Segmentation Helps Understanding: Mask-Infused Vision-Language Pre-training for 3D Medical Images

Yuqi Hu [1] **Xufang Luo** [2] **Zilong Wang** [2] **Dongsheng Li** [2] **Lili Qiu** [2]

Abstract

Pretraining effective 3D medical image encoders is crucial for downstream tasks such as diagnosis and prognosis. Existing vision-language methods learn global semantics from paired radiology reports but often miss fine-grained cues like small lesions. We introduce `SegVL`, a unified contrastive learning framework that integrates segmentation masks into vision-language pretraining. `SegVL` aligns voxel-level features with segmentation labels using mask names as textual anchors and enhances image-text contrast via segmentation-informed features. A Tversky loss addresses class imbalance, and a lightweight decoder preserves encoder capacity. Experiments show `SegVL` outperforms prior methods on multiple classification and segmentation benchmarks, highlighting the complementary strengths of segmentation and language supervision.

1. Introduction

Medical vision-language pretraining (Med-VLP) aims to learn generalizable image representations by aligning medical images with paired radiology reports. While early Med-VLP efforts focused on 2D modalities like chest X-rays (Zhang et al., 2022; Boecking et al., 2022; Wu et al., 2023), recent work has extended to 3D CT volumes (Hamamci et al., 2024; Blankemeier et al., 2024), which offer richer anatomical context but introduce challenges due to their higher dimensionality and the sparse distribution of clinically relevant signals. In 3D scans, for example, small lesions such as lung nodules may occupy less than 0.005% of the volume, making them hard to capture via global image-text alignment.

To address this, recent methods have moved toward finer-grained alignment. CT-FM (Pai et al., 2025) introduces patch-level contrastive learning, MG-3D (Ni et al., 2024) aligns CTs with individual report sentences, and fVLM (Shui et al., 2025) uses an external segmentation model and LLM to align organ-level sub-volumes with organ descriptions. These approaches improve supervision granularity but often rely on complex pipelines or external tools. Moreover, alignment typically stops at the sub-volume level; no prior work has tackled voxel-level alignment.

Meanwhile, segmentation data—despite being unpaired with text—provide dense voxel-level labels and are widely available. Prior work like MedicalNet (Chen et al., 2019) uses segmentation for pretraining, and large-scale models such as SAM (Kirillov et al., 2023) and MedSAM (Zhu et al., 2024) leverage segmentation masks for dense prediction. However, these approaches focus solely on segmentation and often employ heavy decoders, which may absorb important features and hinder encoder generalization.

In this work, we propose `SegVL`, a simple yet effective framework that directly integrates segmentation supervision into Med-VLP via voxel-mask contrastive learning. By aligning voxel embeddings with segmentation mask names and using a lightweight decoder, `SegVL` enhances fine-grained representation learning without increasing architectural complexity.

2. Methodology

Overall, `SegVL` is a kind of VLP framework, whose innovation lies in infusing segmentation data into the pretraining process through a unified contrastive learning approach. In this section, we first detail the encoding of 3D medical images and their corresponding textual reports, which forms the basis of our VLP framework. Subsequently, in the second subsection, we elaborate on how we effectively incorporate segmentation data by adapting the existing text encoder and employing a contrastive learning strategy. Finally, the third subsection describes our VL pretraining component, which is further augmented by the inclusion of segmentation information to facilitate a more comprehensive understand-

[1]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China [2]Microsoft Research Asia, Shanghai, China. Correspondence to: Xufang Luo <xufang.luo@microsoft.com>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

ing of the visual and textual modalities.

2.1. Image and Report Feature Extraction

Given a 3D medical volume $V_i \in \mathbb{R}^{H\times W\times D}$ and its paired report R_i, we extract their features using separate encoders.

3D Image Encoding A 3D Vision Transformer $f_v(\cdot)$ encodes V_i into spatial tokens:

$$\mathbf{X}_i = f_v(V_i) \in \mathbb{R}^{h\times w\times d\times d_f}, \tag{1}$$

where h, w, d are the downsampled dimensions and d_f is the feature size. Global average pooling followed by an MLP projection p_{img} yields the volume-level embedding:

$$\mathbf{z}_i^{\text{img}} = p_{\text{img}}(\text{AvgPool}(\mathbf{X}_i)) \in \mathbb{R}^{d_f}.$$

Report Encoding We use a transformer-based text encoder $f_t(\cdot)$ (initialized from BioViL (Boecking et al., 2022)) to embed the report:

$$\mathbf{T}_i = f_t(R_i) \in \mathbb{R}^{L_r\times d_r}, \tag{2}$$

where L_r is the token length and d_r is the embedding size. The [CLS] token is projected via an MLP p_{rep} to obtain the report-level representation:

$$\mathbf{z}_i^{\text{rep}} = p_{\text{rep}}\left(T_i^{[\text{CLS}]}\right) \in \mathbb{R}^{d_f}, \quad \text{where} \quad T_i^{[\text{CLS}]} \in \mathbb{R}^{d_r}.$$

2.2. Voxel-Mask Contrastive Learning for Infusing Segmentation Data

To enable fine-grained supervision from segmentation data, we design a voxel-mask contrastive learning framework that aligns voxel embeddings with textual embeddings of mask names using a shared text encoder. Instead of predicting masks via a decoder, we treat voxel embeddings as contrastive queries and segmentation label names as anchors in the shared embedding space. This encourages the encoder to learn semantically meaningful voxel-level features from segmentation data.

