

VOLUMETRIC MEDICAL IMAGING

A foundation model approach to chest CT classification

A single, reusable volumetric representation learned from chest CT studies paired with their radiology reports, **specialized to a panel of clinically relevant findings**

AUTHOR

Neurologic AI

PUBLISHED

May 2026

MODALITY

Chest CT

ABSTRACT

Chest computed tomography (CT) is one of the most information-dense examinations in radiology: a single study can comprise hundreds of cross-sectional slices, each of which must be reviewed for dozens of possible findings ranging from incidental lung nodules to acute, time-sensitive conditions such as

pulmonary embolism. This paper describes a foundation-model approach to chest CT classification that learns a single, reusable volumetric representation from a large corpus of real-world chest CT studies paired with their radiology reports, and then specializes that representation to a panel of clinically relevant findings through a two-stage training pipeline. We detail the composition and curation of a dataset of approximately 640,000 chest CT studies (over 81.5 million total images) drawn from outpatient imaging centers across multiple U.S. regions, and describe a training methodology comprising self-supervised, contrastive vision-language pretraining on 3D CT volumes and paired radiology reports, followed by per-finding binary classification fine-tuning. We further present the resulting model's validation performance across 13 clinically relevant findings, measured by area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (PR-AUC). The results indicate that a volumetric, report-grounded pretraining strategy yields strong discrimination even for findings with very low positive prevalence in the training data, supporting the case for a single reusable 3D backbone that can be adapted efficiently to new chest CT findings.

01

Introduction

Chest CT is substantially more information-dense than planar chest X-ray: a single study routinely contains hundreds of thin cross-sectional slices spanning the entire thorax, and a radiologist must scroll through this full volume while actively searching for a wide range of findings, from common conditions such as emphysema and cardiomegaly to acute, time-critical conditions such as pulmonary embolism, acute rib fracture, and pneumothorax. Because many of these findings are visually subtle, localized to a small number of slices, or simply rare in absolute terms, they are disproportionately vulnerable to being overlooked during a fast-paced manual read, especially in high-volume outpatient and emergency settings.

This creates two related challenges for computer-aided detection.

FIRST

The sheer dimensionality of a chest CT volume makes per-slice or per-finding manual labeling far more expensive than for 2D radiographs, so building a large, densely labeled dataset for every clinically desired finding is impractical.

SECOND

Several of the most clinically consequential findings, such as acute pulmonary embolism or acute rib fracture, are also the rarest in real-world data, which starves finding-specific classifiers of positive examples if each classifier is trained independently from scratch.

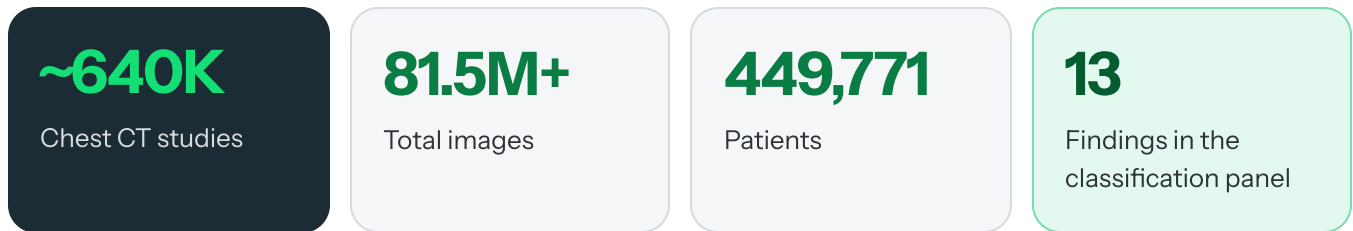
The work described in this paper addresses both challenges with a single design choice: rather than training a separate 3D classifier from scratch for every finding, we first learn a general-purpose volumetric representation of chest CT studies using a self-supervised, contrastive vision-language objective that pairs each 3D CT volume with its own radiology report. Because this pretraining signal comes from reports that are already generated as part of routine clinical care, it can draw on a much larger pool of studies than any single hand-labeled classification task, and it produces a representation that already encodes a broad vocabulary of radiological findings before any task-specific fine-tuning begins.

