Anatomy Contextualized Adaption of CT Foundation Models

Roshan Kenia[1,2], Stephanie L McNamara[3], and William Lotter[2,4,5]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[2] Department of Data Science, Dana-Farber Cancer Institute, Boston, MA
[3] Department of Radiology, Massachusetts General Hospital, Boston, MA
[4] Department of Pathology, Brigham and Women's Hospital, Boston, MA
[5] Department of Pathology, Harvard Medical School, Boston, MA

Abstract. CT vision-language foundation models have demonstrated promising performance across downstream tasks, but are typically trained with whole-volume representations that dilute fine-grained anatomical signals. Fine-grained vision-language pre-training addresses this by aligning anatomy-level visual features with anatomy-specific text, but in doing so discards the global context that whole-volume models provide. Furthermore, existing fine-grained approaches train from scratch, making them computationally expensive. We introduce Anatomy Contextualized Adaptation (ACA), a lightweight framework that adapts frozen CT foundation model representations for anatomy-level vision-language alignment while enhancing global contextualization. ACA uses TotalSegmentator to decompose CT volumes into anatomy-level embeddings, which are refined via a transformer that captures cross-anatomy relationships, and aligned to both per-anatomy and scan-level text extracted from radiology reports. Evaluated on Merlin and CT-RATE, ACA consistently outperforms both the frozen foundation model baselines and existing fine-grained methods in zero-shot finding classification, while requiring less than one hour of training once embeddings are cached. The attention weights learned by ACA's inter-anatomy transformer additionally indicate plausible cross-anatomy context routing. Altogether, these results support ACA as a lightweight approach for adapting CT foundation models to anatomically grounded vision-language alignment while preserving and enhancing global anatomical context[6].

Keywords: CT foundation models · vision-language pre-training · fine-grained alignment

1 Introduction

CT foundation models [32, 34, 35, 40] trained on large-scale data have demonstrated promising performance across a wide range of downstream tasks, including zero-shot disease classification via vision-language contrastive learning. However, these models are typically trained using whole-volume CT representations,

[6] Code is available at https://github.com/lotterlab/ACA

condensing a 3D volume into a single embedding for pre-text training [3, 4, 13]. Conversely, fine-grained vision-language pre-training (FVLP) has emerged as a strategy to align anatomy-specific visual embeddings with corresponding textual descriptions [5, 20, 26, 39]. These methods can recover fine-grained signal, but in doing so discard the global context that whole-volume models provide. Furthermore, existing FVLP approaches have relied on training the entire model from scratch, which is computationally expensive and can be difficult to scale.

We introduce *Anatomy Contextualized Adaption* (ACA), a framework for adapting pretrained CT foundation models to anatomy-level vision-language alignment while enhancing global context. ACA combines the strengths of both whole-volume foundation models and FVLP: the broad, frozen representations of pretrained CT foundation models are adapted using lightweight trainable modules to produce anatomy-level embeddings, which are then contextualized across anatomies using transformer-based attention. The training loss includes both anatomy-level and scan-level vision-language alignment. Across two datasets and two base foundation models, ACA boosts zero-shot diagnostic performance and its learned attention indicates plausible cross-anatomy context routing.

2 Related Work

CT Foundation Models. Recent work has explored large-scale pre-training of CT models that can be adapted to a variety of clinical tasks. Vision-language pretraining approaches, building on image-text contrastive learning [25,29,33], align CT volumes with paired radiology reports [3, 6]. CT-CLIP [13] and Merlin [4] represent the dominant paradigm, training contrastively on chest/abdominal CT volumes paired with free-text reports, enabling zero-shot abnormality detection at supervised-level performance. Vision-only approaches such as CT-FM [24] and 3DINO [37] leverage self-supervised contrastive [7, 10, 15] and masked image modeling [2, 14, 36] objectives on unlabeled CT volumes, demonstrating strong transfer to segmentation and classification tasks. Segmentation-oriented foundation models such as VISTA3D [17] and SAM-Med3D [28], together with large-scale anatomical segmentation frameworks such as TotalSegmentator [30], enable generalized anatomical understanding across dozens to hundreds of anatomical structures in 3D CT imaging. Task-specific multimodal models including M3FM [23] and LCTfound [11] incorporate structured clinical data and imaging to address lung cancer screening and other downstream workflows. Given the emergent zero-shot capabilities of vision-language approaches, we focus on adapting Merlin and CT-CLIP while leveraging TotalSegmentator to facilitate fine-grained modeling.

Fine-grained Vision-language Pre-training. Rather than aligning entire CT volumes with full radiology reports, fine-grained vision-language pretraining (FVLP) aligns individual anatomical regions with their corresponding report descriptions [19, 22, 26, 27, 39]. fVLM [26] explicitly decomposes CT volumes using TotalSegmentator [30] and matches anatomy-level visual tokens to anatomy-specific text via cross-attention. ViSD-Boost [5] identifies a semantic density gap

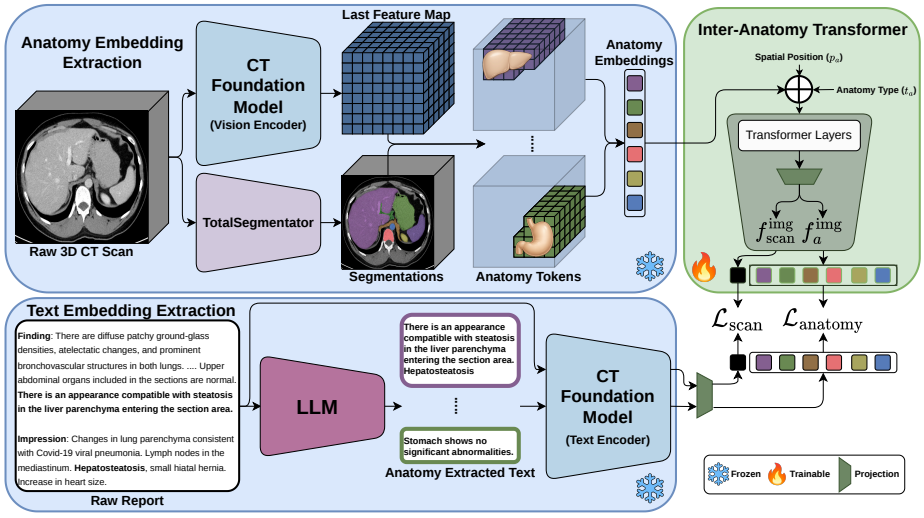

Fig. 1: Overview of ACA. Anatomy-level visual embeddings are constructed from a frozen CT foundation model using TotalSegmentator segmentations, while anatomy-level text embeddings are extracted from radiology reports using an LLM and the corresponding frozen text encoder. An inter-anatomy transformer, augmented with spatial position and anatomy type embeddings, contextualizes the anatomy embeddings across the full set of structures. The resulting embeddings are aligned to per-anatomy text via $\mathcal{L}_{\text{anatomy}}$, while their mean-pooled scan-level embedding is aligned to the full report via $\mathcal{L}_{\text{scan}}$.

between low signal-to-noise visual representations and information-dense diagnostic reports, and addresses it through disease-level visual contrastive learning and anatomical normality modeling.

Despite these advances, existing FVLP methods treat each anatomy independently during both training and inference, ignoring the inter-anatomy context that is often essential for accurate diagnosis [31]. For example, organomegaly (abnormal enlargement of an organ; e.g., hepatomegaly, splenomegaly) is best assessed relative to surrounding structures and body habitus. Our work addresses this gap by introducing an inter-anatomy transformer that contextualizes anatomy embeddings across all present anatomies before alignment, enabling richer and more clinically grounded representations.

CT Foundation Model Datasets. The datasets used to train the Merlin and CT-CLIP foundation models have been publicly released. The Merlin [4] dataset is a large-scale abdominal CT cohort comprising 25,494 CT scans paired with radiology reports from 18,317 unique patients, collected at Stanford University Medical Center, with annotations spanning 30 abdominal findings. CT-CLIP was developed using CT-RATE [13], a chest CT dataset comprising 50,188 reconstructed 3D volumes from 25,692 scans of 21,304 unique patients, each paired

with a radiology report and 18 annotated abnormality labels extracted via an automated text classifier.

3 Methodology: Anatomy Contextualized Adaptation

ACA is illustrated in Figure 1 and consists of three core components. First, anatomy-level visual embeddings are extracted from a frozen CT foundation model by pooling its feature maps based on TotalSegmentator segmentations, while anatomy-level text embeddings are extracted from the corresponding radiology report using an LLM. Second, an inter-anatomy transformer, augmented with learned spatial position and anatomy type embeddings, contextualizes each anatomy's embedding using the full set of structures present in the scan. Finally, the contextualized anatomy embeddings are projected into a shared contrastive space and aligned to their corresponding per-anatomy text, while a mean-pooled scan-level embedding is aligned to the full report embedding to provide a complementary global supervision signal. We describe each component in detail below.

3.1 Embedding Construction

Anatomy Embedding Extraction. To construct anatomy-specific representations, we first apply TotalSegmentator [30] to segment each CT volume into 44 anatomical structures. The 44 structures are based on condensing TotalSegmentator's original 117 classes into a set of related anatomical groups (e.g., left kidney and right kidney $\rightarrow$ kidney), as summarized in Appendix Table 3.

