Learning to Read Where to Look: Disease-Aware Vision–Language Pretraining for 3D CT

Simon Ging[1,2][0000−0002−5159−9792], Philipp Arnold[3][0000−0003−0686−7361], Sebastian Walter[4][0009−0006−2613−3209], Hani Alnahas[1], Hannah Bast[4][0000−0003−1213−6776], Elmar Kotter[3][0000−0001−9022−6000], Jiancheng Yang[5,6][0000−0003−4455−7145], Behzad Bozorgtabar[2][0000−0002−5759−4896], and Thomas Brox[1][0000−0002−6282−8861]

[1] Computer Vision Group, University of Freiburg, Germany
gings@cs.uni-freiburg.de
[2] Adaptive & Agentic AI (A3) Lab, Aarhus University, Denmark
[3] Department of Radiology, Medical Center – University of Freiburg, Germany
[4] Chair of Algorithms and Data Structures, University of Freiburg, Germany
[5] ELLIS Institute Finland
[6] School of Electrical Engineering, Aalto University, Finland

Abstract. Recent 3D CT vision–language models align volumes with reports via contrastive pretraining, but typically rely on limited public data and provide only coarse global supervision. We train a 3D CT vision–language model on 98k report–volume pairs (50k patients) collected at a single hospital, combined with public datasets, using SigLIP-style contrastive pretraining together with prompt-based disease supervision in the shared vision–text embedding space. On CT-RATE, our model achieves state-of-the-art text-to-image retrieval (R@10 31.5 vs. 22.2) and competitive disease classification (AUC 83.8 vs. 83.8), with consistent results on Rad-ChestCT (AUC 77.0 vs. 77.3). We further observe that radiologists routinely reference specific images within their reports (e.g., "series X, image Y"), linking textual descriptions to precise axial locations. We automatically mine 262k such snippet–slice pairs and introduce the task of intra-scan snippet localization—predicting the axial depth referred to by a text snippet—reducing mean absolute error to 36.3 mm at 12 mm feature resolution, compared with 67.0 mm for the best baseline. Adding this localization objective leaves retrieval and classification broadly unchanged within confidence bounds, yielding a single unified model for retrieval, classification, and intra-scan grounding. Code and models are available at https://github.com/lmb-freiburg/radfinder.

Keywords: 3D CT · Vision–language pretraining · Contrastive learning

1 Introduction

The volume of radiological imaging continues to grow, with CT examinations increasing steadily, while the number of radiologists has not kept pace [14]. Each

CT study requires careful analysis: a radiologist reviews hundreds of axial slices, identifies findings, and dictates a free-text report. Foundation models that learn general-purpose representations from large-scale imaging data offer a path toward assisting radiologists across diverse tasks—from automated retrieval and triage to report generation—without task-specific annotation.

Recent 3D CT vision–language models (VLMs) learn such representations by aligning CT volumes with radiology reports via contrastive pretraining. CT-CLIP [6] pioneered this direction on the public CT-RATE dataset of 50k chest CT volumes. SPECTRE [3] scaled self-supervised and cross-modal pretraining to 230k volumes and achieved strong retrieval results, while other approaches leverage structured entity extraction [10] or hierarchical vision architectures [1]. These methods align entire volumes with entire reports, providing only coarse, global supervision. Other approaches improve supervision granularity by aligning at the organ level: MedVista3D [11] and fVLM [16] both use an external segmentation model to identify anatomical regions and contrast each region's visual features against LLM-generated region descriptions. While effective, this requires a pretrained segmentation model and curated region-level text that may not capture the specific findings a radiologist highlights.

In this work, we train *RadFinder*, a 3D CT VLM on 98k report–volume pairs (50k patients) from a single hospital, combined with public datasets. The model uses SigLIP [20] contrastive pretraining on full reports with pretrained medical vision and text encoders, and integrates structured disease labels as text prompts within the contrastive objective (Sec. 3.2). RadFinder achieves state-of-the-art text-to-image retrieval and competitive disease classification on CT-RATE and Rad-ChestCT [4], demonstrating strong cross-dataset transfer.

