When Do Cheap Probes Predict Expensive Training? Probing 3D-CT Encoders for Text Generation

Anonymous authors
Paper under double-blind review

Abstract

Building a 3D CT vision language model begins with a choice of which image encoder to build on. Today that choice is made by fine-tuning every candidate through the full language model and comparing downstream scores, an enormously expensive search. A cheap probe on the encoder's representation promises a way out, but whether it forecasts the expensive outcome has never been tested. We test this with **CheapCT** on report generation and on MeasureVQA, a new VQA dataset we build. MeasureVQA scores the outcome one capability at a time, its answers measured from segmentation masks and Hounsfield units. Report generation scores the whole report at once and reflects mostly disease. CheapCT forecasts expensive training across every capability. The forecast survives changing the probe readout and the language-model backbone. The rank agreement between CheapCT and fine-tuning stays high throughout, from $\rho = 0.90$ to 0.97. Used to choose an encoder, CheapCT picks one nearly as good as the best while fine-tuning a single candidate, at orders of magnitude less compute. We release the code and MeasureVQA at `https://anonymous.4open.science/r/CheapCT-anon-33B5`.

1 Introduction

Computed tomography is a workhorse of modern diagnostic imaging, with about 93 million examinations performed in the United States in 2023 (Smith-Bindman et al., 2025), and a single chest CT is a dense 3D volume that a radiologist reads slice by slice for many distinct findings. Automating that reading with a vision language model would place a second reader in every clinic, and the past two years have produced a steady stream of 3D CT models that write reports and answer questions about a scan (Hamamci et al., 2024; 2026; Bai et al., 2024; Wu et al., 2025; Blankemeier et al., 2024; Xin et al., 2026). These models share one recipe, the visual instruction tuning introduced for natural images (Liu et al., 2023). A pretrained 3D image encoder turns the volume into a set of visual tokens, a projector maps those tokens into a large language model, and the model is fine-tuned to produce text. The encoder strongly shapes the final quality, since the pretrained representation a model starts from is a major driver of how well it transfers to a downstream task (Kornblith et al., 2019; Ericsson et al., 2021; Li et al., 2024). Today the encoder is chosen by fine-tuning every candidate through the full model and comparing downstream scores, a search costly enough to dominate the research cycle. A cheap probe trains a lightweight readout on a representation to measure what it encodes (Alain & Bengio, 2018). It offers a way to select a 3D CT encoder without full vision-language fine-tuning. In neural architecture search, low-cost proxies predict how candidate networks rank after full training (Abdelfattah et al., 2021). Whether a representation probe can similarly predict downstream performance after full fine-tuning in 3D CT vision language models remains unexplored.

Report generation is a widely studied task (Hamamci et al., 2024), and in 3D CT vision language modeling it is the one that draws on a model's capabilities most comprehensively (Wang et al., 2025). But a report is scored by a single number that reflects mostly disease, so against it alone a probe can be judged only in aggregate. The capabilities we mean are the five a radiologist reads separately: size, density, location, texture, and disease (Aerts et al., 2014). To resolve the probe's forecast into these, we build MeasureVQA, a dataset whose answers are measured from segmentation masks and

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

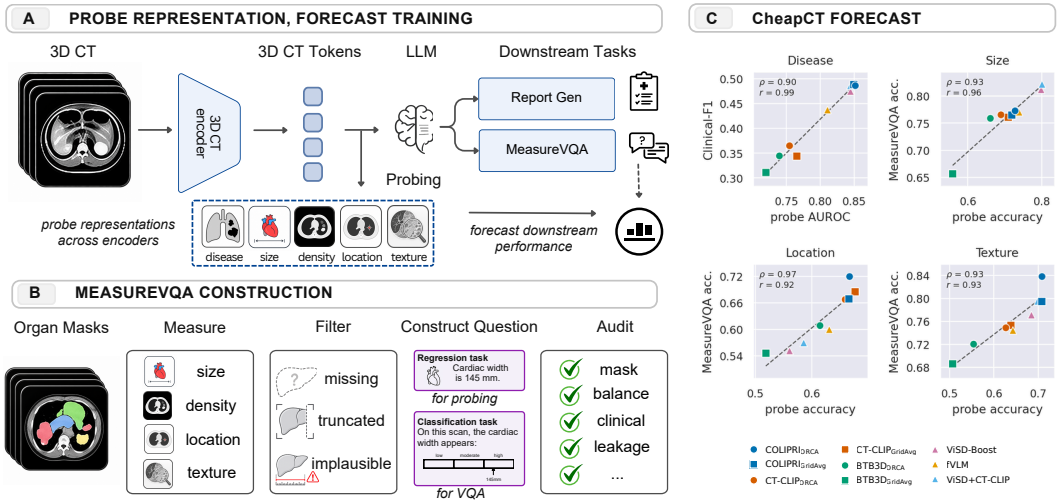


Figure 1: **(A)** A language model is fine-tuned on 3D-CT encoder tokens. A cheap probe reads the same representation and forecasts that outcome without the fine-tuning. **(B)** MeasureVQA construction, from organ masks and Hounsfield units. **(C)** Per capability, the probe score against the downstream score it forecasts across the encoders, on CT-RATE with Llama-3.1-8B. The probe **CheapCT** forecasts the expensive outcome.

Hounsfield units. Each question targets one capability, so a density probe is graded against a density question. We measure the probe against report generation and MeasureVQA in parallel.

