DATA SCALING LAWS FOR RADIOLOGY FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation vision encoders such as CLIP and DINOv2, trained on web-scale data, exhibit strong transfer performance across tasks and datasets. However, medical imaging foundation models remain constrained by smaller datasets, limiting our understanding of how data scale and pretraining paradigms affect performance in this setting. In this work, we systematically study continual pretraining of two vision encoders, MedImageInsight (MI2) and RAD-DINO representing the two major encoder paradigms CLIP and DINOv2, on up to 3.5M chest x-rays from a single institution, holding compute and evaluation protocols constant. We evaluate on classification (radiology findings, lines and tubes), segmentation (lines and tubes), and radiology report generation. While prior work has primarily focused on tasks related to radiology findings, we include lines and tubes tasks to counterbalance this bias and evaluate a model's ability to extract features that preserve continuity along elongated structures. Our experiments show that MI2 scales more effectively for finding-related tasks, while RAD-DINO is stronger on tube-related tasks. Surprisingly, continually pretraining MI2 with both reports and structured labels using UniCL improves performance, underscoring the value of structured supervision at scale. We further show that for some tasks, as few as 30k in-domain samples are sufficient to surpass open-weights foundation models. These results highlight the utility of center-specific continual pretraining, enabling medical institutions to derive significant performance gains by utilizing in-domain data.

1 INTRODUCTION

Foundation models have shown strong potential in computer vision by leveraging large-scale pretraining for broad adaptability. Models trained on massive datasets like LAION-5B (Schuhmann et al., 2022), with billions of image–text pairs, achieve impressive zero-shot and few-shot performance on tasks such as ImageNet classification (Radford et al., 2021). Two main pretraining paradigms dominate: image–text contrastive learning (e.g., CLIP (Radford et al., 2021)) and image-only self-supervised learning (e.g., DINOv2 (Oquab et al., 2024)). These differ in inputs, scalability, and downstream performance: CLIP excels at classification and retrieval, while DINOv2 performs better on segmentation and detection (Bolya et al., 2025; Cherti et al., 2023; Jiang et al., 2024; Tong et al., 2024). In recent years, medical researchers have increasingly adopted foundation models to boost performance across diverse clinical tasks and address data scarcity and annotation challenges (Codella et al., 2024; Pérez-García et al., 2025b; Lin et al., 2023; Zhang et al., 2025; Zedda et al., 2025; Moutakanni et al., 2024). Unlike general-domain datasets, chest X-ray (CXR) datasets typically contain only hundreds of thousands to a few million images, raising questions about how well general-domain pretraining insights transfer. We conduct a controlled comparison of two leading CXR pretraining approaches: MedImageInsight (MI2), which uses CLIP-style image–text contrastive learning (Codella et al., 2024), and RAD-DINO, based on DINOv2-style image-only self-supervision (Pérez-García et al., 2025b). Both provide open weights and have shown state-of-the-art performance on public benchmarks. Our study leverages INST-CXR-BENCH, a large internal dataset of 4M de-identified CXR–report pairs, enabling control over data source and distribution, an advantage over prior work that compares models trained on heterogeneous datasets with varying compute budgets (Codella et al., 2024; Pérez-García et al., 2025b; Huang et al., 2021; Zhang et al., 2022; Bannur, Shruthi et al., 2023). To ensure fairness, we use identical computational resources for both models. In our experiments, Section 5, we continually pretrain with both approaches on INST-CXR-BENCH and evaluate performance across multiple tasks and pretraining dataset sizes.

Our evaluation covers three task categories: classification, segmentation, and report generation. While prior work has focused mainly on radiological findings, we extend this by adding tasks on lines and tubes (l&t) to probe learning of curve-continuity features, structures that preserve continuity along elongated objects (Section 3.2). We extract findings and l&t labels from radiology reports using GPT. These labels primarily serve evaluation but also enable extending MI2 pretraining from CLIP to UniCL (Yang et al., 2022), integrating structured labels with image–text contrastive learning. To support robust analysis, we construct a large test set of 400k samples from INST-CXR-BENCH, capturing long-tail findings and a diverse patient population. We establish scaling laws by analyzing performance across varying pretraining dataset sizes (Kaplan et al., 2020). For classification on INST-CXR-BENCH, as few as 30k samples of continual pretraining can surpass open-weight baselines. MI2 scales more effectively than RAD-DINO for findings classification, while both models show similar trends on l&t classes. Notably, adding structured labels via UniCL significantly boosts MI2, an unexpected result given millions of image–report pairs. These patterns align with report generation experiments, Section 5.3, where we pair vision encoders with a Vicuna-13B LLaVA model (Liu et al., 2023; Bannur et al., 2024; Hyland et al., 2024). Beyond our INST-CXR-BENCH dataset, we also evaluate the continually pretrained models on publicly available benchmarks. For findings classification, our updated models are on-par with or surpass the original open-weights models. For l&t segmentation, RAD-DINO and both versions of MI2 outperform the original open-weights models.

Our findings reveal nuanced trade-offs between CLIP-style and DINOv2-style pretraining in medical imaging: 1. MI2 performs better on findings-related tasks; 2. RAD-DINO excels on l&t; and 3. Adding label supervision via UniCL significantly improves MI2 performance on l&t. More broadly, medical foundation models would benefit from training and evaluation on substantially larger and more diverse datasets than are common today. This need for scale is amplified by center-specific factors, including: (i) variability in image characteristics from scanners, protocols, and resolution; (ii) population-level differences such as age and ethnicity; and (iii) label distribution shifts, including rare conditions and reporting styles. In this context, continually pretraining center-specific foundation models on in-domain data, even with as few as 30k samples, can outperform open-weight models, underscoring current limitations in generalization of CXR foundation vision encoders.

2 Related work

Recent work has explored scaling laws for vision transformers (ViTs) (Zhai et al., 2022) and self-supervised pretraining methods in general domains (Cherti et al., 2023). For instance, Fan et al. (2025) show DINOv2 scales more favorably than CLIP with respect to both dataset size and model capacity at large scales involving billions of samples and parameters. However, these studies are primarily conducted at internet-scale datasets and with billion-parameter models, whereas medical imaging pretraining typically operates in a very different regime: model sizes of 0.3B parameters (i.e., ViT-L scale) or fewer and datasets that are several orders of magnitude smaller. To our knowledge, we are the first to systematically study the scaling behavior of pretraining vision encoders on CXR datasets up to millions of samples. Existing work on scaling in medical imaging has largely focused on supervised learning. For example, Cho et al. (2016) studied the scaling of convolutional neural networks trained with supervised learning on limited medical data. Xu et al. (2023) as well as Sellergren et al. (2022) explore the scaling behavior of a linear findings classifier applied to a frozen vision encoder backbone, and also perform end-to-end fine-tuning with the same (unfrozen) encoder. In contrast, our work examines the effect of dataset size in the context of vision encoder pretraining using modern transformer-based architectures and a large-scale, single-modality medical image dataset.

