_{\pm 0.90}$	$7.04_{\pm 0.91}$
Lung opacity	—	$5.20_{\pm 0.79}$	$12.41_{\pm 1.17}$	$8.92_{\pm 1.01}$	$33.76_{\pm 1.68}$	$12.01_{\pm 1.16}$
Pulmonary fibrotic sequela	—	$9.15_{\pm 1.03}$	$18.56_{\pm 1.38}$	$16.75_{\pm 1.33}$	$19.81_{\pm 1.42}$	$17.01_{\pm 1.34}$
Pleural effusion	—	$11.15_{\pm 1.12}$	$71.44_{\pm 1.61}$	$82.86_{\pm 1.34}$	$18.20_{\pm 1.37}$	$75.39_{\pm 1.53}$
Mosaic attenuation pattern	—	$13.33_{\pm 1.21}$	$42.18_{\pm 1.76}$	$43.80_{\pm 1.76}$	$43.47_{\pm 1.76}$	$42.71_{\pm 1.76}$
Peribronchial thickening	—	$0.00_{\pm 0.07}$	$96.28_{\pm 0.67}$	$92.07_{\pm 0.96}$	$17.87_{\pm 1.36}$	$89.80_{\pm 1.08}$
Consolidation	—	$1.81_{\pm 0.47}$	$18.03_{\pm 1.37}$	$6.09_{\pm 0.85}$	$26.95_{\pm 1.58}$	$5.63_{\pm 0.82}$
Bronchiectasis	—	$0.00_{\pm 0.07}$	$5.26_{\pm 0.79}$	$0.46_{\pm 0.24}$	$2.80_{\pm 0.59}$	$2.11_{\pm 0.51}$
Interlobular septal thickening	—	$0.00_{\pm 0.07}$	$98.29_{\pm 0.46}$	$98.32_{\pm 0.46}$	$0.00_{\pm 0.07}$	$89.87_{\pm 1.07}$
Macro	—	$\mathbf{7.22}_{\pm 0.22}$	$\mathbf{39.06}_{\pm 0.41}$	$\mathbf{37.03}_{\pm 0.40}$	$\mathbf{24.68}_{\pm 0.36}$	$\mathbf{36.13}_{\pm 0.40}$
AUROC						
Medical material	$0.561_{\pm 0.036}$	$0.564_{\pm 0.037}$	$0.504_{\pm 0.034}$	$0.480_{\pm 0.035}$	$0.503_{\pm 0.034}$	$0.540_{\pm 0.034}$
Arterial wall calcification	$0.562_{\pm 0.022}$	$0.518_{\pm 0.024}$	$0.467_{\pm 0.023}$	$0.501_{\pm 0.023}$	$0.501_{\pm 0.024}$	$0.516_{\pm 0.023}$
Cardiomegaly	$0.546_{\pm 0.034}$	$0.503_{\pm 0.036}$	$0.491_{\pm 0.032}$	$0.511_{\pm 0.029}$	$0.534_{\pm 0.034}$	$0.500_{\pm 0.031}$
Pericardial effusion	$0.495_{\pm 0.043}$	$0.472_{\pm 0.043}$	$0.528_{\pm 0.038}$	$0.508_{\pm 0.040}$	$0.605_{\pm 0.039}$	$0.545_{\pm 0.039}$
Coronary artery wall calcification	$0.496_{\pm 0.023}$	$0.455_{\pm 0.024}$	$0.477_{\pm 0.024}$	$0.507_{\pm 0.023}$	$0.487_{\pm 0.024}$	$0.531_{\pm 0.023}$
Hiatal hernia	$0.512_{\pm 0.031}$	$0.492_{\pm 0.030}$	$0.486_{\pm 0.030}$	$0.496_{\pm 0.032}$	$0.514_{\pm 0.030}$	$0.514_{\pm 0.030}$
Lymphadenopathy	$0.563_{\pm 0.024}$	$0.557_{\pm 0.024}$	$0.508_{\pm 0.023}$	$0.533_{\pm 0.024}$	$0.504_{\pm 0.026}$	$0.505_{\pm 0.022}$
Emphysema	$0.652_{\pm 0.025}$	$0.637_{\pm 0.024}$	$0.501_{\pm 0.025}$	$0.515_{\pm 0.027}$	$0.481_{\pm 0.026}$	$0.529_{\pm 0.025}$
Atelectasis	$0.584_{\pm 0.023}$	$0.561_{\pm 0.024}$	$0.508_{\pm 0.026}$	$0.514_{\pm 0.023}$	$0.518_{\pm 0.026}$	$0.527_{\pm 0.025}$
Lung nodule	$0.456_{\pm 0.020}$	$0.443_{\pm 0.020}$	$0.498_{\pm 0.020}$	$0.521_{\pm 0.020}$	$0.531_{\pm 0.021}$	$0.519_{\pm 0.021}$
Lung opacity	$0.625_{\pm 0.021}$	$0.624_{\pm 0.020}$	$0.496_{\pm 0.021}$	$0.485_{\pm 0.021}$	$0.495_{\pm 0.020}$	$0.493_{\pm 0.021}$
Pulmonary fibrotic sequela	$0.545_{\pm 0.022}$	$0.540_{\pm 0.023}$	$0.496_{\pm 0.022}$	$0.511_{\pm 0.022}$	$0.485_{\pm 0.022}$	$0.497_{\pm 0.022}$
Pleural effusion	$0.710_{\pm 0.022}$	$0.702_{\pm 0.023}$	$0.495_{\pm 0.029}$	$0.545_{\pm 0.032}$	$0.605_{\pm 0.033}$	$0.549_{\pm 0.032}$
Mosaic attenuation pattern	$0.467_{\pm 0.038}$	$0.437_{\pm 0.037}$	$0.540_{\pm 0.034}$	$0.486_{\pm 0.039}$	$0.546_{\pm 0.037}$	$0.511_{\pm 0.039}$
Peribronchial thickening	$0.614_{\pm 0.036}$	$0.650_{\pm 0.034}$	$0.471_{\pm 0.033}$	$0.528_{\pm 0.034}$	$0.417_{\pm 0.034}$	$0.596_{\pm 0.031}$
Consolidation	$0.681_{\pm 0.025}$	$0.659_{\pm 0.027}$	$0.512_{\pm 0.027}$	$0.505_{\pm 0.025}$	$0.559_{\pm 0.026}$	$0.534_{\pm 0.027}$
Bronchiectasis	$0.589_{\pm 0.033}$	$0.580_{\pm 0.035}$	$0.506_{\pm 0.032}$	$0.532_{\pm 0.032}$	$0.543_{\pm 0.032}$	$0.522_{\pm 0.033}$
Interlobular septal thickening	$0.682_{\pm 0.030}$	$0.665_{\pm 0.031}$	$0.513_{\pm 0.035}$	$0.503_{\pm 0.036}$	$0.472_{\pm 0.037}$	$0.527_{\pm 0.035}$
Macro	$\mathbf{0.574}_{\pm 0.010}$	$\mathbf{0.559}_{\pm 0.009}$	$\mathbf{0.500}_{\pm 0.009}$	$\mathbf{0.510}_{\pm 0.009}$	$\mathbf{0.517}_{\pm 0.008}$	$\mathbf{0.525}_{\pm 0.009}$