We propose **CheapCT**, a cheap probe that forecasts expensive training. We test it at scale, probing nine encoder variants on CT-RATE (Hamamci et al., 2026) and a range of readout configurations, each probe trained in minutes. We carry the same variants through full fine-tuning on four language-model backbones. We then extend the same test to a second corpus, Merlin (Blankemeier et al., 2024). The forecast holds across a wide band of readouts, and CheapCT is the position-aware one, strong across the board and the only kind that recovers spatial location, which a position-blind probe misses. We make three contributions:

- We give the first evidence that a cheap probe on a pretrained encoder's representation can forecast the outcome of expensive 3D CT vision language fine-tuning, one capability at a time, robust to the choice of probe readout and language-model backbone.
- We build and release MeasureVQA, a 3D CT VQA whose answers are measured from segmentation masks and Hounsfield units and split into five clinical capabilities, each audited end to end.
- Used as an encoder-selection rule, CheapCT (choosing an encoder by its probe score) recovers most of what fine-tuning every candidate would find at a fraction of the cost, with controls that keep the claim falsifiable.

2 Related Work

3D-CT vision language models. 3D-CT vision language models are built on a pretrained image encoder, and encoders now span every pretraining paradigm, self-supervised (Zhou et al., 2021; Wu et al., 2024; Pai et al., 2025), supervised (Li et al., 2024), and contrastive image-text (Hamamci et al., 2026; Blankemeier et al., 2024). A growing set of models fine-tune such an encoder to write reports and answer questions (Wu et al., 2025; Bai et al., 2024; Hamamci et al., 2024; Xin et al., 2026), so which encoder to build on is a consequential choice, yet an encoder's downstream value can be judged today only by fine-tuning a full model on it. A separate line grounds these outputs in anatomy, as in RadGenome-Chest CT (Zhang et al., 2024), whose region labels come from reports. However, we measure our answers from the images instead, so each answer is a physical measurement rather than a report annotation.

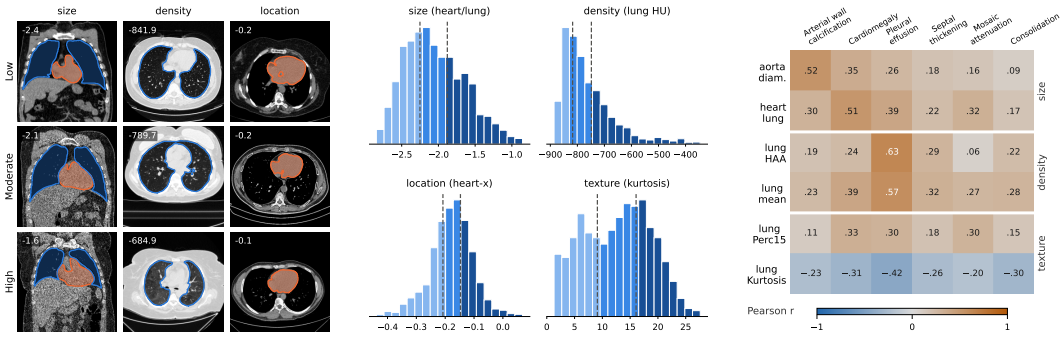
Representation probing. A linear or attention probe reads a task off a representation and is a standard, inexpensive gauge of what that representation encodes (Alain & Bengio, 2018). For model selection, probe and upstream accuracy correlate with full fine-tuning transfer (Kornblith et al., 2019), though the correlation weakens for self-supervised models and off-distribution tasks (Ericsson et al., 2021; Baharoon et al., 2024). A related line forecasts transfer without any fine-tuning, scoring a model from its features alone (Nguyen et al., 2020; You et al., 2021). We bring this idea to 3D-CT text generation and, unlike prior work, test when a cheap probe forecasts training one capability at a time rather than assume it transfers.

Performance prediction in AutoML. Low-cost proxies rank candidate networks in a single pass, without training each one (Abdelfattah et al., 2021; Mellor et al., 2021), and scaling studies predict a large model's behavior from a small proxy run. Closest to us, probing predicts fine-tuning performance in NLP (Zhu et al., 2022). We bring this proxy role to encoder selection for 3D-CT vision language models, and add a capability-resolved test that survives a change of readout and of backbone.

3 The MeasureVQA Dataset

3.1 Design principles

A generated report collapses many kinds of visual information into one aggregate score that reflects mostly disease, so it cannot separately show how well an encoder captures finer, capability-level information. We organize this information into five capabilities that a radiologist assesses separately. They are size, density, location, texture, and disease. We build a dataset that scores each on its own. We define size through organ volumes, diameters, and proportions, density through tissue attenuation, location through relative spatial relationships, and texture through the shape of intensity distributions. Within each capability, guided by the quantitative-CT literature and clinical measurement standards, we select attributes that are reliably measurable, clinically meaningful, and well separated from one another. All measurements are computed from organ masks and the Hounsfield values within them, not from reports or a language model. Disease labels, in contrast, come from the reports. These clinically distinct but related dimensions organize the attributes and questions of the dataset.



(a) Measured from masks (b) Balanced classes (c) Clinical validity

Figure 2: **(a)** CT-RATE slices per class, with the organ mask and the measured value. **(b)** Per-attribute value distributions by class, dashed lines the frozen tertile cutpoints. **(c)** Signed Pearson r between each measurement and the abnormalities.