This shared representation is then reused as a backbone for a family of lightweight, per-finding binary classifiers, each adapted efficiently using low-rank adaptation (LoRA) rather than full model retraining. This design is particularly well suited to chest CT, where positive examples for the rarest but most clinically urgent findings can number only in the low thousands: by starting from a representation that already understands chest anatomy and reporting language, each per-finding classifier requires far less labeled data to reach strong discrimination than it would starting from a generic, non-volumetric, or non-radiological prior. The remainder of this paper describes the dataset underlying this pipeline, the two-stage training methodology in detail, and the resulting validation performance across the full finding panel.

Data

2.1 Description and size

The dataset underlying this work consists of approximately 640 thousand chest CT studies, comprising over 81.5 million total images. Each study is associated with one or more image series captured from different acquisition views; the dataset used for classification is restricted to series acquired in the axial view.

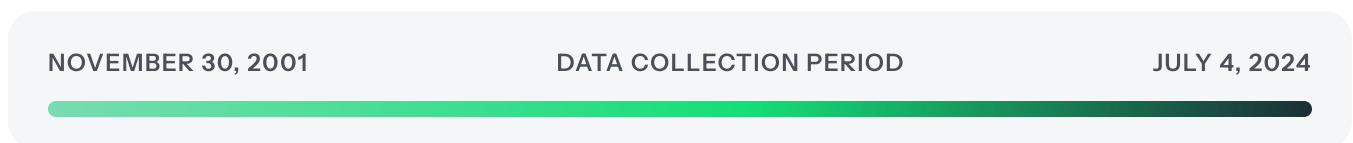


2.2 Source

The dataset is a real-world dataset originating from a single U.S.-based data provider that owns and operates a network of over 400 outpatient imaging centers spanning the East, West, and Southeastern United States.

2.3 Acquisition method

The dataset is composed of proprietary, retrospective data obtained through direct data-use agreements with healthcare providers and data brokers. It includes imaging studies collected during the course of routine clinical care and acquired across a broad array of scanner hardware vendors. Data collection spans the period from November 30, 2001 through July 4, 2024.



2.4 Disease conditions

The chest CT foundation model is trained on a broad range of radiological chest findings so that it generalizes across many downstream classification tasks. The classification panel built on top of this backbone targets 13 key findings prioritized for their clinical relevance,

spanning 640,261 studies from 449,771 patients. Table 1 summarizes the training data distribution across these findings.

Finding	Data distribution (studies)	Relative volume
Lung nodule or mass	350,387	
Air space opacity	254,381	
Aortic atherosclerosis	190,656	
COPD / emphysema	136,022	
Cardiomegaly	45,213	
Pleural effusion	37,764	
Chronic vertebral compression fracture	26,843	
Aortic aneurysm	23,952	
Aortic valve calcification	21,819	
Acute rib fracture	2,428	
Acute pulmonary embolism	1,928	
Acute vertebral fracture	1,802	
Pneumothorax	1,084	

Table 1. Training data distribution by finding. Bars are drawn to scale against the largest finding.

2.5 Patient demographics

Table 2 summarizes the patient demographic distribution across the dataset (percentages are of the total patient population, $n = 449,771$).

Demographic	Category	Data distribution (patients)
Sex	Male	205,370 (45.66%)
	Female	244,211 (54.30%)
	Other / Unknown	190 (0.04%)
Age	0–17 years	1,893 (0.42%)
	18–29 years	10,516 (2.34%)
	30–49 years	65,657 (14.60%)
	50–69 years	220,992 (49.13%)
	70–89 years	144,738 (32.18%)
	90+ years	5,960 (1.33%)
	Other / Unknown	15 (0.003%)
Race	White	242,833 (53.99%)
	Black or African American	54,607 (12.14%)
	Asian	29,146 (6.48%)
	American Indian or Alaska Native	2,005 (0.45%)
	Native Hawaiian or Other Pacific Islander	1,875 (0.42%)
	Other race	9,906 (2.20%)
	Multiple races	2,551 (0.57%)
	Unknown or declined to specify	106,843 (23.75%)

Demographic	Category	Data distribution (patients)
Ethnicity	Not Hispanic or Latino	262,393 (58.34%)
	Hispanic or Latino	37,590 (8.36%)
	Unknown or declined to specify	149,788 (33.30%)

Table 2. Patient demographic data distribution.

Structured data for height and weight were not available for most studies and are therefore not included in this analysis. While these characteristics may influence disease prevalence or image acquisition in certain contexts, their impact on model performance could not be formally assessed given the missing metadata.