Table 20: **Per-abnormality 1-bit results on CT-RATE for Hulu-Med 7B (N=3,039).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Abnormality	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
Medical material	—	$18.36_{\pm 1.38}$	$38.01_{\pm 1.73}$	$38.80_{\pm 1.73}$	$37.97_{\pm 1.73}$	$38.37_{\pm 1.73}$
Arterial wall calcification	—	$9.74_{\pm 1.05}$	$60.15_{\pm 1.74}$	$90.62_{\pm 1.04}$	$12.90_{\pm 1.19}$	$90.06_{\pm 1.06}$
Cardiomegaly	—	$0.99_{\pm 0.35}$	$85.36_{\pm 1.26}$	$84.40_{\pm 1.29}$	$21.85_{\pm 1.47}$	$88.25_{\pm 1.14}$
Pericardial effusion	—	$1.22_{\pm 0.39}$	$76.74_{\pm 1.50}$	$89.63_{\pm 1.08}$	$8.49_{\pm 0.99}$	$95.33_{\pm 0.75}$
Coronary artery wall calcification	—	$5.20_{\pm 0.79}$	$55.48_{\pm 1.77}$	$91.94_{\pm 0.97}$	$26.59_{\pm 1.57}$	$84.93_{\pm 1.27}$
Hiatal hernia	—	$3.62_{\pm 0.66}$	$40.54_{\pm 1.75}$	$86.67_{\pm 1.21}$	$70.02_{\pm 1.63}$	$79.57_{\pm 1.43}$
Lymphadenopathy	—	$7.86_{\pm 0.96}$	$18.72_{\pm 1.39}$	$16.45_{\pm 1.32}$	$58.90_{\pm 1.75}$	$23.72_{\pm 1.51}$
Emphysema	—	$21.29_{\pm 1.46}$	$37.64_{\pm 1.72}$	$34.19_{\pm 1.69}$	$51.60_{\pm 1.78}$	$34.58_{\pm 1.69}$
Atelectasis	—	$17.01_{\pm 1.34}$	$32.41_{\pm 1.66}$	$32.74_{\pm 1.67}$	$49.85_{\pm 1.78}$	$32.77_{\pm 1.67}$
Lung nodule	—	$25.34_{\pm 1.55}$	$59.95_{\pm 1.74}$	$28.66_{\pm 1.61}$	$68.31_{\pm 1.65}$	$27.77_{\pm 1.59}$
Lung opacity	—	$11.58_{\pm 1.14}$	$45.34_{\pm 1.77}$	$49.98_{\pm 1.78}$	$41.59_{\pm 1.75}$	$49.39_{\pm 1.78}$
Pulmonary fibrotic sequela	—	$20.27_{\pm 1.43}$	$40.14_{\pm 1.74}$	$40.38_{\pm 1.74}$	$40.67_{\pm 1.75}$	$40.41_{\pm 1.74}$
Pleural effusion	—	$15.24_{\pm 1.28}$	$66.70_{\pm 1.68}$	$30.83_{\pm 1.64}$	$71.17_{\pm 1.61}$	$24.35_{\pm 1.53}$
Mosaic attenuation pattern	—	$5.66_{\pm 0.82}$	$69.07_{\pm 1.64}$	$67.62_{\pm 1.66}$	$69.04_{\pm 1.64}$	$68.31_{\pm 1.65}$
Peribronchial thickening	—	$0.30_{\pm 0.19}$	$51.00_{\pm 1.78}$	$97.56_{\pm 0.55}$	$32.74_{\pm 1.67}$	$90.33_{\pm 1.05}$
Consolidation	—	$0.63_{\pm 0.28}$	$51.53_{\pm 1.78}$	$9.25_{\pm 1.03}$	$59.43_{\pm 1.75}$	$9.35_{\pm 1.04}$
Bronchiectasis	—	$22.74_{\pm 1.49}$	$39.59_{\pm 1.74}$	$39.32_{\pm 1.74}$	$30.14_{\pm 1.63}$	$39.65_{\pm 1.74}$
Interlobular septal thickening	—	$0.92_{\pm 0.34}$	$84.70_{\pm 1.28}$	$90.10_{\pm 1.06}$	$67.85_{\pm 1.66}$	$95.23_{\pm 0.76}$
Macro	—	$\mathbf{10.44}_{\pm 0.26}$	$\mathbf{52.95}_{\pm 0.42}$	$\mathbf{56.62}_{\pm 0.42}$	$\mathbf{45.51}_{\pm 0.42}$	$\mathbf{56.24}_{\pm 0.42}$
AUROC						
Medical material	$0.554_{\pm 0.035}$	$0.544_{\pm 0.035}$	$0.511_{\pm 0.034}$	$0.505_{\pm 0.033}$	$0.535_{\pm 0.032}$	$0.474_{\pm 0.033}$
Arterial wall calcification	$0.537_{\pm 0.025}$	$0.545_{\pm 0.023}$	$0.488_{\pm 0.025}$	$0.512_{\pm 0.024}$	$0.477_{\pm 0.021}$	$0.480_{\pm 0.023}$
Cardiomegaly	$0.707_{\pm 0.032}$	$0.723_{\pm 0.031}$	$0.524_{\pm 0.034}$	$0.502_{\pm 0.034}$	$0.525_{\pm 0.036}$	$0.480_{\pm 0.032}$
Pericardial effusion	$0.574_{\pm 0.044}$	$0.593_{\pm 0.043}$	$0.541_{\pm 0.038}$	$0.456_{\pm 0.040}$	$0.554_{\pm 0.035}$	$0.539_{\pm 0.039}$
Coronary artery wall calcification	$0.546_{\pm 0.026}$	$0.547_{\pm 0.025}$	$0.468_{\pm 0.025}$	$0.526_{\pm 0.023}$	$0.467_{\pm 0.024}$	$0.478_{\pm 0.023}$
Hiatal hernia	$0.521_{\pm 0.032}$	$0.508_{\pm 0.030}$	$0.492_{\pm 0.030}$	$0.490_{\pm 0.029}$	$0.509_{\pm 0.031}$	$0.507_{\pm 0.029}$
Lymphadenopathy	$0.595_{\pm 0.024}$	$0.580_{\pm 0.023}$	$0.496_{\pm 0.022}$	$0.514_{\pm 0.023}$	$0.526_{\pm 0.023}$	$0.534_{\pm 0.024}$
Emphysema	$0.534_{\pm 0.027}$	$0.533_{\pm 0.028}$	$0.480_{\pm 0.027}$	$0.521_{\pm 0.026}$	$0.472_{\pm 0.026}$	$0.493_{\pm 0.026}$
Atelectasis	$0.562_{\pm 0.025}$	$0.584_{\pm 0.024}$	$0.443_{\pm 0.024}$	$0.499_{\pm 0.024}$	$0.445_{\pm 0.025}$	$0.523_{\pm 0.025}$
Lung nodule	$0.496_{\pm 0.021}$	$0.472_{\pm 0.020}$	$0.456_{\pm 0.020}$	$0.505_{\pm 0.020}$	$0.462_{\pm 0.020}$	$0.504_{\pm 0.021}$
Lung opacity	$0.628_{\pm 0.022}$	$0.618_{\pm 0.022}$	$0.514_{\pm 0.022}$	$0.504_{\pm 0.021}$	$0.520_{\pm 0.022}$	$0.494_{\pm 0.021}$
Pulmonary fibrotic sequela	$0.524_{\pm 0.023}$	$0.534_{\pm 0.023}$	$0.511_{\pm 0.022}$	$0.494_{\pm 0.023}$	$0.502_{\pm 0.022}$	$0.517_{\pm 0.021}$
Pleural effusion	$0.782_{\pm 0.020}$	$0.770_{\pm 0.022}$	$0.476_{\pm 0.032}$	$0.456_{\pm 0.029}$	$0.573_{\pm 0.032}$	$0.503_{\pm 0.031}$
Mosaic attenuation pattern	$0.517_{\pm 0.038}$	$0.515_{\pm 0.040}$	$0.508_{\pm 0.035}$	$0.529_{\pm 0.036}$	$0.421_{\pm 0.033}$	$0.467_{\pm 0.038}$
Peribronchial thickening	$0.492_{\pm 0.035}$	$0.507_{\pm 0.034}$	$0.408_{\pm 0.033}$	$0.491_{\pm 0.032}$	$0.423_{\pm 0.035}$	$0.547_{\pm 0.031}$
Consolidation	$0.645_{\pm 0.022}$	$0.637_{\pm 0.024}$	$0.518_{\pm 0.026}$	$0.463_{\pm 0.026}$	$0.504_{\pm 0.027}$	$0.500_{\pm 0.026}$
Bronchiectasis	$0.525_{\pm 0.038}$	$0.515_{\pm 0.037}$	$0.451_{\pm 0.034}$	$0.512_{\pm 0.034}$	$0.461_{\pm 0.034}$	$0.519_{\pm 0.032}$
Interlobular septal thickening	$0.545_{\pm 0.041}$	$0.542_{\pm 0.043}$	$0.456_{\pm 0.037}$	$0.491_{\pm 0.037}$	$0.490_{\pm 0.042}$	$0.541_{\pm 0.037}$
Macro	$\mathbf{0.571}_{\pm 0.010}$	$\mathbf{0.570}_{\pm 0.010}$	$\mathbf{0.486}_{\pm 0.009}$	$\mathbf{0.498}_{\pm 0.009}$	$\mathbf{0.493}_{\pm 0.009}$	$\mathbf{0.506}_{\pm 0.009}$