Several recent efforts have implicitly compared the performance of DINOv2 and CLIP in medical domains. However, these comparisons are often confounded by differences in pretraining data and potentially compute budget. For instance, RAD-DINO and MI2 both evaluate multiple pretrained models, but the models are trained on different datasets, making direct performance comparisons difficult. Moreover, MI2 focuses on classification and retrieval tasks, which are known to favor CLIP-style contrastive learning approaches, potentially biasing conclusions. Our study addresses this by fairly comparing models trained with CLIP, DINOv2, and UniCL under controlled compute budgets, using consistent data sources, and a variety of tasks.

Last, the importance of considering layer-wise differences in representation quality has been highlighted in recent studies such as Bolya et al. (2025), which showed that different layers of a vision transformer can capture different types of features and exhibit variable downstream performance. However, this perspective has not been thoroughly explored in the medical domain. We are the first to incorporate this consideration into a systematic comparison of medical vision encoders, revealing insights that are potentially obscured when evaluating only the final layer representations.

3 Method

3.1 MedImageInsight

MedImageInsight (MI2) is a CLIP-style contrastive pretraining approach built on the Unified Contrastive Learning (UniCL (Yang et al., 2022)) framework. The open-weights MI2 vision encoder was trained on approximately 500k CXR image-text and image-label pairs plus 3.3M samples from various other medical imaging modalities. The open-weights model version of MI2 will be abbreviated with MI2 OWM throughout the paper. MI2 replaces the standard ViT backbone with a dual-attention ViT (DAViT) (Ding et al., 2022), a hierarchical vision transformer that is claimed to be better suited for medical imaging tasks, particularly given the limited size of domain-specific datasets. CLIP-like models jointly train an vision encoder and a text encoder by projecting both modalities into a shared feature space. A contrastive loss, following the InfoNCE formulation (Oord et al., 2019), aligns matched image–text pairs while pushing apart unmatched ones. UniCL extends the CLIP framework to support image–label contrastive learning. In MI2, structured categorical labels (e.g., disease annotations) are used as input to the text encoder, in the same way as radiology reports. The labels are represented by their category names or a list of names. These labels are tokenized and embedded by the text encoder, allowing the model to learn from both image–text and image–label pairs. Empirically, MI2 outperforms RAD-DINO and other CXR foundation vision encoders (Zhang et al., 2025; Moor et al., 2023) across a wide range of tasks including classification, retrieval, and findings generation, establishing it as the current state-of-the-art on finding-related tasks. This aligns with broader findings in the literature suggesting that CLIP-style models tend to excel at classification and retrieval, while DINOv2-style models may offer advantages in tasks that require dense outputs like segmentation (Jiang et al., 2024). We used the open-weights MI2 weights as a starting point for continual pretraining as described here: `https://techcommunity.microsoft.com/blog/healthcareandlifesciencesblog/discovering-the-power-of-finetuning-medimageinsight-on-your-data/4395057`

3.2 RAD-DINO

RAD-DINO (Pérez-García et al., 2025b) is a self-supervised image-only pretraining approach for CXRs based on the DINOv2 (Oquab et al., 2024) approach. The open-weights RAD-DINO vision encoder was trained on ∼840k frontal and lateral CXRs with slight adjustments of the DINOv2 augmentations to be more suitable for CXRs. The open-weights model version of RAD-DINO will be abbreviated with RAD-DINO OWM throughout the paper. RAD-DINO inherits the core architectural and training principles of DINOv2, including self-distillation with ViT backbones (Caron et al., 2021), and masked image modeling in the style of iBOT (Zhou et al., 2022). There are two ViTs, the student and teacher networks. During training, multiple augmented views of each CXR are generated using radiology-specific transformations such as larger crop sizes and less severe blurring. There are three different parts of the loss function: (i) cross-entropy loss between the teacher's and student's CLS token, (ii) masked image loss where a subset of image patches is masked, and the student is trained to match the teacher's representations of the masked tokens, (iii) the so-called KoLeo regularizer that encourages optimal use of the feature space. The teacher is updated via an exponential moving average of the student instead of gradient descent. RAD-DINO uses a ViT-Base model for both the student and teacher. At inference time, only the teacher network is used. At the time of its release, RAD-DINO outperformed both purely image-trained and image-text contrastive models (Bannur, Shruthi et al., 2023; Zhang et al., 2025; Tiu et al., 2022; Zhou et al., 2023) across a wide range of tasks, including findings classification, metadata classification, segmentation, and report generation. These results challenge the assumption that supervision via radiology reports is necessary for training high-performing vision encoders. The RAD-DINO checkpoint (including the DINO heads) is available at: `https://huggingface.co/microsoft/RAD-DINO`. We use the DINOv2 codebase for continual pretraining: `https://github.com/facebookresearch/dinov2`.

Throughout Section 5, RAD-DINO demonstrates strong performance on tasks involving lines and tubes (l&t), performing on par with MI2 even in l&t classification. This is especially notable given the well-established advantage of CLIP-pretrained models over DINOv2 on classification tasks (Bolya et al., 2025). We hypothesize that RAD-DINO benefits from self-distillation with masked multi-view objectives, which encourages the learning of curve-continuity features (CCF), features that preserve continuity along elongated structures such as l&t, see Figure 1. These features are particularly well-suited for tasks like tube tip localization and segmentation, where even small discontinuities can result in significant penalties. In contrast, several aspects of CLIP may inhibit the learning of CCF. First, CLIP aligns global features from the image and text encoders, which may fail to capture fine-grained structural details (Huang et al., 2021). Second, chest X-ray reports often omit or only sparsely mention medical devices, frequently lacking the detailed descriptions needed for robust alignment.

3.3 Continual pretraining and scaling laws

Previous work has shown that the performance of large models improves predictably with scale, following power-law relationships with respect to model size, dataset size, and compute budget (Kaplan et al., 2020). This has enabled performance extrapolation from early training curves, providing a framework to guide the development of increasingly capable models. In particular, dataset size has been identified as a dominant factor in scaling performance. Bansal et al. (2022) argue that dataset size contributes more significantly than architecture or model size in the domain of neural machine translation. Similar findings in the vision domain confirm the critical role of data quantity in driving performance gains (Zhai et al., 2022). In this work, we focus specifically on dataset size scaling laws while keeping compute and model size fixed. This decision is motivated by two key factors: (i) we use pretrained RAD-DINO and MI2 checkpoints as our starting point, which constrains our ability to vary model size; and (ii) prior work consistently demonstrates that data quantity is the most influential factor in driving performance improvements. When applicable (e.g., Section 5.1.1), we fit a power-law of the form $f(x) = \alpha x^k$, where $f(x)$ is a performance metric (e.g., AUPRC) obtained by evaluating a frozen encoder on a downstream task, pretrained on a dataset of size x. However, recent studies caution against overgeneralizing the predictive power of scaling laws. For instance, Caballero et al. (2023) and Alabdulmohsin et al. (2022) show that power-law behavior is often confined to a narrow region of the parameter space, with saturation effects becoming apparent at larger scales. Similarly, Lourie et al. (2025) demonstrate that downstream performance may exhibit emergent behavior, saturation, or even inverse scaling, where increased scale degrades performance potentially due to a distribution shift and catastrophic forgetting.