As a supporting measure of dataset diversity, patient and facility locations were also analyzed using zip-code-derived Rural-Urban Commuting Area (RUCA) codes, aggregated to the three-digit ZIP3 level and classified as urban, suburban/mixed, or rural based on the dominant category within each ZIP3 region. Table 3 summarizes this distribution. Zip-code-based analysis is used here as supporting context for dataset diversity rather than a primary measure of representativeness; evaluation of generalizability continues to focus primarily on clinically relevant variables such as acquisition characteristics and available patient demographics.

Category	Group	Data distribution
Patient distribution	Urban	406,299 (90.33%)
	Suburban / Mixed	40,157 (8.93%)
	Rural	455 (0.10%)
	Unknown	2,860 (0.64%)
Facility distribution	Urban	302 (87.03%)
	Suburban / Mixed	16 (4.61%)
	Rural	0 (0.00%)
	Unknown	29 (8.36%)

Table 3. Patient and facility zip-code (urban/rural) data distribution.

Given the breadth and scale of the dataset, the available demographic, geographic, and acquisition-level metadata support an assessment of dataset diversity and model generalizability, though subgroup analyses remain limited to variables with sufficient structured metadata availability. As detailed further in Section 4, training data is concentrated in urban and suburban outpatient settings, which constrains conclusions about generalization to rural or acute-care populations.

2.6 Acquisition sites and equipment

Studies were labeled with scanner manufacturer information derived from DICOM metadata, spanning six major hardware vendors; unknown, mislabeled, or less common scanners were grouped into an “Other” category, representing 0.08% of all data. Table 4 summarizes acquisition equipment distribution, and Table 5 summarizes the contributing acquisition site, whose specific identity has been withheld pursuant to confidentiality obligations.

Scanner manufacturer	Data distribution (studies)	Share
GE Healthcare	533,296 (83.29%)	
Siemens Healthineers	71,609 (11.18%)	
Philips	19,259 (3.01%)	
Canon Medical Systems	10,014 (1.56%)	
Hitachi	4,707 (0.74%)	
Fujifilm	864 (0.13%)	
Other / Unknown	512 (0.08%)	

Table 4. Data acquisition equipment distribution.

Acquisition site	Data distribution (studies)
Data Site 1 (outpatient)	640,261

Table 5. Data acquisition sites.

2.7 Preprocessing

The following preprocessing steps are applied to all images prior to training:

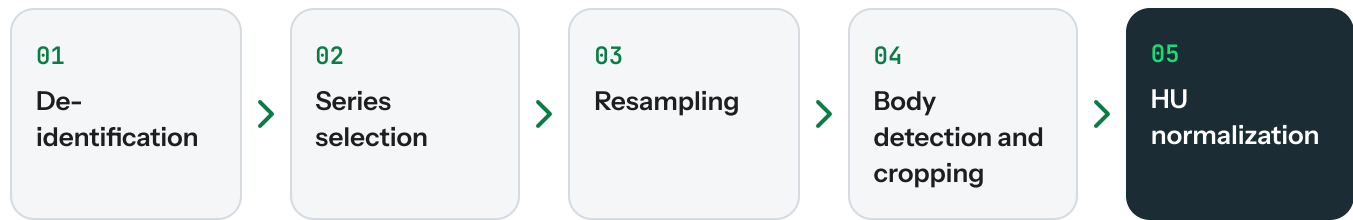


Figure 1. Preprocessing steps applied to all images prior to training.

- **De-identification:** images are de-identified using customized third-party OCR solutions to identify and redact embedded text that has not been explicitly whitelisted.
- **Series selection:** the optimal CT series per study is selected using quality-control filters: slice thickness $\leq 5.0\text{mm}$, coverage extent 100 to 500mm, z-spacing variability (IQR/median) $< 10\%$, and axial orientation (cosine similarity > 0.98). When multiple series pass QC, reconstruction kernels are prioritized in the order HARD $>$ STANDARD $>$ SOFT $>$ UNKNOWN.
- **Resampling:** each DICOM series is resampled to $0.7 \times 0.7\text{mm}$ in-plane spacing at a fixed 512×512 pixel output using B-spline interpolation. The z-axis is resampled to a target spacing of 2.5mm. Hounsfield Unit (HU) values are clamped to $[-2048, 4095]$ to prevent interpolation artifacts.
- **Body detection and cropping:** a body mask is computed using an HU threshold of -350 with a 5mm margin. The volume is cropped to the detected body region, slices with insufficient body content are pruned, and smart center-padding produces the final 512×512 output with the body centered.
- **HU normalization:** pixel values are windowed to the lung range $[-1000, 1000]$ HU and linearly scaled to $[0, 1]$.