Table 21: **Per-abnormality 1-bit results on CT-RATE for Hulu-Med 32B (N=3,039).** Top panel: macro prediction-flip rate (%) versus bf16. Bottom panel: AUROC. Each entry is the point estimate followed by the 95% CI half-width in subscript. The macro row of each panel is bolded.

Abnormality	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Pred-flip rate (%) vs bf16						
Medical material	—	$32.97_{\pm 1.67}$	$36.69_{\pm 1.71}$	$49.16_{\pm 1.78}$	$32.08_{\pm 1.66}$	$49.00_{\pm 1.78}$
Arterial wall calcification	—	$0.36_{\pm 0.21}$	$23.86_{\pm 1.52}$	$98.26_{\pm 0.46}$	$27.57_{\pm 1.59}$	$79.66_{\pm 1.43}$
Cardiomegaly	—	$13.92_{\pm 1.23}$	$29.94_{\pm 1.63}$	$27.74_{\pm 1.59}$	$49.75_{\pm 1.78}$	$28.00_{\pm 1.60}$
Pericardial effusion	—	$0.00_{\pm 0.07}$	$9.05_{\pm 1.02}$	$97.89_{\pm 0.51}$	$15.53_{\pm 1.29}$	$43.99_{\pm 1.76}$
Coronary artery wall calcification	—	$11.94_{\pm 1.15}$	$24.88_{\pm 1.54}$	$71.21_{\pm 1.61}$	$20.63_{\pm 1.44}$	$62.78_{\pm 1.72}$
Hiatal hernia	—	$0.72_{\pm 0.30}$	$3.36_{\pm 0.64}$	$71.60_{\pm 1.60}$	$11.94_{\pm 1.15}$	$56.96_{\pm 1.76}$
Lymphadenopathy	—	$5.36_{\pm 0.80}$	$40.44_{\pm 1.74}$	$72.16_{\pm 1.59}$	$48.60_{\pm 1.78}$	$67.59_{\pm 1.66}$
Emphysema	—	$3.19_{\pm 0.62}$	$20.66_{\pm 1.44}$	$68.48_{\pm 1.65}$	$46.20_{\pm 1.77}$	$62.49_{\pm 1.72}$
Atelectasis	—	$17.97_{\pm 1.37}$	$43.76_{\pm 1.76}$	$26.82_{\pm 1.58}$	$51.66_{\pm 1.78}$	$30.54_{\pm 1.64}$
Lung nodule	—	$39.39_{\pm 1.74}$	$41.03_{\pm 1.75}$	$50.74_{\pm 1.78}$	$55.71_{\pm 1.77}$	$46.03_{\pm 1.77}$
Lung opacity	—	$17.87_{\pm 1.36}$	$30.44_{\pm 1.64}$	$44.88_{\pm 1.77}$	$31.16_{\pm 1.65}$	$43.83_{\pm 1.76}$
Pulmonary fibrotic sequela	—	$0.30_{\pm 0.19}$	$5.76_{\pm 0.83}$	$94.24_{\pm 0.83}$	$20.90_{\pm 1.45}$	$66.83_{\pm 1.67}$
Pleural effusion	—	$16.72_{\pm 1.33}$	$41.82_{\pm 1.75}$	$72.92_{\pm 1.58}$	$53.21_{\pm 1.77}$	$63.64_{\pm 1.71}$
Mosaic attenuation pattern	—	$1.32_{\pm 0.41}$	$5.73_{\pm 0.83}$	$68.34_{\pm 1.65}$	$11.65_{\pm 1.14}$	$68.90_{\pm 1.65}$
Peribronchial thickening	—	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$97.27_{\pm 0.58}$	$0.56_{\pm 0.27}$	$29.88_{\pm 1.63}$
Consolidation	—	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.79_{\pm 0.31}$	$0.00_{\pm 0.07}$	$3.32_{\pm 0.64}$
Bronchiectasis	—	$18.76_{\pm 1.39}$	$28.79_{\pm 1.61}$	$61.30_{\pm 1.73}$	$59.26_{\pm 1.75}$	$58.28_{\pm 1.75}$
Interlobular septal thickening	—	$0.00_{\pm 0.07}$	$14.05_{\pm 1.24}$	$97.76_{\pm 0.53}$	$18.03_{\pm 1.37}$	$46.69_{\pm 1.77}$
Macro	—	$\mathbf{10.04}_{\pm 0.25}$	$\mathbf{22.24}_{\pm 0.35}$	$\mathbf{65.09}_{\pm 0.40}$	$\mathbf{30.80}_{\pm 0.39}$	$\mathbf{50.47}_{\pm 0.42}$
AUROC						
Medical material	$0.575_{\pm 0.037}$	$0.566_{\pm 0.035}$	$0.500_{\pm 0.034}$	$0.502_{\pm 0.033}$	$0.467_{\pm 0.032}$	$0.528_{\pm 0.034}$
Arterial wall calcification	$0.474_{\pm 0.024}$	$0.431_{\pm 0.023}$	$0.410_{\pm 0.022}$	$0.501_{\pm 0.023}$	$0.405_{\pm 0.022}$	$0.504_{\pm 0.023}$
Cardiomegaly	$0.745_{\pm 0.026}$	$0.651_{\pm 0.030}$	$0.594_{\pm 0.032}$	$0.490_{\pm 0.035}$	$0.425_{\pm 0.030}$	$0.539_{\pm 0.035}$
Pericardial effusion	$0.548_{\pm 0.045}$	$0.517_{\pm 0.044}$	$0.467_{\pm 0.043}$	$0.495_{\pm 0.041}$	$0.406_{\pm 0.039}$	$0.560_{\pm 0.040}$
Coronary artery wall calcification	$0.504_{\pm 0.026}$	$0.456_{\pm 0.025}$	$0.422_{\pm 0.024}$	$0.479_{\pm 0.024}$	$0.383_{\pm 0.024}$	$0.511_{\pm 0.024}$
Hiatal hernia	$0.508_{\pm 0.030}$	$0.495_{\pm 0.032}$	$0.471_{\pm 0.029}$	$0.486_{\pm 0.029}$	$0.464_{\pm 0.026}$	$0.496_{\pm 0.031}$
Lymphadenopathy	$0.494_{\pm 0.023}$	$0.465_{\pm 0.023}$	$0.500_{\pm 0.024}$	$0.508_{\pm 0.023}$	$0.458_{\pm 0.023}$	$0.521_{\pm 0.023}$
Emphysema	$0.469_{\pm 0.027}$	$0.468_{\pm 0.027}$	$0.474_{\pm 0.027}$	$0.474_{\pm 0.025}$	$0.465_{\pm 0.027}$	$0.496_{\pm 0.025}$
Atelectasis	$0.500_{\pm 0.026}$	$0.505_{\pm 0.027}$	$0.457_{\pm 0.023}$	$0.484_{\pm 0.026}$	$0.415_{\pm 0.024}$	$0.490_{\pm 0.025}$
Lung nodule	$0.470_{\pm 0.019}$	$0.494_{\pm 0.021}$	$0.478_{\pm 0.021}$	$0.476_{\pm 0.021}$	$0.467_{\pm 0.021}$	$0.470_{\pm 0.020}$
Lung opacity	$0.421_{\pm 0.021}$	$0.469_{\pm 0.021}$	$0.544_{\pm 0.021}$	$0.497_{\pm 0.021}$	$0.505_{\pm 0.022}$	$0.532_{\pm 0.021}$
Pulmonary fibrotic sequela	$0.485_{\pm 0.024}$	$0.487_{\pm 0.022}$	$0.473_{\pm 0.023}$	$0.487_{\pm 0.023}$	$0.487_{\pm 0.021}$	$0.515_{\pm 0.022}$
Pleural effusion	$0.707_{\pm 0.029}$	$0.700_{\pm 0.027}$	$0.635_{\pm 0.031}$	$0.497_{\pm 0.030}$	$0.521_{\pm 0.032}$	$0.553_{\pm 0.032}$
Mosaic attenuation pattern	$0.487_{\pm 0.036}$	$0.473_{\pm 0.036}$	$0.427_{\pm 0.036}$	$0.517_{\pm 0.036}$	$0.378_{\pm 0.030}$	$0.451_{\pm 0.035}$
Peribronchial thickening	$0.443_{\pm 0.036}$	$0.425_{\pm 0.036}$	$0.397_{\pm 0.033}$	$0.482_{\pm 0.032}$	$0.387_{\pm 0.027}$	$0.478_{\pm 0.033}$
Consolidation	$0.596_{\pm 0.025}$	$0.548_{\pm 0.026}$	$0.590_{\pm 0.027}$	$0.520_{\pm 0.027}$	$0.574_{\pm 0.026}$	$0.528_{\pm 0.027}$
Bronchiectasis	$0.460_{\pm 0.035}$	$0.450_{\pm 0.032}$	$0.454_{\pm 0.034}$	$0.486_{\pm 0.034}$	$0.503_{\pm 0.036}$	$0.464_{\pm 0.034}$
Interlobular septal thickening	$0.528_{\pm 0.043}$	$0.495_{\pm 0.043}$	$0.433_{\pm 0.038}$	$0.463_{\pm 0.037}$	$0.399_{\pm 0.037}$	$0.508_{\pm 0.039}$
Macro	$\mathbf{0.523}_{\pm 0.011}$	$\mathbf{0.505}_{\pm 0.010}$	$\mathbf{0.485}_{\pm 0.010}$	$\mathbf{0.491}_{\pm 0.008}$	$\mathbf{0.450}_{\pm 0.010}$	$\mathbf{0.508}_{\pm 0.010}$