4 INST-CXR-BENCH DATASET CREATION

For pretraining and evaluation we use a large internal dataset consists of 3.1M CXR studies sourced from WITHHELD FOR REVIEW, with approximately 23% of the studies containing at least one line or tube. The data was split on a patient level into 80% for training, 10% for validation, and 10% for testing. Each study contains longitudinal information, incorporating current frontal and lateral images, prior frontal images, prior reports, and clinical context such as indication and comparison sections. From the 3.1M studies we create a dataset INST-CXR-BENCH containing approximately 4 million CXR images and associated reports, where the same patient can contribute multiple images and reports. The images are divided into frontal (62%) and lateral (38%) images. Patient sex is divided into three categories: 'Male' (49%), 'Female' (48%), and 'Other' (2%). Patient ethnicity consists of seven categories: 'White' (87%), 'Black or African American' (3%), 'Asian' (1%), 'Asian - Far East' (1%), 'Native American and Pacific Islander' (1%), and 'Asian - Indian Subcontinent' (<1%). Patient age has the following distribution: '<20' (2%), '20-30' (7%), '30-40' (9%), '40-50' (13%), '50-60' (22%), '60-70' (22%), '70-80' (17%), '80-89' (7%), '89+' (1%). The images and reports were created between 2013 and 2023: '2007-2012' (35%), '2013-2017' (25%), '2018-2023' (37%). There are 16 different departments that ordered CXRs, the major six are 'Internal Medicine' (14%), 'Emergency Medicine' (13%), 'Cardiovascular Diseases' (9%), 'Radiology' (9%), 'Family Medicine' (8%), 'General Practice' (6%). The patients are grouped into 'inpatient' (58%) and 'outpatient '(40%). The images were acquired by scanner from 19 different manufacturers, the major seven are: 'FUJIFILM Corporation' (38%), 'Carestream Health' (24%), 'GE Healthcare' (15%), 'SIEMENS' (8%), 'Philips' (6%), 'Canon Inc.' (3%). CXR DICOM images were converted to PNG and resized to 518px using B-spline interpolation with antialiasing. To ensure patient privacy, white boxes were overlaid on the images to mask identifying information such as text, facial features, or other visual elements that could potentially reveal a patient's identity. Intensities were normalized to an 8-bit range. GPT-4o (OpenAI, 2024) was used to parse and clean the reports into structured JSON, handling inconsistent formatting, duplications, and artifacts of the EHR storing process. Each frontal image was linked to corresponding lateral and prior images, when available. A de-duplication step retained one image per type (frontal, lateral, prior frontal) per visit, prioritizing original images with complete metadata. Last, Fastdup[1] was used to detect and remove ∼6% of outlier images, such as blank or non-chest X-rays.

5 EXPERIMENTS

To better understand how foundation models in medical imaging scale with data, we perform continual pretraining starting from two strong open-weights baselines: MI2 (Codella et al., 2024) and RAD-DINO (Pérez-García et al., 2025b). We progressively pretrain both models on five strictly nested subsets of INST-CXR-BENCH: 30k, 50k, 100k, 1M, and 3.5M image-report pairs. Each model is trained for equal wall-clock time on four nodes, with eight H100 GPUs per node, with a batch size of 1280 (40 samples per GPU). The training durations for each data size are: 0.33, 0.58, 1.17, 11.67, 40.83 hours, respectively, with a standard deviation of 8% in training time. These durations correspond to 15 epochs of MI2 training, and RAD-DINO is trained for an equivalent amount of time by adjusting its epoch count accordingly. For MI2, we pretrain two variants: Image-report contrastive learning (standard CLIP) and image-report plus image-label following the UniCL approach (Yang et al., 2022), where each image is seen once paired with its report and a second time paired with a label per epoch, if available. We adjust the number of epochs to ensure equal wall-clock time. As a case study, we use tube presence labels (e.g., "Nasogastric Tube, Endotracheal Tube") extracted via GPT (Appendix A.7); 23% of CXRs include at least one line or tube. All MI2 text inputs include a view-position prefix. For 30k, 50k, and 100k, we create three random dataset subsets resulting in three encoders per pretraining approach. For 1M and 3.5M, we train one encoder each. Results are compared against the original open-weight models (OWM) (Figures 2–5). Throughout the following section, we clearly distinguish which results

[0][1] `https://www.visual-layer.com/`

are statistically significant, and whenever we label a result as significant, the claim is supported by the significance test described in Appendix A.8.

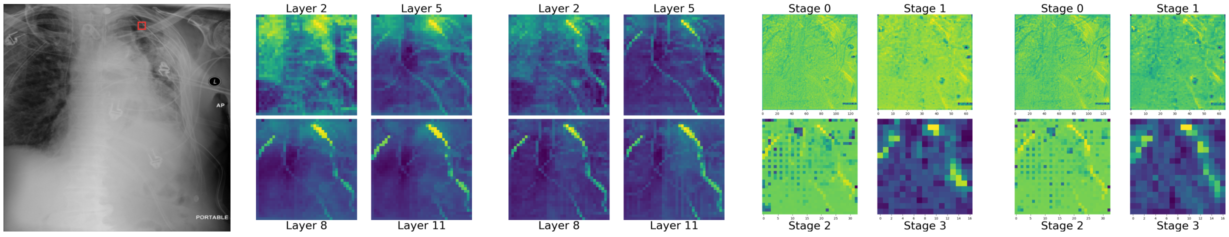
5.1 Classification

To compare embedding quality, we run classification experiments using frozen backbones, isolating feature quality without fine-tuning. Our goal is to evaluate MI2 and RAD-DINO under identical downstream conditions. We avoid using the [CLS] token, as prior work (Bolya et al., 2025) shows different layers encode different information. Instead, we extract features from four layers per encoder. MI2, based on DAViT, outputs multi-scale features, so we apply convolutions and linear projections to unify dimensions to 33×33×1024 (third block size). While RAD-DINO uses a ViT-B backbone where all feature maps have identical shapes, we still apply the same projection strategy for fairness, using layers 2, 5, 8, and 11 (zero-indexed). Details are in Table 1. Projected features from the four blocks are concatenated (4096 dims for MI2, 3072 for RAD-DINO), then reduced via attention pooling to a single token per task, followed by a linear classifier. We train for 25 epochs with batch size 512. Each experiment is repeated three times: for 30k, 50k, and 100k using different encoders and seeds; for 1M and 3.5M using the same encoder with three seeds.

5.1.1 Findings classification on INST-CXR-BENCH-FIND-CLASS

We use a 2M subset of frontal CXRs from INST-CXR-BENCH (75%/25% train/test images) for training and evaluating findings classification models. Subsequently, we will call the subset INST-CXR-BENCH-FIND-CLASS. We choose 19 findings categories covering a wide range of appearances (some are more diffuse/texture like, others are more localized/shape like) and areas of a CXR (from the esophagus to the diaphragm). Each finding has at least 10k examples in the train set. All labels are extracted from the paired reports as described in Appendix A.7. Table 2 contains the list of all findings we consider and their prevalences. Note: While we are varying the number of pretraining samples, the amount of samples to train the classification model is always the same.

