Cheap Probes Predict Expensive Training in 3D-CT Vision–Language Models

Renjie Liang
University of Florida, Gainesville, FL, USA

Abstract

Picking the frozen image encoder for a 3D CT vision–language model (VLM), together with the token-compression scheme on top of it, is a search over many candidates. There are several encoders, several ways to compress their tokens, and several token budgets, and the combinations grow fast. Comparing them the usual way means fine-tuning a large language model (LLM) on each combination, and running the whole sweep this way needs far more compute than most groups can spend. We ask whether a cheap probe on the encoder's cached embeddings can stand in for that comparison. We build an image-grounded probing benchmark over (encoder $\times$ compression) cells, with clinical attribute families and two validation gates, scale-sanity and probe-separability, that keep each attribute well-scaled and decodable. These gates are the main methodological contribution. On this benchmark we compare a range of read-out heads, and in a preliminary study we pair each probe with its matched downstream task. The early signal is encouraging: the cheap probe orders the candidates in close agreement with expensive fine-tuning, at about $r \approx 0.95$ on the cells measured so far. We read this as an ordinal claim, a ranking predictor rather than an exact estimate, and we are explicit about where it stays preliminary. If it holds up, encoder and compression choices can be screened in minutes with frozen-token probes, with full training spent only on the finalists.

1 Introduction

Modern 3D CT vision–language models follow a standard recipe. A frozen 3D image encoder turns a volume into a grid of visual tokens, an optional compression module shortens that grid, a projector maps the tokens into an LLM's embedding space, and the LLM is fine-tuned to generate reports or answer questions. Two early choices dominate final quality: which encoder, and which compression. Both are usually settled by brute force, fine-tuning the LLM on every candidate and comparing downstream metrics, at days of GPU time per configuration.

A probe is a lightweight read-out, a closed-form ridge regressor or a small attention head, trained on the frozen embeddings to predict a clinical attribute. It costs orders of magnitude less than fine-tuning an 8B-parameter LLM. If probe quality tracked downstream quality, one could screen encoders and compression schemes with probes and pay for full training only on the survivors. This paper makes that claim concrete and gives preliminary evidence for it, in three parts:

1. **How the benchmark is built** (Section 3): a grid of *(encoder $\times$ compression)* cells, 15 image-grounded regression attributes plus 18 disease findings, two question types, two validation gates (scale-sanity and probe-separability), explicit input normalization, and a six-dimension probe-audit rubric.

2. **Many probing models** (Section 4): ten read-out heads compared on the *same* frozen tokens. Most heads are strong; one over-parameterized head is pathological.

3. **Cheap predicts expensive** (Section 5): under this probing setup, disease-probe AUROC predicts report-generation clinical micro-F1 at $r = 0.95$ / $\rho = 0.89$ over six matched cells, robustly across read-out choice. The claim is *ordinal*, and we disclose where within-encoder rankings disagree.

This is a stake-a-claim marker: concise, honest, preliminary. We want to show that frozen-token probing is a usable stand-in for 3D-CT encoder and compression selection, and to give the community a read-out comparison and an audit rubric that keep such claims falsifiable.

2 Related Work

3D CT vision–language models. CT-CLIP and CT-CHAT [1, 2] pair a contrastive 3D encoder with an LLM for report generation; CT-Agent [3] adds anatomy-aware retrieval; BTB3D [4] replaces contrastive tokens with a reconstruction-based tokenizer. These systems differ chiefly in their encoder and in how its many tokens are compressed before the LLM, yet the choice is validated only by expensive end-to-end training. We ask whether it can instead be predicted by probing.

Probing frozen representations. Linear and attention probes are a standard tool for asking *what* a representation encodes [5, 6], with control tasks and selectivity introduced to guard against the probe doing the work [7]. Most studies stop at "the information is present." We push further to ask whether probe scores are *predictive* of downstream quality, and make that question falsifiable with an audit rubric, negative controls, and noise floors.