For each CT volume, we pass the scan through a frozen foundation model backbone to obtain intermediate spatial feature representations. We experiment with two foundation models: Merlin, which uses an I3D ResNet152 [8] encoder, and CT-CLIP, which uses a CT-ViT [12] encoder. Merlin uses a $224 \times 224 \times 160$ voxel input, and generates a final feature map of size $2048 \times 10 \times 7 \times 7$. CT-CLIP uses a $480 \times 480 \times 240$ voxel input, and produces a feature map of size $512 \times 24 \times 24 \times 24$. Both models subsequently pool these features to generate a scan-level representation for downstream tasks, whereas we construct anatomy-level features from them.

To obtain anatomy-level representations, we project the TotalSegmentator segmentation masks into the spatial resolution of the extracted feature maps. Specifically, we apply a non-overlapping max-pool over each binary organ mask using a kernel matched to the backbone's effective patch size ($32 \times 32 \times 16$) voxels for Merlin and ($40 \times 40 \times 20$) voxels for CT-CLIP), yielding a discrete patch-presence grid at the feature map resolution. For each of the 44 anatomical structures, we extract all features in which the structure occupies a patch across the last feature map of the backbone. We subsequently mean-pool these features into a single structure embedding. Structures for which TotalSegmentator produces no non-empty segmentation mask in a given scan are considered absent and excluded from that scan's input sequence. For each anatomy a with N_a final layer extracted features $\{v_1, \ldots, v_{N_a}\} \in \mathbb{R}^d$, we denote the anatomy embedding

as $\bar{v}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} v_i$, where $d = 2048$ for the Merlin backbone and $d = 512$ for CT-CLIP.

Each embedding is subsequently ℓ_2-normalized. In addition, we compute inter-anatomy spatial position encodings for each anatomical structure as a 44-dimensional vector, where the g-th entry is the ℓ_2 distance between the structure's voxel-space centroid and the centroid of structure g, normalized by the volume diagonal length $\sqrt{D^2 + H^2 + W^2}$, where D, H, W correspond to the number of voxels along the depth, height, and width, respectively. These encodings capture the geometric relationships between anatomical structures within a volume and are used as positional signals in downstream models.

Text Embedding Extraction. For the text modality, anatomy-specific findings are extracted from radiology reports using Qwen3-4B-Instruct [38] (LLM in Figure 1), following a similar strategy to [26]. These findings are subsequently encoded using the corresponding frozen Merlin or CT-CLIP text encoder to produce anatomy-level text embeddings. Both the prompts used for anatomy-specific finding extraction and an example of the resulting anatomy-specific findings are shown in Appendix Figures 5 and 6, respectively. A report-level text embedding is also generated by passing the full report through the text encoder.

3.2 Inter-Anatomy Transformer

To enable contextual reasoning across anatomical structures within a scan, we pass all present anatomy embeddings jointly through a transformer encoder. Prior to transformer input, each anatomy embedding is projected and combined with two learned tokens. The first consists of an MLP applied to a positional encoding vector $p_a \in \mathbb{R}^{44}$, which encodes the normalized distances from anatomy a to all other structures in the taxonomy as described in Section 3.1. The second token is a learnable anatomy type embedding t_a specific to each of the 44 anatomy groups. The final anatomy embedding for transformer input, h_a, is computed as follows:

$$h_a = \bar{v}_a W_{\text{vis}} + \text{MLP}(p_a) + t_a \tag{1}$$

The sequence $\{h_a\}_{a \in \mathcal{P}}$, where $\mathcal{P}$ denotes the set of anatomical structures present in the scan, is passed through an L-layer transformer encoder with pre-layer normalization, producing a contextualized embedding for each anatomy.

3.3 Combined Anatomy-Level and Global Report Loss

We supervise the inter-anatomy transformer with two complementary loss terms operating at different granularities: an anatomy-level contrastive loss, $\mathcal{L}_{\text{anatomy}}$, that aligns each anatomy's contextualized embedding to its corresponding text description, and a scan-level report loss, $\mathcal{L}_{\text{scan}}$, that grounds the full anatomy sequence to the global radiology report.

To enable the contrastive losses, the anatomy-level image and text embeddings are first projected to a shared d'-dimensional contrastive space. For both

embedding types, a two-layer projection head is used, consisting of a linear layer, GELU activation, layer normalization, a second linear layer, and ℓ_2 normalization in sequence. We denote the final anatomy-level image and text embeddings as f_a^{img} and f_a^{txt} respectively. To enable the scan-level loss, a scan-level image embedding is computed as the ℓ_2-normalized mean over anatomy embeddings: $f_{\text{scan}}^{\text{img}} = \ell_2\Big(\frac{1}{|\mathcal{P}|}\sum_{a\in\mathcal{P}} f_a^{\text{img}}\Big)$. The scan-level text embedding $f_{\text{scan}}^{\text{txt}}$ is obtained by passing the full radiology report through the frozen text encoder and then projecting to the shared contrastive space using the same text projection head as the per-anatomy text embeddings.

Anatomy-Level Loss. For each anatomical structure s, we collect the N_s instances present across the batch and compute a symmetric image-text contrastive loss over them. The standard contrastive loss uses a hard diagonal objective wherein matching image-text pairs are treated as positives and all other combinations are treated as negatives. To handle the prevalence of false negatives in this formulation, where anatomical structures from different patients may be semantically identical yet penalized as different, we follow [26] and construct a soft target matrix $T \in [0,1]^{N_s\times N_s}$, where i and j index instances of structure s:

$$T_{ij} = \mathbf{1}[i=j] + \mathbf{1}[\text{normal}(i)\wedge\text{normal}(j)] + \mathbf{1}[\text{patient}(i)=\text{patient}(j),\ i\neq j] \quad (2)$$

where normal$(\cdot)$ are binary normality labels extracted from radiology reports using Qwen3-4B-Instruct (see Figure 5), and each row is normalized to sum to one. Anatomies where all instances are normal are skipped entirely, focusing capacity on pathological variation. The per-anatomy loss is then:

$$\mathcal{L}_s = -\frac{1}{2N_s}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s} T_{ij}\left[\log\frac{e^{f_i^{\text{img}}\cdot f_j^{\text{txt}}/\tau}}{\sum_k e^{f_i^{\text{img}}\cdot f_k^{\text{txt}}/\tau}} + \log\frac{e^{f_i^{\text{txt}}\cdot f_j^{\text{img}}/\tau}}{\sum_k e^{f_i^{\text{txt}}\cdot f_k^{\text{img}}/\tau}}\right] \quad (3)$$

where f_i^{img} and f_i^{txt} are the projected and ℓ_2-normalized visual and text embeddings for instance i, and τ is a learnable temperature initialized at 0.07 and clamped to $[0.001, 0.5]$. As seen in Equation 4, the total anatomy-level loss, $\mathcal{L}_{\text{anatomy}}$, sums over all active structures $\mathcal{S}$, defined as those that appear in at least one scan in the batch and have at least one abnormal instance. Scans with no present anatomies are excluded from all loss terms.

$$\mathcal{L}_{\text{anatomy}} = \sum_{s\in\mathcal{S}} \mathcal{L}_s \quad (4)$$

Scan-Level Report Loss. While $\mathcal{L}_{\text{anatomy}}$ aligns each anatomy independently to short descriptive text, it provides no signal connecting the anatomy sequence as a whole to the broader clinical context of the scan. To bridge this gap, we include a scan-level contrastive loss, $\mathcal{L}_{\text{scan}}$, that aligns the scan-level visual embedding, $f_{\text{scan}}^{\text{img}}$, to the text embedding of the full radiology report, $f_{\text{scan}}^{\text{text}}$. Unlike the anatomy-level loss, each scan is matched only to its own report, so a hard diagonal objective is used:

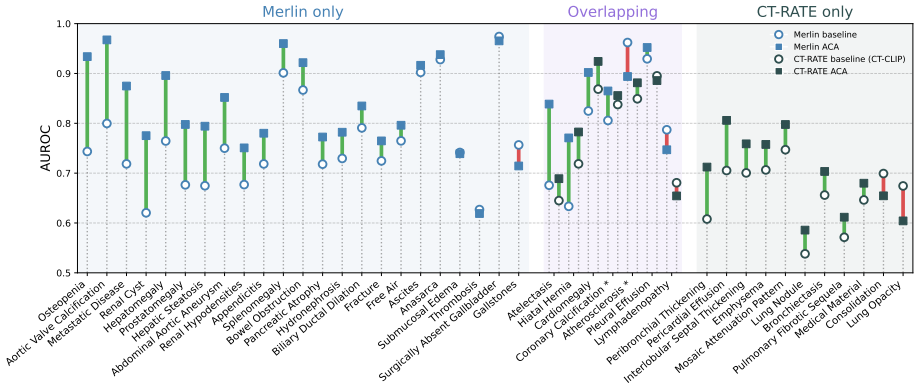

Fig. 2: Per-finding AUROC comparison between the original models and the ACA anatomy-guided model for In-Distribution zero-shot evaluation. Green lines represent improvement over the baseline, Red lines represent regression. * These findings' names from CT-RATE are simplified to match with Merlin: Coronary Artery Wall Calcification = Coronary Calcification, Arterial Wall Calcification = Atherosclerosis.