D.3 Video-MME

Tables 22–25 report per-task-type accuracy at 1-bit on Video-MME (no-subtitle) for the four general-purpose VLMs. The twelve task types follow the official Video-MME taxonomy [13], covering temporal, spatial, attribute, action, object, and OCR perception, counting, four reasoning categories, and information synopsis. The Overall row at the bottom of each table reports per-question accuracy across all 2,700 questions and reproduces the value listed in the corresponding column of Table 5.

Table 22: **Per-task-type 1-bit accuracy (%) on Video-MME (no-subtitle) for Qwen2.5-VL-7B (N=2,700).** Each entry is the point estimate followed by the 95% CI half-width in subscript. The Overall row is bolded.

Task type	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Temporal Perception	$65.45_{\pm 12.20}$	$60.00_{\pm 12.53}$	$12.73_{\pm 8.86}$	$16.36_{\pm 9.70}$	$25.45_{\pm 11.24}$	$36.36_{\pm 12.32}$
Spatial Perception	$64.81_{\pm 12.35}$	$62.96_{\pm 12.47}$	$24.07_{\pm 11.16}$	$12.96_{\pm 9.00}$	$29.63_{\pm 11.84}$	$44.44_{\pm 12.81}$
Attribute Perception	$74.32_{\pm 5.71}$	$70.27_{\pm 5.98}$	$24.77_{\pm 5.65}$	$12.61_{\pm 4.38}$	$18.47_{\pm 5.08}$	$45.50_{\pm 6.50}$
Action Recognition	$56.87_{\pm 5.46}$	$54.63_{\pm 5.48}$	$19.49_{\pm 4.38}$	$15.34_{\pm 3.99}$	$20.45_{\pm 4.46}$	$37.70_{\pm 5.34}$
Object Recognition	$64.41_{\pm 4.96}$	$60.73_{\pm 5.06}$	$24.58_{\pm 4.47}$	$15.54_{\pm 3.77}$	$19.21_{\pm 4.09}$	$38.98_{\pm 5.05}$
OCR Problems	$66.19_{\pm 7.77}$	$60.43_{\pm 8.03}$	$22.30_{\pm 6.87}$	$10.79_{\pm 5.19}$	$22.30_{\pm 6.87}$	$39.57_{\pm 8.02}$
Counting Problem	$39.18_{\pm 5.80}$	$36.57_{\pm 5.73}$	$19.03_{\pm 4.68}$	$8.96_{\pm 3.45}$	$17.16_{\pm 4.50}$	$26.49_{\pm 5.25}$
Temporal Reasoning	$43.50_{\pm 7.23}$	$36.72_{\pm 7.04}$	$14.69_{\pm 5.21}$	$6.78_{\pm 3.78}$	$19.77_{\pm 5.84}$	$25.42_{\pm 6.37}$
Spatial Reasoning	$78.57_{\pm 10.55}$	$76.79_{\pm 10.84}$	$32.14_{\pm 11.89}$	$16.07_{\pm 9.56}$	$19.64_{\pm 10.25}$	$67.86_{\pm 11.89}$
Action Reasoning	$55.79_{\pm 5.73}$	$54.04_{\pm 5.75}$	$23.16_{\pm 4.88}$	$11.58_{\pm 3.73}$	$21.40_{\pm 4.75}$	$42.46_{\pm 5.70}$
Object Reasoning	$56.39_{\pm 4.55}$	$55.73_{\pm 4.55}$	$22.47_{\pm 3.83}$	$10.57_{\pm 2.83}$	$16.08_{\pm 3.37}$	$36.34_{\pm 4.41}$
Information Synopsis	$75.85_{\pm 4.65}$	$73.37_{\pm 4.80}$	$32.51_{\pm 5.08}$	$11.46_{\pm 3.48}$	$22.29_{\pm 4.53}$	$52.32_{\pm 5.41}$
Overall	$\mathbf{60.00}_{\pm 1.85}$	$\mathbf{57.15}_{\pm 1.87}$	$\mathbf{23.04}_{\pm 1.59}$	$\mathbf{12.04}_{\pm 1.23}$	$\mathbf{19.70}_{\pm 1.50}$	$\mathbf{39.44}_{\pm 1.84}$

Table 23: **Per-task-type 1-bit accuracy (%) on Video-MME (no-subtitle) for Qwen3-VL-8B (N=2,700).** Each entry is the point estimate followed by the 95% CI half-width in subscript. The Overall row is bolded.