During training, additional augmentations (implemented using the MONAI library) are applied to enhance model robustness and generalization, differing between the pretraining and fine-tuning stages:

FOUNDATION MODEL PRETRAINING STAGE

- **Resizing:** volumes are resized to a base spatial size of $320 \times 320 \times 128$ (H $\times$ W $\times$ D).
- **Random spatial cropping:** a $256 \times 256 \times 128$ crop is extracted from a random location within the resized volume.
- **Random flips:** three independent per-axis flips are applied with probabilities of 0.5 (H-axis), 0.5 (W-axis), and 0.3 (D-axis).

BINARY CLASSIFICATION FINE-TUNING STAGE

- **Random 3D rotation (20% probability):** independent rotations along the x, y, and z axes, with angles uniformly sampled from $[-10^\circ, +10^\circ]$ for each axis.
- **Center spatial crop (100% probability):** a 448×448 center crop extracted in the x-y plane while preserving full depth along the z-axis, with padding applied when input dimensions are smaller than 448 pixels.
- **Resizing (100% probability):** volumes resampled to a fixed size of $384 \times 384 \times 128$ in height, width, and depth.
- **Random color jitter (30% probability):** one volumetric intensity augmentation randomly selected from gamma contrast adjustment (gamma sampled from $[0.5, 4.5]$), Gaussian sharpening, Gaussian smoothing, or additive Gaussian noise with standard deviation 0.1.
- **Random intensity scaling (30% probability):** the full volume multiplied by a random factor uniformly sampled from $[0.98, 1.02]$ ($\pm 2\%$ intensity scaling).
- **Random intensity shift (30% probability):** a random offset uniformly sampled from $[-0.02, +0.02]$ added to all voxels.
- **Intensity clamping (100% probability):** voxel values clipped to $[0.0, 1.0]$ after intensity augmentations, to remove out-of-range values introduced during augmentation.

2.8 Dataset splitting

All training data were split into training, validation, and test sets using an 81/10/9 ratio, with partitioning performed at the patient level so that all studies and images belonging to a given patient remain within a single subset. This prevents data leakage across splits and avoids scenarios where a single patient's data could appear in more than one partition, which would compromise the independence of model evaluation.



Figure 2. The 81/10/9 training, validation, and test split.

During the fine-tuning stage, both the training and validation datasets were curated to maintain a balanced distribution of positive and negative cases per finding, allowing the model to learn effectively across the full spectrum of conditions. The test dataset, by contrast, was constructed to reflect the real-world prevalence of findings, providing a more clinically realistic assessment of generalization. Table 6 summarizes the total number of studies assigned to each partition, and Table 7 summarizes the number of positive cases per finding across the three splits.

Partition	Total studies
Training	520,108
Validation (hyperparameter tuning)	60,862
Test (final model evaluation)	59,291

Table 6. Training / validation / test split for the classification fine-tuning stage.

Finding	# Train positive	# Val positive	# Test positive
Air space opacity	206,654	23,942	23,785
Acute pulmonary embolism	1,533	186	209
Acute rib fracture	1,942	254	232
Acute vertebral fracture	1,499	150	153
Aortic aneurysm	19,312	2,354	2,286
Aortic atherosclerosis	154,771	18,385	17,500
Aortic valve calcification	17,641	2,092	2,086
Cardiomegaly	36,720	4,271	4,222

Finding	# Train positive	# Val positive	# Test positive
Chronic vertebral compression fracture	21,901	2,501	2,441
COPD / emphysema	110,785	12,939	12,298
Lung nodule or mass	285,030	33,108	32,249
Pleural effusion	30,444	3,759	3,561
Pneumothorax	882	101	101

Table 7. Number of positive samples per finding across data splits.

The dataset used for classification is restricted to qualifying axial series identified by the series-selection quality-control step described in Section 2.7; each qualifying series within a study is classified independently.

03

Model training

The classification system was developed using a two-stage training pipeline that supports both generic volumetric representation learning and task-specific performance. The first stage pretrains a vision-language foundation model to learn generalizable, radiologically meaningful representations directly from chest CT volumes and their paired radiology reports; the second stage refines these representations through supervised fine-tuning for individual findings.