$$\mathcal{L}_{\text{scan}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{e^{f^{\text{img}}_{\text{scan},i}\cdot f^{\text{txt}}_{\text{scan},i}/\tau_{\text{scan}}}}{\sum_{k=1}^{B} e^{f^{\text{img}}_{\text{scan},i}\cdot f^{\text{txt}}_{\text{scan},k}/\tau_{\text{scan}}}} + \log\frac{e^{f^{\text{txt}}_{\text{scan},i}\cdot f^{\text{img}}_{\text{scan},i}/\tau_{\text{scan}}}}{\sum_{k=1}^{B} e^{f^{\text{txt}}_{\text{scan},i}\cdot f^{\text{img}}_{\text{scan},k}/\tau_{\text{scan}}}}\right] \tag{5}$$

where B is the number of scans in the batch with at least one present anatomy, and τ_{scan} is a separate learnable temperature initialized at 0.07 and clamped to $[0.001, 0.5]$. Because $f^{\text{img}}_{\text{scan}}$ is a mean over the same f^{img}_{a} embeddings optimized by $\mathcal{L}_{\text{anatomy}}$, the scan-level signal propagates directly through the visual projection head, jointly shaping a space where individual embeddings are discriminative at the anatomy level and coherent at the scan level.

Total Loss. The combined objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{anatomy}} + \lambda\mathcal{L}_{\text{scan}} \tag{6}$$

We refer to this full model as **ACA**. We use $\lambda = 1$ in our experiments to provide a simple balance of local and global signals, while also performing two ablations: **ACA w/o $\mathcal{L}_{\text{scan}}$**, which uses only $\mathcal{L}_{\text{anatomy}}$, and **ACA w/o $\mathcal{L}_{\text{anatomy}}$**, which uses only $\mathcal{L}_{\text{scan}}$.

4 Experiments

4.1 Datasets and Training Setup

We train on two CT datasets using their respective frozen foundation model backbones. For Merlin [4], we use the creator-defined splits of 15,314 training,

5,055 validation, and 5,125 test scans. For CT-RATE [13], we partition the training data into 37,545 training and 9,598 validation scans, and evaluate on the 3,039-scan test set defined by the dataset creators. For each dataset, we train ACA and each baseline independently using AdamW with learning rate 1×10^{-4}, batch size 64, and 50 epochs, selecting the best model for evaluation based on validation loss. Full hyperparameters are reported in Appendix Tables 5 and 6.

4.2 Evaluation

We evaluate each model on the held-out test split defined by the organizers of each dataset. We focus on zero-shot finding classification, where the findings present in each dataset are detailed in Appendix Tables 7 and 8. Each finding is treated as binary classification (positive or negative) at the scan-level. Unlike Merlin [4], we do not subsample negative samples to match the number of positives and evaluate on all samples for a finding.

To evaluate similarity with the model's scan-level visual embedding, we associate each finding with a set of positive prompts describing the pathology and negative prompts describing normal appearance (see Appendix Figure 7). These are encoded by the frozen foundation model text encoder used by the model, averaged within each class (positive or negative), and projected through the model's trained text projection head to obtain f^+ and $f^- \in \mathbb{R}^{d'}$. The model's scan-level score for the finding (defined in Appendix B) is computed as the difference in cosine similarity between the model's scan-level image representation ($f^{\text{img}}_{\text{scan}}$) and each class embedding (f^+ and f^-), with all embeddings ℓ_2 normalized. In the case where no structures are detected, the scan embedding defaults to a zero vector, which scores 0 against all finding prompts and predicts negative. Further details are provided in Appendix B.

We evaluate each model under in-distribution and out-of-distribution settings. In the in-distribution setting, models are trained and tested on the same dataset using the corresponding foundation model (e.g., Merlin train $\rightarrow$ Merlin test). Out-of-distribution tests generalization across datasets and base foundation model (e.g., Merlin train $\rightarrow$ CT-RATE test). As the Merlin and CT-RATE datasets cover different anatomies and finding sets, out-of-distribution evaluation is restricted to the 7 findings shared between the datasets (see Figure 2).

4.3 Baselines

We compare to several baselines to highlight the effects of each modeling aspect of ACA. Each baseline uses the same prompting strategy to generate positive and negative text embeddings for each finding, using the corresponding text encoder associated with the model.

Global VLMs. We assess the baseline performance of Merlin and CT-CLIP, using the original, fixed scan-level vision embeddings for zero-shot similarity.

Fine-grained adaptation. To isolate the effect of fine-grained adaptation on top of frozen foundation model embeddings, we evaluate an **MLP** baseline that does not include the inter-anatomy transformer in ACA. In this baseline,

each anatomy's pre-computed mean embedding is passed directly through a two-layer projection head consisting of a linear layer, GELU activation, layer normalization, and a second linear layer, with ℓ_2-normalization applied to the output. No inter-anatomy context is used. Each anatomy is projected independently and the model is trained with the $\mathcal{L}_{\text{anatomy}}$ loss alone.

We additionally compare to two existing fine-grained architectures, using these approaches to generate anatomy-level embeddings on top of the foundation model features. This adaptation enables a compute-matched comparison, but means these baselines do not necessarily reflect the performance of the originally published methods which rely on end-to-end training (see Limitations). The first is the anatomy query pooling mechanism from **fVLM** [26]. For the Merlin backbone, features from all four I3D ResNet152 layers are concatenated to form 3840-dimensional multi-scale representations per patch. For the CT-CLIP CT-ViT backbone, features from the last layer of the spatial transformer and the last layer of the causal transformer are concatenated to form 1024-dimensional representations. A learned anatomy-specific query vector then attends over these features via cross-attention to produce a single anatomy embedding, which is projected and aligned to the corresponding per-anatomy text using $\mathcal{L}_{\text{anatomy}}$.

The second architecture is an adaptation of **ViSD-Boost** [5], which amplifies disease signals by modeling the normal distribution of each anatomy in latent space via a VQ-VAE. Because our framework operates on frozen embeddings, we adapt the method to three training stages for fair comparison. First, we perform anatomy-level contrastive alignment using the fVLM query pooling mechanism, with normal–normal anatomy pairs upweighted in the contrastive targets, rather than a vision-only pre-training stage. Second, we train the Transformer-based VQ-VAE exclusively on normal anatomy instances to learn a shared anatomy-conditioned codebook of healthy representations. Finally, with the VQ-VAE frozen, the original multi-scale anatomy tokens are fused with their VQ-VAE reconstruction through a residual projection, with the hybrid embedding aligned to per-anatomy text as other fine-grained baselines via $\mathcal{L}_{\text{anatomy}}$ rather than binary positive/negative prompts. The disease-level contrastive pre-training stage and momentum encoder from the original method are omitted as they require end-to-end vision encoder training.

Global adaptation. To isolate the effect of anatomy-level, fine-grained training in ACA, we define a **Spatial Transformer** baseline that bypasses anatomy segmentation entirely, but includes a module analogous to the inter-anatomy transformer in ACA. In this baseline, the final feature map of the frozen backbone is depth mean-pooled and reshaped into a sequence of spatial patch features, which are processed by a transformer encoder. The mean of all spatial patch outputs is then projected through a two-layer projection head and used as the scan-level image representation, followed by radiology report alignment using $\mathcal{L}_{\text{scan}}$.

Table 1: AUROC comparison across MERLIN and CT-RATE datasets. † These methods were adapted to operate on frozen embeddings. * Denotes out-of-distribution evaluation restricted to the 7 shared findings, where models trained on CT-RATE are evaluated on Merlin* and models trained on Merlin are evaluated on CT-RATE*. Per finding results can be found in Appendix Tables 11, 12, 13, and 14.

	In-distribution		Out-of-distribution		
Model	Merlin	CT-RATE	Merlin*	CT-RATE*	Average
Global VLM					
Merlin [4]	0.7729±0.0062	-	-	0.6919±0.0058	
CT-CLIP [13]	-	0.7082±0.0037	0.5822±0.0121	-	
Global Adaptation					
Spatial Transformer	0.7840±0.0055	0.6467±0.0043	0.5605±0.0126	0.7003±0.0056	0.6729
Fine-grained Adaptation					
MLP	<u>0.7981</u>±0.0057	0.6129±0.0041	0.5699±0.0121	0.7389±0.0048	0.6800
fVLM† [26]	0.7699±0.0058	<u>0.6842</u>±0.0040	0.5587±0.0124	0.7036±0.0049	0.6791
ViSD-Boost† [5]	0.7803±0.0059	0.6511±0.0042	<u>0.6139</u>±0.0115	**0.7413**±0.0048	<u>0.6967</u>
Global + Fine-grained Adaptation					
ACA	**0.8213**±0.0052	**0.7311**±0.0036	**0.6723**±0.0117	<u>0.7400</u>±0.0049	**0.7412**

Table 2: Ablation results for ACA across the Merlin and CT-RATE datasets, comparing loss components ($\mathcal{L}_{\text{anatomy}}$, $\mathcal{L}_{\text{scan}}$) and anatomy-guided inference-time pooling (Section B). Merlin* and CT-RATE* denote out-of-distribution evaluation.