3.2 Dataset construction

MeasureVQA is constructed from two public 3D CT datasets. CT-RATE provides non-contrast chest CT and Merlin contrast-enhanced abdominal CT. The two sources cover different body regions and acquisition conditions, allowing the same types of clinical capability to be constructed in the chest and abdomen. We follow their existing training and validation splits. Figure 1(B) sketches the construction pipeline, from organ masks to audited questions. For each CT volume we use TotalSegmentator (Wasserthal et al., 2023) organ masks, from which we extract continuous

measurements using mask geometry and the Hounsfield values within the masks. These include organ volumes and ratios, anatomical diameters, normalized positions, tissue attenuation, intensity percentiles, and distribution shape. We then filter unreliable values caused by missing, truncated, or erroneous segmentations using organ presence, physical mask size, Hounsfield composition, and anatomical plausibility. These rules use neither disease labels nor the extremity of a value in the population distribution. Filtering is applied per attribute, so an unreliable organ removes only its own measurements rather than the whole volume. The resulting filtered measurements form the basis for the audit and the questions. Figure 2(a) shows a representative slice per class, the value read off the mask.

3.3 Dataset audit

We audit the dataset with statistical and analytic checks on the masks, measurements, and reports, independent of the probes and encoders we evaluate. The masks are reliable. Each organ is present in about 98% of CT-RATE volumes and over 98% of Merlin volumes, and guards withhold rather than corrupt the rare failures. Repeat measurements agree closely, for example a test-retest of 0.969 for vertebral bone density. Figure 2(b) plots the value distributions against their frozen cutpoints. The classes are balanced, every tertile holding between 0.30 and 0.39 of the mass and the reference letter uniform to within a third of a percentage point, so a baseline that always answers the most frequent letter scores at chance.

Across capabilities, each measurement's extreme decile is enriched for the disease it implies. Cardiomegaly shows an $11\times$ lift and splenomegaly a $70\times$ lift. This is convergent evidence from the masks and the reports. Figure 2(c) shows this per measurement, texture the exception because disease flattens its intensity histogram. For shortcuts, a size proxy built from raw organ voxel count stays near its baseline on most density, location, and texture attributes, so organ scale alone does not explain them, and the few measurable couplings are flagged. The full audit, covering mask quality, thresholds and balance, and shortcuts, is in Appendix B.

Table 1: MeasureVQA statistics by corpus and capability. Reliability is the reproducibility of a measurement across reconstructions of the same scan. Question counts combine training and validation and exclude open-value questions. Chance is the random-guessing accuracy.

Corpus	Capability	Attributes	Volumes	Coverage (%)	Reliability	Questions	Avg. words	Chance
CT-RATE	Size	5	49,371	98.4	0.965	246,504	9.0	0.367
	Density	4	49,411	98.5	0.956	195,892	9.8	0.374
	Location	3	49,382	98.4	0.949	148,001	10.0	0.333
	Texture	3	43,359	86.4	0.592	130,075	11.0	0.389
	Disease	18	50,188	100.0	—	903,384	7.5	0.500
Merlin	Size	3	20,356	100.0	—	60,060	8.7	0.444
	Density	3	20,360	100.0	—	60,409	9.0	0.388
	Location	2	19,977	98.1	—	39,954	13.0	0.333
	Disease	30	20,189	99.2	—	136,052	7.1	0.500
Total		**71**	**70,548**			**1,920,331**		

3.4 VQA and statistics

After auditing, each measurement is released in two forms that share one underlying value, an open-value question that requests the exact number and a discretized letter question. The letter question is built by cutting most measurements into low, moderate, and high tertiles and a few into two classes, a clinical threshold replacing the tertile only where the measured quantity matches the guideline and its split stays balanced enough. The cutpoints are frozen on the training split and applied unchanged to validation, and the label is then rendered as a multiple-choice question from a fixed template with options shuffled so answer position carries no signal. Disease is asked instead as a presence question with report-derived labels. Because both forms come from the same number, a question and its value are matched by construction. MeasureVQA draws on $47{,}149$ training and $3{,}039$ validation volumes from CT-RATE and $15{,}306$ and $5{,}054$ from Merlin, posing about 1.92M letter-scored questions and a further 0.45M open-value questions, summarised in Table 1.

4 METHOD

4.1 THE REPRESENTATION PROBE

We propose a lightweight probe that measures how much of a capability an encoder's representation carries. The probe takes the same visual tokens the language model would receive, pools them into a single vector, and predicts that capability's attributes, a regression for the continuous measurements and a classification for disease. It never trains the language model. A small readout, a linear pool, a gated attention, or a small transformer, trains in minutes on a single GPU. We vary the pooling mechanism, its capacity, and the input normalization across a family of readouts. The headline, which we call CheapCT, is a position-aware attention pool and the strongest all-round reader in the family. Each probe reports one score per capability on the validation split. The full readout family and its training protocol are in Appendix C.3.

4.2 THE TESTBED

Encoders. The encoder is the axis we vary. We compare six pretrained 3D-CT encoders on CT-RATE, spanning the pretraining paradigms in use today. They are the contrastive image-text encoders COLIPRI (Wald et al., 2026) and CT-CLIP (Hamamci et al., 2026), the discrete visual tokenizer BTB3D (Hamamci et al., 2025), and the organ-aware encoders ViSD (Cao et al., 2025), fVLM (Shui et al., 2025), and a combination of ViSD and CT-CLIP (Liang et al., 2026a). A dense grid carries more tokens than a language model can accept, so we reduce it to a common budget by one of two standard aggregations, an organ-centroid pooling (ORCA) (Liang et al., 2026b) or a plain grid average (GridAvg), and take each (encoder, aggregation) pair as a representation of its own. Encoders that already emit organ-level tokens enter directly. This yields nine representation variants, the points our correlations run over.

Downstream tasks. The expensive outcome is a language model fine-tuned on an encoder to produce text. Report generation draws on every capability at once but scores mostly disease, and MeasureVQA supplies the four measurement capabilities one question at a time, so the two image-grounded tasks together expose all five.