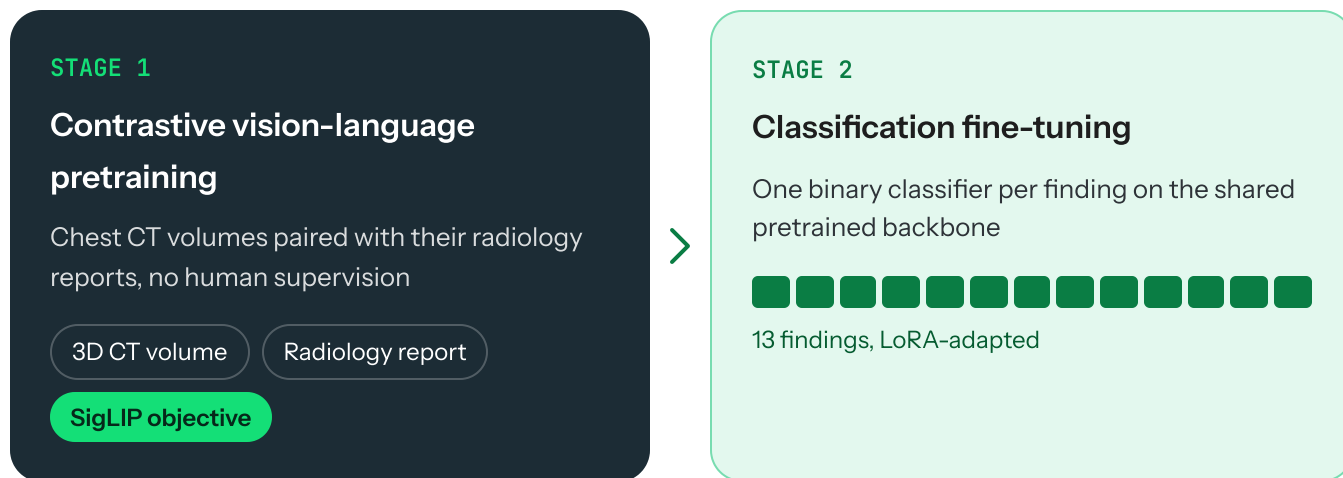


Figure 3. The two-stage training pipeline: a reusable volumetric backbone, then per-finding classifiers.

3.1 Stage 1: self-supervised, contrastive vision-language pretraining

The objective of this stage is to pretrain a Vision Transformer (ViT) backbone using multimodal chest CT data, consisting of 3D CT volumes paired with corresponding radiology reports. The model is pretrained without human supervision using an image-text alignment objective inspired by the CLIP framework, learning general-purpose, radiologically meaningful representations prior to supervised fine-tuning.