Cheap proxies for model selection. Transferability-estimation scores (LEEP [8], LogME [9], H-score [10], NCE [11]) rank pre-trained models from frozen features, but target single-number classification transfer. Our proxy is domain-specific: a predictor for 3D CT encoder and compression selection whose predictand is free-text report generation. The closest medical probing benchmark [12] is 2D chest X-ray, and studies neither 3D CT, token compression, nor the link from probe to generative downstream quality.

3D CT VQA. Report-derived CT VQA (e.g. M3D-VQA [13]) has an LLM generate questions and answers from diagnostic reports. Our VQA gold answers come straight from the image (masks and Hounsfield units), carry no report or LLM noise, and are exactly reproducible, so the probe target and the VQA target share one identical label. We use it as an image-grounded probe target [14], not a standalone leaderboard. It speaks to the field's gap of scoring *where* and *how much*, not only *what* [15].

3 Part 1: Benchmark construction

The benchmark probes *what a token compressor preserves*. It has four parts: a grid of (encoder $\times$ compression) cells, a set of image-grounded clinical attributes organized into families, two question types per attribute, and its methodology contribution, *two validation gates* that every attribute must pass before it becomes a question. The frozen tokens of each cell also feed the expensive path (projector + LLM fine-tune, $\sim$1 GPU-day/cell), against which predictive validity is later measured. Cells are built on CT-RATE non-contrast chest CT [1].

Cells: encoder $\times$ compression. A *cell* is one (encoder, compression) configuration; the encoder is frozen throughout. **Grid encoders** produce a dense token grid: CoLiPri [16] and CT-CLIP [17] are contrastive (768- and 512-dim tokens); BTB3D [4] is a reconstruction tokenizer with low-dimensional (18-d) codebook tokens. **Organ encoders** (VISD-Boost [18], fVLM [19]) emit a

handful of organ-level tokens. On top of each grid encoder we apply one of three compressions: **none** (native tokens), **uniform_pool** (average-pool to a fixed budget of 216 tokens, preserving per-token dimension), and **adp** (a learned pack-then-linear bottleneck). The pool column fixes the token budget across grid encoders so within-column comparisons isolate the encoder from token count.

Attribute families: 15 regression attributes + 18 disease findings. Every attribute is *image-grounded*. It is measured straight from the CT image (segmentation masks or Hounsfield units), never from a report, so labels carry no report or LLM noise and are exactly reproducible. Fifteen continuous attributes span four families: **size** (5: cardiothoracic ratio, aorta-to-heart and IVC-to-aorta ratios, aortic diameter in mm, heart width in mm); **density** (4: aortic-wall calcification, vertebral bone density, mean lung density, lung high-attenuation fraction); **radiomics** (3: lung and vertebral first-order kurtosis, emphysema index `lung_Perc15`); and **location** (3: left–right lung log-ratio, cardiac lateral position, IVC axial level). A fifth family is the **18 CT-RATE disease findings** (prevalence 8–45%), extracted with RadBERT [20]. Segmentation-derived attributes use TotalSegmentator masks [21]; density and radiomics use per-organ HU statistics, which stay stable on non-contrast CT-RATE.

Two question types. Each regression attribute is emitted in two forms as metadata, and the final per-attribute form is chosen by clinical meaning and budget. **Catalog** is a letter-MCQ. It is either a 3-class *tertile* bucketing (cut-points fixed on the *train* split, so there is no validation leakage, and saved to `tertile_cutpoints.json`, chance $= 0.333$), or, for the three attributes that pass the scale gate below, a *clinical-threshold* split (chance $= 0.5$). Option order is randomized per record (seeded by question id) to defeat positional shortcuts. **Value** is open numeric generation ("answer with a number only"), emitted only for interpretable units (mm, HU, or a unitless number) and scored by MAE and within-tolerance accuracy. Disease findings are presence Yes/No questions.