In Figure 2 (left), we observe clear power-law behavior when plotting classification performance against dataset size on a log scale. The scaling trends appear linear, indicating predictable and consistent gains as more data is used for continual pretraining. In agreement with (Bolya et al., 2025), MI2 significantly outperforms RAD-DINO on this classification task starting at the pretraining dataset size of 100k. Different MI2 variants (CLIP vs UniCL) perform comparably, suggesting that the additional supervision using tube presence labels has no negative effect on findings classification. Comparing the power law fits in Figure 2 (left) we find that MI2 scales about three times better than RAD-DINO. In addition, we find that already with 100k images, continual pretraining can significantly outperform public foundation model checkpoints, highlighting the value of domain-specific adaptation. While average AUPRC improvements may appear modest, class-specific gains can be substantial. In Figure 2 (right), we show scaling laws for the binary task of rib fracture classification. MI2 continually pretrained with 3.5M images shows an improvement of 6% compared to MI2 open-weights model, also see Table 2. For MI2, 30k samples of INST-CXR-BENCH are sufficient to significantly outperform the open-weights model. In contrast, for RAD-DINO, 100k samples are needed to outperform the open-weights model. In Table 2, we compare models trained with all of INST-CXR-BENCH's data (3.5M samples) vs the open-weights models. We find that for all 20 binary tasks both variants of MI2 pretrained



(a) Original image (b) RAD-DINO 30k (c) RAD-DINO 3.5M (d) MI2 reports 30k (e) MI2 reports 3.5M

Figure 1: We visualize the cosine similarity maps between the patch marked with a red box (tip of a chest tube) and all other patches extracted for four layers from RAD-DINO and MI2, each pretrained with either 30k or 3.5M images. We argue that an ideal feature map should: (i) highlight all patches belonging to the chest tube on the right side of the image, (ii) highlight all the patches belonging to the chest tube on the left side of the image, (iii) should not highlight any other tubes or structures. For both models pretrained with 3.5M the last feature maps are the closest to the ideal feature map described above, all other feature maps seem to highlight additional tubes and structures. While more prevalent in MI2, at 30K both models miss patches associated with chest tubes, underscoring the benefits of large pretraining datasets.

on INST-CXR-BENCH outperform the open-weights model of MI2 as well as RAD-DINO. The most noticeable improvements (greater than 5%) are observed in the binary classification tasks of detecting pneumothorax, enlarged pulmonary artery, and rib fracture. For findings classification (see also Section 5.1.3), we identify several classes with low AUPRC, likely due to noisy labels. We attribute this noise to three main sources: (i) inter-reader variation among radiologists, (ii) inaccurate original reports, and (iii) errors introduced during the GPT-based extraction of structured labels (see Appendix A.7). In Appendix A.9, we report all results from section stratified by various metadata variables.

5.1.2 Tube presence classification on INST-CXR-BENCH-TUBE-CLASS

We use a 1M subset of frontal CXRs from INST-CXR-BENCH (90%/10% train/test images) for training and evaluating tube presence classification models. Subsequently, we will call this subset INST-CXR-BENCH-TUBE-CLASS. L&t prevalences are ranging from 0.56% to 14.03%, see Table 3 for exact numbers. All labels are extracted from the paired reports as described in Appendix A.7. Note: While we are varying the number of pretraining samples, the amount of samples to train the classification model is always the same.

RAD-DINO consistently has a higher or on par average AUPRC compared to MI2 trained solely on reports. Only at 1M, the performance is comparable, see Figure 3. Overall, MI2 seems to saturate faster with increasing pretraining data than RAD-DINO, a similar observation was made in Fan et al. (2025). Furthermore, RAD-DINO and MI2 trained on INST-CXR-BENCH begin to significantly outperform open-weights models at 100k pretraining samples. Adding tube presence labels to report-based MI2 improves the performance of MI2, surpassing RAD-DINO on average AUPRC. MI2 trained with both reports and tube presence labels outperforms the open-weights MI2 model at around 50k samples, significantly earlier than MI2 trained only with reports. It is important to note that for the average AUPRC none of the three pretraining methods significantly outperforms the other two. While the average AUPRC gains reported in Figure 3 (left) are modest, we observe more substantial improvements for less prevalent and harder-to-detect tube types, such as intra-aortic balloon pumps (small structure) and mediastinal drains (often obscured by the spine), see Figure 3 (right) and Table 3. For the mediastinal drain in particular (Figure 3 right), MI2 trained with both reports and tube presence labels significantly outperforms RAD-DINO starting at a pretraining dataset size of 30k and MI2 trained without tube presence labels starting at a pretraining dataset size of 100k, highlighting the benefit of incorporating GPT-extracted labels during pretraining. In Table 3 we compare models trained with all of INST-CXR-BENCH (3.5M samples) vs the open-weights models. We find that for all 11 binary l&t detection tasks MI2 trained with reports and tube presence labels is outperforming all other models or is on par with RAD-DINO.

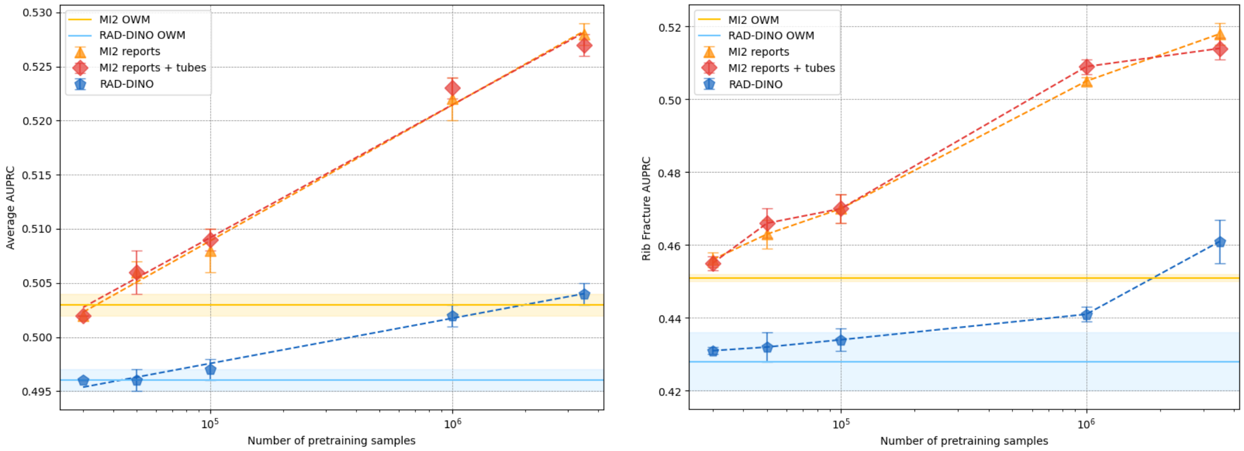


Figure 2: Findings classification performance on INST-CXR-BENCH-FIND-CLASS as a function of vision encoder pretraining with increasing sample sizes from INST-CXR-BENCH. Left: AUPRC averaged across 20 finding tasks. Both MI2 models have slope k=0.012 and intercept α=0.447, RAD-DINO has slope k=0.004 and intercept α=0.466, i.e., MI2 scales about three times better than RAD-DINO. Right: AUPRC for finding rib fracture (prevalance 2.3%), which shows the greatest improvement when pretrained with 3.5M samples.