Voxel Embeddings Given the 3D tokens $\mathbf{X}_i$ from Eq. 1, we apply a lightweight MLP segmentation head $f_{\text{seg}}(\cdot)$ followed by a reshape operation to obtain voxel embeddings:

$$\begin{aligned}\hat{\mathbf{X}}_i &= f_{\text{seg}}(\mathbf{X}_i) \in \mathbb{R}^{h\times w\times d\times d_{\text{mlp}}}\\ \mathbf{V}_i &= \text{reshape}(\hat{\mathbf{X}}_i) \in \mathbb{R}^{H\times W\times D\times n_{\text{logits}}},\end{aligned} \tag{3}$$

where each voxel in the original volume has an embedding of size $\mathbb{R}^{n_{\text{logits}}}$. We emphasize the use of a lightweight MLP head instead of a heavy decoder, as commonly seen in segmentation models (Li et al., 2023; Hatamizadeh et al., 2022). This design preserves encoder capacity by preventing the decoder from absorbing key information, aligning with our goal of learning transferable image representations (see Appendix. 3).

Mask Name Embeddings To encode class names, we use prompts of the form $\mathbf{p}_c =$ "This is ⟨mask name⟩" for each of the C segmentation classes. These prompts are fed into the same text encoder $f_t(\cdot)$ used for reports (Eq. 2):

$$\mathbf{M} = f_t(\mathbf{p}_c) \in \mathbb{R}^{C\times L\times d_f},$$

where L is the token length and d_f the embedding size. We extract [CLS] tokens and project them via a text head $f_{\text{text}}(\cdot)$ to obtain final mask name embeddings:

$$\mathbf{M}^{\text{mask}} = f_{\text{text}}(\mathbf{M}) \in \mathbb{R}^{C\times n_{\text{logits}}},$$

where n_{logits} is the output dimension.

Voxel-Mask Contrastive Learning We align voxel embeddings with mask name embeddings using contrastive learning. Voxels labeled as foreground serve as positives, background as negatives. While InfoNCE loss (Oord et al., 2018) is a natural choice, it underperforms under extreme class imbalance (see Section 3). We instead adopt a Tversky-based loss.

We first compute cosine similarity between each voxel embedding $\mathbf{V}_i$ and each mask embedding $\mathbf{M}_j^{\text{mask}}$:

$$P_{i,j} = \frac{1}{2}\left(\frac{\mathbf{V}_i \cdot {\mathbf{M}_j^{\text{mask}}}^\top}{\|\mathbf{V}_i\|\|\mathbf{M}_j^{\text{mask}}\|} + 1\right) \in \mathbb{R}^C, \tag{4}$$

where $P_{i,j}$ is the predicted similarity in [0,1] for voxel i and class j.

The final voxel-mask contrastive loss is:

$$\mathcal{L}_{\text{VM}} = \frac{1}{C}\sum_{j=1}^{C}\frac{\sum_{i=1}^{N}\left(\alpha\hat{y}_{i,j}(1-P_{i,j}) + \beta(1-\hat{y}_{i,j})P_{i,j}\right)}{\sum_{i=1}^{N}\left(\hat{y}_{i,j}(1-P_{i,j}) + (1-\hat{y}_{i,j})P_{i,j} + \hat{y}_{i,j}P_{i,j}\right)} \tag{5}$$

where $\hat{y}_{i,j}$ is the one-hot label for voxel i and class j, and $N = H\times W\times D$ is the number of voxels. α and β balance false positives and false negatives.

2.3. Image-Text Contrastive Learning with Visual Enhancement from Segmentation

We now detail the image-report contrastive learning component within `SegVL`, with a particular emphasis on how we leverage segmentation information to further enhance the image features. Specifically, tokens learned by the segmentation head (i.e., $\hat{\mathbf{X}}_i$ in Eq. 3) are passed through an MLP (f_{VE}) and then aggregated using average pooling to obtain a single feature vector:

$$\mathbf{z}_i^{\text{seg}} = \text{AvgPool}(f_{\text{VE}}(\hat{\mathbf{X}}_i)) \in \mathbb{R}^{d_f}.$$

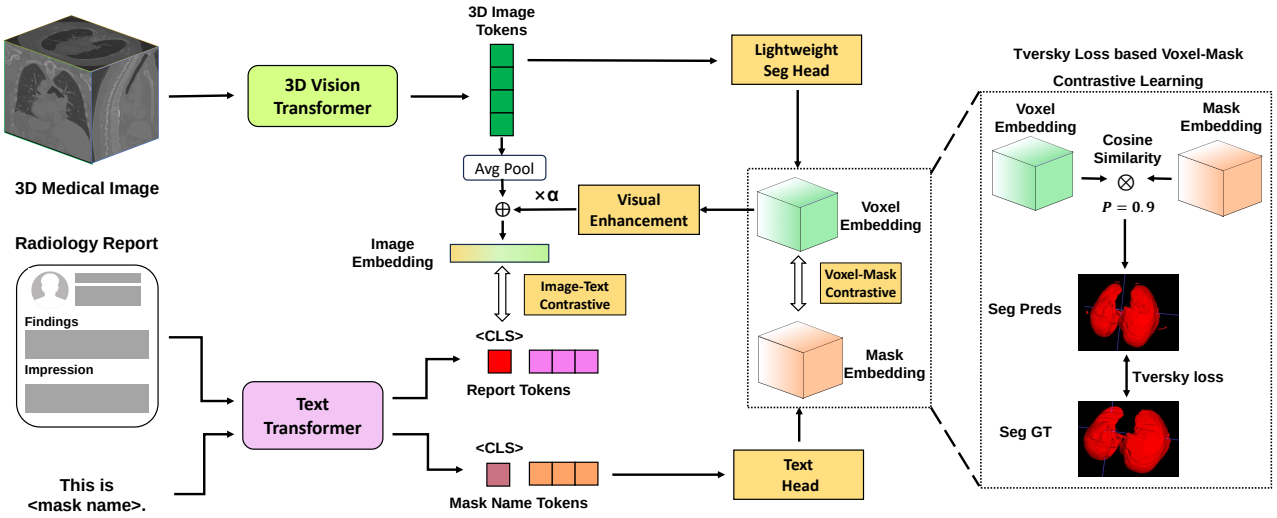

Figure 1. Architecture of our SegVL framework. The model jointly leverages image-report and image-segmentation data via two contrastive objectives. A 3D vision transformer encodes volumes into tokens and a lightweight segmentation head convert tokens into voxel embeddings. Voxel embeddings are supervised by voxel-mask contrastive learning using Tversky loss to handle class imbalance. Meanwhile, segmentation features are fused into global image embeddings through a visual enhancement module, which are contrasted with report embeddings to learn both high-level semantics and fine-grained anatomical cues.