The two validation gates (methodology contribution). An image-grounded attribute is useful only if its labels are *well-scaled* and its buckets are *decodable*. Two gates enforce both before any question ships, and they are the core of the construction methodology.

Gate 1: scale sanity (anti-degenerate). We apply each proposed clinical threshold to the ground truth and inspect the class balance, and reject any threshold that produces a near-degenerate split (say, 99%-one-class). Only **3 of the candidate clinical catalogs survive**: aortic dilation (`aorta_diameter_mm`, real <40 mm), emphysema (`lung_Perc15`, real <-950 HU), and aortic calcification (`hu_aorta_calc`, zero-inflated, so present/absent). Osteoporosis (`vert_median`) and cardiomegaly (`sz_heart_lung`) *fail*: their ground truth is a whole-vertebra median or a log-ratio, not calibrated to the clinical units, so they fall back to data-driven tertiles rather than ship an ill-scaled question.

Gate 2: probe-separability. A strong CoLiPri linear probe must separate each attribute's tertile buckets above chance (0.333). If a strong probe cannot separate an attribute at that granularity, the attribute is ill-posed, and we drop or coarsen it. Figure 1 shows the result. **13 of 15 attributes are well-posed** (separability 0.58–0.87). The two location targets `lung_LR_logratio` and `heart_x` are *marginal* (≈ 0.45); we flag them and keep them as honest "spatial-position canaries," the axis a merge-based compressor is expected to preserve least, rather than quietly drop them.

Downstream pairing. Each family is paired with the *same* capability of a downstream task trained with the full LLM (two-stage: projector warm-up, then LoRA). **Disease → report generation**, scored by clinical micro-F1 (RadBERT-extracted findings, generated vs. reference). **Regression**

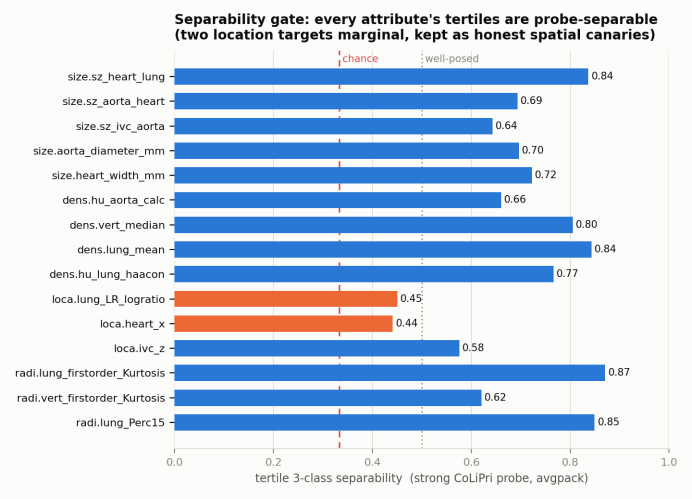

Figure 1: **Separability gate.** Tertile 3-class separability of a strong CoLiPri linear probe on each of the 15 regression attributes (avgpack tokens). Thirteen attributes clear the well-posed line; the two marginal location targets (orange) sit near chance and are kept, flagged, as spatial-position canaries. A bucketing a strong probe cannot separate is dropped or coarsened.

attributes $\rightarrow$ **VQA**, where probe and VQA share the identical image-derived tertile/threshold label, so the two scores are directly comparable. Predictive validity is then the cross-cell rank correlation between the cheap probe score and the expensive downstream score, computed one capability at a time.

Input normalization. Encoders differ by an order of magnitude in embedding scale (per-channel std ≈ 0.04 for CT-CLIP vs. ≈ 0.94 for BTB3D). A light read-out with weight decay then quietly penalizes low-variance encoders unless the input is normalized to the encoder's scale. We treat normalization as an explicit variable: ZCA-whitening (fit once on training features) for the low-dimensional variants, and per-token LayerNorm for the high-dimensional learned-projector variant, with raw features as an ablation. Normalization sharpens the cross-encoder ranking rather than distorting it. On CT-CLIP a mean read-out decodes 0.60 AUROC raw and 0.72 under ZCA, which cleanly separates it from BTB3D.