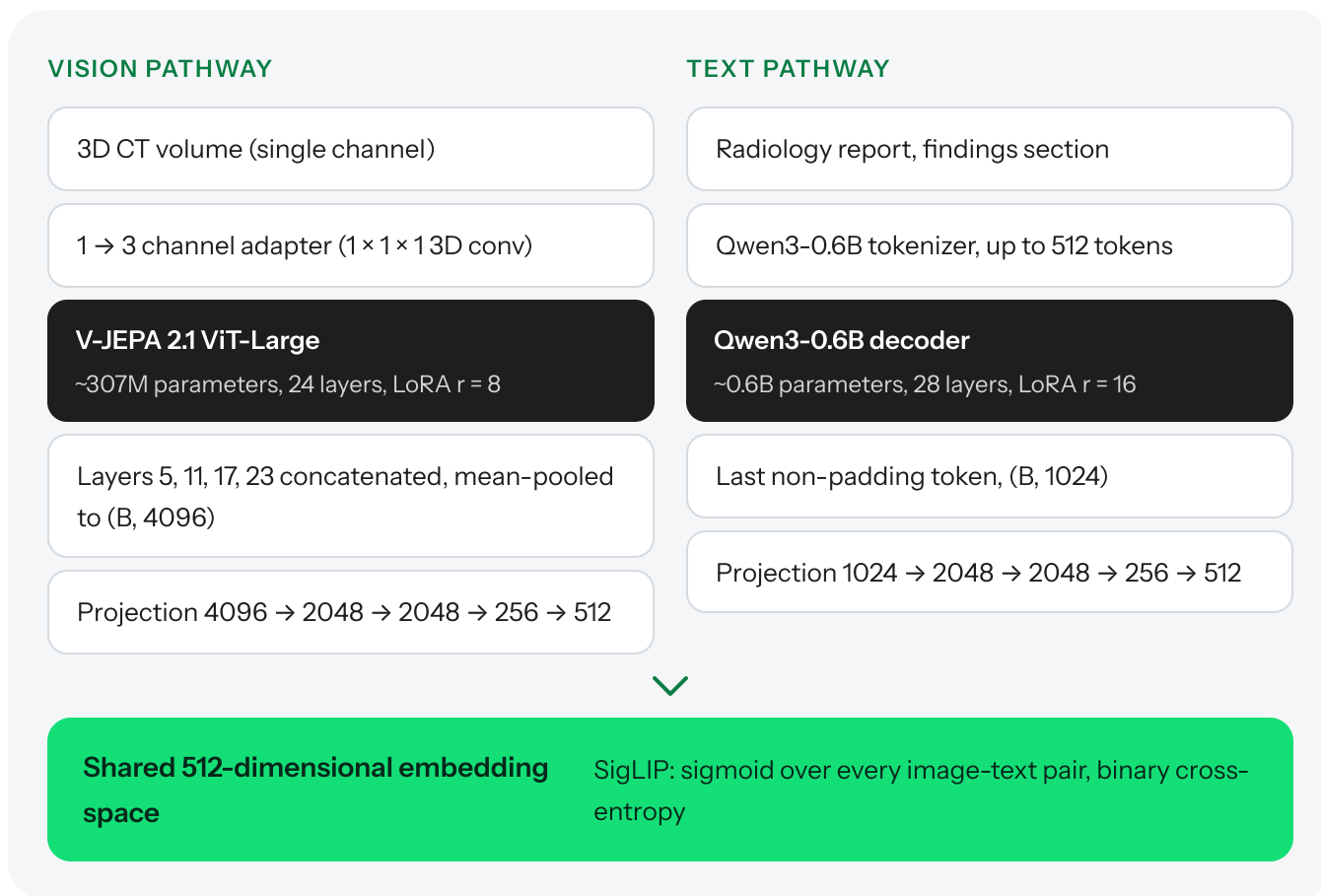


Figure 4. Stage 1 architecture: each CT volume and its report are encoded independently and aligned in a shared embedding space.

VISION ENCODER

- **Backbone:** built on top of V-JEPA 2.1 [1], a 3D vision transformer developed by Meta AI. The base architecture is ViT-Large with approximately 307M parameters, comprising 24 transformer layers, each with 1024 hidden dimensions and 16 attention heads. The encoder uses a spatial patch size of 16×16 and a temporal tubelet size of 2, processing 64 frames per scan. Weights are initialized from the public V-JEPA 2.1 pretrained checkpoint.
- **Channel adaptation:** a learned $1 \rightarrow 3$ channel adapter (a 3D convolution with kernel size $1 \times 1 \times 1$) is inserted before the patch embedding layer to map the single-channel CT volume into a 3-channel volume compatible with the pretrained encoder. Its weights are initialized randomly and trained jointly with the rest of the vision pathway.
- **Depth resampling:** after channel adaptation, the input is resampled along the z-dimension whenever it does not match the number of slices expected by the model.

- **Hierarchical feature extraction:** features are extracted from four intermediate transformer layers (layers 5, 11, 17, and 23) and concatenated along the feature dimension to produce a multi-scale representation. Each layer outputs (B, N, 1024) patch tokens, concatenated to (B, N, 4096) and mean-pooled over spatial tokens to produce the final (B, 4096) vision embedding, capturing both low-level texture (early layers) and high-level semantics (deep layers).
- **Vision projection head:** a 4-layer MLP with dimensions $4096 \rightarrow 2048 \rightarrow 2048 \rightarrow 256 \rightarrow 512$, consisting of two hidden blocks (each a linear layer followed by a GELU activation), a linear layer projecting to a 256-dimensional bottleneck, and a final weight-normalized linear layer (without bias) producing the 512-dimensional output embedding.

TEXT ENCODER

- **Backbone:** built on top of Qwen3-0.6B, a decoder-only transformer language model with approximately 0.6B parameters, comprising 28 transformer layers with hidden dimension 1024 and 16 attention heads (using grouped-query attention). It uses rotary positional embeddings (RoPE), RMSNorm, and SwiGLU MLPs, and accepts input sequences up to 512 tokens. Weights are initialized from the publicly released Qwen3-0.6B pretrained checkpoint.
- **Sequence pooling:** because Qwen3 is decoder-only, the hidden state at the last non-padding token position is taken as the report-level embedding, analogous to how a [CLS] token is used in encoder-only models; through causal attention, this token attends to all earlier positions and aggregates the full sequence context, yielding a single (B, 1024) embedding per report.
- **Text projection head:** a 4-layer MLP with dimensions $1024 \rightarrow 2048 \rightarrow 2048 \rightarrow 256 \rightarrow 512$, structured identically to the vision projection head.