Backbones. Because the outcome could depend on the language model as much as on the encoder, we run four backbones, Llama-3.1-8B, the larger Llama-3.3-70B, Gemma-3-27B, and its medically pretrained sibling MedGemma-27B, covering different model sizes and degrees of domain specialization. Each encoder is attached to a backbone by the same lightweight projector and adapter, trained in stages and held fixed so that from one encoder to the next only the encoder changes. Every encoder is carried through full fine-tuning on every backbone, so each backbone is an independent test of whether the one probe score forecasts training.

4.3 MEASURING FORECAST AGREEMENT

For a single encoder, each capability gives a probe score, AUROC for disease and R^2 (variance explained) for the four continuous measurements, set against the downstream score for the same capability. Disease is scored on report generation by CE-F1 (clinical-efficacy F1) (Yan et al., 2022) and GREEN (Ostmeier et al., 2024), and each measurement capability by the accuracy of its MeasureVQA questions. The probe predicts the very value each measurement question is built from, so the two are matched by construction.

Forecast agreement comes from varying the encoder. Holding a capability fixed, each encoder gives a probe score and a paired downstream score, and we correlate the two across the encoders. We summarise each pairing with three standard correlation coefficients, all running from -1 to 1. Spearman ρ (Spearman, 1987) is the rank correlation, high when the probe orders the encoders as training does. Pearson r (Pearson, 1895) is the raw-score correlation, high when probe and downstream scores line up tightly enough to forecast from. Weighted Kendall τ (Vigna, 2015) is a rank correlation that weights the top of the order, the region model selection cares about, since a practitioner fine-tunes only the encoders the probe ranks first. Where these are high the cheap probe stands in for the expensive run. We count a capability as forecastable only where this holds on each of the four backbones, whose downstream scores the probe never sees, and survives the controls that rule out artifacts.

	Report generation					MeasureVQA							
	Disease					Size		Density		Location		Texture	
Encoder	CE-F1	GREEN	BLEU-4	ROUGE	CheapCT	LLM	CheapCT	LLM	CheapCT	LLM	CheapCT	LLM	CheapCT
Noise floor	0.20	0.12	0.08	0.18	–	0.45	–	0.40	–	0.33	–	0.52	–
Shuffle floor	0.28	0.29	0.18	0.29	–	0.45	–	0.38	–	0.33	–	0.50	–
$\text{BTB3D}_{\text{GridAvg}}$	0.31	0.25	0.19	0.31	0.72	0.66	0.56	0.65	0.52	0.55	0.52	0.69	0.51
$\text{BTB3D}_{\text{ORCA}}$	0.34	0.30	0.20	0.31	0.74	0.76	0.66	0.70	0.53	0.61	0.61	0.72	0.55
$\text{COLIPRI}_{\text{GridAvg}}$	0.49	0.33	0.21	0.32	0.85	0.77	0.72	0.81	0.74	0.67	0.67	0.79	0.71
$\text{COLIPRI}_{\text{ORCA}}$	0.49	0.35	0.20	0.32	0.85	0.77	0.73	0.84	0.75	0.72	0.67	0.84	0.71
$\text{CT-CLIP}_{\text{GridAvg}}$	0.34	0.31	0.20	0.32	0.77	0.76	0.71	0.73	0.58	0.68	0.68	0.75	0.64
$\text{CT-CLIP}_{\text{ORCA}}$	0.37	0.30	0.19	0.31	0.75	0.77	0.69	0.71	0.57	0.67	0.66	0.75	0.63
fVLM	0.44	0.32	0.21	0.32	0.81	0.77	0.74	0.72	0.67	0.60	0.63	0.74	0.64
ViSD-Boost	0.47	0.31	0.21	0.32	0.84	0.81	0.80	0.78	0.71	0.55	0.56	0.77	0.68
ViSD+CT-CLIP	0.49	0.32	0.21	0.33	0.84	0.82	0.80	0.81	0.70	0.57	0.59	0.79	0.70
Spearman ρ	0.90	0.97	0.62	0.60	–	0.93	–	0.93	–	0.97	–	0.93	–
Pearson r	0.99	0.83	0.73	0.76	–	0.96	–	0.91	–	0.92	–	0.93	–
Weighted Kendall τ	0.73	0.94	0.32	0.32	–	0.89	–	0.84	–	0.83	–	0.87	–

Table 2: The downstream LLM score beside the CheapCT probe score on CT-RATE with Llama-3.1-8B. The CheapCT probe score is AUROC for disease and tertile accuracy for the measurements. The two *floor* rows are averaged over the encoders. CheapCT has no floor.

5 Experiments

5.1 Forecast agreement

Table 2 reports the downstream score each of the nine encoders reaches after fine-tuning on CT-RATE with Llama-3.1-8B, beside its CheapCT score. For each capability, we correlate the two across the encoders as a rank and a linear agreement, Spearman ρ and Pearson r, plotted in Figure 1. Training curves are in Appendix D.4.

Report generation. CE-F1 separates the encoders widely. CheapCT tracks it tightly, at $\rho = 0.90$ and $r = 0.99$, the top three encoders falling within 0.003, which holds ρ just below r. GREEN judges the whole report, well beyond the eighteen disease labels, yet CheapCT forecasts it about as tightly as CE-F1, at $\rho = 0.97$ and $r = 0.83$. The surface n-gram metrics, BLEU-4 and ROUGE-L, score word overlap instead. They stay nearly flat across the encoders, so they carry no forecast and measure fluency rather than clinical correctness.