5.1.3 Findings classification on holdout dataset VinDR

In addition to the in-domain classification experiments in Section 5.1.1 and Section 5.1.2, we evaluate finding classification performance on a public holdout dataset, called VinDR (Nguyen et al., 2022), comprising 9k CXR images

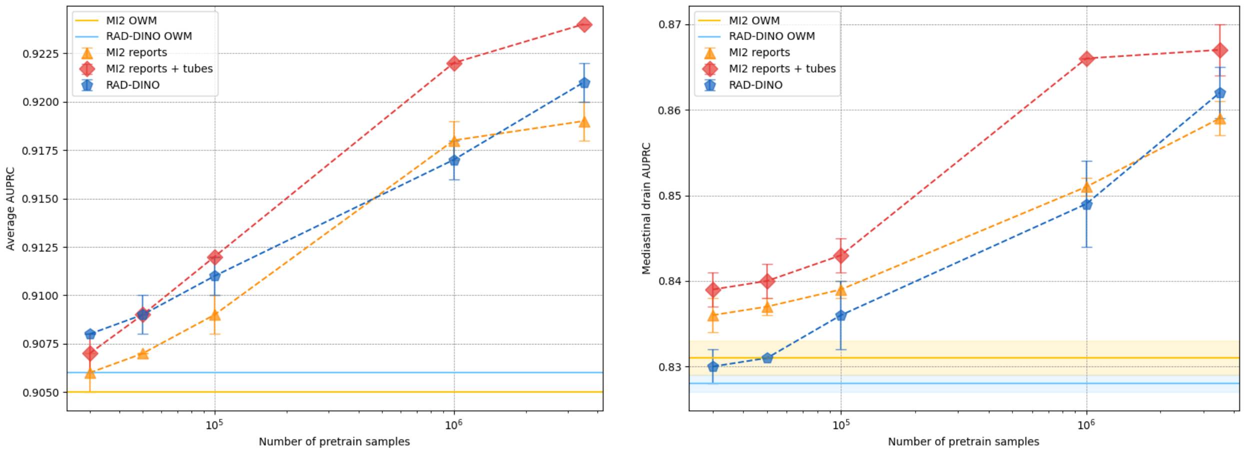


Figure 3: Tube presence classification on INST-CXR-BENCH-TUBE-CLASS as a function of vision encoder pretraining with increasing samples sizes from INST-CXR-BENCH. Left: AUPRC averaged across 11 lines and tubes tasks. Right: AUPRC for the tube mediastinal drain (Prevalance 5.6%). Task with the highest improvement when pretrained with 3.5M samples.

(90%/10% train/test). Note: While we are varying the number of pretraining samples, the amount of samples to train the classification model is always the same. Due to the small size of the dataset and its potential domain mismatch with our training data, clear performance scaling trends are difficult to establish, see Section 3.3. This reflects a broader challenge in medical imaging: small public benchmarks often provide limited insight into the generalization capabilities of foundation models. In Figure 4 (left), RAD-DINO exhibits a reversed scaling trend. As discussed in Section 3.3, this can occur when the target data distribution aligns more closely with that of a pretrained model. In this case, the open-weights RAD-DINO encoder appears better suited to the VinDR dataset, likely due to favorable pretraining data distribution and checkpoint selection. At larger scales, we observe signs of catastrophic forgetting, further suggesting a distribution shift. Both versions of MI2 are more in line with the expected trend: more pretraining data leads to better models. In contrast to the experiments in Section 5.1.1, we need substantially more samples (>100k) to outperform the open-weights models. We also observe a notable effect when tube presence labels are included alongside reports during MI2 pretraining. For example, we find significant improvements for cardiomegaly and pleural thickening across all pretrain dataset sizes. We hypothesize that this effect is due to the implicit clinical context conveyed by the presence of medical tubes. Such devices often indicate severe illness and correlate with specific pathologies. Including tube presence labels may help the model distinguish between disease-related and device-induced visual features, with tube presence potentially acting as a proxy for disease severity. However, since we do not observe the same behavior in Figure 2, this may also reflect shortcut learning, likely due to the small dataset size (Pérez-García et al., 2025a; Geirhos et al., 2020). Another issue related to limited data can be seen in Table 4, where we compare models trained on the full INST-CXR-BENCH (3.5M samples) to open-weights models. Surprisingly, MI2, even when pretrained on 3.5M (image, report, tube presence label) samples, does not consistently outperform the RAD-DINO open-weights model, contrary to the trend in all of our other experiments, suggesting benchmark saturation.

5.2 Lines and tubes segmentation on holdout dataset RANZCR-CLiP

Since INST-CXR-BENCH does not contain segmentation masks, we train and evaluate l&t segmentation models on a public holdout dataset, called RANZCR-CLiP (Tang et al., 2021), consisting of 17k CXR (75%/25% train/test). Note: While we are varying the number of pretraining samples, the amount of samples to train the segmentation model is always the same. We use the same feature pyramid architecture as in the classification setup, see Section 5.1. For MI2, additional upsampling layers are applied to match the spatial resolution of RAD-DINO feature maps. A linear segmentation head followed by upsampling to the original image size (518×518) is applied. All segmentation models are trained using Dice loss. For evaluating tube-like structures, we prioritize the Hausdorff distance as the primary metric due to its sensitivity to spatial localization. In Figure 6, we additionally provide the scaling curves for the DICE metric. In Figure 4 (right), we see that RAD-DINO significantly outperforms MI2 when MI2 is pretrained using report-only supervision for all pretraining dataset sizes but 30k and 100k, which aligns with expectations for segmentation tasks (Bolya et al., 2025). In general, scaling trends are inconsistent below 100k pretraining samples, likely due to the small size of the segmentation benchmark. However, incorporating tube presence labels during MI2 pretraining (via UniCL) leads to significant performance gains, surpassing the RAD-DINO open-weights model for all but one pretraining dataset size and closing the gap to the continually pretrained RAD-DINO across almost all pretraining dataset sizes. This suggests that the added labels provide valuable spatial context during contrastive learning. At the full 3.5M scale RAD-DINO significantly outperforms all other models. Extrapolating the performance trend of RAD-DINO pretrained on 100k, 1M, and 3.5M samples from INST-CXR-BENCH, we conclude that DINOv2 scales more effectively than CLIP/UniCL for l&t segmentation tasks, which is in agreement with (Fan et al., 2025). Last, in Table 5, we compare models trained on the full INST-CXR-BENCH dataset (3.5M samples) with open-weights models. RAD-DINO pretrained on 3.5M CXRs from INST-CXR-BENCH significantly outperforms all other models across all l&t types.