Data processing pipeline: all input 3D CT volumes undergo intensity normalization and geometric augmentation using a MONAI-based pipeline, as described in Section 2.7. Text reports are preprocessed to extract the findings section and tokenized using the Qwen3-0.6B tokenizer with a maximum sequence length of 512 tokens.

Approach: the model is pretrained via contrastive vision-language alignment on paired (CT volume, radiology findings) examples. Each CT volume and its corresponding report are encoded independently by the vision and text pathways and projected into a shared 512-dimensional embedding space; the two encoders are trained jointly so that the embedding of each CT scan is pulled toward the embedding of its own report and pushed away from the

embeddings of other reports in the batch (and vice versa). The vision encoder backbone is updated using LoRA modules with rank $r = 8$, scaling factor $\alpha = 16.0$, and dropout = 0.1; the channel adapter and vision projection head are trained end-to-end. The text encoder backbone is updated using LoRA modules with rank $r = 16$, $\alpha = 32.0$, and dropout = 0.05, while the text projection head is trained end-to-end.

VISION LORA

$r = 8$ $\alpha = 16.0$ dropout = 0.1

TEXT LORA

$r = 16$ $\alpha = 32.0$ dropout = 0.05

Loss function: the model is trained using the SigLIP objective, a sigmoid-based contrastive loss for vision-language pretraining. Unlike the standard CLIP objective, which applies softmax normalization across image-to-text and text-to-image similarity scores, SigLIP formulates alignment as independent pairwise classification over image-text pairs: each true (CT volume, report) pair is treated as a positive example, while all other image-text combinations within the same batch are treated as negative examples. Image and text embeddings are normalized and projected into the shared 512-dimensional space; for each pair, a similarity score is computed via dot product, followed by a learnable temperature scaling factor and a learnable bias term, and a sigmoid function produces an independent match prediction. The training loss is binary cross-entropy over all positive and negative pairwise predictions within the batch.

STANDARD CLIP

Softmax normalization across image-to-text and text-to-image similarity scores

SIGLIP (USED HERE)

Independent pairwise classification: dot product, learnable temperature and bias, sigmoid match prediction

Figure 5. How SigLIP differs from the standard CLIP objective.

3.2 Stage 2: classification fine-tuning

The objective of this stage is to adapt the pretrained chest CT foundation model for binary and multi-class classification of individual radiological findings. The pretrained representations are optimized using labeled chest CT studies to distinguish the presence versus absence of each target finding, enabling task-specific prediction while leveraging the generic features learned during pretraining.

Base model: the backbone uses the pretrained ViT-Large vision encoder from the foundation model pretraining stage. During fine-tuning, the backbone is updated using LoRA adapters with rank $r = 8$ and scaling factor $\alpha = 16$, while the channel adapter is fully fine-tuned.

MLP classification head: a task-specific MLP classification head is appended to the pretrained vision encoder to convert the learned chest CT representation into binary class predictions. The head consists of two intermediate fully connected layers with output dimensions of 512 and 64, respectively, each followed by a SiLU activation and dropout regularization with probability 0.5. The final linear layer maps the 64-dimensional feature representation to two output logits, corresponding to the negative and positive classes for the target finding. The MLP head is fully fine-tuned during supervised training, enabling task-specific adaptation on top of the pretrained foundation model features.

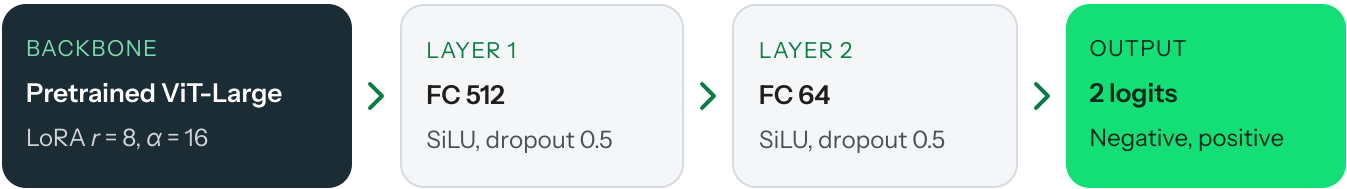


Figure 6. The per-finding MLP classification head on the pretrained vision encoder.