	In-distribution		Out-of-distribution		
Model	Merlin	CT-RATE	Merlin*	CT-RATE*	Average
ACA w/o $\mathcal{L}_{\text{anatomy}}$	0.7941±0.0057	0.7117±0.0037	0.6130±0.0116	0.7143±0.0059	0.7083
ACA w/o $\mathcal{L}_{\text{scan}}$	0.8108±0.0053	0.6935±0.0040	0.6364±0.0120	0.7068±0.0045	0.7119
+ *Anatomy-Guided*	<u>0.8222</u>±0.0052	<u>0.7365</u>±0.0036	0.6399±0.0120	<u>0.7451</u>±0.0045	0.7359
ACA	0.8213±0.0052	0.7311±0.0036	**0.6723**±0.0117	0.7400±0.0049	<u>0.7412</u>
+ *Anatomy-Guided*	**0.8372**±0.0048	**0.7413**±0.0035	<u>0.6718</u>±0.0118	**0.7530**±0.0048	**0.7508**

5 Results

We evaluate ACA on two CT datasets, Merlin and CT-RATE, using their respective frozen foundation models, Merlin and CT-CLIP, as backbones. For ACA and each baseline (Section 4.3), we assess zero-shot classification performance in both in-distribution (i.e., train/test on same dataset) and out-of-distribution (i.e., train/test on different datasets) settings (Section 4.2). We report AUROC as our evaluation metric, as it is threshold-independent and robust to the class imbalance present across both datasets, where most findings have substantially fewer positive than negative cases (Appendix Tables 9 and 10). Table 1 reports macro-average AUROC across findings for all models, datasets, and distribution settings. To quantify uncertainty, we estimate the mean and standard deviation of macro-average AUROC over 1000 bootstrap resamples.

ACA outperforms both global and fine-grained baselines. ACA improves over both global VLM foundation models in every setting, raising in-distribution average AUROC from 0.7729 to 0.8213 on Merlin and from 0.7082 to 0.7311 on CT-RATE, with larger gains out-of-distribution (0.5822 to 0.6723 on Merlin*, 0.6919 to 0.7400 on CT-RATE*; note that out-of-distribution evaluation is restricted to the 7 shared findings, see also Appendix Tables 11, 12). ACA likewise outperforms the fine-grained baselines, with an average boost in AUROC across settings of 0.061, 0.062, and 0.044 compared to MLP, fVLM, and ViSD-Boost, respectively. Gains are observed up to 11.8 points in-distribution (CT-RATE, vs. MLP) and 11.4 points out-of-distribution (Merlin*, vs. fVLM); the one exception is a near-tie with MLP and ViSD-Boost on CT-RATE*, which we revisit below. Furthermore, ACA outperforms the purely global adaptation strategy (Spatial Transformer) in each setting, with an average increase of 0.068 AUROC.

Figure 2 illustrates the performance changes per finding compared to the original global VLM models, with finding-level results for all models contained in Appendix Tables 11, 12, 13, and 14. The boosts compared to Merlin and CT-RATE highlight the benefits of structured, anatomy-level modeling, where the findings with the largest performance boosts are often subtle, single organ pathologies (e.g., aortic valve calcification, renal cyst, hepatomegaly), that might get washed out in global, scan-level embedding. Conversely, more diffuse or non-anatomy specific findings show less benefits (e.g., free air, submucosal edema, thrombosis). Compared to fine-grained adaptation alone, the benefits of the global inter-anatomy transformer are apparent in findings such as arterial wall calcification (e.g., 0.13 average AUROC boost (Table 12)), that can occur in any artery and thus requires integrating features throughout the scan, as well as organomegaly findings.

Ablation of combined loss function. Along with the MLP and Spatial Transformer baselines, which ablate ACA's global and anatomy-level architectural components respectively, we performed ablations of ACA's loss function itself (Table 2). Removing either the anatomy-level loss (ACA w/o $\mathcal{L}_{\text{anatomy}}$) or the scan-level loss (ACA w/o $\mathcal{L}_{\text{scan}}$) decreases performance (-0.033 and -0.029 average AUROC, respectively), indicating that both anatomy-level and scan-level supervision is important even when using the ACA architecture fixed.

Anatomy-guided pooling. As the results thus far have used ACA's scan-level representation, consisting of mean pooling over the contextualized anatomy embeddings, we also explored an inference-time strategy that assigns different pooling weights depending on the tested finding (see Appendix B). Upweighting the pooling of anatomical structures relevant to each finding (Table 2) further improves AUROC for both ACA and ACA w/o $\mathcal{L}_{\text{scan}}$ in most settings. For ACA, anatomy-guided pooling raises in-distribution AUROC from 0.8213 to 0.8372 on Merlin and from 0.7311 to 0.7413 on CT-RATE, and out-of-distribution AUROC from 0.7400 to 0.7530 on CT-RATE*, while leaving Merlin* essentially unchanged (0.6723 vs. 0.6718). For ACA w/o $\mathcal{L}_{\text{scan}}$, the effect is larger, in-distribution AUROC improves from 0.6935 to 0.7365 on CT-RATE, and out-

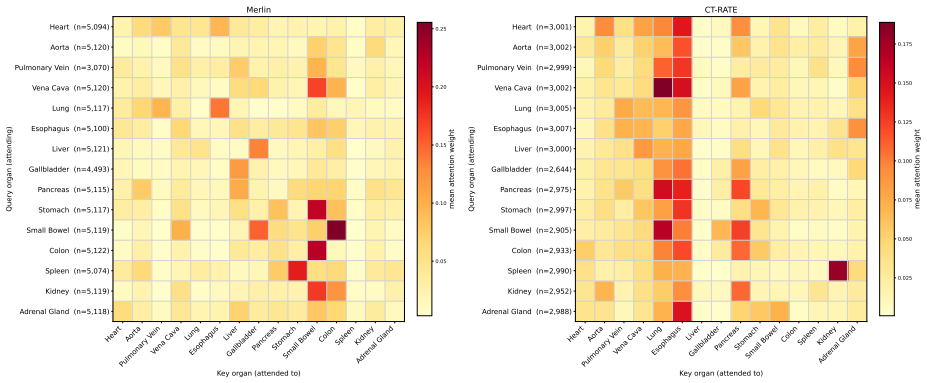

Fig. 3: Each heatmap shows the mean attention weights between fifteen major anatomical structures, averaged over all transformer layers, attention heads, and test scans. Left: ACA model trained on Merlin and evaluated on the Merlin test set. Right: ACA model trained on CT-RATE and evaluated on the CT-RATE validation set. Sample counts per query anatomy reflect each respective dataset.

of-distribution AUROC improves on both Merlin* (0.6364 vs. 0.6399) and CT-RATE* (0.7068 vs. 0.7451). These results suggest that restricting attention to clinically relevant anatomy is most useful when the model lacks scan-level supervision to otherwise aggregate global context.

Learned Anatomy Associations. Using the inter-anatomy transformer, we visualize the attention to understand any inter-anatomy relationships the model implicitly learns. Figure 3 shows the average attention weights between key organs across the test set for ACA models trained on Merlin (left) and CT-RATE (right). We find that both models learn anatomically plausible cross-anatomy associations that emerge purely from contrastive training, and the specific associations each model learns track the anatomical scope of its training dataset. Merlin is an abdominal CT cohort with findings concentrated in abdominal and GI pathology (Appendix Table 10), which is reflected in its strongest learned associations. The largest attention weights are amongst the stomach, small bowel, and colon, reflecting their continuity along the GI tract. Additionally, the liver and gallbladder attend to one another with the highest weight in each anatomy's row, mirroring their biliary connection. CT-RATE, by contrast, is a chest CT dataset with findings concentrated in cardiopulmonary pathology (Appendix Table 9), which is reflected in the strong attention weights directed to the lung and esophagus in the CT-RATE ACA model. The spleen also attends most strongly to the kidney and the kidney to the pancreas, structures that sit directly adjacent to one another behind the abdominal cavity.

Another notable observation is that self-attention along the diagonal is relatively weak in both datasets. We hypothesize that the fine-grained adaptation in ACA, built upon frozen foundation models, already provides strong anatomy-specific representations, reducing the need for the transformer to rein-

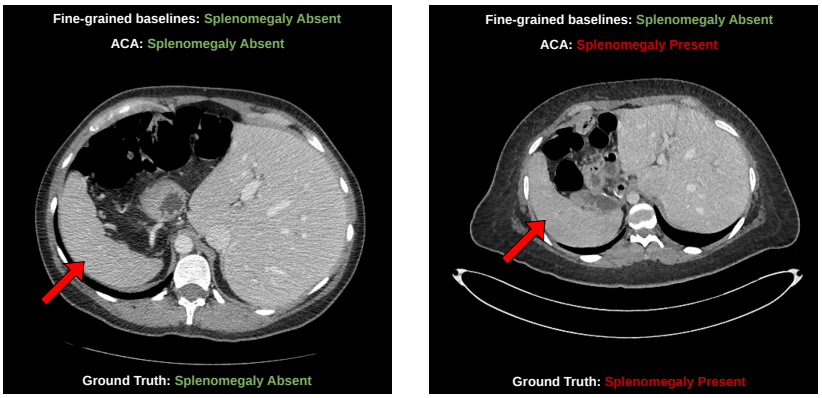

Fig. 4: Two CT scans from the Merlin dataset. The right patient has splenomegaly and the left does not. Despite the spleens appearing of comparable size in isolation, ACA correctly identifies the right patient as positive while all fine-grained baselines predict negative. The relative proportions of surrounding organs provide a discriminative signal that single-organ embeddings cannot capture.

force anatomy identity through self-attention. Instead, the adaptation module appears to focus on modeling inter-anatomy relationships, consistent with prior work viewing attention as a mechanism for information routing and interaction modeling rather than self-reinforcement [1].