We additionally observe that radiology reports contain a largely untapped form of local supervision: radiologists frequently reference specific slices within a scan—for example, "*hepatic lesion, see series 4, image 38*". We mine 262k such snippet–slice pairs and propose the task of *intra-scan snippet localization*: given a text snippet, predict the axial depth it refers to within a CT volume. While slice-level matching from reports has been explored at small scale [8], we formulate this in 3D feature space and achieve results substantially better than simple baselines. Training localization jointly with global objectives does not degrade retrieval or classification, yielding a unified model for all three tasks.

Contributions.

- We build a large-scale 3D report–volume dataset (98k pairs, 50k patients) from clinical routine at a single hospital and automatically mine 262k snippet–slice pairs that provide weak local supervision without additional annotation.
- Using this local supervision signal, we propose the task of intra-scan snippet localization, establish baselines, and show that localization can be trained jointly with global objectives without degrading retrieval or classification.
- We train a 3D CT VLM by combining contrastive pretraining on reports with prompt-based disease label training and snippet localization. Our model achieves state-of-the-art retrieval and competitive disease classification on ex-

ternal benchmarks, demonstrating strong cross-dataset transfer. Code and pretrained models will be made publicly available.

2 Dataset

We collect 97,760 report–volume pairs from 50,474 patients at a single hospital, spanning 13 years of clinical practice. The dataset covers chest (46%), abdomen (22%), and combined chest–abdomen (33%) studies. The radiologist report for the study contains detailed findings and a summary impression, and we select the single largest axial series from each study. In-plane resolution and slice thickness medians are 0.71 mm and 3.0 mm, respectively.

Snippet mining. Radiologists frequently reference specific slices when dictating reports, where *series* identifies the volume and *image* the axial slice number (e.g., "hepatic lesion, see series 4, image 38" or "pulmonary nodule in the right lower lobe (3/72)"). We extract these references via pattern matching heuristics. Each snippet associates a short textual finding with a precise axial position in the scan, providing weak local supervision without additional annotation. To verify our pipeline, we compare against manually extracted references from 100 reports and find that our heuristics achieve 99.4% precision and 90.2% recall (F1 94.6%). To ensure correct spatial alignment, we keep only slices where slicing the 3D volume at the referenced position produces identical content to the original 2D image file stored by the scanner. After filtering, this yields 261,800 snippet–slice pairs across all scans, 2.7 per scan on average.

Text processing. We anonymize all text by removing patient and physician identifiers via rule-based matching and converting absolute dates into relative references (e.g., "Spine injury 5 years ago"). Reports are translated from German to English using Gemma 3 27B [5]; all training is conducted on the English translations. We apply the RATE protocol from Pillar-0 [1], using LLM-based question answering to classify each report into 93 chest and 226 abdomen binary findings across 30 organ categories; these structured labels serve as input for prompt-based training (Sec. 3.2). Data is split 80/10/10 by randomized patient IDs, with no patient overlap with public evaluation datasets.

3 Method

3.1 Global Vision–Language Pretraining

We initialize from the pretrained SPECTRE [3] model, which uses a two-stage vision encoder: a ViT-Large local backbone processes each volume in separate $128 \times 128 \times 32$-voxel windows at $0.75 \times 0.75 \times 3.0\,\mathrm{mm}^3$ spacing, and a 4-layer global feature combiner aggregates the window-level representations. The text encoder is Qwen3-Embedding [12] (0.6B) with LoRA adapters. SigLIP [20] projection heads map both modalities to a 512-dim shared space. We freeze the local vision backbone and fine-tune all other modules with SigLIP contrastive loss. Unlike SPECTRE, which crops to a fixed grid, we process full volumes with variable