MeasureVQA. Size, density, and texture are forecast at $\rho = 0.93$, and location, the spatial capability, at $\rho = 0.97$. The answer is a single letter, so the accuracy measures the capability directly, with little generation noise between the encoder and the score. CheapCT predicts the very measurement each question is built from, so the probe and downstream share one target.

Selection quality. We also frame the forecast in the terms model selection uses. The Weighted Kendall τ in Table 2 stays high on the measurements and dips on disease, where the leading encoders tie, and a permutation test rejects chance at $p < 0.001$, with a wide interval at nine encoders. Appendix D.1 gives the credibility checks. Choosing the encoder CheapCT ranks first recovers 95 to 100% of the best downstream score, the exact best on three of the five capabilities. CheapCT, trained in minutes, thus forecasts the expensive outcome well enough to select an encoder with.

5.2 Across backbones

CheapCT forecasts fine-tuning across language-model backbones. We carry the same nine encoders through full fine-tuning on four backbones and evaluate the forecast. Figure 3 gives the cross-encoder ρ per backbone. On report generation CheapCT tracks CE-F1 on all four backbones, at $\rho = 0.83$ to 0.93. On MeasureVQA it tracks the measurements on both backbones we fine-tune it for, from $\rho = 0.70$ on location under MedGemma to $\rho = 0.98$ on texture.

The backbones differ in what they do best. Llama-3.1-8B is strongest on CE-F1, marginally over Llama-3.3-70B, and the two Gemma models trail. Llama-3.3-70B is the clear best on GREEN. The split fits their sizes. The small model already reads disease, which CE-F1 scores. The large model's broader capability shows on GREEN, which judges the whole report. CheapCT tracks GREEN on

three of the four backbones, from $\rho = 0.77$ to 0.97. On Gemma-3-27B it is weaker, at $\rho = 0.50$. Gemma is a general instruct model, and its reports arrive with conversational preambles and non-radiology formatting. We regenerate them with an explicit style instruction before scoring. We attribute this to its instruct prior leaking into the output, strong enough to shape the report in ways the holistic GREEN judge penalizes, even though the clinical labels CE-F1 scores stay well ordered.

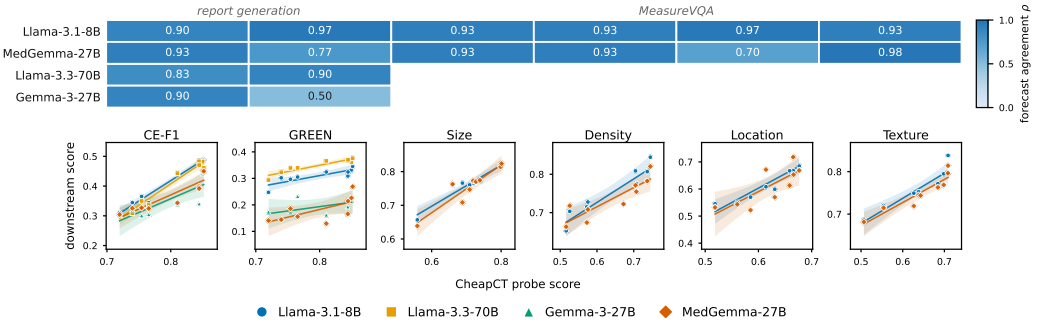
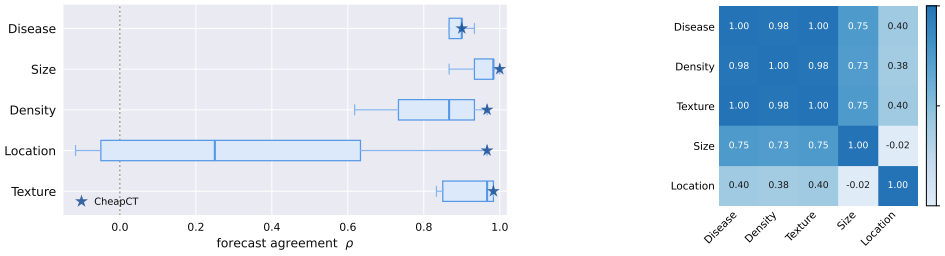


Figure 3: Cross-backbone forecast agreement on CT-RATE. Top, ρ per backbone and target. Bottom, CheapCT against the downstream score per capability, one point per encoder.

5.3 Robustness to the readout

The forecast does not depend on the readout hyperparameters. We sweep the readout over a band of thirteen choices, varying its capacity, its pooling type, and the input normalization, shown in Figure 4a. For four of the five capabilities the verdict does not move. Size, density, texture, and disease stay forecast across the band, their correlation never falling below $\rho = 0.57$. For these the forecast is a property of the representation, not of the probe that reads it. Location is the exception, and an informative one. It is near $\rho = 0$ for any permutation-invariant pool, a mean, a max, or a top-k, and it climbs to $\rho = 0.96$ once the readout is position-aware. Location is a spatial relation, and the fine-tuned model can read it because it consumes the tokens as an ordered sequence. A permutation-invariant probe discards exactly that position signal, so it cannot forecast what the model does with location, while a position-aware probe keeps it and can. This is why CheapCT is the position-aware pool, the strongest all-round probe in the family and best or near-best on every capability. It also stays at a moderate capacity. The highest-capacity readouts, a four-layer transformer and the dual pool, extract less rather than more, the signature of a probe fitting the task instead of the representation.



(a) Readout robustness

(b) Capability coupling

Figure 4: (a) For each capability, a box spanning the forecast agreement ρ over the thirteen readouts. (b) The cross-encoder Spearman ρ between the CheapCT probe scores of each capability pair, on the Llama-3.1-8B backbone.