Contextual Anatomy Reasoning. Many findings in the dataset require cross-anatomy context to detect reliably. A clear case study is organomegaly, where the challenge is not detecting a structure's presence but judging its relative size. For instance, a spleen of a given absolute cross-sectional area may be pathological in one patient and entirely normal in another, depending on body habitus and the proportions of surrounding structures. Figure 4 illustrates this, where two scans in the Merlin test set contain spleens of similar cross-section area, yet only one carries a splenomegaly label. Baselines that embed each organ independently have no mechanism to make this comparison, as each anatomy's representation is formed without reference to neighboring structures. ACA's inter-anatomy transformer, by attending jointly over all organ tokens in the same forward pass, can represent the relative size of each structure with respect to the others. The contrastive objective then ties this relational representation to anatomy-level text that naturally expresses such comparisons (*"the spleen is enlarged"*), grounding the model in contextual reasoning and facilitating a correct splenomegaly prediction for the example on the right at inference.

6 Discussion

ACA adapts frozen CT foundation model representations to provide structured, anatomy-level embeddings while facilitating scan-level, contextual reasoning with a lightweight inter-anatomy transformer. ACA consistently improves zero-shot

finding classification over global vision-language models and existing fine-grained adaptation methods, in-distribution and out-of-distribution across two large-scale datasets. Ablations confirm that that both core components, anatomical decomposition and scan-level report supervision, contribute independently to this improvement, and that restricting pooling to clinically relevant anatomical structures at inference time can provide an additional low-cost gain. The attention weights learned by the inter-anatomy transformer also reflect plausible associations and underlying dataset characteristics.

Computational Benefits of Adaptation. ACA is designed to adapt frozen foundation model representations with lightweight trainable modules, avoiding the computational cost of full end-to-end retraining that has traditionally been performed for fine-grained modeling [5, 20, 26]. A natural alternative is to fine-tune the backbone directly, which would significantly increase compute but could further improve performance. To explore this potential, we developed a LoRA-finetuned [18] variant of the full Merlin-based ACA model, trained end-to-end using the same anatomy decomposition and contrastive objective described in Section 3.3. The LoRA variant achieves an average AUROC of 0.8159 on the Merlin test set, compared to 0.8081 for the original ACA model, a gap of just 0.008. This suggests that the lightweight ACA adaptation of frozen foundation model representations recovers the majority of the benefit of full end-to-end training at a fraction of the computational cost, consistent with a broader trend in parameter-efficient adaptation [9,16,21]. A further practical advantage of this design is its modularity: because the adaptation module is decoupled from the backbone, it can be applied to any CT foundation model at low additional cost.

Limitations. Our comparisons to fVLM [26] and ViSD-Boost [5] are reimplementations adapted to operate on frozen foundation model embeddings, rather than the original end-to-end trained methods. This isolates each method's alignment strategy under a matched, frozen-embedding setting, but means our results support a narrower claim than outperforming fVLM and ViSD-Boost as originally published. A full end-to-end retraining of these methods would be needed to compare against their originally reported performance. Additionally, ACA's anatomy decomposition depends on TotalSegmentator's vocabulary, so structures outside it are not directly represented. Per-anatomy findings and normality labels used during training are extracted automatically with an LLM rather than verified by radiologists, which may introduce label noise, though all evaluations use each dataset's ground-truth scan-level labels. Our evaluation is limited to two datasets and backbones, with out-of-distribution comparisons restricted to the 7 findings shared between Merlin and CT-RATE, so broader generalization remains untested. Finally, this work addresses only zero-shot finding classification, and extending ACA to tasks such as segmentation, outcome prediction, and report generation is left to future work.

Conclusion. ACA combines fine-grained alignment with global contextualization for CT vision-language modeling by adapting existing foundation models rather than training from scratch. We see this as a practical path toward improving representation learning in a compute-efficient manner.

Acknowledgements

W.L. gratefully acknowledges funding support from the Ellison Foundation.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 4190–4197 (2020)
2. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Beeche, C., Kim, J., Tavolinejad, H., Zhao, B., Sharma, R., Duda, J., Gee, J., Dako, F., Verma, A., Morse, C., et al.: A pan-organ vision-language model for generalizable 3d ct representations. medRxiv (2025)
4. Blankemeier, L., Kumar, A., Cohen, J.P., Liu, J., Liu, L., Van Veen, D., Gardezi, S.J.S., Yu, H., Paschali, M., Chen, Z., et al.: Merlin: a computed tomography vision–language foundation model and dataset. Nature pp. 1–11 (2026)
5. Cao, W., Zhang, J., Shui, Z., Wang, S., Chen, Z., Li, X., Lu, L., Ye, X., Zhang, Q., Liang, T., et al.: Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23041–23050 (2025)
6. Cao, W., Zhang, J., Xia, Y., Mok, T.C., Li, Z., Ye, X., Lu, L., Zheng, J., Tang, Y., Zhang, L.: Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11238–11247 (2024)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
9. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems **35**, 16664–16678 (2022)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
11. Gao, Z., Zhang, G., Liang, H., Liu, J., Ma, L., Wang, T., Guo, Y., Chen, Y., Yan, Z., Chen, X., et al.: A lung ct vision foundation model facilitating disease diagnosis and medical imaging. Nature Communications (2025)
12. Hamamci, I.E., Er, S., Sekuboyina, A., Simsar, E., Tezcan, A., Simsek, A.G., Esirgun, S.N., Almas, F., Doğan, I., Dasdelen, M.F., et al.: Generatect: Text-conditional generation of 3d chest ct volumes. In: European Conference on Computer Vision. pp. 126–143. Springer (2024)
13. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., et al.: Generalist foundation models from a multimodal dataset for 3d computed tomography. Nature Biomedical Engineering pp. 1–19 (2026)

14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
16. He, X., Li, C., Zhang, P., Yang, J., Wang, X.E.: Parameter-efficient model adaptation for vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 817–825 (2023)
17. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., et al.: Vista3d: A unified segmentation foundation model for 3d medical imaging. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20863–20873 (2025)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
19. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3942–3951 (2021)
20. Lin, J., Xia, Y., Zhang, J., Yan, K., Cao, K., Lu, L., Luo, J., Zhang, L.: Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. arXiv preprint arXiv:2404.15272 (2024)
21. Liu, Y.C., Ma, C.Y., Tian, J., He, Z., Kira, Z.: Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks. Advances in neural information processing systems **35**, 36889–36901 (2022)
22. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Joint learning of localized representations from medical images and reports. In: European conference on computer vision. pp. 685–701. Springer (2022)
23. Niu, C., Lyu, Q., Carothers, C.D., Kaviani, P., Tan, J., Yan, P., Kalra, M.K., Whitlow, C.T., Wang, G.: Medical multimodal multitask foundation model for lung cancer screening. Nature Communications **16**(1), 1523 (2025)
24. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.: Vision foundation models for computed tomography. arXiv preprint arXiv:2501.09001 (2025)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
26. Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., Lu, L., Yang, L., Ye, X., Liang, T., Zhang, Q., et al.: Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. arXiv preprint arXiv:2501.14548 (2025)
27. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in neural information processing systems **35**, 33536–33549 (2022)
28. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: a vision foundation model for general-purpose segmentation on volumetric medical images. IEEE Transactions on Neural Networks and Learning Systems (2025)
29. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)

30. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
31. Wen, J.: Biological age shows that no organ system is an island. Nature **4**, 1182–1183 (2024)
32. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
33. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21372–21383 (2023)
34. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22873–22882 (2024)
35. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European Conference on Computer Vision. pp. 558–575. Springer (2022)
36. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
37. Xu, T., Hosseini, S., Anderson, C., Rinaldi, A., Krishnan, R.G., Martel, A.L., Goubran, M.: A generalizable 3d framework and model for self-supervised learning in medical imaging. npj Digital Medicine **8**(1), 639 (2025)
38. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
39. Zhang, H., Liu, Y., Dai, D., Yang, J., Liu, Q., Xie, Y., Wang, P.: Ca-gcl: Cross-anatomy global-local contrastive learning for robust 3d medical image understanding. arXiv preprint arXiv:2605.13544 (2026)
40. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical image analysis **67**, 101840 (2021)

A Preprocessing

Table 3: Mapping of TotalSegmentator labels to grouped anatomical structures for fine-grained modeling.