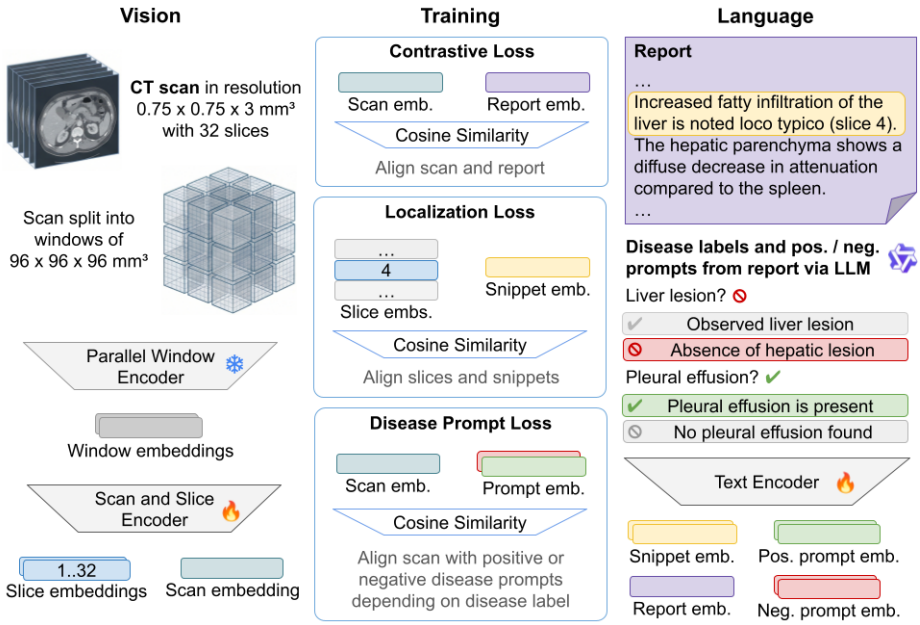

Fig. 1. Overview of the RadFinder architecture and training pipeline. **Left (Vision):** A 3D CT scan is processed by a frozen parallel window encoder, followed by a trainable encoder to extract both a global volume embedding and local slice-level embeddings. **Right (Language):** A trainable text encoder processes full radiology reports, localized text snippets (e.g., the sentence that references the fourth slice out of the 32 slices in this example scan), and LLM-extracted positive/negative disease prompts. **Center (Training):** The model is optimized via three contrastive objectives in a shared embedding space: a global Contrastive Loss aligning the full scan and report, an intra-scan Localization Loss aligning text snippets with their specific slices, and a Disease Prompt Loss aligning the global scan with corresponding disease prompt descriptions

input shapes. We train on up to four datasets totaling 159k report–volume pairs: *RefCT* (internal, 78k in the training split), CT-RATE [6] (47k), Merlin [2] (15k), and INSPECT [7] (19k). During training, we apply three text augmentations, each with probability 0.2: replacing the full report with concatenated organ-level descriptions from the RATE pipeline [1], removing historic comparisons ("compared to prior exam…") via LLM parsing following CT-RATE [6], or dropping the *findings* section.

3.2 Prompt-Based Disease Label Training

To improve text-based disease classification, we augment the contrastive loss with disease label prompts. During training, labels are represented as text prompts and supervised via a BCE loss on the cosine similarity difference between the volume embedding and positive/negative prompt embeddings. We do not train

a separate classification head; supervision is applied entirely in the shared vision–text embedding space.

Labels. Using the RATE protocol [1], we extract binary findings from each report via LLM-based question answering with Qwen3-30B-A3B [12] (93 chest and 226 abdomen findings). We additionally derive binary labels for the 18 CT-RATE disease classes by mapping the 93 RATE chest findings.

Prompts. For each finding q, we construct three positive and three negative text prompt variants (e.g., "Pleural effusion is present.", "No pleural effusion is identified."). During training, one variant is sampled randomly; at inference, the three embeddings are averaged.