The resulting pooled feature vector $\mathbf{z}_{\text{seg}}$ is combined with the original image embedding $\mathbf{z}_i^{\text{img}}$ as:

$$\mathbf{z}_i^{\text{img,seg}} = \mathbf{z}_i^{\text{img}} + \lambda \mathbf{z}_i^{\text{seg}},$$

where λ is a learnable weight initialized with 0 that controls the contribution of the segmentation features to the final enhanced image embedding. This weighted sum enhances the image embedding, allowing it to better capture both the global structure and local details, improving the image-text alignment in the contrastive learning framework.

The overall loss function is a weighted combine of Image-Text (denoted as $\mathcal{L}_{\text{IT}}$) and Voxel-Mask ($\mathcal{L}_{\text{VM}}$ in Eq. 5) contrastive learning loss:

$$\mathcal{L}_{\texttt{SegVL}} = \mathcal{L}_{\text{IT}}(\mathbf{z}_i^{\text{img,seg}}, \mathbf{z}_i^{\text{rep}}) + \alpha_{\text{seg}} \mathcal{L}_{\text{VM}}.$$

Here, we use InfoNCE loss (Oord et al., 2018) to implement $\mathcal{L}_{\text{IT}}$ and α_{seg} is a hyperparameter for balance.

3. Experiment

Experimental Setting To evaluate the effectiveness of SegVL, we conduct pretraining on the CT-RATE (Hamamci et al., 2024) and RadGenome-ChestCT (Zhang et al., 2024) datasets. CT-RATE provides over 50,000 chest CT volumes paired with radiology reports, while RadGenome augments these with segmentation masks spanning 197 anatomical categories. We select six major categories for segmentation supervision. Preprocessing follows CT-RATE protocol.

We assess generalization on two task types. For classification, we use four datasets: CT-RATE (internal), RadChest-CT (Draelos et al., 2021), CC-CCII (Zhang et al., 2020), and RICORD (Tsai et al., 2020), covering thoracic abnormalities and COVID-19. For segmentation, we evaluate on BTCV (Landman et al., 2015), TotalSegmentator (Wasserthal et al., 2023), and MSD (Antonelli et al., 2022), focusing on anatomical structure parsing and tumor localization.

3.1. Results on Classification Tasks

On multi-disease classification tasks in CT-RATE and RadChest-CT, as shown in Table. 1, SegVL consistently outperforms prior methods under both linear probing and finetuning. Compared to CT-CLIP, it improves AUC by +4.4% and +5.5% on CT-RATE, and by +4.9% and +6.6% on RadChest-CT. These gains highlight the effectiveness of incorporating segmentation data during pretraining, which enhances the encoder's ability to capture fine-grained features beyond global image-text alignment. SegVL also surpasses fVLM, achieving new state-of-the-art AUC scores across both datasets and settings.

On CC-CCII and RICORD, which require fine-grained discrimination for COVID-19 diagnosis, SegVL shows strong generalization in Table. 1. It improves over CT-CLIP by up to +4.1% AUC and surpasses fVLM across both datasets. Finetuning further boosts performance, with SegVL reaching 0.936 AUC on CC-CCII and 0.912 on RICORD, demon-

Table 1. AUC comparison of SegVL and baselines on downstream classification tasks under **linear probing** and **finetuning**.

Method	Linear Probing (AUC)				Finetuning (AUC)			
	CT-RATE	RadChest-CT	CC-CCII	RICORD	CT-RATE	RadChest-CT	CC-CCII	RICORD
CT-CLIP (Hamamci et al., 2024)	0.749	0.653	0.865	0.846	0.756	0.650	0.920	0.863
Merlin (Blankemeier et al., 2024)	0.770	0.677	0.877	0.854	0.762	0.694	0.919	0.879
UniMiSS (Xie et al., 2022)	\	\	0.841	0.862	\	\	0.913	0.891
T3D (Liu et al., 2023)	0.775	\	\	\	0.802	\	0.927	\
fVLM (Shui et al., 2025)	0.783	0.697	0.871	0.858	0.794	0.704	0.926	0.885
SegVL (ours)	**0.793**	**0.702**	**0.893**	**0.887**	**0.811**	**0.716**	**0.936**	**0.912**

Table 2. Dice score comparison of SegVL and baselines on downstream segmentation tasks under **linear probing** and **finetuning**.

Method	Linear Probing				Finetuning			
	TotalSegmentor	MSD-Lung	BTCV	Mean	TotalSegmentor	MSD-Lung	BTCV	Mean
CT-CLIP (Hamamci et al., 2024)	0.852	0.637	0.845	0.778	0.867	0.684	0.854	0.802
Merlin (Blankemeier et al., 2024)	0.859	0.645	0.848	0.784	0.870	0.689	0.855	0.805
UniMiSS (Xie et al., 2022)	0.854	0.642	0.838	0.778	0.868	0.677	0.850	0.798
fVLM (Shui et al., 2025)	0.861	0.639	0.849	0.783	0.874	0.686	0.861	0.807
SegVL (ours)	**0.878**	**0.675**	**0.863**	**0.805**	**0.889**	**0.733**	**0.872**	**0.831**

strating its transferability and adaptability to complex downstream targets.

3.2. Results on Semantic Segmentation Tasks

SegVL achieves state-of-the-art Dice scores across all segmentation benchmarks under both linear probing and finetuning, as shown in Table. 1. On TotalSegmentator and MSD-Lung, it consistently outperforms CT-CLIP and fVLM, demonstrating the advantage of voxel-level contrastive supervision in capturing both large anatomical structures and small lesion regions. For example, SegVL improves Dice of lung tumor segmentation by +4.5% on MSD-Lung under finetuning compared to CT-CLIP, showing superior fine-grained feature learning.