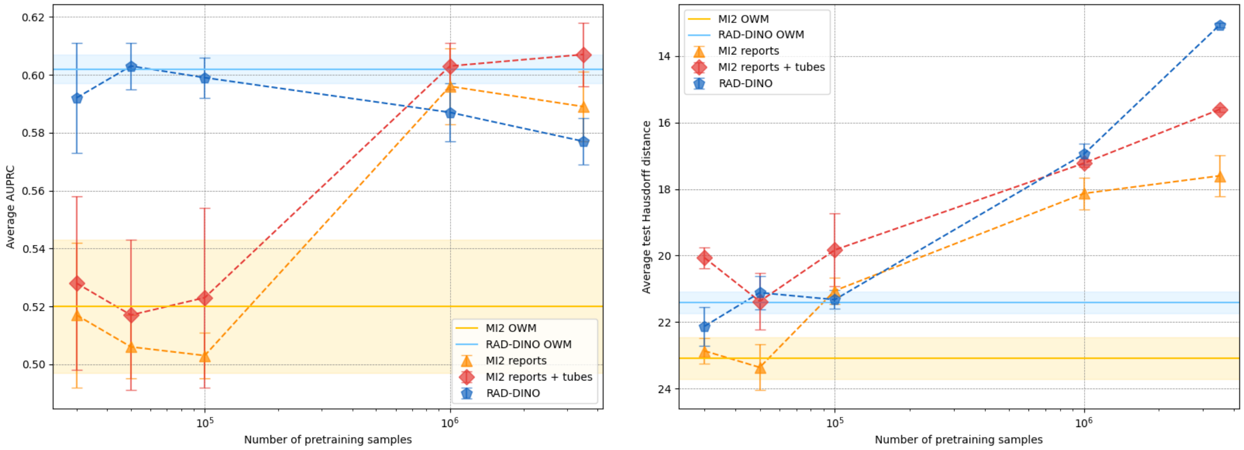


Figure 4: Performance of encoders as a function of vision encoder pretraining with increasing sample sizes from INST-CXR-BENCH. Left: Findings classification on holdout dataset VinDR, AUPRC averaged across seven findings. Right: Lines and tubes segmentation on holdout dataset RANZCR-CLiP, Hausdorff distance averaged across four l&t.

5.3 Report generation on INST-CXR-BENCH-REPORT-GEN

We compare three vision encoder pretraining approaches: RAD-DINO, MI2 reports, and MI2 reports + tube labels for report generation using the MAIRA-2 13B framework (Hyland et al., 2024; Bannur et al., 2024). Training and evaluation use 2.5M studies from INST-CXR-BENCH (subset: INST-CXR-BENCH-REPORT-GEN). Inputs include current frontal and lateral views, prior frontal view, clinical indication, comparison sections, and the full prior report. While encoder pretraining size varies, the report generation training set remains fixed. Unlike earlier experiments (Section 5.1, 5.2), we train one MAIRA-2 model per pretraining size and method due to the high cost of training a 13B LLM. To still measure experiment variability, the results in Figure 5 and Table 6 report medians and 95% CIs from 500 bootstrap samples. The performance of MAIRA-2 is assessed using four metrics for natural language generation (NLG) and clinical efficacy (CE): ROUGE-L (NLG) (Lin & Och, 2004), CheXbert Macro F1-14 (CE) (Smit et al., 2020), RadFact Logical F1 (CE) (Bannur et al., 2024), and a novel CE metric, called Incorrect Placement F1, measuring detection of misplaced lines/tubes, a clinically critical task requiring curve-continuity features (Section 3.2). Due to compute-heavy metrics (especially CheXbert and RadFact), inference uses a 40k-study subset, each study retaining one frontal, one lateral, and one prior frontal image per case. The resulting test dataset is designed to reflect the real-world distribution of l&t encountered in an ICU setting. Given the high cost of MAIRA-based generation, we ran ablations to find the best layer(s) to extract features. We compared the four-layer combination from Section 5.1 with final-layer features and found the latter performed best for both RAD-DINO and MI2. Notably, MI2's last layer uses only 289 image tokens versus 1369 for RAD-DINO, reducing training time by 75% and inference time by 33%. Across our experiments, the largest performance differences among pretraining methods appear in CheXbert Macro F1-14 and Incorrect Placement F1 (Figure 5). Consistent with Section 5.1, MI2 significantly outperforms RAD-DINO on CheXbert Macro F1-14, which strongly correlates with findings classification. Notably, performance curves for all three strategies show parallel trends, suggesting a shared scaling exponent k but different scaling multipliers α. As expected, the two MI2 variants (with and without tube-label supervision) show no significant difference across scales. Compared to open-weight models, both RAD-DINO and MI2 only surpass baseline performance when trained on 1M or more examples. For Incorrect Placement F1, RAD-DINO significantly outperforms MI2 trained solely with report supervision, consistent with Section 5.2. We attribute this to radiology reports often lacking detailed descriptions of l&t placements, including tip positions. However, adding explicit tube presence labels to MI2 pretraining narrows the gap. Unlike the near-linear scaling seen with CheXbert Macro F1-14, Incorrect Placement F1 saturates quickly, likely due to the low (∼12%) prevalence of incorrectly placed tubes in INST-CXR-BENCH-REPORT-GEN. For this task, RAD-DINO exceeds open-weights models with just 100k samples, whereas MI2 requires ≥1M to consistently outperform open weights. In Table 6, we compare models trained on all INST-CXR-BENCH data (3.5M samples) to open weights. For ROUGE-L, CheXbert Macro F1-14, and Incorrect Placement F1, models pretrained on INST-CXR-BENCH significantly outperform their open-weight counterparts. For RadFact, both RAD-DINO versions are on par. For ROUGE-L, CheXbert Macro F1-14, and RadFact, MI2 outperforms RAD-DINO, likely because these metrics correlate with findings classification. Only for Incorrect Placement F1 do RAD-DINO models, both open-weight and continually pretrained, outperform MI2.

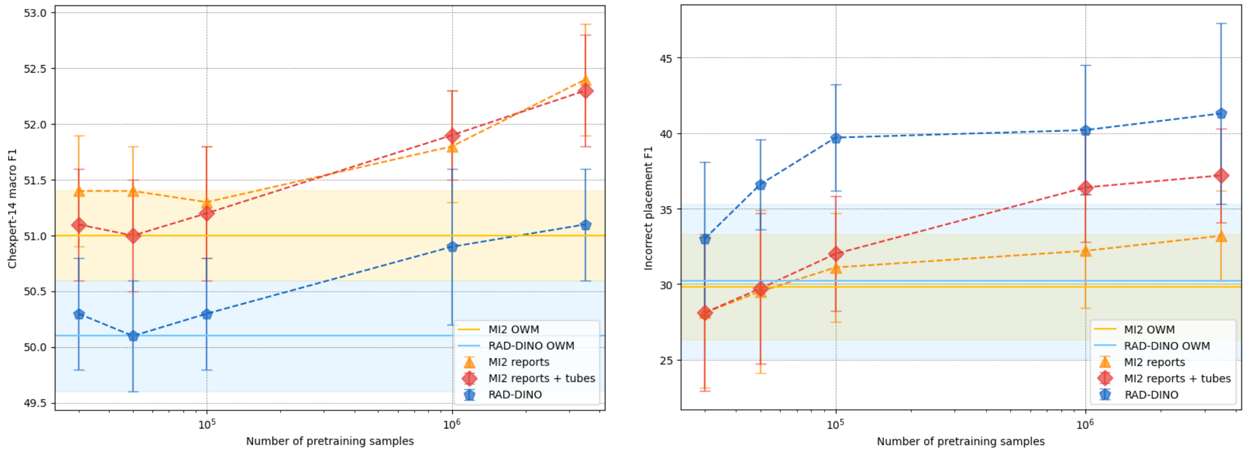


Figure 5: Report generation performance on INST-CXR-BENCH-REPORT-GEN as a function of vision encoder pretraining with increasing sample sizes INST-CXR-BENCH. Left: CheXbert Macro F1 averaged across 14 findings. Right: Incorrect Placement F1.