5.4 Capability coupling

An encoder that reads one capability well tends to capture the others well. The tendency is not uniform. Figure 4b correlates the CheapCT probe scores across the nine encoders, one capability

pair at a time. Disease, density, and texture move together almost perfectly, ρ above 0.98, and size tracks them at $\rho \approx 0.75$. These four share one axis of encoder quality. Location does not join it. Its rank is near-independent of the rest, $\rho = 0.40$ with disease and -0.02 with size, and a principal-components analysis agrees, one component for the shared axis and a second, about a fifth of the variance, for location alone. The downstream scores give the same two-axis structure. This is a second sense in which location stands apart. A different pool is needed to extract it, and a different property of the encoder carries it, so a single quality score would rank it wrong. The forecast has to be made one capability at a time.

5.5 The floor control

A probe forecasts a downstream task only to the extent that the fine-tuned model itself reads the image. We re-score each trained model with its CT tokens replaced two ways. *Noise* substitutes per-channel Gaussian tokens matched to the real tokens' mean and variance but without spatial structure. *Shuffle* substitutes the real tokens of a randomly paired volume, a valid scan for the wrong patient. The score that survives is what the task solves without the image, and the gap to the real score is the image gain.

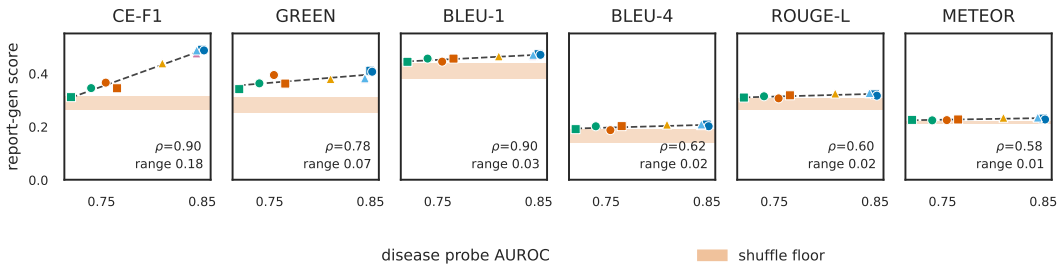


Figure 5: Each report-generation metric against the CheapCT disease probe, on CT-RATE with Llama-3.1-8B.

For report generation the metrics split. The clinical ones collapse. CE-F1 drops from 0.42 with the real image to 0.20 under noise and 0.28 under shuffle, since the report is specific to the patient and cannot be produced without reading the scan, and the clinical judge GREEN moves the same way. This gap is an average, wide for the strong encoders and narrow for the weakest, whose report barely beats a mismatched scan. The surface n-gram metrics do not collapse. BLEU-4 and ROUGE-L score a report from a mismatched scan almost as high as the real one, and across the encoders they span two points where CE-F1 spans eighteen, plotted in Figure 5. A metric that cannot tell a correct report from a plausible wrong one measures fluency, not clinical correctness. We treat these as a fluency floor and rank clinical quality by CE-F1, GREEN, and CheapCT.

MeasureVQA collapses as sharply, each capability falling to the noise and shuffle floors in Table 2. Its question is identical across volumes, so the only thing separating a right answer from a wrong one is the measurement in the scan. The collapse is also a check on the benchmark. Shuffled options and image-only answers leave a mismatched volume no textual shortcut, and the fall to the floor confirms none was left open. Report generation and MeasureVQA both read the image, so the correlations reported above rest on tasks that genuinely use the scan. Per-encoder floors are in Appendix D.2.

5.6 Generality to Merlin Dataset

We also validate the finding on Merlin, a contrast-enhanced abdominal corpus, with four encoder cells. That is too few for a stable correlation, so Figure 6 shows the rank pattern. The cells are sorted by CheapCT. CheapCT tracks density on both backbones, and the gain over the shuffle floor stays wide. Location is the exception, where some cells barely train and CheapCT over-predicts them. Size already sits high on the abdomen for both backbones, so it does not separate. GREEN and RadGraph sit just above their shuffle floor in a narrow band, so their ranking says little.



Figure 6: Forecast agreement on Merlin. Rows are encoder cells sorted by CheapCT. Each capability gives the CheapCT score beside the fine-tuned score on Llama-8B and MedGemma, the shuffle floor beneath.

5.7 Cost

On a single B200 GPU, a CheapCT probe trains in well under a minute, 14 seconds on $\text{COLIPRI}_{\text{ORCA}}$ and 49 on ViSD+CT-CLIP. Fine-tuning and evaluating one language model on the same encoder takes hours. Report generation runs 10.7 GPU-hours on $\text{COLIPRI}_{\text{ORCA}}$ and 14.9 on ViSD+CT-CLIP, MeasureVQA 9.3 and 6.6. Either training costs hundreds to thousands of times more than the CheapCT probe that forecasts it.

6 Discussion

Clinical evaluation. CE-F1 scores a report against a fixed, limited set of abnormality labels. GREEN judges more of the report and reflects a broader clinical signal. CheapCT forecasts the narrow, label-based metric consistently. The broad judge is a noisier target and harder to forecast reliably. Making that forecast dependable is left to future work.

The encoder ceiling. The encoder is the high-leverage choice. Across the four backbones the absolute scores move, but the encoder ranking CheapCT recovers does not. What a 3D-CT vision language model can do is set largely by the encoder it starts from, not by the decoder placed on top. The encoder appears to bound what the system can reach, so building and improving encoders should return more than scaling the language model. CheapCT makes that selection cheap.