On BTCV, which tests cross-region generalization beyond the thoracic domain, SegVL achieves the best performance in both settings. These results indicate that the segmentation-aware pretraining introduces transferable structural cues that generalize well to out-of-distribution anatomy, even without paired text supervision.

3.3. Zero-shot Segmentation Results

Our voxel-mask contrastive formulation naturally enables zero-shot segmentation. Given a volume and a target anatomical label, SegVL computes voxel-wise similarity scores $P_{i,j}$ between each voxel embedding $\mathbf{V}_i$ and the corresponding mask name embedding $\mathbf{M}_{\text{mask},j}$ via Eq. 4, producing probability maps without further training.

As shown in Figure 2, despite using a lightweight segmentation head, our model produces predictions that closely match the expected anatomical regions. This indicates that the voxel and mask embeddings are well aligned via contrastive learning. Notably, for fine-grained targets like nodules and effusions, the model accurately highlights relevant regions, indicating it has learned transferable voxel-level features.

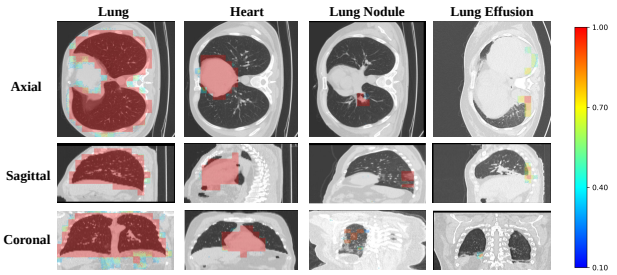

Figure 2. **Visualization of predicted probability maps from voxel-mask contrastive learning.** SegVL predicts segmentation for Lung, Heart, Nodule, and Effusion across three views. Warmer colors indicate higher voxel-wise similarity to corresponding mask embeddings.

4. Conclusion

We present SegVL, a unified contrastive learning framework that incorporates segmentation supervision into vision-language pretraining for 3D medical images. By introducing voxel-level contrastive learning between voxel embeddings and segmentation prompts, and enhancing image-text contrast through segmentation-aware fusion, our model captures fine-grained anatomical features that improve performance across both classification and segmentation tasks. Extensive experiments show that SegVL outperforms existing 3D MedVLP methods, especially in fine-grained recognition scenarios.

Limitations and Future Work. This study uses a limited set of segmentation classes due to annotation and resource

constraints. Future work will expand the anatomical vocabulary and explore semi-supervised strategies, such as consistency-based objectives, to better leverage limited segmentation data.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. The medical segmentation decathlon. *Nature communications*, 13(1): 4128, 2022.

Blankemeier, L., Cohen, J. P., Kumar, A., Van Veen, D., Gardezi, S. J. S., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Truyts, C., et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pp. rs–3, 2024.

Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Chen, S., Ma, K., and Zheng, Y. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Dao, T. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Draelos, R. L., Dov, D., Mazurowski, M. A., Lo, J. Y., Henao, R., Rubin, G. D., and Carin, L. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857, 2021.

Gugger, S., Debut, L., Wolf, T., Schmid, P., Mueller, Z., Mangrulkar, S., Sun, M., and Bossan, B. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

Hamamci, I. E., Er, S., Almas, F., Simsek, A. G., Esirgun, S. N., Dogan, I., Dasdelen, M. F., Wittmann, B., Simsar, E., Simsar, M., et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*, 2024.

Hatamizadeh, A., Xu, Z., Yang, D., Li, W., Roth, H., and Xu, D. Unetformer: A unified vision transformer model and pre-training framework for 3d medical image segmentation. *arXiv preprint arXiv:2204.00631*, 2022.

He, X., Yang, Y., Jiang, X., Luo, X., Hu, H., Zhao, S., Li, D., Yang, Y., and Qiu, L. Unified medical image pre-training in language-guided common semantic space. In *European Conference on Computer Vision*, pp. 123–139. Springer, 2024.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., and Klein, A. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, pp. 12. Munich, Germany, 2015.

Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., and Hong, Q. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107, 2023.

Liu, C., Ouyang, C., Chen, Y., Quilodrán-Casas, C. C., Ma, L., Fu, J., Guo, Y., Shah, A., Bai, W., and Arcucci, R. T3d: Towards 3d medical image understanding through vision-language pre-training. *arXiv preprint arXiv:2312.01529*, 2023.

Liu, C., Cheng, S., Shi, M., Shah, A., Bai, W., and Arcucci, R. Imitate: Clinical prior guided hierarchical vision-language pre-training. *IEEE Transactions on Medical Imaging*, 2024.

Ni, X., Wu, L., Zhuang, J., Wang, Q., Wu, M., Vardhanabhuti, V., Zhang, L., Gao, H., and Chen, H. Mg-3d: Multi-grained knowledge-enhanced 3d medical vision-language pre-training. *arXiv preprint arXiv:2412.05876*, 2024.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B. H., Fedorov, A., Mak, R. H., and Aerts, H. J. Vision foundation models for computed tomography. *arXiv preprint arXiv:2501.09001*, 2025.

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.

Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M. E., Geurts, B., et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

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.

Tsai, E., Simpson, S., Lungren, M., Hershman, M., Roshkovan, L., Colak, E., Erickson, B., Shih, G., Stein, A., Kalpathy-Cramer, J., et al. Data from the medical imaging data resource center-rsna international covid radiology database release 1a-chest ct covid+(midrc-ricord-1a). *The Cancer Imaging Archive*, 2020.

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.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. Medklip: Medical knowledge enhanced language-image pre-training. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Wu, L., Zhuang, J., and 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.

Xie, Y., Zhang, J., Xia, Y., and Wu, Q. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pp. 558–575. Springer, 2022.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433, 2020.

Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., and Xie, W. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Zhou, H.-Y., Lian, C., Wang, L., and Yu, Y. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*, 2023.