Idx	TotalSegmentator Name	Grouping Name	Grp
1	spleen	spleen	1
2	kidney_right	kidney	2
3	kidney_left	kidney	2
4	gallbladder	gallbladder	3
5	liver	liver	4
6	stomach	stomach	5
7	pancreas	pancreas	6
8	adrenal_gland_right	adrenal_gland	7
9	adrenal_gland_left	adrenal_gland	7
10	lung_upper_lobe_left	lung	8
11	lung_lower_lobe_left	lung	8
12	lung_upper_lobe_right	lung	8
13	lung_middle_lobe_right	lung	8
14	lung_lower_lobe_right	lung	8
15	esophagus	esophagus	9
16	trachea	trachea	10
17	thyroid_gland	thyroid_gland	11
18	small_bowel	small_bowel	12
19	duodenum	small_bowel	12
20	colon	colon	13
21	urinary_bladder	urinary_bladder	14
22	prostate	prostate	15
23	kidney_cyst_left	kidney	2
24	kidney_cyst_right	kidney	2
25	sacrum	sacrum	16
26	vertebrae_S1	sacrum	16
27	vertebrae_L5	lumbar_vertebrae	17
28	vertebrae_L4	lumbar_vertebrae	17
29	vertebrae_L3	lumbar_vertebrae	17
30	vertebrae_L2	lumbar_vertebrae	17
31	vertebrae_L1	lumbar_vertebrae	17
32	vertebrae_T12	thoracic_vertebrae	18
33	vertebrae_T11	thoracic_vertebrae	18
34	vertebrae_T10	thoracic_vertebrae	18
35	vertebrae_T9	thoracic_vertebrae	18
36	vertebrae_T8	thoracic_vertebrae	18
37	vertebrae_T7	thoracic_vertebrae	18
38	vertebrae_T6	thoracic_vertebrae	18
39	vertebrae_T5	thoracic_vertebrae	18
40	vertebrae_T4	thoracic_vertebrae	18
41	vertebrae_T3	thoracic_vertebrae	18
42	vertebrae_T2	thoracic_vertebrae	18

Idx	TotalSegmentator Name	Grouping Name	Grp
43	vertebrae_T1	thoracic_vertebrae	18
44	vertebrae_C7	cervical_vertebrae	19
45	vertebrae_C6	cervical_vertebrae	19
46	vertebrae_C5	cervical_vertebrae	19
47	vertebrae_C4	cervical_vertebrae	19
48	vertebrae_C3	cervical_vertebrae	19
49	vertebrae_C2	cervical_vertebrae	19
50	vertebrae_C1	cervical_vertebrae	19
51	heart	heart	20
52	aorta	aorta	21
53	pulmonary_vein	pulmonary_vein	22
54	brachiocephalic_trunk	brachiocephalic_trunk	23
55	subclavian_artery_right	subclavian_artery	24
56	subclavian_artery_left	subclavian_artery	24
57	common_carotid_artery_right	common_carotid_artery	25
58	common_carotid_artery_left	common_carotid_artery	25
59	brachiocephalic_vein_left	brachiocephalic_vein	26
60	brachiocephalic_vein_right	brachiocephalic_vein	26
61	atrial_appendage_left	heart	20
62	superior_vena_cava	vena_cava	27
63	inferior_vena_cava	vena_cava	27
64	portal_vein_and_splenic_vein	portal_vein_and_splenic_vein	28
65	iliac_artery_left	iliac_artery	29
66	iliac_artery_right	iliac_artery	29
67	iliac_vena_left	iliac_vena	30
68	iliac_vena_right	iliac_vena	30
69	humerus_left	humerus	31
70	humerus_right	humerus	31
71	scapula_left	scapula	32
72	scapula_right	scapula	32
73	clavicula_left	clavicula	33
74	clavicula_right	clavicula	33
75	femur_left	femur	34
76	femur_right	femur	34
77	hip_left	hip	35
78	hip_right	hip	35
79	spinal_cord	spinal_cord	36
80	gluteus_maximus_left	gluteus	37
81	gluteus_maximus_right	gluteus	37
82	gluteus_medius_left	gluteus	37
83	gluteus_medius_right	gluteus	37
84	gluteus_minimus_left	gluteus	37
85	gluteus_minimus_right	gluteus	37
86	autochthon_left	autochthon	38
87	autochthon_right	autochthon	38
88	iliopsoas_left	iliopsoas	39
89	iliopsoas_right	iliopsoas	39
90	brain	brain	40

Idx	TotalSegmentator Name	Grouping Name	Grp
91	skull	skull	41
92	rib_left_1	rib	42
93	rib_left_2	rib	42
94	rib_left_3	rib	42
95	rib_left_4	rib	42
96	rib_left_5	rib	42
97	rib_left_6	rib	42
98	rib_left_7	rib	42
99	rib_left_8	rib	42
100	rib_left_9	rib	42
101	rib_left_10	rib	42
102	rib_left_11	rib	42
103	rib_left_12	rib	42
104	rib_right_1	rib	42
105	rib_right_2	rib	42
106	rib_right_3	rib	42
107	rib_right_4	rib	42
108	rib_right_5	rib	42
109	rib_right_6	rib	42
110	rib_right_7	rib	42
111	rib_right_8	rib	42
112	rib_right_9	rib	42
113	rib_right_10	rib	42
114	rib_right_11	rib	42
115	rib_right_12	rib	42
116	sternum	sternum	43
117	costal_cartilages	costal_cartilages	44

B Zero-Shot Scoring and Anatomy-Guided Evaluation

For all models, a scan-level score is computed by averaging projected anatomy embeddings across all present anatomical structures $\mathcal{P}$ and taking the difference in cosine similarity to each class:

$$c_{\text{mean}} = \cos\left(\frac{1}{|\mathcal{P}|}\sum_{a\in\mathcal{P}} f_a^{\text{img}},\ f^+\right) - \cos\left(\frac{1}{|\mathcal{P}|}\sum_{a\in\mathcal{P}} f_a^{\text{img}},\ f^-\right) \tag{7}$$

A finding is predicted as positive when $c_{\text{mean}} > 0$. For ACA, we additionally evaluate an anatomy-guided scoring variant that exploits the per-anatomy structure of the model's representations. Rather than pooling across all present anatomical structures, the scan embedding for each finding c is computed by restricting to an anatomically relevant subset $\mathcal{P}_c$:

$$c_{\text{guided}} = \cos\left(\frac{1}{|\mathcal{P}_c|}\sum_{a\in\mathcal{P}_c} f_a^{\text{img}},\ f^+\right) - \cos\left(\frac{1}{|\mathcal{P}_c|}\sum_{a\in\mathcal{P}_c} f_a^{\text{img}},\ f^-\right) \tag{8}$$

The mapping from finding to subset is contained in Table 4. For findings without a well-defined anatomy subset (e.g., lymphadenopathy, free air), $\mathcal{P}_c$ falls back to all present structures. The final anatomy-guided score is a linear blend with the mean-pool score c_{mean}:

$$c = \alpha \cdot c_{\text{guided}} + (1 - \alpha) \cdot c_{\text{mean}} \tag{9}$$

where $\alpha \in \{0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$ is selected on the validation set by maximizing macro-average AUROC, and the same value is applied at test time for both **ACA (Anatomy-Guided)** and **ACA w/o $\mathcal{L}_{\text{scan}}$ (Anatomy-Guided)**.

C Implementation Details

All models are trained with the same optimization setup. Table 5 lists the shared training hyperparameters, and Table 6 lists the architecture hyperparameters for each model. Where Merlin and CT-RATE differ due to their respective backbone output dimensions, both values are shown as Merlin / CT-RATE.

LLM Prompt Templates for Anatomy Information Extraction

Mention Detection Prompt

```
{section}
You are a professional radiologist. Please determine if the
anatomy ({anatomy}{alias_clause}) is mentioned in this CT
image report.
Please answer directly with "Yes" or "No".
```

Information Extraction Prompt

```
{section} You are a professional radiologist. Please extract the
descriptive information about the specific anatomy
({anatomy}{alias_clause}) from this CT image diagnostic report.

Please follow these guidelines:
1. Precise extraction: Extract the descriptive information
   directly related to {anatomy} from the report.
2. Specify anatomical details: If the report mentions specific
   areas, parts, or anatomical details of {anatomy}, make sure to
   include this information in the description.
3. Concise and clear: Directly extract the report content,
   avoiding unnecessary explanations or background.
4. Format requirement: Return the information in the format
       "{anatomy}: descriptive information"
   ensuring {anatomy} is used as the unified prefix.
   Even if the anatomy has multiple independent parts or
   lateral characteristics, treat it as a single anatomy
   and return one comprehensive description.
```

Abnormality Classification Prompt

```
{description}
You are a professional radiologist. Based only on the description
above, is the {anatomy} abnormal?

Answer directly with:
"Yes" = abnormal finding present
"No" = normal or incidental finding only
```

Fig. 5: Prompt templates used for structured anatomy-level information extraction from CT radiology reports. Since reports describe findings globally rather than per organ, a three-stage pipeline is applied. The **mention detection** prompt screens whether a given anatomy is discussed at all, avoiding spurious extractions for absent organs. The **information extraction** prompt isolates organ-specific descriptive text, using alias hints and formatting constraints to produce a clean, unified description across substructures. The **abnormality classification** prompt labels the result as normal or abnormal, providing the supervision signal for false negative reduction during training. In all templates, `{section}` is the raw report text, `{alias_clause}` optionally appends clinical synonyms (e.g., *also known as splenic* for the spleen), and `{description}` is the organ-specific text extracted by the preceding stage. Full implementation details are provided in `preprocess_reports.py` in the given code.