Probe-audit rubric. A predictive-validity claim is only as trustworthy as the probes behind it, so we audit every probe along six dimensions [7, 5]: (1) **label validity** (trustworthy, leakage-free target); (2) **input validity** (same volumes and preprocessing across cells); (3) **read-out strength** (a studied variable, since a head that is too strong measures the information ceiling, not what the LLM can use); (4) **metric & null** (appropriate metric, explicit chance level, negative controls); (5) **construct validity** (does the probe's capability match the paired downstream metric); (6) **robustness** (sensitivity to read-out, compression, seeds). The two gates above fill in the label-validity and construct-validity dimensions, and the rubric doubles as a reusable protocol for probing studies. Appendix A shows representative label-validation visualizations from the construction pipeline.

From construction to validation. Parts 4 and 5 are a *preliminary validation of this probing methodology*. Their correlations were computed on an earlier label revision of the same probing pipeline, not on the exact gated 15-attribute / 18-finding benchmark above. We therefore read them

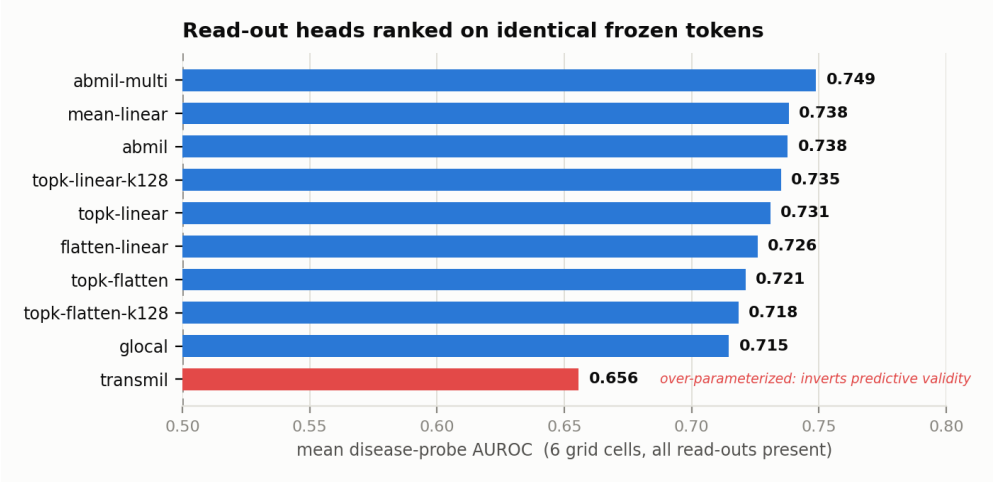

Figure 2: **Ten read-out heads on identical frozen tokens.** Mean disease-probe AUROC per head, averaged over the six grid cells (btb3d/colipri/ct_clip $\times$ adp/pool) for which every head is instantiated. Heads form a broad strong plateau (0.72–0.75); only the over-parameterized TransMIL collapses toward chance and, as Section 5 shows, inverts the probe→downstream correlation.

as evidence for the *methodology*, that cheap frozen-token probes predict expensive fine-tuning under this setup, not as measurements on the rigorous benchmark itself.

4 Part 2: Comparing ten read-out heads

A probe's read-out head is a design choice, not a given, so we compare ten heads on the *same* cached embeddings: attention pooling (ABMIL, ABMIL-multi), simple pools (mean-linear, top-k-linear at two k), flatten variants (flatten-linear, top-k-flatten), a global-local head (GLocal), and TransMIL. Figure 2 ranks them by mean disease-probe AUROC over the six grid cells where all heads are present.