Loss. Let $\mathbf{z}$ denote the L2-normalized image embedding and $\mathbf{p}_q^+, \mathbf{p}_q^-$ the positive and negative prompt embeddings for finding q. We classify each finding via the logit difference scaled by the shared SigLIP temperature τ:

$$\mathcal{L}_{\text{prompt}} = \frac{1}{|\mathcal{M}|} \sum_{q \in \mathcal{M}} w_q \Big(-\alpha_q \, y_q \, \log \sigma(x_q) - (1 - y_q) \log\big(1 - \sigma(x_q)\big) \Big) \tag{1}$$

$$x_q = (\mathbf{z}^\top \mathbf{p}_q^+ - \mathbf{z}^\top \mathbf{p}_q^-)/\tau, \qquad \alpha_q = \min(n_q^+/n_q^-, 20) \tag{2}$$

where $y_q \in \{0, 1\}$ is the binary finding label, σ is the logistic sigmoid, $\mathcal{M}$ is the set of valid image–question pairs, w_q upweights the 18 CT-RATE classes to balance them against the larger RATE label set, and α_q handles per-question label imbalance given n_q^+ positive and n_q^- negative training examples. The prompt loss is weighted by λ relative to the SigLIP loss.

The idea of using label text as prompts in contrastive training originates in CXR-CLIP [19] and UniCL [18]. The closest prior work in 3D CT either extracts free-text disease descriptions from reports with GPT-4 for contrastive training (BrgSA [9]) or trains an explicit classification head on LLM-extracted labels before contrastive alignment (MPS-CT [10]). Our approach uses structured binary labels directly as text prompts within the contrastive objective.

3.3 Intra-Scan Localization Loss

To encourage depth-aware representations, we introduce an intra-scan localization objective. Given a text snippet describing a radiological finding and the corresponding CT volume, the model classifies which depth position along the axial axis the snippet refers to. Crucially, negatives come from other depth positions *within the same scan*, avoiding cross-sample false negatives that arise when different patients share similar pathology.

Depth features. Let $\mathbf{z}_d \in \mathbb{R}^E$ denote the L2-normalized image embedding at depth position $d \in \{1, \ldots, D\}$, with one position covering 12mm in the axial plane, obtained by processing backbone patch features through the global feature combiner, averaging over the coronal and sagittal dimensions, and projecting through the SigLIP head. Let $\mathbf{t} \in \mathbb{R}^E$ denote the L2-normalized projected text embedding of the snippet.

Gaussian soft target. Each snippet is associated with a specific axial slice, whose physical position determines a ground-truth depth index $d^* \in \{1, \dots, D\}$ in the feature grid. Since annotations are inherently imprecise—a finding typically extends to neighboring slices—we define a one-hot indicator $\mathbf{m} \in \{0,1\}^D$ with $m_{d^*} = 1$ and convolve it with a normalized 1-D Gaussian kernel g of standard deviation σ=2:

$$\tilde{\mathbf{m}} = \frac{(\mathbf{m} * g)}{\|\mathbf{m} * g\|_1}, \qquad g_k = \frac{1}{Z} \exp\Big(-\frac{k^2}{2\sigma^2}\Big), \quad |k| \le \lceil 3\sigma \rceil. \tag{3}$$

Loss. We compute cosine-similarity logits between the snippet embedding and each depth feature, scaled by a temperature τ=0.1, and minimize the cross-entropy against the soft target $\tilde{\mathbf{m}}$:

$$\ell_d = \mathbf{z}_d^\top \mathbf{t}/\tau, \qquad \mathcal{L}_{\text{loc}} = -\sum_{d=1}^{D} \tilde{m}_d \log \frac{\exp(\ell_d)}{\sum_{d'} \exp(\ell_{d'})}. \tag{4}$$

The total training objective combines the global report–volume contrastive loss $\mathcal{L}_{\text{global}}$, the prompt loss $\lambda\,\mathcal{L}_{\text{prompt}}$ (Eq. 1), and the localization loss $\beta\,\mathcal{L}_{\text{loc}}$.