Task type	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Temporal Perception	$78.18_{\pm 10.71}$	$74.55_{\pm 11.24}$	$12.73_{\pm 8.86}$	$3.64_{\pm 5.66}$	$23.64_{\pm 10.99}$	$9.09_{\pm 7.81}$
Spatial Perception	$77.78_{\pm 10.87}$	$74.07_{\pm 11.40}$	$11.11_{\pm 8.50}$	$5.56_{\pm 6.60}$	$46.30_{\pm 12.85}$	$3.70_{\pm 5.76}$
Attribute Perception	$83.78_{\pm 4.84}$	$78.38_{\pm 5.39}$	$18.92_{\pm 5.13}$	$3.15_{\pm 2.41}$	$33.33_{\pm 6.16}$	$5.86_{\pm 3.15}$
Action Recognition	$71.25_{\pm 4.99}$	$67.09_{\pm 5.18}$	$16.29_{\pm 4.08}$	$2.56_{\pm 1.83}$	$30.03_{\pm 5.05}$	$5.43_{\pm 2.55}$
Object Recognition	$76.27_{\pm 4.42}$	$71.75_{\pm 4.67}$	$17.80_{\pm 3.98}$	$2.26_{\pm 1.63}$	$34.46_{\pm 4.93}$	$10.45_{\pm 3.20}$
OCR Problems	$77.70_{\pm 6.86}$	$72.66_{\pm 7.34}$	$12.23_{\pm 5.46}$	$0.72_{\pm 1.92}$	$34.53_{\pm 7.81}$	$6.47_{\pm 4.21}$
Counting Problem	$51.49_{\pm 5.94}$	$48.88_{\pm 5.95}$	$11.94_{\pm 3.89}$	$4.85_{\pm 2.63}$	$19.40_{\pm 4.72}$	$8.58_{\pm 3.38}$
Temporal Reasoning	$67.80_{\pm 6.82}$	$55.93_{\pm 7.23}$	$5.08_{\pm 3.34}$	$5.08_{\pm 3.34}$	$10.73_{\pm 4.59}$	$5.08_{\pm 3.34}$
Spatial Reasoning	$80.36_{\pm 10.25}$	$75.00_{\pm 11.09}$	$25.00_{\pm 11.08}$	$3.57_{\pm 5.57}$	$60.71_{\pm 12.39}$	$10.71_{\pm 8.23}$
Action Reasoning	$64.21_{\pm 5.53}$	$60.35_{\pm 5.64}$	$11.23_{\pm 3.67}$	$4.91_{\pm 2.56}$	$15.79_{\pm 4.23}$	$7.37_{\pm 3.06}$
Object Reasoning	$69.60_{\pm 4.22}$	$65.20_{\pm 4.36}$	$13.22_{\pm 3.12}$	$3.08_{\pm 1.63}$	$21.81_{\pm 3.79}$	$6.61_{\pm 2.30}$
Information Synopsis	$86.38_{\pm 3.74}$	$81.73_{\pm 4.20}$	$16.72_{\pm 4.06}$	$3.72_{\pm 2.12}$	$29.10_{\pm 4.93}$	$8.36_{\pm 3.04}$
Overall	$\mathbf{72.33}_{\pm 1.69}$	$\mathbf{67.56}_{\pm 1.77}$	$\mathbf{14.33}_{\pm 1.32}$	$\mathbf{3.44}_{\pm 0.69}$	$\mathbf{26.63}_{\pm 1.67}$	$\mathbf{7.37}_{\pm 0.99}$

Table 24: **Per-task-type 1-bit accuracy (%) on Video-MME (no-subtitle) for LLaVA-OneVision-7B (N=2,700).** Each entry is the point estimate followed by the 95% CI half-width in subscript. The Overall row is bolded.

Task type	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Temporal Perception	$61.82_{\pm 12.44}$	$54.55_{\pm 12.72}$	$25.45_{\pm 11.24}$	$9.09_{\pm 7.81}$	$32.73_{\pm 12.04}$	$18.18_{\pm 10.07}$
Spatial Perception	$61.11_{\pm 12.58}$	$51.85_{\pm 12.88}$	$53.70_{\pm 12.85}$	$11.11_{\pm 8.50}$	$55.56_{\pm 12.81}$	$29.63_{\pm 11.84}$
Attribute Perception	$75.68_{\pm 5.61}$	$69.82_{\pm 6.00}$	$50.90_{\pm 6.52}$	$9.46_{\pm 3.88}$	$39.64_{\pm 6.39}$	$22.07_{\pm 5.43}$
Action Recognition	$54.95_{\pm 5.48}$	$54.31_{\pm 5.48}$	$37.38_{\pm 5.33}$	$7.99_{\pm 3.02}$	$37.06_{\pm 5.32}$	$19.17_{\pm 4.35}$
Object Recognition	$66.95_{\pm 4.88}$	$58.76_{\pm 5.10}$	$47.46_{\pm 5.17}$	$10.17_{\pm 3.16}$	$42.66_{\pm 5.12}$	$20.62_{\pm 4.21}$
OCR Problems	$60.43_{\pm 8.03}$	$54.68_{\pm 8.16}$	$37.41_{\pm 7.94}$	$5.76_{\pm 4.00}$	$42.45_{\pm 8.11}$	$18.71_{\pm 6.45}$
Counting Problem	$37.69_{\pm 5.76}$	$35.07_{\pm 5.68}$	$29.10_{\pm 5.41}$	$4.48_{\pm 2.54}$	$28.73_{\pm 5.39}$	$11.57_{\pm 3.84}$
Temporal Reasoning	$42.94_{\pm 7.21}$	$42.37_{\pm 7.21}$	$21.47_{\pm 6.02}$	$7.34_{\pm 3.91}$	$24.29_{\pm 6.27}$	$14.12_{\pm 5.13}$
Spatial Reasoning	$82.14_{\pm 9.92}$	$78.57_{\pm 10.55}$	$64.29_{\pm 12.18}$	$5.36_{\pm 6.38}$	$58.93_{\pm 12.47}$	$21.43_{\pm 10.55}$
Action Reasoning	$54.04_{\pm 5.75}$	$49.47_{\pm 5.77}$	$32.28_{\pm 5.39}$	$8.42_{\pm 3.25}$	$38.60_{\pm 5.61}$	$21.05_{\pm 4.71}$
Object Reasoning	$56.39_{\pm 4.55}$	$54.63_{\pm 4.56}$	$36.12_{\pm 4.40}$	$5.95_{\pm 2.19}$	$39.43_{\pm 4.48}$	$17.40_{\pm 3.49}$
Information Synopsis	$73.99_{\pm 4.77}$	$72.14_{\pm 4.86}$	$51.70_{\pm 5.41}$	$8.05_{\pm 2.99}$	$54.49_{\pm 5.40}$	$21.98_{\pm 4.50}$
Overall	$\mathbf{59.26}_{\pm 1.85}$	$\mathbf{55.63}_{\pm 1.87}$	$\mathbf{39.56}_{\pm 1.84}$	$\mathbf{7.63}_{\pm 1.00}$	$\mathbf{40.00}_{\pm 1.85}$	$\mathbf{18.96}_{\pm 1.48}$

Table 25: **Per-task-type 1-bit accuracy (%) on Video-MME (no-subtitle) for VideoLLaMA3-7B (N=2,700).** Each entry is the point estimate followed by the 95% CI half-width in subscript. The Overall row is bolded.