Zhu, J., Hamdi, A., Qi, Y., Jin, Y., and Wu, J. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

A. Implementation Details

A.1. Ablation Studies

We conduct an ablation study on CT-RATE classification and BTCV segmentation tasks to analyze the impact of segmentation supervision, loss function design, and decoder complexity, as shown in Table 3. Models in all ablation studies are pretrained on $1/10$ data fold of CT-RATE and RadGenome-Chest.

Effect of segmentation supervision. Comparing the baseline without segmentation (row 1) and the model trained with segmentation data (row 2), we observe substantial gains in both classification and segmentation. Specifically, segmentation-augmented training improves classification AUC from 0.744 to 0.795 in the finetuning setting, and boosts segmentation Dice from 0.843 to 0.868. This highlights the benefit of introducing spatial supervision to guide the model toward fine-grained anatomical regions relevant to diagnosis.

Impact of segmentation loss. We compare Tversky, Dice, and InfoNCE as segmentation losses. The Tversky loss achieves the best performance across both classification and segmentation, reaching 0.803 classification AUC and 0.868 Dice. This is likely due to its ability to better handle foreground-background imbalance, particularly for underrepresented structures such as lung nodules or pleural effusions. In contrast, InfoNCE—which contrasts each voxel embedding with all class prompts—amplifies the effect of class imbalance and fails to capture subtle targets, leading to significantly degraded segmentation performance.

Effect of decoder complexity. Using a heavy CNN decoder (row 4) results in lower performance compared to a light MLP decoder (row 6), under otherwise similar settings. This suggests that complex decoders tend to absorb fine-grained information into themselves, limiting the encoder's capacity to learn generalizable and transferable representations. In contrast, the lightweight decoder promotes better encoder learning and improves both classification and segmentation outcomes.

A.2. Pre-training Dataset Settings

The segmentation classes used for pre-training in RadGenome-ChestCT dataset include lung, heart, tracheobronchial tree, cardiovascular, lung nodule and lung effusion.

Due to the low-quality of lung nodule masks in RadGenome-ChestCT dataset, these masks are refined using trained nnUNet (Isensee et al., 2021) model on LUNA (Setio et al., 2017) dataset.

Additionally, due to the high false positive rate in lung effusion labels, we discard all corresponding label masks for CT volumes with negative label of "no pleural effusion," replacing them with all-background masks.

The tracheobronchial tree mask is constructed by merging the "trachea" and "bronchi" masks. The cardiovascular mask is created by combining anatomical structures including the aorta, aortic arch, brachiocephalic trunk, brachiocephalic vein, carotid artery, common carotid artery, heart ascending aorta, heart atrium, heart ventricle, heart, inferior vena cava, internal carotid artery, and internal jugular vein.

A.3. Training Details

Initialization and Training Stages. The image encoder is randomly initialized, while the text encoder is initialized from BioViL (Boecking et al., 2022). The training consists of two stages. The first stage focuses on learning high-level semantic representations, where α_{seg} in Section 2.3 is set to 0 and the segmentation/text heads remain frozen. The second stage incorporates segmentation supervision to enhance fine-grained feature learning, with α_{seg} set to 0.75. The fusion weight λ (Section 2.3) is initialized to 0 at the beginning of the second stage.

Model Architecture The text encoder is a transformer-based model following (Hamamci et al., 2024). The image encoder utilizes a ViT-Base with patch size of (20, 20, 10) in each dimension.

Optimizing Hyper-parameters. Each stage is trained for 10^6 steps using the AdamW optimizer and a cosine learning rate schedule, with the learning rate decaying linearly from 5×10^{-6} to 5×10^{-8}. The batch size is set to 32 in the first stage and 16 in the second. Gradient accumulation with a step size of 2 is used in both stages to balance GPU memory constraints.

Loss Hyper-parameters. For segmentation loss, we use the Tversky loss with class-specific (α, β) weights to handle severe class imbalance: $(0.3, 0.7)$ for lung nodules, $(0.4, 0.6)$ for lung effusion, and $(0.5, 0.5)$ for other categories.

Data Augmentations. We follow fVLM (Shui et al., 2025) and apply standard augmentations including random flipping and random spatial shifting during the second stage. The total training time on the CT-RATE and RadGenome datasets is approximately 4 days for each stage using 4 NVIDIA A100 GPUs (80GB) with mixed-precision training enabled.

Transformers Implementations. Due to the long sequence length of volumetric tokens, we adopt PyTorch's FlashAttention v2 (Dao, 2024) for the vision transformer to reduce memory usage without precision loss.

Table 3. Ablation studies of segmentation task, loss designs and decoder designs. VE denotes the "Visual Enhancement" from segmentation, which is proposed in Section 2.3.

Seg	VE	Seg Loss	Decoder	CT-RATE(Cls)			BTCV(Seg)	
				Zero-shot	Lipro	Finetune	Lipro	Finetune
×	×	\	Light MLP	0.672	0.683	0.744	0.839	0.843
✓	×	Tversky	Light MLP	0.707	0.760	0.795	<u>0.862</u>	<u>0.866</u>
✓	✓	InfoNCE	Light MLP	0.676	0.688	0.742	0.845	0.847
✓	✓	Tversky	Heavy CNN	0.737	0.752	0.775	0.854	0.859
✓	✓	Dice	Light MLP	0.735	<u>0.763</u>	<u>0.797</u>	0.857	0.858
✓	✓	Tversky	Light MLP	**0.742**	**0.776**	**0.803**	**0.864**	**0.868**

A.4. Downstream Dataset and Evaluation Tasks

CT-RATE (Hamamci et al., 2024) (Internal Classification) The CT-RATE internal validation dataset consists of 3039 volumes from 1304 patients. The evaluated classes for classification follows (Hamamci et al., 2024), including 18 classes of Medical material, Arterial wall calcification, Cardiomegaly, Pericardial effusion, Coronary artery wall calcification, Hiatal hernia, Lymphadenopathy, Emphysema, Atelectasis, Lung nodule, Lung opacity, Pulmonary fibrotic sequela, Pleural effusion, Mosaic attenuation pattern, Peribronchial thickening, Consolidation, Bronchiectasis, Interlobular septal thickening.