Table 1. Text-to-image retrieval on CT-RATE, Rad-ChestCT (RC) and Merlin validation sets. Training datasets: A: RefCT (ours, 78k), B: CT-RATE [6] (47k), C: Merlin [2] (15k), D: INSPECT [7] (19k). *fVLM additionally trains on MedVL-CT69 [16] (272k). We highlight **values within the confidence interval of the highest score** separately for the comparison of our main model with baselines (top) and our model with our ablations (bottom).

Method	Data				CT-RATE				RC		Merlin		R@1	
	A	B	C	D	R@10		MAP@5		AUC		AUC		Findings		Impr.	
Random Chance					0.3		0.1		50.0		50.0		0.8		0.8	
CT-CLIP [6]	✓				5.0		68.3		73.1		62.9		—		—	
MedVista3D-ViT [11]	✓				10.7		—		77.8		—		—		—	
MedVista3D-UniMISS	✓				8.7		—		78.2		—		—		—	
Merlin [2]		✓			2.7		62.6		72.8		64.4		59.4		19.4	
fVLM* [16]	✓				—		—		77.8		68.0		—		—	
BrgSA [9]	✓				22.2		70.4		82.9		74.2		—		—	
MPS-CT [10]	✓				16.3		71.3		**83.8**		**77.3**		—		—	
SPECTRE [3]	✓	✓	✓		18.2	1.4	**76.8**	1.7	54.3	1.1	55.7	0.6	44.5	0.7	29.2	0.6
RadFinder (Ours)	✓	✓	✓	✓	**31.5**	1.6	**78.0**	1.7	**83.8**	0.7	**77.0**	0.5	**69.0**	0.9	**40.3**	0.7
No localization loss	✓	✓	✓	✓	**31.5**	1.6	**78.2**	1.7	83.6	0.7	76.9	0.5	69.3	0.9	40.2	0.6
Global loss only	✓	✓	✓	✓	29.4	1.6	**76.6**	1.7	56.9	1.1	62.8	0.6	**70.7**	0.7	**42.2**	0.7
Prompt loss only	✓	✓	✓	✓	5.6	0.8	**77.9**	1.7	**84.8**	0.7	**78.6**	0.5	13.1	0.6	10.9	0.4
RefCT dataset	✓				26.3	1.6	**76.7**	1.7	80.4	0.8	77.7	0.6	60.5	0.9	35.8	0.7
Public datasets, no loc.	✓	✓	✓		26.9	1.6	**77.1**	1.7	83.2	0.7	73.6	0.6	67.5	0.9	40.2	0.8

Table 2. Results for snippet localization on the RefCT test set. *SigLIP2 finetuned on RefCT snippet–slice pairs via contrastive loss.

Method	MAE (mm)		<6mm		<18mm		<30mm	
Random slice	126.9	5.7	4.7	1.1	11.8	1.6	17.8	2.0
Middle slice	95.8	3.6	4.6	1.2	13.0	1.8	20.8	2.2
CLIP ViT-B/32 [13]	124.6	5.5	3.2	0.8	9.4	1.4	14.8	1.7
BiomedCLIP [21]	86.6	5.2	8.3	1.3	20.2	1.9	30.5	2.2
MedSigLIP-448 [15]	75.6	5.0	9.7	1.4	26.2	2.1	40.1	2.3
SigLIP2 [17]	82.4	4.7	7.8	1.3	19.3	1.8	31.5	2.3
SigLIP2 finetuned*	67.0	4.9	17.4	1.8	35.8	2.4	48.7	2.3
RadFinder (Ours)	**36.3**	2.9	**20.3**	2.0	**45.3**	2.5	**61.8**	2.4
Loc. loss only	**36.0**	2.9	**19.9**	2.0	**45.2**	2.5	**62.7**	2.4

4 Results

Implementation details. We fine-tune for 10 epochs with AdamW (lr 2×10^{-4}, 1 epoch linear warmup, cosine decay to 1×10^{-6}) and a batch size of 4096 on a single NVIDIA H100 (96 GB, 32h). We use loss weights $\lambda = 8$ and $\beta = 1$. We report 95% bootstrap confidence intervals (B=10,000) over test set samples.