Task type	bf16	TACT-KV	TQ	KIVI	VidKV	AKVQ-VL
Temporal Perception	$65.45_{\pm 12.20}$	$65.45_{\pm 12.20}$	$38.18_{\pm 12.44}$	$9.09_{\pm 7.81}$	$40.00_{\pm 12.53}$	$12.73_{\pm 8.86}$
Spatial Perception	$64.81_{\pm 12.35}$	$62.96_{\pm 12.47}$	$51.85_{\pm 12.88}$	$9.26_{\pm 7.95}$	$48.15_{\pm 12.88}$	$22.22_{\pm 10.87}$
Attribute Perception	$76.58_{\pm 5.55}$	$72.07_{\pm 5.87}$	$44.59_{\pm 6.48}$	$9.91_{\pm 3.96}$	$40.99_{\pm 6.42}$	$13.06_{\pm 4.44}$
Action Recognition	$67.09_{\pm 5.18}$	$64.54_{\pm 5.27}$	$41.53_{\pm 5.43}$	$5.75_{\pm 2.62}$	$39.62_{\pm 5.39}$	$13.42_{\pm 3.78}$
Object Recognition	$71.19_{\pm 4.70}$	$66.95_{\pm 4.88}$	$40.68_{\pm 5.09}$	$8.47_{\pm 2.92}$	$39.27_{\pm 5.06}$	$14.69_{\pm 3.69}$
OCR Problems	$73.38_{\pm 7.27}$	$65.47_{\pm 7.81}$	$42.45_{\pm 8.11}$	$5.76_{\pm 4.00}$	$38.13_{\pm 7.97}$	$13.67_{\pm 5.71}$
Counting Problem	$44.78_{\pm 5.91}$	$42.54_{\pm 5.88}$	$29.48_{\pm 5.43}$	$8.96_{\pm 3.45}$	$34.33_{\pm 5.65}$	$10.45_{\pm 3.68}$
Temporal Reasoning	$50.28_{\pm 7.29}$	$45.20_{\pm 7.25}$	$24.86_{\pm 6.32}$	$6.21_{\pm 3.63}$	$22.03_{\pm 6.07}$	$7.91_{\pm 4.04}$
Spatial Reasoning	$82.14_{\pm 9.92}$	$75.00_{\pm 11.09}$	$58.93_{\pm 12.47}$	$3.57_{\pm 5.57}$	$62.50_{\pm 12.29}$	$17.86_{\pm 9.92}$
Action Reasoning	$58.95_{\pm 5.68}$	$59.65_{\pm 5.66}$	$38.95_{\pm 5.62}$	$8.07_{\pm 3.19}$	$37.89_{\pm 5.59}$	$10.53_{\pm 3.58}$
Object Reasoning	$60.35_{\pm 4.48}$	$57.27_{\pm 4.53}$	$39.87_{\pm 4.48}$	$6.39_{\pm 2.27}$	$40.09_{\pm 4.49}$	$11.01_{\pm 2.89}$
Information Synopsis	$79.88_{\pm 4.37}$	$76.78_{\pm 4.59}$	$53.56_{\pm 5.41}$	$7.43_{\pm 2.89}$	$56.97_{\pm 5.36}$	$13.31_{\pm 3.71}$
Overall	$\mathbf{65.19}_{\pm 1.80}$	$\mathbf{62.00}_{\pm 1.83}$	$\mathbf{40.81}_{\pm 1.85}$	$\mathbf{7.44}_{\pm 0.99}$	$\mathbf{40.56}_{\pm 1.85}$	$\mathbf{12.44}_{\pm 1.24}$

Table 26: **Macro prediction-flip rate (%) relative to bf16 on RSNA-ICH and CT-RATE at matched visual-cache footprints.** Retention budgets match the measured footprint of TACT-KV at a nominal 1-bit rate. Entries include Wald 95% CI half-widths. **Bold** and underline mark the lowest and second-lowest values in each column within a model.

Method	RSNA-ICH	CT-RATE
MedGemma 1.5 4B		
TACT-KV (ours)	$\underline{3.78}_{\pm 0.28}$	$\underline{4.69}_{\pm 0.18}$
VL-Cache	$\mathbf{2.77}_{\pm 0.24}$	$\mathbf{4.21}_{\pm 0.17}$
FlashCache	$8.27_{\pm 0.40}$	$9.35_{\pm 0.24}$
MedGemma 4B		
TACT-KV (ours)	$\underline{2.37}_{\pm 0.22}$	$\mathbf{2.95}_{\pm 0.14}$
VL-Cache	$\mathbf{2.17}_{\pm 0.21}$	$\underline{3.58}_{\pm 0.16}$
FlashCache	$7.77_{\pm 0.39}$	$8.58_{\pm 0.23}$
MedGemma 27B		
TACT-KV (ours)	$0.37_{\pm 0.09}$	$\mathbf{2.63}_{\pm 0.13}$
VL-Cache	$\underline{0.33}_{\pm 0.08}$	$\underline{3.57}_{\pm 0.16}$
FlashCache	$\mathbf{0.23}_{\pm 0.07}$	$6.02_{\pm 0.20}$
Hulu-Med 7B		
TACT-KV (ours)	$\mathbf{6.08}_{\pm 0.35}$	$\mathbf{10.44}_{\pm 0.26}$
VL-Cache	$\underline{13.49}_{\pm 0.50}$	$\underline{15.91}_{\pm 0.31}$
FlashCache	$14.01_{\pm 0.51}$	$19.41_{\pm 0.33}$
Hulu-Med 4B		
TACT-KV (ours)	$\mathbf{8.53}_{\pm 0.41}$	$\mathbf{7.22}_{\pm 0.22}$
VL-Cache	$10.30_{\pm 0.44}$	$\underline{11.60}_{\pm 0.27}$
FlashCache	$\underline{8.84}_{\pm 0.41}$	$14.72_{\pm 0.30}$
Hulu-Med 32B		
TACT-KV (ours)	$\underline{9.13}_{\pm 0.42}$	$\mathbf{10.04}_{\pm 0.25}$
VL-Cache	$11.53_{\pm 0.47}$	$\underline{11.27}_{\pm 0.27}$
FlashCache	$\mathbf{8.85}_{\pm 0.41}$	$13.17_{\pm 0.28}$

E Additional Baseline Comparisons

We compare VL-Cache [38] and FlashCache [43] at the measured visual-cache footprint of TACT-KV at a nominal 1-bit rate. Their retention budgets are set separately for each model and dataset. Table 26 reports macro prediction-flip rates relative to bf16. Both methods retain non-visual tokens at bf16 and compress the visual cache after prefill. VL-Cache and FlashCache achieve lower flip rates in five of the twelve model and dataset pairs. TACT-KV has the lowest flip rate in the other seven.

E.1 Implementation Details

All baselines use the same visual inputs and prompts as TACT-KV. Non-visual tokens and newly generated tokens remain at bf16. KVQuant uses calibrated nonuniform codebooks, pre-RoPE key quantization, and a 1% outlier fraction stored at fp16. We construct codebooks at 1, 2, 3, and 4 bits. KVTC uses the public implementation, with PCA bases fitted per layer and attention head, adaptive bit allocation, affine scalar quantization, and DEFLATE entropy coding. Four sink tokens and the most recent 128 tokens remain at bf16. A nominal budget of b bits corresponds to an average of b bits per PCA coefficient before entropy coding.

Table 27: **Theorem 1 closure on twelve VLM-dataset pairs at 2- and 4-bit average rates.** For each pair we report the closed-form prediction $1/\gamma_K$ and the empirical ratios at each rate. **Bold** marks closure within 30% of $1/\gamma_K$, <u>underline</u> within 50%.

	RSNA-ICH			CT-RATE		
Model	$1/\gamma_K$	2-bit	4-bit	$1/\gamma_K$	2-bit	4-bit
MedGemma 1.5 4B	3.58	**2.97**	**2.65**	5.85	<u>8.76</u>	**7.02**
MedGemma 4B	3.06	**2.84**	**2.45**	5.00	<u>7.05</u>	**5.80**
MedGemma 27B	3.77	**2.82**	<u>2.61</u>	5.56	**6.08**	**5.42**
Hulu-Med 7B	4.85	**4.01**	**3.61**	7.58	<u>10.35</u>	**8.45**
Hulu-Med 4B	3.95	**3.23**	**3.04**	6.17	<u>9.01</u>	**7.99**
Hulu-Med 32B	4.46	**3.75**	**3.36**	6.58	**7.51**	**6.43**

Calibration uses general web text. We use 80 windows of 2,048 tokens, with 16 windows used to fit KVQuant codebooks and 512-token windows for its Fisher estimates. The calibration data are shared across bit budgets. VL-Cache allocates retention budgets using attention scores. FlashCache uses a token-axis DCT and outlier retention, with its frequency parameter set to 0.2.

F Closure at Higher Bit Rates

We extend the closure check in Table 2 to 2- and 4-bit average rates for all twelve VLM-dataset pairs. Table 27 reports the median K-cache reconstruction-MSE ratio $\mathcal{D}_{\text{uniform}}/\mathcal{D}_{\text{water-fill}}$ across study-layer measurements, with the transform and quantizer held fixed. Eight of twelve pairs are within 30% of the prediction at 2 bits, and eleven are within 30% at 4 bits. The remaining 4-bit pair is within 50%.