Rad-ChestCT (Draelos et al., 2021) (External Classification) The Rad-ChestCT dataset consists of 3630 CT volumes. The evaluated classes for classification follows (Hamamci et al., 2024), including 16 classes of Medical material, Calcification, Cardiomegaly, Pericardial effusion, Hiatal hernia, Lymphadenopathy, Emphysema, Atelectasis, Lung nodule, Lung opacity, Pulmonary fibrotic sequela, Pleural effusion, Peribronchial thickening, Consolidation, Bronchiectasis and Interlobular septal thickening.

CC-CCII (Zhang et al., 2020) (External Classification) The CC-CCII dataset consists of 3,993 scans from 2,698 patients, and we perform classification following the settings of (He et al., 2024). The downstream task is to classify each volume into three categories: novel coronavirus pneumonia(NCP), common pneumonia (CP), and normal (Normal).

RICORD (Tsai et al., 2020) (External Classification) The RICORD dataset comprises 182 training volumes and 45 testing volumes. Following the protocol of UnimiSS (Xie et al., 2022), we formulate the classification task as a binary prediction of COVID-19 positivity.

BTCV (Landman et al., 2015) (External Segmentation) The BTCV dataset comprises abdomen CT volumes and segmentations of multiple organs. Following the settings of UnimiSS (Xie et al., 2022), we divide the dataset into 24 training volumes, 6 validation volumes and 20 online testing volumes. All the labels of training volumes are used for linear-probing and finetuning settings.

MSD-Lung (Antonelli et al., 2022) (External Segmentation) The Medical Segmentation Decathlon (MSD) challenge dataset comprises CT volumes with lesion and its segmentation. We use the task 6 split of MSD challenge dataset to focus on the lung tumor segmentation on chest CT. The dataset comprises 63 training volumes with lung tumor annotations and 32 testing volumes. We follow the evaluation settings in VoCo (Wu et al., 2024).

TotalSegmentor (Wasserthal et al., 2023) (External Segmentation) We use the organ subset of the TotalSegmentor dataset for evaluation. The organ subset include segmentations of Spleen, Left & Right Kidney, Gallbladder, Liver, Stomach, Pancreas, Left & Right Adrenal Gland, Lobes of Left & Right Lung, Esophagus, Trachea, Thyroid Gland, Small Bowel, Duodenum, Colon, Urinary Bladder, Prostate, Left & Right Kidney Cyst. We split the dataset into 928 training, 52 validation, and 248 testing volumes.

RadGenome (Zhang et al., 2024) (Zero-Shot Segmentation) We perform zero-shot segmentation on the RadGenome validation set for qualitative visualization. The evaluated classes include lung, heart, lung nodule, and lung effusion. Heatmaps are generated to visualize voxel-level segmentation predictions for each class.

A.5. More Training Details

Sampling Different Types of Training Dataset As we use both image-report and image-segmentation dataset for pre-training, the sampling and balancing strategy between two types of dataset become crucial. We perform gradient accumulation of 2 steps, which contain one step of training on image-report dataset and one step of training on image-segmentation dataset. For the image-report dataset, the segmentation head is also used for visual enhancement on image latents.

Implementations of CLIP Loss on Multi-GPU The official CT-CLIP (Hamamci et al., 2024) code does not support multi-GPU training for the CLIP loss, as it fails to perform contrastive learning across image/text embeddings from different GPUs. We address this issue by adapting the original CLIP (Radford et al., 2021) implementation to HuggingFace Accelerate library (Gugger et al., 2022) in all our experiments, including ablation studies. This limitation has also been resolved in the fVLM (Shui et al., 2025) implementation, which is adapted from a different codebase.

A.6. More Network Architecture Details

Vision Encoder. We adopt a clean `ViT-Base` architecture as our 3D vision encoder, which directly applies standard Transformer blocks over volumetric patches. Specifically, the input CT volume is tokenized into non-overlapping $20 \times 20 \times 10$ patches, resulting in volumetric tokens. Unlike CT-ViT in (Hamamci et al., 2024), we remove the preceding convolutional layers and the large MLP expansion module, preserving a pure ViT design for better modularity and generalization. The model comprises 12 Transformer layers, each with 12 attention heads and hidden size 768, matching the original ViT-Base(Dosovitskiy et al., 2020) configuration.

Text and Segmentation Heads. Both the text projection head and segmentation head used for contrastive learning are implemented as lightweight 2-layer MLPs. Each MLP comprises a linear projection, followed by a LeakyReLU activation, LayerNorm, and a final linear layer to produce the embedding. These heads are designed to be parameter-efficient and compatible with the contrastive objectives.

Segmentation Decoder for Fine-tuning. When fine-tuning on downstream segmentation tasks, we replace the 2-layer segmentation MLP head with a small transposed convolutional decoder following UniMiSS (Xie et al., 2022). This adjustment introduces spatial positional bias into the predictions, which is important for accurate voxel-level decoding. Without this change, we observe that shared MLP heads tend to generate overly uniform predictions across all voxels in a token due to the lack of location-specific modeling.

A.7. More Visualizations of Zero-Shot Segmentations

To supplement our analysis of voxel-mask contrastive learning, we include additional zero-shot segmentation results in Figure 3–5. The visualizations display predicted probability maps obtained from cosine similarity between voxel and mask embeddings. These examples further confirm that our method produces semantically meaningful and spatially precise segmentations, even for subtle or fine-grained structures, without relying on heavy decoder designs.

A.8. Performance Comparison with Standard Deviation

We provide the performance comparisons with standard deviation in our main experiments of downstream tasks evaluation. The deviation is obtained from five repeated experiments with different random seeds.