Two findings stand out. First, **read-out choice matters little among sensible heads**. Nine of the ten heads sit in a narrow 0.72–0.75 band, with attention pooling (ABMIL-multi 0.749) and a plain mean-linear pool (0.738) roughly tied at the top, so almost any competent head recovers the encoder-side information. Second, **one head is pathological**. TransMIL collapses to 0.656, and (Section 5) it is the only head whose probe→downstream correlation *inverts*. That is itself diagnostic. An over-parameterized read-out measures the representation's information ceiling rather than the usable information the LLM can exploit, the exact failure the rubric's "read-out strength" dimension warns about. In practice, then, screen with a strong but modest head, attention pooling or a mean/top-k linear pool, and distrust any head whose in-sample strength does not carry over into predictive validity.

5 Part 3: Cheap probes predict expensive training

The central result. We pair the disease probe (ABMIL AUROC) with report-generation clinical micro-F1 over six matched cells (CoLiPri / CT-CLIP / BTB3D under *adp*, *pool*, or *none*). Figure 3 plots the two. The relationship is strong and monotone, at **Pearson** $r = 0.95$**, Spearman** $\rho = 0.89$. A minute-long probe on cached embeddings orders the cells almost exactly as a GPU-day-per-cell report-generation training run would, and it even tracks the *magnitude* of the downstream metric. Used as a selection rule, the top-probe cell recovers essentially the best encoder at two-to-three orders of magnitude less compute. Probes train in seconds to minutes; fine-tuning one cell costs

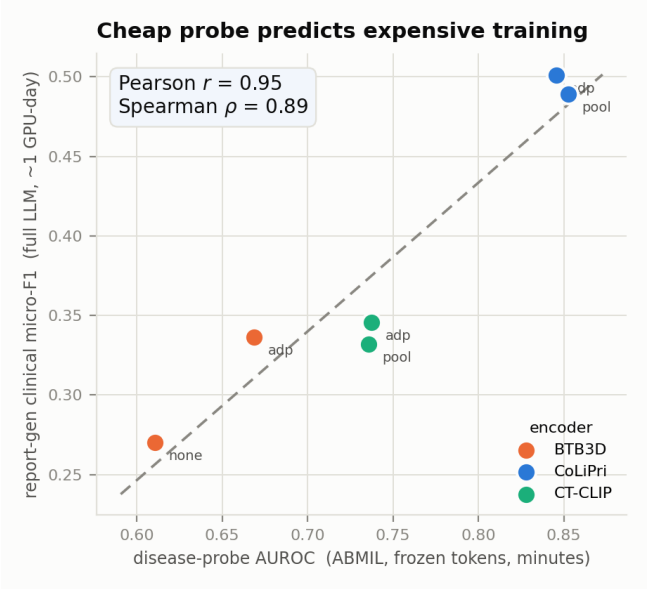

Figure 3: **The money figure.** Cheap disease-probe AUROC (ABMIL, frozen tokens, minutes) vs. report-generation clinical micro-F1 after full LLM training ($\sim$1 GPU-day), over the six matched cells. Pearson $r = 0.95$, Spearman $\rho = 0.89$. Marker color is the encoder; text labels give the compression.

Table 1: **Predictive validity is not an artifact of one read-out.** Cross-cell correlation between disease-probe AUROC and report-generation clinical micro-F1 (six matched cells), under each read-out head. Every strong head agrees; only the over-parameterized TransMIL inverts.

Read-out	ρ	r	**Read-out**	ρ	r
ABMIL	0.89	0.95	top-k-flatten-k128	0.77	0.93
mean-linear	0.89	0.95	top-k-flatten	0.77	0.91
top-k-linear-k128	0.89	0.95	flatten-linear	0.66	0.94
top-k-linear	0.89	0.94	GLocal	0.60	0.74
ABMIL-multi	0.77	0.94	TransMIL	-0.30	-0.03

about a GPU-day.