Datasets and protocols. We evaluate on three external benchmarks. CT-RATE [6] provides 1564 reports paired with 3039 volumes (multiple reconstructions per study). We report text-to-image retrieval (R@10) over the full 3039-volume pool, volume-to-volume retrieval (MAP@5) using disease-label IoU as relevance, and disease classification (AUC) over 18 pathology classes. For the disease class prediction, we use the 7 prompt templates (e.g., "{a} is present.") proposed by CT-RATE and classify by softmax over cosine similarities between the volume embedding and the averaged positive/negative prompt embeddings. Rad-ChestCT [4] provides 3630 chest CT scans with disease labels (no reports); we evaluate AUC with the same prompt protocol. Merlin [2] provides 5125 report–volume pairs; we report R@1 separately for findings and impressions sections ($R@1_{find}$, $R@1_{impr}$) with the protocol by the original authors, which averages the metric over 100 trials with pools of 128 randomly sampled volumes. RefCT (internal) is used to evaluate snippet localization at 12 mm axial resolution. The test set contains 1,564 volumes with 1 snippet–slice pair each.

Retrieval (Tab. 1). Training on RefCT alone already surpasses all published baselines on CT-RATE retrieval (R@10 26.3 vs. 22.2 for BrgSA)—demonstrating strong cross-dataset transfer. Adding public datasets further improves retrieval to R@10 31.5. On the Merlin dataset, our model achieves $R@1_{find}$ 69.0, surpassing the original Merlin model (59.4).

Disease classification (Tab. 1). Without prompt-based training, our model achieves only AUC ∼57 on CT-RATE—comparable to SPECTRE and well below methods that incorporate disease knowledge. Adding prompts yields AUC 83.8 on CT-RATE and 77.0 on Rad-ChestCT, matching MPS-CT [10] (83.8 / 77.3)

within confidence bounds. Training on RefCT alone already reaches AUC 80.4 / 77.7, showing that disease knowledge transfers from our internal data. Conversely, training with prompt loss only yields the highest classification AUC (84.8) but sacrifices retrieval (R@10 5.6), as expected when the model sees only binary disease labels without free-text reports. The combined model achieves both: strong retrieval and competitive disease classification.

Public datasets (Tab. 1, last row). Training our method only on public datasets reaches R@10 26.9 / AUC 83.2 on CT-RATE and AUC 73.6 on Rad-ChestCT, showing that both retrieval and disease classification benefit significantly from our training setup compared to the original SPECTRE method. By adding our internal dataset, we gain R@10 4.6 points on CT-RATE and AUC 3.4 points on Rad-ChestCT; the contributions of our method and our internal dataset are complementary.

Localization (Tab. 2). Naive baselines (middle slice, random slice) achieve MAE 96 and 127 mm. Pretrained 2D VLMs reduce this to 87 mm (BiomedCLIP [21]) and 76 mm (MedSigLIP [15]); finetuning SigLIP2 [17] on our snippet–slice pairs reaches 67 mm. RadFinder achieves MAE 36.3 mm (20.3% within 6 mm), nearly halving the error of the best baseline. Training with localization loss alone yields comparable results (MAE 36.0 mm), confirming that the global objectives do not interfere with local grounding. Conversely, adding localization to the full model leaves retrieval and classification unchanged within confidence bounds (Tab. 1, RadFinder vs. No loc.), yielding a single unified model for all three tasks.

5 Conclusion

We presented RadFinder, a 3D CT vision–language model that combines contrastive pretraining on reports with prompt-based disease label supervision, achieving state-of-the-art retrieval and competitive disease classification on external benchmarks. We additionally showed that snippet–slice references mined from radiology reports provide effective local supervision for intra-scan localization, and that this task can be trained jointly with global objectives in a single unified model without degrading retrieval or classification.

Limitations. The localization resolution is limited to 12 mm along the slice axis, which may be insufficient for precisely localizing small findings. Furthermore, our snippet mining relies on explicit slice references in reports, which limits coverage to institutions where such references are part of the reporting practice.