Example Extracted Anatomy-Level Text and Normality Labels

Kidney (abnormal)

"In both kidneys included in the examination, appearances evaluated in favor of multiple cysts are observed."

Heart (abnormal)

"Heart size increased. Calcific atheroma plaques are noted in the coronary arteries, associated with the heart's supply vessels. No pericardial effusion or pericardial thickness increase was observed."

Lung (abnormal)

"Minimal bronchiectatic changes and peribronchial thickness increases are observed at the level of the hilum of both lungs. A sequela calcific pulmonary nodule is present in the posterobasal segment of the right lung lower lobe."

Spleen (normal)

"Spleen shows no significant abnormalities."

All remaining structures (normal)

e.g., "Liver shows no significant abnormalities.", "Stomach shows no significant abnormalities.", "Pancreas shows no significant abnormalities.", etc.

Fig. 6: Example anatomy-specific findings extracted from a radiology report using Qwen3-4B-Instruct, along with the corresponding binary normality label used in the soft-target contrastive loss described in Section 3.3. Anatomical structures with no reported findings receive a default normal description.

Example Zero-Shot Finding Prompts

Atelectasis

- *Positive:* "atelectasis", "bibasilar atelectasis", "subsegmental atelectasis", "lobar atelectasis"
- *Negative:* "no atelectasis", "lungs are clear", "no airspace opacities or atelectasis"

Cardiomegaly

- *Positive:* "cardiomegaly", "enlarged heart", "cardiac silhouette is enlarged", "moderate cardiomegaly"
- *Negative:* "no cardiomegaly", "heart is normal in size", "cardiac silhouette is normal"

Hiatal Hernia

- *Positive:* "hiatal hernia", "sliding hiatal hernia", "paraesophageal hernia"
- *Negative:* "no hiatal hernia", "hiatus is normal", "no herniation through the diaphragmatic hiatus"

Fig. 7: Example positive and negative text prompts used for zero-shot finding classification. For each finding, a set of positive prompts describing the pathology and negative prompts describing normal appearance are encoded by the frozen text encoder and averaged to obtain class-level embeddings f^+ and f^-.

Finding	Anatomy Subset
Atelectasis	Lung
Pleural Effusion	Lung
Emphysema	Lung
Lung Nodule	Lung
Lung Opacity	Lung
Pulmonary Fibrotic Sequela	Lung
Mosaic Attenuation Pattern	Lung
Consolidation	Lung
Bronchiectasis	Lung
Interlobular Septal Thickening	Lung
Peribronchial Thickening	Lung, Trachea
Cardiomegaly	Heart
Pericardial Effusion	Heart
Coronary Calcification	Heart
Arterial Wall Calcification	Aorta, Iliac, Subclavian, & Common Carotid Arteries
Coronary Artery Wall Calcification	Heart, Aorta
Aortic Valve Calcification	Heart, Aorta
Abdominal Aortic Aneurysm	Aorta
Atherosclerosis	Aorta, Iliac Artery
Hiatal Hernia	Stomach, Esophagus
Hepatomegaly	Liver
Hepatic Steatosis	Liver
Biliary Ductal Dilation	Liver, Gallbladder
Gallstones	Gallbladder
Surgically Absent Gallbladder	Gallbladder
Splenomegaly	Spleen
Renal Cyst	Kidney
Renal Hypodensities	Kidney
Hydronephrosis	Kidney
Pancreatic Atrophy	Pancreas
Prostatomegaly	Prostate
Submucosal Edema	Small Bowel, Colon
Bowel Obstruction	Small Bowel, Colon
Appendicitis	Colon
Thrombosis	Portal/Splenic Vein, Vena Cava, Iliac Vein
Metastatic Disease	Liver, Lung, Rib, Lumbar Vertebrae, Thoracic Vertebrae
Osteopenia	Lumbar Vertebrae, Thoracic Vertebrae, Rib
Fracture	Rib, Lumbar Vertebrae, Thoracic Vertebrae
Ascites	Liver, Spleen
Anasarca	All present structures
Lymphadenopathy	All present structures
Free Air	All present structures
Medical Material	All present structures

Table 4: Anatomy subsets used for anatomy-guided zero-shot scoring per finding. Findings evaluated on both Merlin and CT-RATE datasets use the respective dataset's assignment.

Hyperparameter	**Value**
Batch size	64
Epochs	50
Early stopping patience	5
Learning rate	1×10^{-4}
Weight decay	0.05
Warmup fraction	0.05
Gradient clip	1.0
Projection hidden dim	512
Projection output dim	256
Temperature init (τ)	0.07

Table 5: Training hyperparameters shared across all models and both datasets.

Model	**Hyperparameter**	**Value**
MLP	Visual input dim	2048 / 512
fVLM	Multi-scale input dim	3840 / 1024
	Cross-attention heads	4
	Cross-attention dropout	0.1
Spatial Transformer	Spatial features	49 / 144
	Token input dim	2048 / 512
	Transformer hidden dim	1024
	Transformer layers	2
	Transformer heads	4
	Dropout	0.1
ACA w/o $\mathcal{L}_{\text{scan}}$, ACA w/o $\mathcal{L}_{\text{anatomy}}$ ACA	Visual input dim	2048 / 512
	Transformer hidden dim	1024
	Positional encoding dim	44
	Transformer layers	2
	Transformer heads	4
	Dropout	0.1
ViSD-Boost	Multi-scale input dim	3840 / 1024
	Cross-attention heads	4
	Cross-attention dropout	0.1
	VQ-VAE codebook size	512
	VQ-VAE embedding dim	512
	VQ-VAE Transformer layers	1
	VQ-VAE Transformer heads	8
	VQ commitment cost	0.25

Table 6: Architecture hyperparameters per model. Where values differ between the Merlin and CT-RATE backbones, both are shown as Merlin / CT-RATE.

Table 7: Merlin: positive finding counts per split (unlabeled excluded).

Finding	Train	Val	Test
Atelectasis	2665	902	903
Surgically Absent Gallbladder	2515	805	843
Atherosclerosis	2493	778	879
Pleural Effusion	2484	794	834
Renal Cyst	1650	544	546
Ascites	1048	366	365
Anasarca	1020	318	352
Hiatal Hernia	886	280	312
Hepatic Steatosis	793	250	272
Gallstones	478	148	148
Fracture	470	175	177
Pancreatic Atrophy	453	118	161
Osteopenia	376	119	147
Submucosal Edema	359	146	144
Cardiomegaly	352	123	119
Splenomegaly	324	109	141
Prostatomegaly	306	115	81
Renal Hypodensities	302	122	122
Hydronephrosis	266	72	112
Thrombosis	247	80	77
Bowel Obstruction	211	64	72
Aortic Valve Calcification	206	61	72
Coronary Calcification	166	57	63
Hepatomegaly	165	44	57
Biliary Ductal Dilation	110	33	52
Appendicitis	106	32	43
Lymphadenopathy	104	28	49
Free Air	100	37	45
Metastatic Disease	83	24	46
Abdominal Aortic Aneurysm	71	16	17

Table 8: CT-RATE: positive finding counts per split.

Finding	Train	Val	Test
Lung nodule	17032	4349	1361
Lung opacity	13845	3575	1184
Arterial wall calcification	10697	2679	867
Pulmonary fibrotic sequela	10127	2461	831
Atelectasis	9843	2418	713
Lymphadenopathy	9684	2534	789
Coronary artery wall calcification	9634	2390	765
Emphysema	7291	1831	600
Consolidation	6617	1701	581
Hiatal hernia	5414	1337	417
Medical material	4668	1150	313
Pleural effusion	4582	1122	376
Cardiomegaly	4259	1049	325
Peribronchial thickening	3960	1013	355
Bronchiectasis	3825	907	330
Interlobular septal thickening	2976	769	249
Mosaic attenuation pattern	2871	767	253
Pericardial effusion	2712	700	226

Finding	Positive	Negative
CT-RATE In-Distribution		
Medical Material	313	2726
Arterial Wall Calcification	867	2172
Cardiomegaly	325	2714
Pericardial Effusion	226	2813
Coronary Artery Wall Calcification	765	2274
Hiatal Hernia	417	2622
Lymphadenopathy	789	2250
Emphysema	600	2439
Atelectasis	713	2326
Lung Nodule	1361	1678
Lung Opacity	1184	1855
Pulmonary Fibrotic Sequela	831	2208
Pleural Effusion	376	2663
Mosaic Attenuation Pattern	253	2786
Peribronchial Thickening	355	2684
Consolidation	581	2458
Bronchiectasis	330	2709
Interlobular Septal Thickening	249	2790
CT-RATE Out-of-Distribution		
Atelectasis	713	2326
Cardiomegaly	325	2714
Hiatal Hernia	417	2622
Lymphadenopathy	789	2250
Pleural Effusion	376	2663
Arterial Wall Calcification	867	2172
Coronary Artery Wall Calcification	765	2274

Table 9: Number of positive and sampled negative cases for each finding evaluated in the in-distribution and out-of-distribution settings on the CT-RATE dataset.

