CheXanatomy: Anatomy–Aware Vision–Language Modeling for Chest Radiographs

Sergios Gatidis[1,2], Curtis Langlotz[1,2], and Christian Bluethgen[1,2]

[1] Stanford Center for Artificial Intelligence in Medicine and Imaging, Stanford University, Palo Alto, CA, USA
[2] Department of Radiology, Stanford University, Stanford, CA, USA
sgatidis@stanford.edu

Abstract. Vision–language models (VLMs) pretrained on large-scale image–text pairs demonstrate strong image-level understanding, but are primarily optimized for global alignment and do not explicitly encode fine-grained anatomical structure, limiting their suitability for spatially precise tasks such as segmentation.
We introduce CheXanatomy, a framework that integrates explicit anatomical knowledge into a pretrained VLM through autoregressive token-space supervision. Instead of adding task-specific decoder heads, the model is trained to generate anatomical segmentation masks via next-token prediction. To enable scalable supervision, we synthesize realistic chest radiographs from CT volumes and forward-project CT segmentation labels to obtain anatomically consistent 2D masks.
We evaluate the approach on synthetic and real chest radiographs against a U-Net baseline, including ablations on model scale, input resolution, and vision encoder fine-tuning. Autoregressive anatomical supervision achieves performance comparable to specialized convolutional models in-distribution and demonstrates improved geometric robustness under domain shift to real CXR data. In addition, anatomy-pretrained models exhibit improved sample efficiency when adapting to novel localization tasks under limited supervision. Larger models and higher input image resolution improve performance, while vision encoder fine-tuning has limited effect.
These results show that embedding anatomical structure directly into the generative objective promotes spatially grounded representations and supports anatomy–aware medical vision–language modeling.

Keywords: Vision–Language Models · Anatomical Segmentation · Chest Radiographs.

1 Introduction

Accurate anatomical segmentation in chest radiographs (CXR) remains challenging due to projection geometry, overlapping structures, and variability in acquisition protocols. Convolutional models such as U-Net achieve strong in-distribution performance [18,10] but rely on dense pixel supervision and often

degrade under domain shift. In contrast, vision–language models (VLMs) pretrained on image–text pairs learn transferable semantic representations [16,25,22], yet their training objectives emphasize global alignment and do not explicitly encode fine-grained anatomical structure.

We propose to integrate explicit anatomy into VLM training. We introduce CheXanatomy, a framework that trains a pretrained VLM to generate anatomical segmentations autoregressively via next-token prediction, without introducing task-specific pixel-space decoder heads. Segmentation is formulated as structured token generation within the standard generative objective.

Because comprehensive CXR annotations are scarce, we generate scalable supervision by projecting 3D CT segmentations into synthetic radiographs. Using TotalSegmentator and differentiable DRR rendering, we obtain anatomically consistent 2D labels without manual CXR annotation [23,6].

Our contributions are threefold: (1) large-scale synthetic AP and lateral radiographs with projected multi-structure labels; (2) autoregressive token-space anatomical supervision of a pretrained VLM via parameter-efficient fine-tuning; (3) comprehensive evaluation on synthetic and real radiographs, including ablations on model scale, input resolution, and vision encoder fine-tuning. Unlike projection-based approaches that train standalone segmentation networks, we use synthetic anatomical supervision to reshape the internal representation of a pretrained VLM within its native generative objective (Fig. 1).

2 Related Work

Anatomical Segmentation on Chest radiographs. Early work established benchmarks for lung, heart, and clavicle segmentation in CXR [5]. Contemporary pipelines rely on U-Net-style architectures and automated configuration strategies [18,10]. These methods require dense pixel supervision and a fixed channel per class and are not language-promptable. Moreover, large-scale multi-structure annotations remain limited. Recent efforts expand coverage via automated or pseudo-labeling approaches such as CheXmask [4].

CT-derived Supervision for Radiographs. To address annotation scarcity, several works project CT volumes into synthetic radiographs using digitally reconstructed radiographs (DRRs) [8]. Differentiable DRR frameworks enable principled projection of 3D segmentations to 2D labels [6]. CT-projected supervision has been used to obtain detailed anatomical masks and improve robustness in CXR segmentation [20,19,3]. Our work extends this paradigm by using projection-based anatomy not to train a standalone segmentation model, but to inject structured anatomical knowledge into a pretrained VLM.