An ordinal claim. Predictive validity here is *ordinal*. The probe's job is to predict the downstream *ranking* of cells, not to reproduce the F1 scale. Spearman $\rho = 0.89$ is the number that matters: a practitioner who trains only the probe-preferred cells will, with high probability, be training the cells that full fine-tuning would also prefer.

Robust to read-out choice. The correlation does not depend on ABMIL. Table 1 reports the same probe$\rightarrow$downstream correlation under each read-out. Every strong head agrees ($\rho = 0.66$–0.89, $r = 0.91$–0.95), and only the pathological TransMIL inverts ($\rho = -0.30$), the same head that collapsed in Figure 2. So predictive validity is a property of the frozen tokens, recovered by any competent head, and not a lucky read-out.

Where the ranking mismatches (honest disclosure). Ordinal prediction is near-perfect *across* encoders but imperfect *within* an encoder's compression variants, and we disclose it. For the strong heads, the within-encoder ranking of compression variants matches downstream for CT-CLIP and BTB3D (adp $>$ pool/none in both probe and F1) but flips for CoLiPri, where the probe ranks pool $>$ adp while report-gen ranks adp $>$ pool. This is a tie, not a real inversion. The two CoLiPri variants are statistically indistinguishable in both probe AUROC (0.853 vs. 0.846) and F1 (0.489 vs. 0.501), so the cross-encoder signal, the one you would act on, is unaffected. The probe reliably picks the right *encoder*, but you should not trust it to break near-ties between an encoder's own near-duplicate compression variants.

Negative control. Probing for *shuffled* disease labels collapses to chance (AUROC 0.51), which confirms the real signal is not a read-out artifact. With the read-out-robustness check and the ordinal framing, the predictive-validity claim survives the obvious attempts to break it.

6 Conclusion

We staked out a probing methodology for 3D-CT encoder and compression selection and gave preliminary evidence for its central promise. We built an image-grounded benchmark over (encoder $\times$ compression) cells, compared ten read-out heads on identical frozen tokens, and paired each probe with the matched downstream task. Two things came out of it: (i) read-out choice matters little among sensible heads, though one over-parameterized head is pathological, and (ii) the cheap disease-probe AUROC predicts report-generation clinical micro-F1 at Pearson $r = 0.95$, Spearman $\rho = 0.89$, robustly across read-outs. Read ordinally, with the probe ranking predicting the downstream ranking, this says encoder and compression choices can be screened with minute-long frozen-token probes instead of a full fine-tuning sweep, and full training paid only for the survivors.

Limitations and ongoing work. This is a preliminary marker. The correlation rests on six matched cells from a single dataset (CT-RATE, single-institution, non-contrast chest CT, model-extracted labels), and it was computed on an earlier label revision of this pipeline, not on the exact 15-attribute / 18-finding gated benchmark of Section 3, so we state it at the methodology level. Within-encoder near-ties are not reliably ordered, and the per-family regression-attribute pairing against the image-grounded VQA is not yet reported here. We are re-running the correlation on the gated benchmark, enlarging the cell grid, adding n=3 seed error bars, replicating across a second LLM backbone, and validating the cut-points with clinician review. External validity to other sites, contrast protocols, and body regions is still to be shown.

References

[1] Hamamci et al. Ct-rate (chest ct + reports). *arXiv preprint arXiv:2403.17834*, 2024. -; first large paired 3D CT-report (25692 studies).

[2] Hamamci et al. A foundation model utilizing chest ct volumes and radiology reports. *arXiv preprint arXiv:2403.17834*, 2024. -; CT-CLIP encoder + LLaMA chat; RG + VQA.

[3] Mao et al. Ct-agent: A multimodal-llm agent for 3d ct (qa + report gen). *arXiv preprint arXiv:2505.16229*, 2025. -; agentic planning + memory retrieval; QA+RG; region-LoRA(verify).