6 Conclusion

Our study demonstrates that continual pretraining of open-weight models on large-scale CXR datasets yields significantly improved vision encoders. This highlights the promise of developing center-specific foundation models, allowing large medical institutions to tailor encoders to their unique patient populations and imaging protocols. We establish clear scaling laws up to 3.5M samples, indicating that existing foundation models such as MedImageInsight and RAD-DINO continue to benefit from additional data. Notably, despite using the same pretraining data and compute budget, these models exhibit complementary strengths: MedImageInsight (CLIP-style) excels at findings-related tasks, whereas RAD-DINO (DINOv2-style) performs better on tube-related tasks. Moreover, incorporating tube presence labels into MedImageInsight pretraining via UniCL closes the performance gap with RAD-DINO, underscoring the importance of structured supervision, even at scale. This demonstrates the value of structured labels extracted by LLMs such as GPT; see Appendix A.7 for a detailed discussion. In our experiments, however, many scaling curves deviate from idealized power-law behavior. In the small-data regime, performance is often noisy, while in the large-data regime, improvements can plateau. Domain shift (e.g., training on data from different hospitals) further complicates these trends. The results in Section 5.1.3 especially highlight the need for larger, multi-center benchmark datasets to effectively compare CXR vision encoders. Overall, our findings suggest that continual pretraining of MI2 using the UniCL framework, combined with automated label extraction, is the most effective strategy for medical centers aiming to train foundation vision encoders on their own data. We also emphasize the importance of a large and diverse test dataset, diverse both in tasks and metadata, to thoroughly evaluate the performance of pretrained vision encoders. Finally, the scaling curves in Figure 2 and Figure 3 indicate that improving the average performance of an vision encoder may require billions of training samples. However, average performance can be misleading, as it aggregates tasks that are nearly saturated with those that could benefit significantly from additional data. This underscores that a brute-force approach of simply collecting more data is not an effective path forward. Instead, efforts should focus on identifying and prioritizing underrepresented or low-performing tasks, potentially through data selection strategies, like active learning (Ren et al., 2021) and data filtering (Vo et al., 2024; Mindermann et al., 2022).

References

Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting Neural Scaling Laws in Language and Vision, November 2022. URL http://arxiv.org/abs/2209.06640. arXiv:2209.06640 [cs].

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. MAIRA-2: Grounded Radiology Report Generation, September 2024. URL http://arxiv.org/abs/2406.04449. arXiv:2406.04449 [cs].

Bannur, Shruthi, Hyland, Stephanie, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C. Castro, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to Exploit Temporal Structure for Biomedical Vision–Language Processing. pp. 15016–15027, 2023.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Maxim Krikun, Colin Cherry, Behnam Neyshabur, and Orhan Firat. Data Scaling Laws in NMT: The Effect of Noise and Architecture, February 2022. URL http://arxiv.org/abs/2202.01994. arXiv:2202.01994 [cs].

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception Encoder: The best visual embeddings are not at the output of the network, April 2025. URL http://arxiv.org/abs/2504.13181. arXiv:2504.13181 [cs].

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken Neural Scaling Laws, July 2023. URL http://arxiv.org/abs/2210.14891. arXiv:2210.14891 [cs].

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. URL http://arxiv.org/abs/2104.14294. arXiv:2104.14294 [cs].

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. pp. 2818–2829, June 2023. doi: 10.1109/CVPR52729.2023.00276. URL http://arxiv.org/abs/2212.07143. arXiv:2212.07143 [cs].

Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?, January 2016. URL http://arxiv.org/abs/1511.06348. arXiv:1511.06348 [cs].

Noel C. F. Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, Hoifung Poon, Stephanie Hyland, Shruthi Bannur, Javier Alvarez-Valle, Xue Li, John Garrett, Alan McMillan, Gaurav Rajguru, Madhu Maddi, Nilesh Vijayrania, Rehaan Bhimai, Nick Mecklenburg, Rupal Jain, Daniel Holstein, Naveen Gaur, Vijay Aski, Jenq-Neng Hwang, Thomas Lin, Ivan Tarapov, Matthew Lungren, and Mu Wei. MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging, October 2024. URL http://arxiv.org/abs/2410.06542. arXiv:2410.06542 [eess].

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. DaViT: Dual Attention Vision Transformers, April 2022. URL http://arxiv.org/abs/2204.03645. arXiv:2204.03645 [cs].

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, and Saining Xie. Scaling Language-Free Visual Representation Learning, April 2025. URL http://arxiv.org/abs/2504.01017. arXiv:2504.01017 [cs].

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://arxiv.org/abs/2004.07780. arXiv:2004.07780 [cs].

SC Huang, L Shen, MP Lungren, and S Yeung. GLoRIA: a multimodal global–local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*, pp. 3942–3951, 2021.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. MAIRA-1: A specialised large multimodal model for radiology report generation, April 2024. URL http://arxiv.org/abs/2311.13668. arXiv:2311.13668 [cs].

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From CLIP to DINO: Visual Encoders Shout in Multi-modal Large Language Models, March 2024. URL http://arxiv.org/abs/2310.08825. arXiv:2310.08825 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs].

Chin-Yew Lin and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 605–612, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1219032. URL https://aclanthology.org/P04-1077/.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-CLIP: Contrastive Language-Image Pre-training using Biomedical Documents, March 2023. URL http://arxiv.org/abs/2303.07240. arXiv:2303.07240 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, December 2023. URL http://arxiv.org/abs/2304.08485. arXiv:2304.08485 [cs].

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling Laws Are Unreliable for Downstream Tasks: A Reality Check, July 2025. URL http://arxiv.org/abs/2507.00885. arXiv:2507.00885 [cs].

Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized Training on Points that are Learnable, Worth Learning, and Not Yet Learnt, June 2022. URL http://arxiv.org/abs/2206.07137. arXiv:2206.07137 [cs] version: 1.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-Flamingo: a Multimodal Medical Few-shot Learner, July 2023. URL http://arxiv.org/abs/2307.15189. arXiv:2307.15189 [cs].