Vision–Language Models and Token-Based Dense Prediction. Vision–language pretraining (e.g., CLIP) aligns images and text via global objectives [16]. Extensions to radiology leverage report supervision but remain largely image-level [25,22]. Promptable segmentation models such as SAM [11] and CLIP-guided

approaches [14,13,12] introduce language-conditioned mask prediction but rely on dedicated decoders.

Recent multimodal foundation models increasingly cast visual tasks as generative prediction problems. Pix2Seq formulates object detection as language modeling over discrete tokens for bounding boxes and class labels [1]. PaliGemma supports bounding box and segmentation outputs through structured token generation within a unified VLM framework [21]. However, explicit anatomical supervision for radiography within autoregressive VLMs remains underexplored.

Our work leverages PaliGemma's native token interface to encode fine-grained anatomy directly into the generative objective using scalable CT-derived supervision.

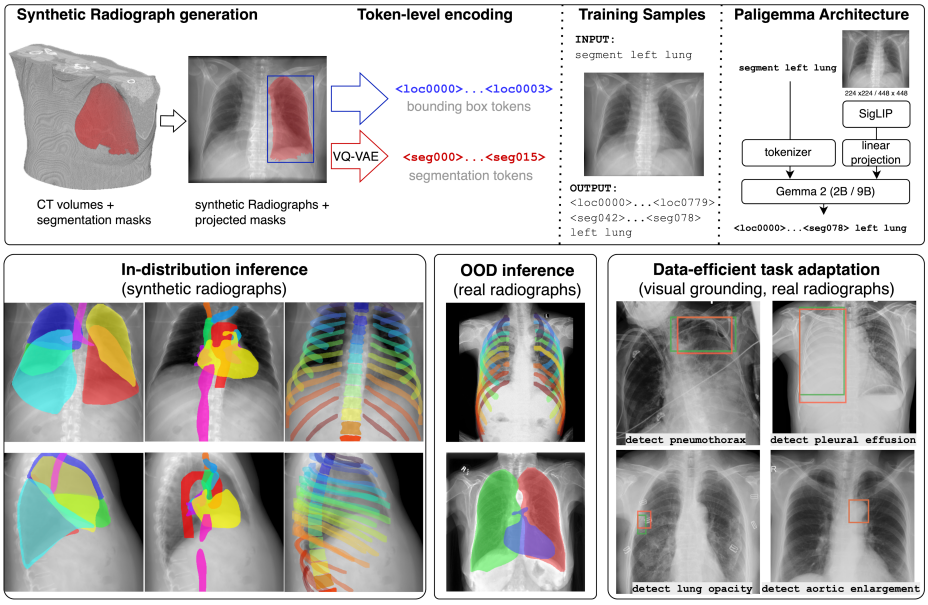

Fig. 1. Overview of **CheXanatomy**. CT volumes with anatomic labels are projected into synthetic CXRs, and bounding boxes and segmentation masks are encoded into structured tokens. A pretrained vision–language model (based on the Paligemma architecture) is fine-tuned to autoregressively generate anatomical bounding boxes and segmentations via next-token prediction. The trained model approaches convolutional baselines in-distribution and demonstrates improved geometric robustness under domain shift, while supporting adaptation to new localization tasks. Unlike conventional segmentation networks, no task-specific decoder heads are introduced, and supervision is applied entirely in token space.

3 Methods

Autoregressive Anatomical Segmentation. We used PaliGemma 2 [21], a pretrained vision–language model that supports structured spatial outputs, including bounding boxes and segmentation masks, through autoregressive next-token prediction. PaliGemma 2 consists of a SigLIP-So400m vision encoder [24] with a linear projection into the language token space and a Gemma 2 [17] large language model for autoregressive prediction. Bounding boxes are represented using four dedicated tokens. Segmentation masks are represented using 16 mask tokens per segment. Mask prediction is implemented through a vector-quantized variational autoencoder (VQ-VAE) trained to encode binary masks into a discrete 16-dimensional latent representation. During training, the model predicts these discrete latent tokens, which are then decoded into segmentation masks using the VQ-VAE decoder. Optimization is performed entirely in token space using standard next-token cross-entropy loss. No pixel-level loss, auxiliary decoder, or additional segmentation objective is introduced. This formulation enables anatomical supervision within the generative training paradigm of the pretrained model.