A.9. Detailed Analysis of Classification Results of Different Labels

We present detailed zero-shot and linear probing classification results on the CT-RATE dataset in Table 6 and Table 7 Overall, our model achieves strong performance across 18 diagnostic categories, demonstrating the generalizability of our pretraining approach.

Two key observations emerge from this analysis. First, the incorporation of segmentation supervision improves the model's understanding of anatomy-related diseases. For instance, we observe notable gains on classes such as *bronchiectasis* and *lung nodule*, which are highly correlated with anatomical structures included in our segmentation vocabulary.

Second, our method exhibits clear advantages in identifying fine-grained patterns. Categories like *lung opacity*, which require subtle feature discrimination, benefit from the detailed spatial information learned through voxel-level contrastive supervision. This suggests that our approach not only enhances high-level semantic alignment but also reinforces the model's sensitivity to nuanced radiological cues.

A.10. Potential Societal Impact.

Our work presents a pre-trained 3D medical image encoder that may benefit a wide range of downstream applications, such as computer-aided diagnosis and clinical decision support. By enabling better understanding of volumetric scans with limited annotations, our approach could help democratize access to high-quality medical AI systems and reduce the burden on radiologists.

However, potential negative impacts should be considered. The pre-trained model may reflect dataset biases, such as under-representation of rare conditions or specific demographic groups, which could lead to reduced accuracy or unintended disparities in clinical settings. Moreover, over-reliance on automated systems without sufficient human oversight may risk diagnostic errors. Careful evaluation and responsible deployment in real-world workflows are necessary to mitigate such risks.

A.11. Licenses

We used existing assets as follows:

- **CT-CLIP** (Hamamci et al., 2024): `https:`

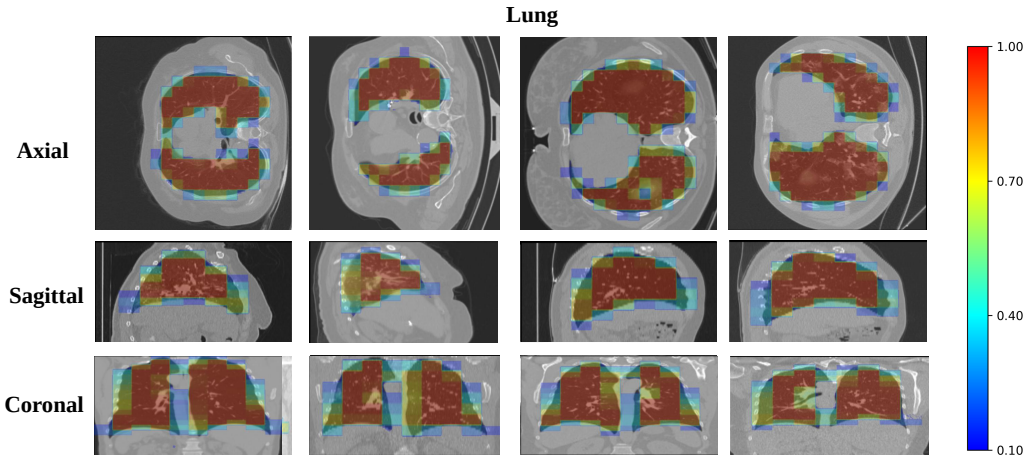

Figure 3. **Additional visualization of lung predictions.** Segmentation heatmaps predicted by `SegVL` for the *lung* class across axial, sagittal, and coronal views. Warmer colors indicate higher predicted probabilities, showing close alignment with lung regions.

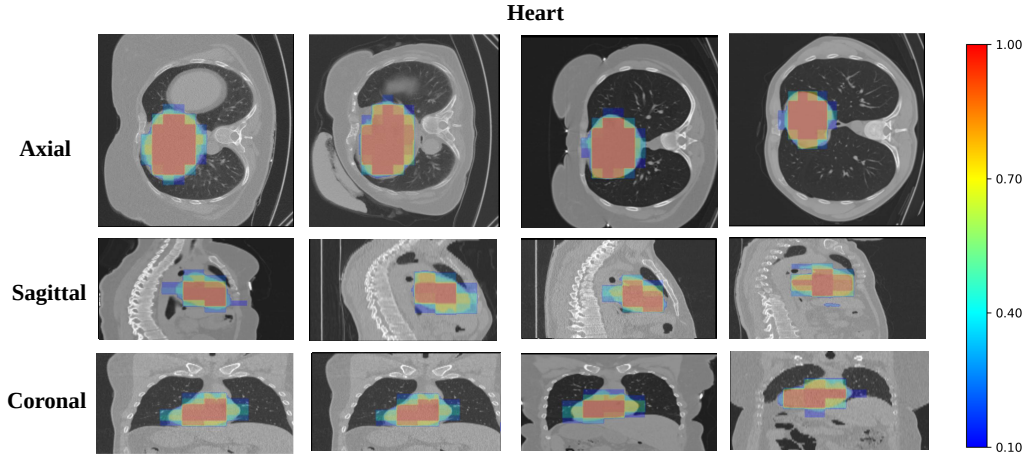

Figure 4. **Additional visualization of heart predictions.** Segmentation heatmaps predicted by `SegVL` for the *heart* class across axial, sagittal, and coronal views. Our model accurately highlights heart regions using only lightweight contrastive supervision.

`//github.com/ibrahimethemhamamci/CT-CLIP`, licensed under CC BY-NC-SA.

- **fVLM** (Shui et al., 2025): `https://github.com/alibaba-damo-academy/fvlm`, no explicit license.
- **UniMiSS** (Xie et al., 2022): `https://github.com/YtongXie/UniMiSS-code`, licensed under MIT.
- **Merlin** (Blankemeier et al., 2024): `https://github.com/StanfordMIMI/Merlin`, licensed under MIT.