[4] Hamamci et al. Better tokens for better 3d: Advancing vl modeling in 3d medical imaging. *arXiv preprint arXiv:2510.20639*, 2025. NeurIPS; learned freq-aware 3D tokenizer; +40% clinical F1; our base lineage.

[5] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

[6] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[7] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019.

[8] Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning (ICML)*, 2020.

[9] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning (ICML)*, 2021.

[10] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

[11] Anh T. Tran, Cuong V. Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[12] Anonymous. Feature quality and adaptability of medical foundation models: A comparative evaluation for radiographic classification and segmentation. *arXiv preprint arXiv:2511.09742*, 2025. 2D chest X-ray; linear probing of 8 foundation encoders.

[13] Bai et al. M3d: Advancing 3d medical image analysis with multi-modal llms. *arXiv preprint arXiv:2404.00578*, 2024. -; generalist 3D MLLM; spatial pooling perceiver; M3D-Data/Bench.

[14] Zhang et al. Radgenome-chest ct: A grounded vision-language dataset. *arXiv preprint arXiv:2404.16754*, 2024. Sci Data; region-grounded reports + masks (665K) on CT-RATE.

[15] Kyung et al. Medregion-ct: Region-focused multimodal llm for 3d ct rg. *arXiv preprint arXiv:2506.23102*, 2025. MICCAI; R2 token pooling + mask-driven extractor + patient attributes.

[16] Wald et al. Comprehensive language-image pre-training for 3d medical image. *arXiv preprint arXiv:2510.15042*, 2025. -; comprehensive LI pretraining; our encoder (3 variants).

[17] Hamamci et al. Ct-clip: Contrastive language-image pretraining for 3d chest ct. *arXiv preprint arXiv:2403.17834*, 2024. -; contrastive 3D CT-text; released CT-RATE; our encoder.

[18] Cao et al. Boosting vision semantic density with anatomy normality modeling. *arXiv preprint arXiv:2508.03742*, 2025. -; anatomy normality modeling for RG.

[19] Shui et al. Large-scale and fine-grained vision-language pre-training for enhanced ct. *arXiv preprint arXiv:2501.14548*, 2025. -; anatomy-level alignment; 69086 patients; our encoder.

[20] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting transformer-based language models to radiology. In *Radiology: Artificial Intelligence*, volume 4, 2022.

[21] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, et al. TotalSegmentator: Robust segmentation of 104 anatomical structures in CT images. *Radiology: Artificial Intelligence*, 5 (5), 2023.

A Benchmark construction and label validation (gallery)

The image-grounded labels are validated at construction time, before any probe or question is generated. Figures 4–7 show representative checks from the shared label-construction pipeline: per-record measurement overlays, target quality-control, the scale-sanity guard distributions, and the confound audit. They illustrate the validation *methodology*. Some panels include organs beyond the chest attribute list, since the same pipeline runs across body regions. Each figure states what it validates.

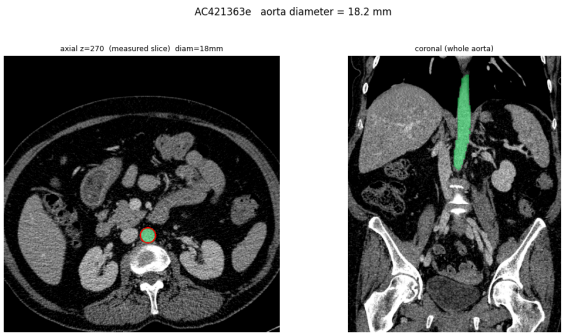

Figure 4: Measurement overlay for the `aorta_diameter_mm` size attribute: the measured axial slice (left) and the whole-aorta segmentation (right) confirm the diameter is read from the correct structure.