From encoder to decoder. CheapCT is a forecast, not a replacement for fine-tuning. It is a cheap first pass that narrows the encoder candidates, and its pick is still confirmed by one fine-tune. It also sees only the encoder. The language model has no such probe, so choosing the decoder still needs a full fine-tune, as our four-backbone comparison did. A natural next step carries the same cheap-proxy idea to the decoder. The goal is to forecast which language model or fine-tuning recipe wins without training each one, as AutoML forecasts trained accuracy from a cheap score.

7 Conclusion

We set out to test whether a cheap probe on a pretrained 3D-CT encoder forecasts the outcome of expensive language-model fine-tuning, and we built MeasureVQA to answer it one capability at a time. Across five clinical capabilities a probe that trains in minutes predicts what full fine-tuning achieves, at orders of magnitude less compute, and it holds across language models and corpora. Spatial location is forecast by a position-aware readout. The forecast holds only where the fine-tuned model actually reads the image, a control we make explicit, so a practitioner can tell before paying whether to trust it. Together, MeasureVQA and CheapCT reframe encoder choice for 3D-CT vision language models as a cheap measurement rather than an expensive search.

AI USE STATEMENT

We used AI tools to help implement the experiment code. We also used AI tools for brainstorming and for language editing that improves readability. We did not use generative AI to design the methodology or the experiments. We did not use it to interpret the results. We did not use it for the literature review or citation selection. The authors reviewed and verified all AI-assisted code and text. We take full responsibility for the final content.

REFERENCES

Mohamed S. Abdelfattah, Abhinav Mehrotra, Lukasz Dudziak, and Nicholas Donald Lane. Zero-cost proxies for lightweight nas. *ArXiv*, abs/2101.08134, 2021. URL https://api.semanticscholar.org/CorpusID:231648113.

Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):4006, 2014.

Arthur S. Agatston, Warren R. Janowitz, Frank J. Hildner, Noel R. Zusmer, Manuel Viamonte, and Robert Detrano. Quantification of coronary artery calcium using ultrafast computed tomography. *Journal of the American College of Cardiology*, 15(4):827–832, 1990. ISSN 0735-1097. doi: 10.1016/0735-1097(90)90282-t. URL https://www.sciencedirect.com/science/article/pii/073510979090282T.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644. Presented at ICLR 2017 Workshop.

Mohammed Baharoon, Waseem Qureshi, Jiahong Ouyang, Yanwu Xu, Abdulrhman Aljouie, and Wei Peng. Evaluating general purpose vision foundation models for medical image analysis: An experimental study of DINOv2 on radiology benchmarks, 2024. URL https://arxiv.org/abs/2312.02366.

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, 2024. URL https://arxiv.org/abs/2404.00578.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

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*, pp. rs–3, 2024.

Cody J. Boyce, Perry J. Pickhardt, David H. Kim, Andrew J. Taylor, Thomas C. Winter, Richard J. Bruce, Mary J. Lindstrom, and J. Louis Hinshaw. Hepatic steatosis (fatty liver disease) in asymptomatic adults identified by unenhanced low-dose CT. *American Journal of Roentgenology*, 194 (3):623–628, 2010. doi: 10.2214/AJR.09.2590. URL https://doi.org/10.2214/AJR.09.2590. PMID: 20173137.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23041–23050, 2025. doi: 10.1109/ICCV51701.2025.02139.

Elliot L. Chaikof, Ronald L. Dalman, Mark K. Eskandari, Benjamin M. Jackson, W. Anthony Lee, M. Ashraf Mansour, Tara M. Mastracci, Matthew Mell, M. Hassan Murad, Louis L. Nguyen, Gustavo S. Oderich, Madhukar S. Patel, Marc L. Schermerhorn, and Benjamin W. Starnes. The society for vascular surgery practice guidelines on the care of patients with an abdominal aortic aneurysm. *Journal of Vascular Surgery*, 67(1):2–77.e2, 2018. doi: 10.1016/j.jvs.2017.10.044. URL https://doi.org/10.1016/j.jvs.2017.10.044.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410–5419, 2021. doi: 10.1109/CVPR46437.2021.00537.

Pierre Alain Gevenois, Viviane de Maertelaer, Paul De Vuyst, Jacques Zanen, and Jean-Claude Yernault. Comparison of computed density and macroscopic morphometry in pulmonary emphysema. *American Journal of Respiratory and Critical Care Medicine*, 152(2):653–657, 1995. doi: 10.1164/ajrccm.152.2.7633722.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated radiology report generation for 3D medical imaging. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 476–486, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72390-2.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 135074–135102. Curran Associates, Inc., 2025. doi: 10.52202/085713-4507. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c4861cc81b768dc36a676fceed8f7672-Paper-Conference.pdf.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, 2026. doi: 10.1038/s41551-025-01599-y. URL https://doi.org/10.1038/s41551-025-01599-y.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Eric M. Isselbacher et al. 2022 ACC/AHA guideline for the diagnosis and management of aortic disease. *Circulation*, 146(24):e334–e482, 2022. doi: 10.1161/CIR.0000000000001106.

Samuel Jang, Peter M. Graffy, Timothy J. Ziemlewicz, Scott J. Lee, Ronald M. Summers, and Perry J. Pickhardt. Opportunistic osteoporosis screening at routine abdominal and thoracic CT: Normative L1 trabecular attenuation values in more than 20 000 adults. *Radiology*, 291(2):360–367, May 2019. ISSN 0033-8419. doi: 10.1148/radiol.2019181648. URL https://doi.org/10.1148/radiol.2019181648.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2019. doi: 10.1109/CVPR.2019.00277.

Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised 3D models transfer to medical imaging tasks? In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun

(eds.), *International Conference on Learning Representations*, volume 2024, pp. 16389–16404, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/47360925c45a166c96f652589265dee9-Paper-Conference.pdf.