Future work. Promising directions include investigating whether local supervision can also improve global representations, collecting data from additional hospitals to study multi-institutional training, evaluating bilingual inference on the original German reports, and improving localization resolution through finer-grained feature maps. Coupling the pretrained encoder with a language model for grounded report generation is another natural extension.

Acknowledgments. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) 417962828, 539134284, and 499552394 –

SFB 1597 – Small Data, as well as through EFRE (FEIH_2698644) and the state of Baden-Württemberg.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, K.K., Liu, L., Lian, L., Nercessian, M., Harguindeguy, N., Wu, Y., Mikhael, P., Lin, G., Sequist, L.V., Fintelmann, F., Darrell, T., Bai, Y., Chung, M., Yala, A.: Pillar-0: A new frontier for radiology foundation models. arXiv preprint arXiv:2511.17803 (2025)
2. Blankemeier, L., Kumar, A., Cohen, J.P., Liu, J., Liu, L., et al.: Merlin: a computed tomography vision–language foundation model and dataset. Nature **652**(8112), 1318–1328 (2026). https://doi.org/10.1038/s41586-026-10181-8
3. Claessens, C., Viviers, C., D'Amicantonio, G., Bondarev, E., van der Sommen, F.: Scaling self-supervised and cross-modal pretraining for volumetric CT transformers. arXiv preprint arXiv:2511.17209 (2025)
4. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., et al.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical Image Analysis **67**, 101857 (2021). https://doi.org/10.1016/j.media.2020.101857
5. Gemma Team: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
6. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., et al.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nature Biomedical Engineering (2026). https://doi.org/10.1038/s41551-025-01599-y
7. Huang, S.C., Huo, Z., Steinberg, E., Chiang, C.C., Langlotz, C., et al.: INSPECT: A multimodal dataset for patient outcome prediction of pulmonary embolisms. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track. vol. 36 (2023)
8. Huemann, Z., Church, S., Warner, J.D., Tran, D., Tie, X., et al.: ConTEXTual Net 3D: Vision-language modeling in PET/CT for visual grounding of positive findings. Journal of Imaging Informatics in Medicine (2026). https://doi.org/10.1007/s10278-026-01879-2
9. Lai, H., Jiang, Z., Yao, Q., Wang, R., He, Z., et al.: Bridged semantic alignment for zero-shot 3D medical image diagnosis. IEEE Journal of Biomedical and Health Informatics pp. 1–14 (2025). https://doi.org/10.1109/JBHI.2025.3629096
10. Li, Y., Lai, H., Zhou, X., Ming, S., Ma, W., et al.: More performant and scalable: Rethinking contrastive vision-language pre-training of radiology in the LLM era. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2025. LNCS, vol. 15966, pp. 348–357. Springer, Cham (2025). https://doi.org/10.1007/978-3-032-04981-0_33
11. Li, Y., Chen, Y., Lai, Y., Zhong, J., Wildman, V., Yang, X.: MedVista3D: Vision-language modeling for reducing diagnostic errors in 3D CT disease detection, understanding and reporting. arXiv preprint arXiv:2509.03800 (2025)

12. Qwen Team: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
14. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ **359**, j4683 (2017). https://doi.org/10.1136/bmj.j4683
15. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., et al.: MedGemma technical report. arXiv preprint arXiv:2507.05201 (2025)
16. Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., et al.: Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding. In: The Thirteenth International Conference on Learning Representations (2025)
17. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., et al.: SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
18. Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., et al.: Unified contrastive learning in image-text-label space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19141–19151. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01857
19. You, K., Gu, J., Ham, J., Park, B., Kim, J., et al.: CXR-CLIP: Toward large scale chest X-ray language-image pre-training. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S.E., Duncan, J., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. LNCS, vol. 14221, pp. 101–111. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43895-0_10
20. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11941–11952. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01100
21. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1), AIoa2400640 (2025). https://doi.org/10.1056/AIoa2400640