Synthetic CXR Generation from CT. Large-scale anatomical annotations for chest radiographs are limited. To address this, we generated synthetic projection radiographs from 3D CT volumes. We used non-contrast CT scans from the CT-RATE dataset [7], consisting of 12,984 training CT volumes and 683 randomly selected test CT volumes. For each CT volume, two digitally reconstructed radiographs were generated: one posterior–anterior projection and one lateral projection. Forward projection was performed using Diff-DRR [6]. This resulted in 25,968 synthetic training images and 1,366 synthetic test images. Multi-structure anatomical segmentations were obtained from each CT volume using the TotalSegmentator [23]. The 3D segmentation labels were projected forward into the 2D projection domain to obtain anatomically consistent segmentation masks aligned with the synthetic radiographs. The binary masks were encoded into the VQ-VAE latent token space required for autoregressive prediction. A sample of the generated data was reviewed by experienced radiologists for plausibility and mask–image alignment.

Training supervision covered 80 anatomical targets spanning thoracic skeletal anatomy, pulmonary structures, cardiac chambers and major vessels, and selected upper abdominal organs visible in projection. Multiple textual synonyms were sampled during training to construct prompts for each anatomical label. To increase robustness, random scaling transformations were applied to synthetic training images to simulate variability in anatomy size and acquisition geometry. This projection-based strategy scales anatomical supervision without requiring manual CXR annotation.

Parameter-Efficient Fine-Tuning. Model adaptation was performed using Low-Rank Adaptation (LoRA) [9]. This enabled parameter-efficient fine-tuning while

keeping most pretrained weights fixed. Two configurations were evaluated: fine-tuning with the vision encoder frozen, and fine-tuning with the vision encoder updated. Experiments were conducted using 3B and 10B parameter models and input resolutions of 224×224 and 448×448. Training was performed on 8 H100 GPUs with batch sizes between 8 (10B parameter model with 448×448 image resolution) and 24 (3B parameter model with 224×224 image resolution).

Convolutional Baseline. As a supervised baseline, we trained a two-dimensional U-Net using the same synthetic training data. The network consisted of four encoder stages. Each stage included two 3×3 convolutional layers followed by batch normalization and ReLU activation, and a 2×2 max pooling layer for downsampling. The number of feature channels doubled at each stage starting from 32, resulting in feature dimensions of 32, 64, 128, and 256, with a bottleneck of 512 channels. The decoder used 2×2 transposed convolutions for upsampling and concatenated skip connections from corresponding encoder features. A final 1×1 convolution produced one output channel per anatomical class (80 in total). The input resolution was 448×448.

Evaluation Data and Metrics. Segmentation performance was evaluated on three datasets: synthetic CT-RATE test projections (1,366 images from posterior–anterior and lateral views), CheXmask [4] (200 randomly selected posterior–anterior radiographs with heart and lung masks), and VinDr-RibCXR [15] (198 posterior–anterior radiographs with bilateral rib segmentations). Segmentation quality was assessed using Dice coefficient, Intersection-over-Union (IoU), Hausdorff distance, and Fourier descriptor distance (FDD). Bounding box localization was evaluated using bounding box IoU.

Ablation Studies. Ablation experiments were performed to assess the impact of model scale, input resolution, and vision encoder fine-tuning. Comparisons were made between 3B and 10B parameter models, between 224×224 and 448×448 input resolutions, and between frozen and fine-tuned vision encoders.

Data-Efficient Transfer to Novel Localization Tasks. To evaluate transfer to novel localization tasks, few-shot experiments were conducted on visual grounding tasks from the RadVLM dataset [2]. Baseline models as well as anatomy-pretrained models were fine-tuned for one epoch using subsets of labeled data corresponding to fractions between 0% and 100% of the available RadVLM training data (n=118,360). Performance on held-out test data (n=22,972) was measured using bounding box Intersection-over-Union and compared to models without anatomical pretraining.

Code Availability. Code is available at:

- `https://github.com/sergiosgatidis/CheXanatomy`
- `https://github.com/sergiosgatidis/CheXsynth`