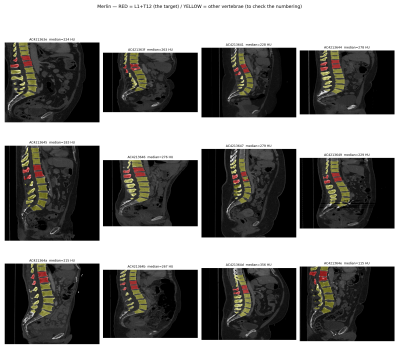

Figure 5: Quality-control for the `vert_median` bone-density attribute: the target vertebrae (red) against neighbours (yellow), with the per-case median HU, confirm the correct level is measured.

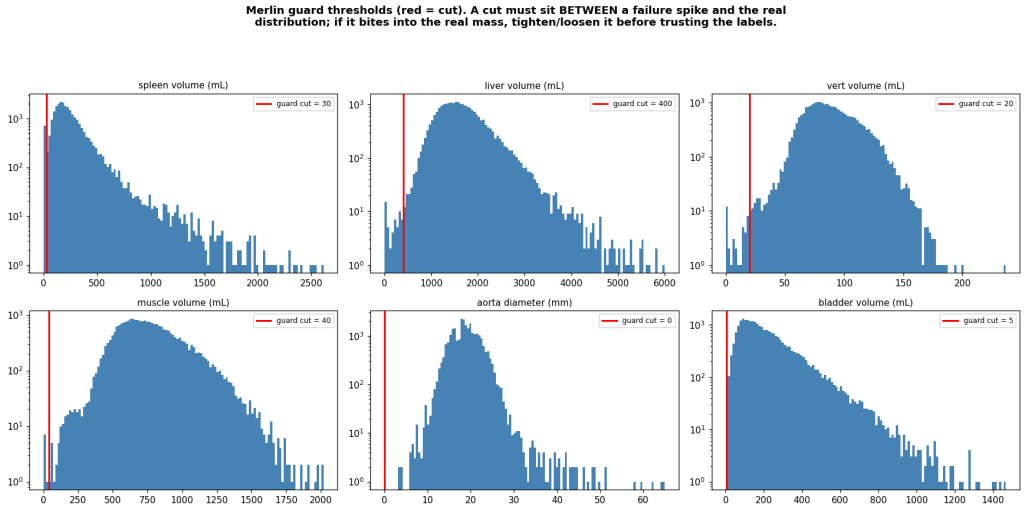

Figure 6: Scale-sanity gate (Gate 1): per-attribute measurement distributions with the guard cut (red). A cut must sit between a failure spike and the real distribution, so no threshold carves a degenerate, near-one-class split.

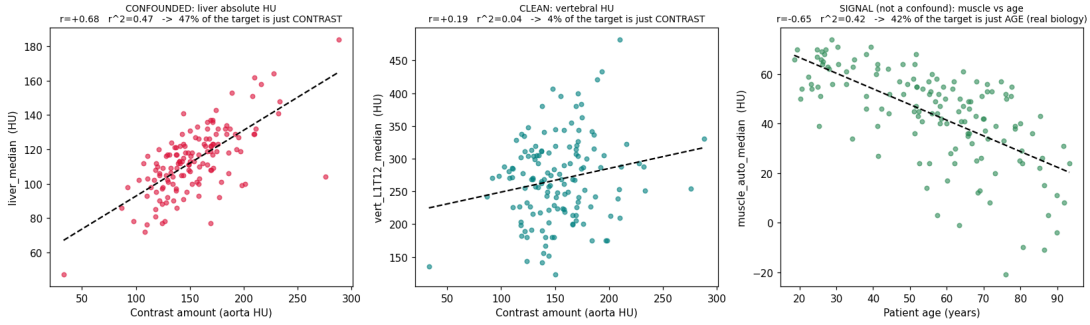

Figure 7: Confound audit: each candidate HU target is regressed against nuisance factors so confounded targets are disclosed or dropped. Left, an absolute-HU target that is largely a contrast effect; middle, a clean target; right, a target driven by real biology (age). This fills in the rubric's confound-control dimension.