Renjie Liang, Yiling Ma, Yang Xing, Zhengkang Fan, Jinqian Pan, Chengkun Sun, Li Li, Kuang Gong, and Jie Xu. Beyond the embedding bottleneck: Adaptive retrieval-augmented 3D CT report generation, 2026a. URL https://arxiv.org/abs/2603.15822.

Renjie Liang, Zijian Xu, Jinqian Pan, Chengkun Sun, Zhengkang Fan, Shawn Li, You Qin, Mei Liu, and Jie Xu. ORCA: ORgan-centroid aggregation for training-free 3D CT visual token compression, 2026b. URL https://arxiv.org/abs/2608.00345.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516.

David A. Lynch, John H. M. Austin, James C. Hogg, Philippe A. Grenier, Hans-Ulrich Kauczor, Alexander A. Bankier, R. Graham Barr, Thomas V. Colby, Jeffrey R. Galvin, Pierre Alain Gevenois, Harvey O. Coxson, Eric A. Hoffman, John D. Newell, Massimo Pistolesi, Edwin K. Silverman, and James D. Crapo. CT-definable subtypes of chronic obstructive pulmonary disease: A statement of the fleischner society. *Radiology*, 277(1):192–205, 2015. doi: 10.1148/radiol.2015141579. URL https://doi.org/10.1148/radiol.2015141579.

Joe Mellor, Jack Turner, {Amos J} Storkey, and {Elliot J} Crowley. Neural architecture search without training. July 2021. doi: 10.48550/arXiv.2006.04647. URL https://icml.cc/. Thirty-eighth International Conference on Machine Learning, ICML 2021 ; Conference date: 18-07-2021 Through 24-07-2021.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/nguyen20b.html.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision foundation models for computed tomography, 2025. URL https://arxiv.org/abs/2501.09001.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. ISSN 03701662. URL http://www.jstor.org/stable/115794.

Perry J. Pickhardt, B. Dustin Pooler, Travis Lauder, Alejandro Munoz del Rio, Richard J. Bruce, and Neil Binkley. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Annals of Internal Medicine*, 158(8):588–595, 2013. doi: 10.7326/0003-4819-158-8-201304160-00003.

P. Prassopoulos, M. Daskalogiannaki, M. Raissaki, A. Hatjidakis, and N. Gourtsoyiannis. Determination of normal splenic volume on computed tomography in relation to age, gender and body habitus. *European Radiology*, 7(2):246–248, 1997. doi: 10.1007/s003300050145. URL https://doi.org/10.1007/s003300050145.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2026. URL https://arxiv.org/abs/2507.05201.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and fine-grained vision-language pretraining for enhanced CT image understanding. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 24094–24107, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3c3b120aae631df27fcd6f94e26615c0-Paper-Conference.pdf.

Rebecca Smith-Bindman, Philip W Chu, Hana Azman Firdaus, Carly Stewart, Matthew Malekhedayat, Susan Alber, Wesley E Bolch, Malini Mahendra, Amy Berrington de González, and Diana L Miglioretti. Projected lifetime cancer risks from current computed tomography imaging. *JAMA Internal Medicine*, 185(6):710–719, 2025. doi: 10.1001/jamainternmed.2025.0505. URL https://doi.org/10.1001/jamainternmed.2025.0505.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987. ISSN 00029556. URL http://www.jstor.org/stable/1422689.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian

Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Krystian Truszkiewicz, Rafał Poręba, and Paweł Gać. Radiological cardiothoracic ratio in evidence-based medicine. *Journal of Clinical Medicine*, 10(9):2016, 2021. ISSN 2077-0383. doi: 10.3390/jcm10092016. URL https://www.mdpi.com/2077-0383/10/9/2016.

Sebastiano Vigna. A weighted correlation index for rankings with ties. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pp. 1166–1176, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741088. URL https://doi.org/10.1145/2736277.2741088.

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel C. F. Codella, Maria Teodora Wetscherek, Klaus H. Maier-Hein, Panagiotis Korfiatis, Valentina Salvatelli, Javier Alvarez-Valle, and Fernando Pérez-García. Comprehensive language-image pre-training for 3D medical image understanding, 2026. URL https://arxiv.org/abs/2510.15042.

Xinyi Wang, Grazziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. A survey of deep-learning-based radiology report generation using multimodal inputs. *Medical Image Analysis*, 103:103627, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103627. URL https://www.sciencedirect.com/science/article/pii/S1361841525001744.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023. doi: 10.1148/ryai.230024. URL https://doi.org/10.1148/ryai.230024.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D and 3D medical data. *Nature Communications*, 16:7866, 2025. doi: 10.1038/s41467-025-62385-7. URL https://doi.org/10.1038/s41467-025-62385-7.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. VoCo: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22873–22882, 2024. doi: 10.1109/CVPR52733.2024.02158.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3DVLM: An efficient vision-language model for 3D medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 30 (3):2524–2536, 2026. doi: 10.1109/JBHI.2025.3604595.

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. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022. doi: 10.1148/ryai.210258. URL https://doi.org/10.1148/ryai.210258.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. RadGenome-Chest CT: A grounded vision-language dataset for chest CT analysis, 2024. URL https://arxiv.org/abs/2404.16754.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Models Genesis. *Medical Image Analysis*, 67:101840, jan 2021. doi: 10.1016/j.media.2020.101840.

Zining Zhu, Soroosh Shahtalebi, and Frank Rudzicz. Predicting fine-tuning performance with probing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11534–11547, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.793. URL https://aclanthology.org/2022.emnlp-main.793/.