Approach: each binary classifier is trained independently for its corresponding visual finding, allowing per-class optimization without interfering with shared learned features. The input is the 3D chest CT scan; the target is a binary label (positive/negative) for the specific finding; the output is class logits (2 units) interpreted as positive versus negative probabilities.

Training for both stages was conducted in a cloud-based environment.

04

Metrics and results

4.1 Validation methodology

The primary training objective during Stage 2 binary classification fine-tuning was optimized using binary cross-entropy loss, appropriate for independent binary classification tasks across multiple radiological findings. Model performance was monitored throughout training using validation ROC-AUC, PR-AUC, and loss curves to assess convergence and

detect overfitting. For each task, the checkpoint with the best validation ROC-AUC was selected for downstream evaluation. These metrics were chosen for their clinical relevance and interpretability across both common and rare findings.

The results reported below are based exclusively on Stage 2 fine-tuning and do not reflect metrics from the pretraining stage.

4.2 Results

Table 8 reports validation ROC-AUC and PR-AUC for each of the 13 findings in the classification panel, using the same finding names as Table 1 and Table 7.

Finding	ROC-AUC	PR-AUC
Acute pulmonary embolism	0.9504	0.2460
Acute rib fracture	0.8976	0.1094
Acute vertebral fracture	0.9294	0.1247
Air space opacity	0.8168	0.7987
Aortic aneurysm	0.9593	0.7397
Aortic atherosclerosis	0.8677	0.9289
Aortic valve calcification	0.9054	0.7567
Cardiomegaly	0.9343	0.6870
Chronic vertebral compression fracture	0.8425	0.4293
COPD / emphysema	0.9258	0.9240
Lung nodule or mass	0.8096	0.8767
Pleural effusion	0.9741	0.8633
Pneumothorax	0.9502	0.5031

Table 8. Validation results by finding. The highest value in each column is shaded.



Figure 7. Validation ROC-AUC and PR-AUC by finding, drawn from Table 8.

4.3 Discussion

8 of 13

Findings with validation ROC-AUC above 0.90

0.9504

ROC-AUC, acute pulmonary embolism

0.9502

ROC-AUC, pneumothorax

Validation ROC-AUC exceeds 0.90 for eight of the 13 findings, including several of the rarest and most acute conditions in the panel: acute pulmonary embolism (0.9504), acute vertebral fracture (0.9294), pneumothorax (0.9502), and aortic aneurysm (0.9593) all reach strong discrimination despite having only a few hundred to a few thousand positive training examples (Table 7). This is consistent with the central premise of the pretraining strategy: because the vision-language backbone already encodes a broad, report-grounded understanding of chest CT anatomy before any finding-specific labels are introduced, each per-finding classifier can reach high discrimination from a comparatively small number of positive examples.

PR-AUC tells a complementary story. For low-prevalence findings such as acute rib fracture (0.1094 PR-AUC despite 0.8976 ROC-AUC) and acute vertebral fracture (0.1247 PR-AUC despite 0.9294 ROC-AUC), the gap between strong ROC-AUC and modest PR-AUC reflects the low base rate of positive cases in the test distribution rather than weak model discrimination; PR-AUC is inherently more sensitive to class imbalance than ROC-AUC. By contrast, high-prevalence findings such as aortic atherosclerosis (0.9289 PR-AUC), COPD/emphysema (0.9240 PR-AUC), and lung nodule or mass (0.8767 PR-AUC) show strong performance on both metrics, as expected given their much larger positive sample sizes.

The two comparatively lower ROC-AUC findings, lung nodule or mass (0.8096) and air space opacity (0.8168), are also the two most prevalent and most visually heterogeneous findings in the panel, spanning a wide range of nodule sizes, densities, and opacity patterns; their PR-AUC values (0.8767 and 0.7987, respectively) indicate that the model retains strong practical utility for these findings even where ROC-AUC is comparatively lower. Taken together, these results support reusing a single volumetric, report-grounded backbone across a broad panel of chest CT findings, including rare and acute conditions that would be difficult to model in isolation given their limited positive sample sizes.

TAKEN TOGETHER

These results support reusing a single volumetric, report-grounded backbone across a broad panel of chest CT findings, including rare and acute conditions that would be difficult to model in isolation given their limited positive sample sizes.

References

1. Mur-Labadia, L., Muckley, M., Bar, A., Assran, M., Sinha, K., Rabbat, M., LeCun, Y., Ballas, N., & Bardes, A. (2026). V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*.