Finding	Positive	Negative
Merlin In-Distribution		
Submucosal Edema	144	162
Renal Hypodensities	122	1355
Aortic Valve Calcification	72	69
Coronary Calcification	63	372
Thrombosis	77	32
Metastatic Disease	46	123
Pancreatic Atrophy	161	3197
Renal Cyst	546	1412
Osteopenia	147	1069
Surgically Absent Gallbladder	843	2186
Atelectasis	903	990
Abdominal Aortic Aneurysm	17	79
Anasarca	352	1992
Hiatal Hernia	312	1847
Lymphadenopathy	49	205
Prostatomegaly	81	1213
Biliary Ductal Dilation	52	211
Cardiomegaly	119	291
Splenomegaly	141	3225
Hepatomegaly	57	1150
Atherosclerosis	879	55
Ascites	365	183
Pleural Effusion	834	556
Hepatic Steatosis	272	27
Appendicitis	43	52
Gallstones	148	248
Hydronephrosis	112	2079
Bowel Obstruction	72	1345
Free Air	45	929
Fracture	177	196
Merlin Out-of-Distribution		
Coronary Calcification	63	372
Atelectasis	903	990
Hiatal Hernia	312	1847
Lymphadenopathy	49	205
Cardiomegaly	119	291
Atherosclerosis	879	55
Pleural Effusion	834	556

Table 10: Number of positive and sampled negative cases for each finding evaluated in the in-distribution and out-of-distribution settings on the Merlin datasets.

Table 11: Per-finding AUROC on the Merlin in-distribution test set. Macro average is computed over all 30 findings. Overlapping Macro average restricts the macro-average to the 7 findings shared with CT-RATE, matching the finding set used for out-of-distribution evaluation (Table 13).

Finding	Merlin	Spatial Transformer	MLP	fVLM	ViSD-Boost	ACA
Submucosal edema	0.7415	0.7309	0.7220	0.6832	0.6923	0.7185
Renal hypodensities	0.6770	0.6731	0.7218	0.7081	0.7405	0.7512
Aortic valve calcification	0.7997	0.8133	0.9506	0.9578	0.9557	0.9637
Coronary calcification	0.8057	0.8198	0.8392	0.8442	0.8469	0.8496
Thrombosis	0.6269	0.5759	0.5481	0.6041	0.6057	0.6175
Metastatic disease	0.7186	0.7363	0.8305	0.8801	0.8643	0.8719
Pancreatic atrophy	0.7180	0.7394	0.7744	0.7861	0.7706	0.7653
Renal cyst	0.6204	0.6281	0.7436	0.6704	0.6754	0.7795
Osteopenia	0.7435	0.7932	0.9199	0.9146	0.8990	0.9279
Surgically absent gallbladder	0.9741	0.9688	0.9214	0.7125	0.8299	0.9794
Atelectasis	0.6757	0.7256	0.7800	0.6530	0.7567	0.8189
Abdominal aortic aneurysm	0.7501	0.8679	0.7791	0.7732	0.7492	0.7826
Anasarca	0.9279	0.9140	0.9002	0.9303	0.9352	0.9378
Hiatal hernia	0.6334	0.6797	0.7721	0.7625	0.7295	0.7604
Lymphadenopathy	0.7870	0.7291	0.6910	0.6541	0.7131	0.7469
Prostatomegaly	0.6765	0.7324	0.7535	0.7425	0.7315	0.7658
Biliary ductal dilation	0.7908	0.8189	0.7831	0.7898	0.8044	0.8181
Cardiomegaly	0.8247	0.8545	0.8801	0.8458	0.8663	0.8925
Splenomegaly	0.9012	0.8836	0.9120	0.9058	0.9197	0.9267
Hepatomegaly	0.7645	0.7486	0.7799	0.8221	0.8318	0.8672
Atherosclerosis	0.9621	0.9602	0.8683	0.8382	0.8808	0.8899
Ascites	0.9022	0.9114	0.8711	0.8351	0.8532	0.9310
Pleural effusion	0.9294	0.9246	0.8691	0.8525	0.9204	0.9455
Hepatic steatosis	0.6747	0.7628	0.7698	0.7809	0.7192	0.6580
Appendicitis	0.7185	0.7123	0.7333	0.5840	0.5012	0.7344
Gallstones	0.7566	0.7221	0.6920	0.6212	0.7251	0.7374
Hydronephrosis	0.7295	0.7304	0.7313	0.7611	0.7499	0.7655
Bowel obstruction	0.8668	0.8888	0.8868	0.8489	0.7493	0.8791
Free air	0.7648	0.7802	0.8249	0.7487	0.7294	0.7961
Fracture	0.7246	0.6947	0.6941	0.5871	0.6622	0.7615
Overlapping Macro average	0.8026	0.8134	0.8143	0.7786	0.8163	0.8434
Macro average	0.7729	0.7840	0.7981	0.7699	0.7803	0.8213

Table 12: Per-finding AUROC on the CT-RATE in-distribution test set. Macro average is computed over all 18 findings. Overlapping Macro average restricts the macro-average to the 7 findings shared with Merlin, matching the finding set used for out-of-distribution evaluation (Table 14).

Finding	CT-CLIP	Spatial Transformer	MLP	fVLM	ViSD-Boost	ACA
Medical material	0.6462	0.6208	0.6252	0.6866	0.6784	0.6798
Arterial wall calcification	0.8492	0.7892	0.6513	0.8135	0.7810	0.8768
Cardiomegaly	0.8686	0.7745	0.6677	0.7815	0.8420	0.8973
Pericardial effusion	0.7053	0.6543	0.6337	0.7394	0.7692	0.7787
Coronary artery wall calcification	0.8377	0.7655	0.7052	0.7953	0.7748	0.8505
Hiatal hernia	0.7187	0.6843	0.6466	0.6532	0.6605	0.7956
Lymphadenopathy	0.6807	0.6226	0.5633	0.6444	0.5585	0.6543
Emphysema	0.7065	0.6256	0.5965	0.6739	0.6715	0.7506
Atelectasis	0.6449	0.6220	0.6058	0.6567	0.6098	0.6679
Lung nodule	0.5381	0.4701	0.5473	0.5283	0.4869	0.5918
Lung opacity	0.6743	0.5990	0.5440	0.6095	0.5814	0.6078
Pulmonary fibrotic sequela	0.5713	0.5197	0.5351	0.5651	0.5246	0.6230
Pleural effusion	0.8952	0.8306	0.6624	0.8485	0.7648	0.8744
Mosaic attenuation pattern	0.7470	0.7179	0.6584	0.7176	0.6818	0.7519
Peribronchial thickening	0.6079	0.5412	0.6506	0.6678	0.5395	0.6735
Consolidation	0.6992	0.6249	0.5302	0.6291	0.6035	0.6442
Bronchiectasis	0.6560	0.5675	0.5815	0.5871	0.5949	0.7182
Interlobular septal thickening	0.7005	0.6100	0.6279	0.7180	0.5973	0.7245
Overlapping Macro average	0.7850	0.7270	0.6432	0.7419	0.7131	0.8024
Macro average	0.7082	0.6467	0.6129	0.6842	0.6511	0.7311

Table 13: Per-finding AUROC on the Merlin out-of-distribution test set.

Finding	CT-CLIP	Spatial Transformer	MLP	fVLM	ViSD-Boost	ACA
Atelectasis	0.5784	0.5903	0.5871	0.5763	0.7383	0.6939
Cardiomegaly	0.6443	0.5604	0.5535	0.5456	0.5789	0.6598
Hiatal hernia	0.5995	0.6398	0.6168	0.5795	0.6553	0.6822
Lymphadenopathy	0.4493	0.4750	0.4812	0.5282	0.5067	0.5687
Pleural effusion	0.5838	0.6409	0.6951	0.6421	0.6765	0.7691
Atherosclerosis	0.5662	0.4954	0.5579	0.5624	0.6448	0.6438
Coronary calcification	0.6541	0.5220	0.4974	0.4770	0.4968	0.6886
Macro average	0.5822	0.5605	0.5699	0.5587	0.6139	0.6723

Table 14: Per-finding AUROC on the CT-RATE out-of-distribution test set.

Finding	Merlin	Spatial Transformer	MLP	fVLM	ViSD-Boost	ACA
Atelectasis	0.5753	0.5856	0.5845	0.4770	0.5951	0.6050
Cardiomegaly	0.7875	0.8044	0.8519	0.8265	0.8583	0.8448
Hiatal hernia	0.6120	0.6192	0.6711	0.6766	0.6487	0.6565
Lymphadenopathy	0.5873	0.5927	0.5350	0.5166	0.5940	0.6119
Pleural effusion	0.9002	0.9118	0.8462	0.8062	0.9024	0.8789
Arterial wall calcification	0.6853	0.7014	0.8409	0.8042	0.7880	0.7853
Coronary artery wall calcification	0.6959	0.6871	0.8424	0.8178	0.8024	0.7979
Macro average	0.6919	0.7003	0.7389	0.7036	0.7413	0.7400