G Reconstruction Error of the Inner Scalar Quantizer

This appendix gives a layer-level mechanistic account of the quantizer attribution result reported in Table 3. We compare three scalar quantizers, gain-shape SQ, symmetric uniform, and NormalFloat, on the actual post-DCT K-cache of MedGemma 1.5 4B at one transformer layer, isolating the inner scalar quantization step from the rest of the TACT-KV pipeline.

Protocol. We sample one RSNA-ICH study at the standard $S = 28$ slice budget, run prefill on MedGemma 1.5 4B, and read out the K-cache at layer 5 as a mid-stack representative. After the 3-axis DCT, the cache produces $S \cdot H \cdot W = 28 \cdot 16 \cdot 16 = 7{,}168$ frequency bins per (batch, head). The model has four KV heads at this layer, so we obtain $4 \times 7{,}168 = 28{,}672$ per-bin coefficient vectors of head dimension $d = 256$. For each bit budget $b \in \{1, 2, 3, 4\}$ we quantize each vector with gain-shape SQ, symmetric uniform with per-vector absmax scaling, and NormalFloat [10], then compute the dequantization error against the original coefficient vector. The DCT and the per-layer bit budget are held fixed across the three quantizers, so the only varying factor is the inner scalar quantizer.

gain-shape SQ's reconstruction MSE is $168\times$ lower than the two naive baselines at 1-bit. Table 28 reports per-quantizer reconstruction MSE at each bit budget. Relative to the per-tensor energy $\|F\|^2/N = 1{,}722.97$, gain-shape SQ's 1-bit reconstruction sits at 0.15% error while Uniform sits at 25.6%. Uniform stays $21\times$ worse than gain-shape SQ at 4-bit, where the model tolerates the residual magnitude and the downstream flip rate of Table 3 converges, but at 1-bit the same factor drives the catastrophic flip-rate degradation reported there.

The MSE gap comes from the bulk of post-DCT coordinates that sit far below the per-vector absmax. Table 29 partitions the 1-bit reconstruction error by $|x|/\text{absmax}$. 60% of all coordinates lie below $0.1\times$ the per-vector absmax, and symmetric uniform reconstructs every one of them at $\pm$absmax, yielding mean absolute error 6.114. gain-shape SQ first extracts the per-vector ℓ_2 norm g and quantizes the rotated unit-vector residual, so the bulk is recovered at MAE 0.507, a $12\times$ improvement on the dominant fraction of the data. The remaining 0.6% of coordinates near the per-vector boundary trade in the opposite direction, where Uniform's $\pm$absmax output happens to be

Table 28: **Per-quantizer reconstruction MSE on real post-DCT K-cache coefficients of MedGemma 1.5 4B.** Mean squared error between the original and dequantized coefficient vectors, computed over 28,672 per-bin vectors at one transformer layer. Lower is better. The first row is the configuration used in TACT-KV.

Quantizer	b=1	b=2	b=3	b=4
gain-shape SQ (TACT-KV)	**2.617**	**0.768**	**0.218**	**0.0621**
Uniform sym	440.708	43.189	6.884	1.288
NormalFloat	440.708	29.126	3.373	0.196

close to truth and gain-shape SQ's residual is comparatively wide. Aggregated over the 7.34 million coordinates, this trade produces the 168× MSE gap of Table 28.

Table 29: 1**-bit reconstruction MAE on MedGemma 1.5 4B post-DCT K-cache, partitioned by per-coordinate magnitude.** Coordinates are binned by $|x|/\text{absmax}$ over the 28,672 per-bin vectors. The first column reports the bin range, the second the coordinate count and its share of the total. The last column reports the multiplicative advantage of gain-shape SQ over Uniform.

$\lvert x\rvert/\text{absmax}$ range	# coordinates (%)	Uniform MAE	gain-shape SQ MAE	SQ advantage
[0.0, 0.1] (the bulk)	4,396,691 (60.0%)	6.114	0.507	12.1×
[0.1, 0.3]	2,047,900 (28.0%)	3.397	0.405	8.4×
[0.3, 0.5]	569,408 (7.8%)	2.445	0.466	5.2×
[0.5, 0.7]	203,768 (2.8%)	1.653	0.567	2.9×
[0.7, 0.9]	76,210 (1.0%)	0.871	0.691	1.3×
[0.9, 1.0]	46,055 (0.6%)	**0.081**	0.991	SQ 12× worse

H Tile-Fused Inverse-DCT Attention Kernel

This appendix describes the fused inverse path referenced in Sections 2.4 and 3.6. The compressed K and V caches stored by TACT-KV live in the frequency domain as a tensor of b_i-bit packed indices and per-bin ℓ_2-norm scalars. Decoding requires (i) dequantizing each frequency bin into the dequantized frequency tensor $\hat{F}$, (ii) applying the inverse three-axis DCT to recover the spatial-grid representation $\hat{X}$, and (iii) running the attention dot product against the current query. A naive serial implementation materializes the full reconstructed key tensor $\hat{X}$ of shape $B \times n_h \times n_v \times d$ and the corresponding value tensor at every decode step before launching attention. This intermediate footprint scales linearly with the decoding batch dimension B and dominates the deployment-time HBM budget at large B, eliminating the iso-HBM headroom recovered by compression.

The Q-projection commutes with the inverse three-axis DCT, so the full reconstructed cache is never needed. The inverse three-axis DCT is a linear map. Writing $\Phi \in \mathbb{R}^{n_v \times n_v}$ for its composed matrix form (separable as $\Phi = D_s^\top \otimes D_h^\top \otimes D_w^\top$) and viewing $\hat{F}, \hat{X}$ as $(n_v \times d)$ matrices for one (batch, head) slot, we have $\hat{X} = \Phi\,\hat{F}$. For a single-row decoder query $Q \in \mathbb{R}^d$, the attention score against the n-th reconstructed key satisfies

$$Q \cdot \hat{X}[n] \;=\; Q \cdot \big(\Phi\,\hat{F}\big)[n] \;=\; \sum_{n'=1}^{n_v} \Phi[n, n']\,\big(Q \cdot \hat{F}[n']\big),$$

so the score equals a linear combination of the scalar projections of Q against the *frequency-domain* bins of $\hat{F}$, weighted by the n-th row of Φ. The same identity applies on the value side with the attention weights in place of Q. The Q-projection and the weights-projection are therefore applied directly to the dequantized but un-transformed $\hat{F}$, producing a scalar per frequency bin, and the inverse three-axis DCT is applied last to those scalars rather than to the full (n_v, d) tensor. The full key and value tensors are never instantiated.

The kernel applies dequantization, the attention dot product, and the inverse DCT inline within each FlashAttention tile. The identity above is realized as a Triton kernel in two stages. The first stage walks the frequency-domain bins in register-resident tiles. For each tile the kernel loads the packed index buffer at the bit width b_i, unpacks the indices, gathers the corresponding gain-shape SQ

codebook entries, multiplies by the per-bin ℓ_2-norm scalars, and accumulates the dot product against the query in fp32. The second stage performs the same per-tile dequantization on the value cache and accumulates the attention-weighted sum. The inverse three-axis DCT closes the kernel by a small matmul on the per-frequency scalars, which inherits the separability of the DCT and reduces to three sequential one-dimensional transforms along the slice, height, and width axes.

Deployment-time HBM under the fused path is proportional to the compressed cache rather than to the reconstructed full cache. The fused path replaces the $\mathcal{O}(B \cdot n_h \cdot n_v \cdot d)$ intermediate of the naive path with a per-tile scratch whose size is bounded by the FlashAttention block configuration and is independent of B. The HBM held during decode is therefore the compressed cache (packed indices plus per-bin ℓ_2-norm scalars) plus a small per-tile transient, which is what enables the iso-HBM decoding-batch multipliers reported in Table 6 for the Hulu-Med family.