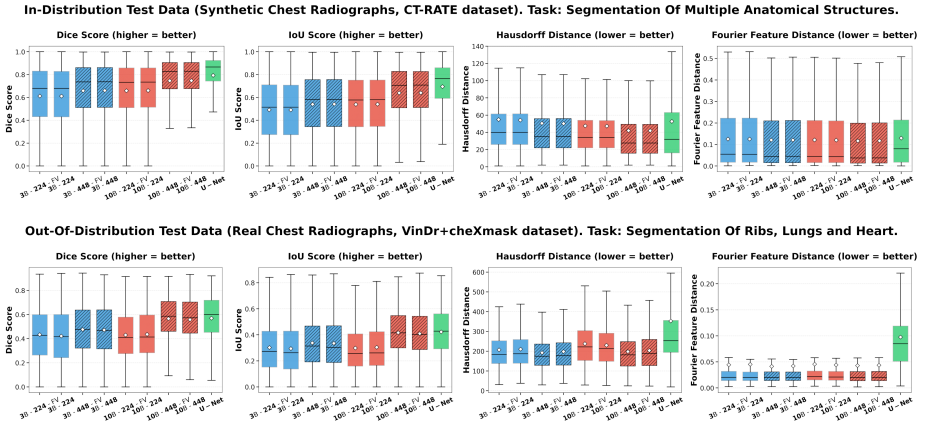

Fig. 2. Segmentation performance on synthetic CT-RATE (top) and combined real-world radiographs from CheXmask and VinDr-RibCXR (bottom). Boxplots show Dice, IoU, Hausdorff distance, and Fourier descriptor distance across model scales and input resolutions. On synthetic data, the best PaliGemma models approach U-Net in overlap metrics and achieve lower boundary and shape errors. Under domain shift, PaliGemma maintains comparable Dice and IoU while consistently reducing boundary errors relative to U-Net. Models are named based on size (3B vs. 10B) and input image size (224×224 vs. 448×448). FV = Vision encoder frozen during training.

4 Results

In-Distribution Performance on synthetic CXRs. Segmentation performance on the held-out synthetic CT-RATE radiographs is summarized in Fig. 2 (top). The best PaliGemma configuration (10B, 448×448) approached the performance of the specialized U-Net baseline in overlap-based metrics, achieving comparable Dice and IoU values. While U-Net achieved slightly higher Dice and IoU overall, the high-resolution 10B PaliGemma model obtained lower Hausdorff distance and lower Fourier descriptor distance, indicating improved boundary alignment and shape consistency. Increasing input resolution consistently improved segmentation performance across both model scales. Fine-tuning the vision encoder had no measurable influence on performance.

Out-of-Distribution Generalization on Real CXRs. Combined results across CheXmask and VinDr-RibCXR are shown in Figs. 2 (bottom) and 3. A consistent and practically relevant pattern emerges under domain shift. The best PaliGemma models achieved Dice and IoU comparable to the U-Net baseline while consistently outperforming it on boundary and shape-based metrics. In particular, Hausdorff distance and Fourier descriptor distance were substantially lower for PaliGemma models compared to U-Net, indicating improved geometric stability and shape fidelity on real radiographs. The 10B model at 448×448 resolution closely matched U-Net in overlap metrics while maintaining superior boundary and shape consistency.

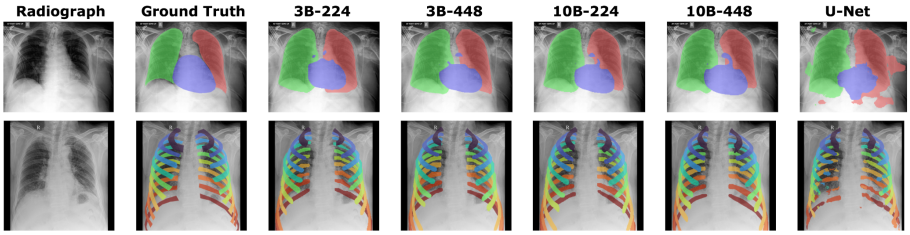

Fig. 3. Representative examples for out-of-distribution segmentation performance of anatomy-trained VLMs and a specialized U-Net compared to ground truth segmentations. Only VLMs with frozen vision encoders during fine-tuning are shown. Top row: CheXmask example (lungs and heart). Bottom row: VinDr-RibCXR example (bilateral rib structures). Compared to U-Net, anatomy-trained VLMs demonstrate more stable geometric structure and reduced boundary irregularities under domain shift.

Data-Efficient Transfer to Novel Localization Tasks. We evaluated whether anatomy pretraining improves adaptation to unseen localization tasks using the RadVLM visual grounding dataset [2]. When fine-tuned with limited or no supervision (0%, 1%, and 3% of the training data), anatomy-pretrained models consistently outperformed their non-pretrained counterparts in bounding box IoU, as shown in Fig. 4. The performance gap was most pronounced in the low-data regime, indicating improved sample efficiency and a stronger spatial prior. As the amount of task-specific training data increased (10% and above), performance between pretrained and non-pretrained models converged. These results suggest that explicit anatomical supervision provides a structured initialization that facilitates rapid adaptation to new spatial tasks, particularly when labeled data are scarce.