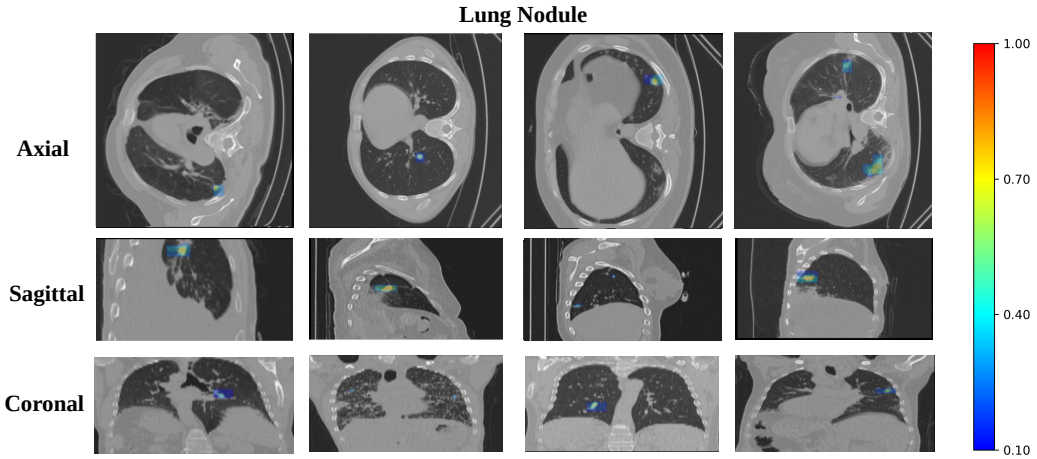

Figure 5. **Additional visualization of lung nodule predictions.** Visualization of voxel-wise prediction maps for the *lung nodule* class. The highlighted small regions reflect our model's ability to localize subtle, fine-grained features via voxel-mask contrastive learning.

Table 4. Mean AUC (↑) and estimated standard deviation on downstream classification under **linear probing** setting.

Model	RadChest-CT	CC-CCII	RICORD
CT-CLIP (Hamamci et al., 2024)	0.653 ± 0.014	0.865 ± 0.009	0.846 ± 0.008
Merlin (Blankemeier et al., 2024)	0.677 ± 0.011	0.877 ± 0.008	0.854 ± 0.007
UniMiSS (Xie et al., 2022)	\	0.841 ± 0.009	0.862 ± 0.008
fVLM (Shui et al., 2025)	0.697 ± 0.010	0.871 ± 0.007	0.858 ± 0.007
SegVL (ours)	$\mathbf{0.702 \pm 0.009}$	$\mathbf{0.893 \pm 0.006}$	$\mathbf{0.887 \pm 0.007}$

Table 5. Comparison of our SegVL with other baselines on downstream classification tasks under **finetuning** setting. Results are reported as AUC (↑) in the format of mean ± std. Values marked with n/a are directly taken from original papers without standard deviation.

Model	RadChest-CT	CC-CCII	RICORD
CT-CLIP (Hamamci et al., 2024)	0.650 ± 0.011	0.920 ± 0.007	0.863 ± 0.006
Merlin (Blankemeier et al., 2024)	0.694 ± 0.012	0.919 ± 0.009	0.879 ± 0.010
T3D (Liu et al., 2023)	\	$0.927 \pm$ n/a	\
UniMiSS (Xie et al., 2022)	\	0.913 ± 0.008	0.891 ± 0.009
MRM (Zhou et al., 2023)	\	$0.880 \pm$ n/a	\
IMITATE (Liu et al., 2024)	\	$0.892 \pm$ n/a	\
fVLM (Shui et al., 2025)	0.704 ± 0.010	0.926 ± 0.007	0.885 ± 0.006
SegVL (ours)	$\mathbf{0.716 \pm 0.008}$	$\mathbf{0.936 \pm 0.006}$	$\mathbf{0.912 \pm 0.005}$

Table 6. Detailed zero-shot classification results on CT-RATE. We report class-wise Precision, AUC, and F1 score for each of the 18 diagnostic categories, along with their mean.

Class	Precision	AUC	F1
Medical material	0.247	0.742	0.700
Arterial wall calcification	0.581	0.890	0.831
Cardiomegaly	0.557	0.880	0.911
Pericardial effusion	0.477	0.891	0.849
Coronary artery wall calcification	0.548	0.839	0.779
Hiatal hernia	0.185	0.738	0.655
Lymphadenopathy	0.416	0.675	0.700
Emphysema	0.358	0.759	0.743
Atelectasis	0.372	0.691	0.705
Lung nodule	0.581	0.702	0.658
Lung opacity	0.522	0.711	0.659
Pulmonary fibrotic sequela	0.359	0.590	0.554
Pleural effusion	0.480	0.931	0.918
Mosaic attenuation pattern	0.132	0.714	0.671
Peribronchial thickening	0.249	0.719	0.698
Consolidation	0.356	0.809	0.759
Bronchiectasis	0.240	0.775	0.751
Interlobular septal thickening	0.218	0.756	0.786
Mean	0.382	0.767	0.740

Table 7. Per-class linear probing results on the CT-RATE dataset. We report class-wise Precision, AUC, and F1 score for each of the 18 diagnostic categories, along with their mean.

Class	Prec	AUC	F1
Medical material	0.286	0.780	0.809
Arterial wall calcification	0.575	0.853	0.847
Cardiomegaly	0.387	0.915	0.924
Pericardial effusion	0.420	0.912	0.958
Coronary artery wall calcification	0.550	0.869	0.876
Hiatal hernia	0.257	0.759	0.865
Lymphadenopathy	0.456	0.799	0.810
Emphysema	0.310	0.720	0.828
Atelectasis	0.480	0.815	0.844
Lung nodule	0.636	0.725	0.791
Lung opacity	0.520	0.749	0.781
Pulmonary fibrotic sequela	0.385	0.720	0.739
Pleural effusion	0.680	0.929	0.943
Mosaic attenuation pattern	0.227	0.754	0.808
Peribronchial thickening	0.184	0.699	0.762
Consolidation	0.374	0.839	0.876
Bronchiectasis	0.216	0.641	0.697
Interlobular septal thickening	0.217	0.791	0.812
Mean	0.398	0.793	0.832