The fused path is implemented for Qwen2 and Qwen3 attention and transfers to other attention classes through the same per-tile principle. The kernel is realized as a per-attention-class wrapper that replaces the model's attention forward on decode steps when the cache is a TACT-KV cache and the layer is full-attention. The Qwen2 and Qwen3 implementations cover the Hulu-Med family of vision-language models on which we report decoding-batch profiling in Section 3.6. The same fusion identity transfers to other attention classes such as Gemma-3, Llama, and Qwen2.5-VL through their corresponding attention forward surfaces.

I Detailed Related Work

This appendix provides extended related work.

LLM-Targeted KV Cache Quantization

KIVI [29] quantizes the K cache per channel along token groups and the V cache per token along head-dimension groups under asymmetric uniform quantization at 2-bit, on the observation that K exhibits channel-wise outliers while V exhibits token-wise magnitude variation. KVQuant [19] combines per-channel quantization on K with pre-RoPE rotation that removes the rotary embedding's effect on the quantization range, a sensitivity-weighted non-uniform codebook derived from a calibration set, and per-vector dense-and-sparse splitting that isolates outliers separately for each vector. TurboQuant [44] rotates each per-token coordinate vector by a random orthonormal matrix that pushes the post-rotation distribution toward an isotropic Gaussian, then applies a Lloyd–Max scalar codebook tuned for that target distribution. Our gain-shape SQ codebook in Section 2.4 inherits the rotation-then-Lloyd–Max pattern of TurboQuant Algorithm 1 and adds gain-shape extraction.

VLM-Targeted KV Cache Compression

AKVQ-VL [35] applies a Walsh–Hadamard transform along the head-dimension of K to suppress channel outliers, then per-channel asymmetric uniform quantization on the rotated K and per-token asymmetric uniform quantization on V, with an adaptive bit budget allocated by Text-Salient Attention (a layer-level selection of attention heads that concentrate on text tokens) and Pivot-Token-Salient Attention (a token-level rule that retains a small set of pivot tokens at higher precision). The published configuration is mixed-precision around an average of 2-bit on the visual cache. VL-Cache [38] is eviction-based rather than quantization-based, allocating a layer-adaptive sparsity budget across decoder layers and a modality-aware token scoring policy that distinguishes visual from text tokens during prefill and decoding, retaining about 10% of the entries at full precision and discarding the rest.

Frequency-Domain Compression

FreqKV [21] targets context-window extension on text-only LLMs and applies a discrete cosine transform along the token-sequence axis. When the cache reaches a fixed maximum size, the method iteratively compresses older chunks in the frequency domain by retaining low-frequency components and discarding high-frequency components, on the observation that context information concentrates in the low-frequency band. Adapting the model to this representation requires a brief fine-tuning pass on a long-context corpus, in contrast to the calibration-free TACT-KV pipeline of Section 2.

FlashCache [43] extracts a low-frequency baseline of the multimodal KV cache by low-pass filtering and identifies outlier KV pairs as those that deviate substantially from this baseline. The compression retains the low-frequency baseline together with the outlier pairs and discards the rest. KVTC [34] is a transform-coding pipeline that applies a calibration-fitted PCA along the head dimension, an adaptive bit allocation across the principal components, and an entropy coder on the resulting indices. KVTC's basis is one-dimensional and is fitted per model from a calibration set, whereas TACT-KV uses a fixed three-axis DCT over the volumetric input geometry without calibration. Fourier-VLM [41] applies a 2-D discrete cosine transform to the vision encoder's output features and discards high-frequency components by a low-pass filter before the compressed visual tokens enter the language-model decoder. The transform is therefore applied on the input side of the LLM and does not modify the per-layer KV cache, which is the target of TACT-KV.

Video-Domain KV Cache Compression

Sali-Cache [32] reuses cache entries across temporally adjacent video frames through optical-flow-based detection of inter-frame redundancy and applies a saliency-guided spatial filter that retains visually significant regions and prunes the rest. LOOK-M [40] is a plug-and-play eviction and merging policy for multimodal long-context inference. It observes that the model's attention during prompt prefill is biased toward text tokens, ranks visual cache entries by cumulative attention score with a text-side bonus, evicts the lowest-ranked entries, and merges the discarded entries into the surviving ones through averaged, pivotal, or similarity-weighted variants. The method is precision-preserving on the surviving entries and requires no fine-tuning. VidKV [36] applies a 1-D fast Fourier transform per token along the head-dimension of the K cache, then 1-bit sign quantization on the real and imaginary parts of the resulting spectrum, with channels classified as anomalous held at 2-bit asymmetric uniform instead of going through the FFT branch. The V cache uses 1.58-bit ternary quantization with per-group absmax scaling along the head-dimension.

J Limitations

Limitations. The volumetric prior we exploit is validated on chest CT (CT-RATE) and head CT (RSNA-ICH); other 3D medical modalities such as MRI, PET, and 3D ultrasound carry the same anatomical-axis structure in principle but are not measured here. The closed form of Theorem 1 holds in the high-rate regime in which every frequency bin receives non-trivial bit allocation, and on the smaller Hulu-Med variants the lowest-rate operating point pushes the empirical ratio outside the band predicted by the theorem (Section 3.3); deriving a low-rate counterpart bound is left to future work. Our pipeline compresses the visual portion of the cache only and leaves text tokens at bf16, so the savings reported here are tied to inputs in which visual tokens dominate; integrating TACT-KV with attention-aware eviction on the text portion is orthogonal and a natural next step. At decode, the gain–shape codebook and the tile-fused inverse path add per-step compute relative to a uniform quantizer with no transform, so the iso-HBM batch capacity reported in Section 3.6 is recovered at the cost of additional FLOPs per decoded token; the trade-off favors HBM-bound deployment and is less attractive when decoding is FLOP-bound. Deployment profiling is reported on a single GPU class, and cross-hardware behavior at smaller HBM budgets is not measured.

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: The abstract and Section 1 state the method, dataset, and main quantitative claims, and these are supported by the experiments in Section 3 and the conclusion.

   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: Section 4 and Appendix J discuss limitations of the theoretical assumptions, modality coverage, and deployment setting.

   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: Section 2.3 states Theorem 1 and its assumptions and provides a proof sketch. Appendix A gives the full proof.

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: Sections 2 and 3 and Appendices C, E, and H describe the compression pipeline, model configurations, evaluation protocols, and baseline settings.

   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: The TACT-KV code is publicly available through the repository linked on the first page. Dataset access and evaluation protocols are described in Appendix C.

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: Section 3.1 and Appendices C and E specify the data splits, preprocessing, prompts, bit budgets, and baseline configurations.

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: Flip rates and multiple-choice accuracies are reported with Wald 95% confidence intervals. AUROC results use 95% bootstrap confidence intervals with 1,000 resamples, as described in Appendix D.

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: Appendix C.2 reports the GPU type, memory capacity, per-run timings, and approximate total compute budget.

   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: The study evaluates existing models on public benchmarks without recruiting participants or conducting clinical interventions.

   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:[Yes]

    Justification: The paper discusses positive clinical impact in the introduction and conclusion, and it also addresses risks and failure possibilities.

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 released asset is a KV-cache compression implementation. The study uses existing datasets and pretrained models.

    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: The datasets, models, and baseline methods are credited in the paper. Appendix C.1 summarizes the dataset licenses and terms of use.

    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: [Yes]

    Justification: The released asset is the TACT-KV implementation. Its design and evaluation protocols are documented in Sections 2 and 3 and Appendices C, E, and H.

    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 study does not involve crowdsourcing or experiments with recruited participants.

    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 experiments use existing public benchmarks and do not involve new data collection, participant recruitment, or clinical intervention.

    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: LLM is not used for core method proposal and creation, LLM is only used for format, writing polish.

    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.