5 Discussion

We demonstrated that explicit anatomical supervision can be integrated into a pretrained vision–language model through autoregressive token prediction. Our results show that token-space supervision enables segmentation performance comparable to a specialized convolutional baseline on synthetic in-distribution data, while improving geometric robustness under domain shift to real radiographs.

On synthetic CT-RATE projections, the best PaliGemma configurations approached U-Net in Dice and IoU and achieved lower boundary and shape errors. Under real-world distribution shift, PaliGemma maintained comparable overlap metrics while consistently reducing Hausdorff and Fourier descriptor distances. These findings indicate that anatomical supervision in token space promotes spatially coherent and geometrically stable representations, even when trained solely on synthetic projections.

Scaling model size and input resolution improved performance, whereas fine-tuning the vision encoder had limited impact, suggesting that anatomical rea-

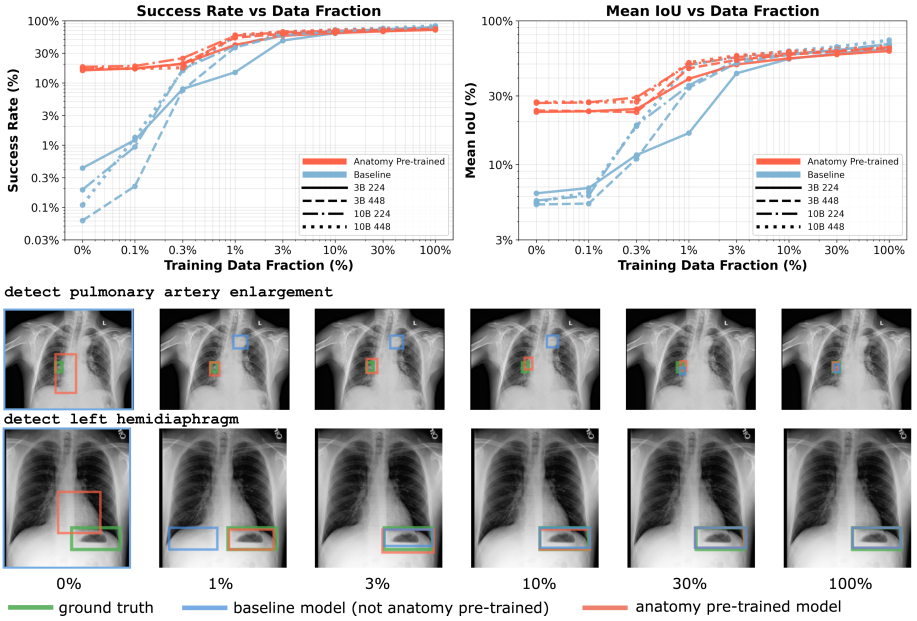

Fig. 4. Performance scaling on the RadVLM grounding benchmark comparing anatomy-pretrained and baseline models across training data fractions. Top: Success rate and mean IoU as a function of available training data (0–100%). Anatomy-pretrained models show clear advantages in the low-data regime (0–3%), with performance converging as more task-specific data becomes available. Bottom: Representative grounding examples for two tasks ("detect pulmonary artery enlargement" and "detect left hemidiaphragm") across data fractions. Green: ground truth, red: anatomy-pretrained model (3B-448), blue: baseline model (3B-448). Anatomy pretraining provides improved localization accuracy and stability when supervision is limited.

soning primarily emerges in the multimodal token space. Few-shot localization experiments further demonstrated improved sample efficiency, with anatomy-pretrained models outperforming baselines in low-data regimes and converging as task-specific supervision increased.

Beyond quantitative performance, the autoregressive VLM formulation offers structural advantages over conventional segmentation networks. The same model supports flexible language prompting, enables integration of new anatomical structures through lightweight fine-tuning without architectural modification, and can be embedded into pipelines requiring spatial grounding, such as grounded report generation or region-level reasoning.

Overall, structured anatomical supervision provides a scalable mechanism for aligning generative medical vision–language models with spatial structure in radiographs.

6 Conclusion

We introduced CheXanatomy, a framework for injecting explicit anatomical knowledge into a pretrained vision–language model through autoregressive token-space supervision. By leveraging scalable CT-derived synthetic radiographs, we trained a VLM to generate anatomical segmentations without task-specific decoder heads.

Our results demonstrate that anatomy-aware token supervision enables competitive in-distribution performance and improved geometric generalization to real radiographs. More broadly, this suggests that spatial grounding can be embedded directly into multimodal foundation models through structured generative supervision.

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

References

1. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A Language Modeling Framework for Object Detection. In: International Conference on Learning Representations (ICLR) (2022), https://arxiv.org/abs/2109.10852
2. Deperrois, N., Matsuo, H., Ruipérez-Campillo, S., Vandenhirtz, M., Laguna, S., Ryser, A., Fujimoto, K., Nishio, M., Sutter, T., Vogt, J.E., Kluckert, J., Frauenfelder, T., Blüthgen, C., Nooralahzadeh, F., Krauthammer, M.: RadVLM: A multitask conversational vision-language model for radiology (2025), https://arxiv.org/abs/2502.03333
3. Dong, Z., Wu, W., Hao, J., Chen, T., Weng, Z., Zhou, B.: AnyCXR: Human anatomy segmentation of chest x-ray at any acquisition position using multi-stage domain randomized synthetic data with imperfect annotations and conditional joint annotation regularization learning (2025), https://arxiv.org/abs/2512.17263
4. Gaggion, N., Mosquera, C., Mansilla, L., Saidman, J.M., Aineseder, M., Milone, D.H., Ferrante, E.: CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. Scientific Data **11**(1), 511 (2024). https://doi.org/10.1038/s41597-024-03358-1
5. van Ginneken, B., Stegmann, M.B., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. Medical Image Analysis **10**(1), 19–40 (2006). https://doi.org/10.1016/j.media.2005.02.002
6. Gopalakrishnan, V., Golland, P.: Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In: Clinical Image-Based Procedures: 11th Workshop, CLIP 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings. pp. 1–11. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-23179-7_1
7. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Durugol, O.F., Wittmann, B., Amiranashvili, T., Simsar, E., Simsar, M., Erdemir, E.B., Alanbay, A., Sekuboyina, A., Lafci, B., Bluethgen, C., Ozdemir,

M.K., Menze, B.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nature Biomedical Engineering (2025). https://doi.org/10.1038/s41551-025-01599-y
8. Hou, B., Zhu, Q., Mathai, T.S., Jin, Q., Lu, Z., Summers, R.M.: Shadow and light: Digitally reconstructed radiographs for disease classification. arXiv preprint arXiv:2406.03688 (2024)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
10. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods (2021). https://doi.org/10.1038/s41592-020-01008-z
11. 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)
12. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: MedCLIP-SAMv2: Towards universal text-driven medical image segmentation. Medical Image Analysis **106**, 103749 (2025). https://doi.org/10.1016/j.media.2025.103749
13. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: CLIP-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
14. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7086–7096 (2022)
15. Nguyen, H.C., Le, T.T., Pham, H.H., Nguyen, H.Q.: VinDr-RibCXR: A benchmark dataset for automatic segmentation and labeling of individual ribs on chest x-rays. arXiv preprint (2021), https://arxiv.org/abs/2107.01327, arXiv:2107.01327
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), https://proceedings.mlr.press/v139/radford21a.html
17. Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Le Lan, C., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., et al.: Gemma 2: Improving open language models at a practical size (2024), https://arxiv.org/abs/2408.00118
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer (2015), https://arxiv.org/abs/1505.04597
19. Seibold, C., Jaus, A., Fink, M.A., Kim, M., Reiß, S., Herrmann, K., Kleesiek, J., Stiefelhagen, R.: Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling (2023), https://arxiv.org/abs/2306.03934
20. Seibold, C.M., Reiß, S., Sarfraz, M.S., Fink, M.A., Mayer, V., Sellner, J., Kim, M.S., Maier-Hein, K.H., Kleesiek, J., Stiefelhagen, R.: Detailed annotations of

chest x-rays via ct projection for report understanding. In: British Machine Vision Conference (BMVC) (2022), https://bmvc2022.mpi-inf.mpg.de/0058.pdf
21. Steiner, A., Susano Pinto, A., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., Qin, S., Ingle, R., Bugliarello, E., Kazemzadeh, S., Mesnard, T., Alabdulmohsin, I., Beyer, L., Zhai, X.: PaliGemma 2: A family of versatile VLMs for transfer. arXiv preprint (2024), https://arxiv.org/abs/2412.03555, arXiv:2412.03555
22. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**, 1399–1406 (2022). https://doi.org/10.1038/s41551-022-00936-9
23. 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)
24. 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 (ICCV). pp. 11975–11986 (October 2023)
25. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the 7th Machine Learning for Healthcare Conference (MLHC). Proceedings of Machine Learning Research, vol. 182, pp. 2–25. PMLR (2022), https://proceedings.mlr.press/v182/zhang22a.html