Théo Moutakanni, Piotr Bojanowski, Guillaume Chassagnon, Céline Hudelot, Armand Joulin, Yann LeCun, Matthew Muckley, Maxime Oquab, Marie-Pierre Revel, and Maria Vakalopoulou. Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning, May 2024. URL http://arxiv.org/abs/2405.01469. arXiv:2405.01469 [cs].

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations, March 2022. URL http://arxiv.org/abs/2012.15029. arXiv:2012.15029 [eess].

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs].

OpenAI. GPT-4o System Card, October 2024. URL http://arxiv.org/abs/2410.21276. arXiv:2410.21276 [cs].

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

Fernando Pérez-García, Sam Bond-Taylor, Pedro P. Sanchez, Boris van Breugel, Daniel C. Castro, Harshita Sharma, Valentina Salvatelli, Maria T. A. Wetscherek, Hannah Richardson, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, Ozan Oktay, and Maximilian Ilse. RadEdit: stress-testing biomedical vision models via diffusion image editing. 15070:358–376, 2025a. doi: 10.1007/978-3-031-73254-6_21. URL http://arxiv.org/abs/2312.12865. arXiv:2312.12865 [cs].

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, January 2025b. ISSN 2522-5839. doi: 10.1038/s42256-024-00965-w. URL http://arxiv.org/abs/2401.10815. arXiv:2401.10815 [cs].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A Survey of Deep Active Learning. *ACM Comput. Surv.*, 54(9):180:1–180:40, October 2021. ISSN 0360-0300. doi: 10.1145/3472291. URL https://doi.org/10.1145/3472291.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models, October 2022. URL http://arxiv.org/abs/2210.08402. arXiv:2210.08402 [cs].

Andrew B. Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Florencia Garcia-Vicente, David Melnick, Yun Liu, Krish Eswaran, Daniel Tse, Neeral Beladia, Dilip Krishnan, and Shravya Shetty. Simplified Transfer Learning for Chest Radiography Models Using Less Data. *Radiology*, 305(2):454–465, November 2022. ISSN 1527-1315. doi: 10.1148/radiol.212482.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT, October 2020. URL http://arxiv.org/abs/2004.09167. arXiv:2004.09167 [cs].

Jennifer S. N. Tang, Jarrel C. Y. Seah, Adil Zia, Jay Gajera, Richard N. Schlegel, Aaron J. N. Wong, Dayu Gai, Shu Su, Tony Bose, Marcus L. Kok, Alex Jarema, George N. Harisis, Chris-Tin Cheng, Helen Kavnoudias, Wayland Wang, Anouk Stein, George Shih, Frank Gaillard, Andrew Dixon, and Meng Law. CLiP, catheter and line position dataset. *Scientific Data*, 8(1):285, October 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-01066-8. URL https://www.nature.com/articles/s41597-021-01066-8.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6 (12):1399–1406, September 2022. ISSN 2157-846X. doi: 10.1038/s41551-022-00936-9. URL https://www.nature.com/articles/s41551-022-00936-9.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, April 2024. URL http://arxiv.org/abs/2401.06209. arXiv:2401.06209 [cs].

Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Automatic Data Curation for Self-Supervised Learning: A Clustering-Based Approach, June 2024. URL http://arxiv.org/abs/2405.15613. arXiv:2405.15613 [cs].

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, Jungyeon Park, Patricia Strachan, Yun Liu, Chuck Lau, Preeti Singh, Christina Chen, Mozziyar Etemadi, Sreenivasa Raju Kalidindi, Yossi Matias, Katherine Chou, Greg S. Corrado, Shravya Shetty, Daniel Tse, Shruthi Prabhakara, Daniel Golden, Rory Pilgrim, Krish Eswaran, and Andrew Sellergren. ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders, September 2023. URL http://arxiv.org/abs/2308.01317. arXiv:2308.01317 [cs].

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified Contrastive Learning in Image-Text-Label Space, April 2022. URL http://arxiv.org/abs/2204.03610. arXiv:2204.03610 [cs].

Luca Zedda, Andrea Loddo, and Cecilia Di Ruberto. Radio DINO: A foundation model for advanced radiomics and AI-driven medical imaging analysis. *Computers in Biology and Medicine*, 195:110583, September 2025. ISSN 00104825. doi: 10.1016/j.compbiomed.2025.110583. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482525009345.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling Vision Transformers, June 2022. URL http://arxiv.org/abs/2106.04560. arXiv:2106.04560 [cs].

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, January 2025. URL http://arxiv.org/abs/2303.00915. arXiv:2303.00915 [cs].

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text, September 2022. URL http://arxiv.org/abs/2010.00747. arXiv:2010.00747 [cs].

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing Radiograph Representation Learning with Masked Record Modeling, February 2023. URL http://arxiv.org/abs/2301.13155. arXiv:2301.13155 [cs].

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer, January 2022. URL http://arxiv.org/abs/2111.07832. arXiv:2111.07832 [cs].

A Appendix

A.1 Comparison of MI2 and RAD-DINO

Table 1: Comparison of MedImageInsight (MI2) and RAD-DINO

	MedImageInsight (MI2)	**RAD-DINO**
Architecture / #parameters	DAViT / 360M	ViT-B / 87M
Training method	UniCL	Dinov2
Training data	CXRs (500k) + other modalities (3.3M)	CXRs (800k)
# of tokens (518×518 input image size)	Block0: 130×130 = 16,900 Block1: 65×65 = 4,225 Block2: 33×33 = 1,098 Block3: 17×17 = 289	Everywhere: 37×37 = 1,369
Token dimension	Block0: 256 Block1: 512 Block2: 1024 Block3: 2048	Everywhere: 768

A.2 Findings classification on INST-CXR-BENCH-FIND-CLASS

In Table 2, we compare models trained with all of INST-CXR-BENCH's data (3.5M samples) vs the open-weights models.

Table 2: Findings classification on the INST-CXR-BENCH-FIND-CLASS subset. Comparison of the open-weights models (OWM) and encoders pretrained with the full INST-CXR-BENCH dataset (3.5M samples).

Finding	HI	ILD	ATL	CAB	PE	PTX	AD	EPA	AC	OA	RF	BWT	HRN	SA	OP	VC	CM	DE	PDE	NF	AVG
Prevalence	1.57%	1.97%	14.34%	1.03%	10.39%	2.61%	0.35%	0.61%	7.60%	8.83%	2.28%	0.60%	0.81%	1.18%	12.85%	2.71%	10.23%	2.10%	2.64%	7.30%	
MI2 OWM	$35.8_{0.5}$	$21.2_{0.3}$	$73.6_{0.1}$	$16.2_{0.1}$	$84.4_{0.1}$	$76.8_{0.1}$	$16.3_{0.3}$	$28.3_{0.4}$	$58.4_{0.3}$	$57.0_{0.2}$	$45.1_{0.1}$	$10.3_{0.1}$	$65.1_{0.3}$	$78.4_{0.1}$	$75.9_{0.0}$	$26.1_{0.1}$	$76.3_{0.1}$	$42.3_{0.2}$	$36.5_{0.2}$	$81.3_{0.1}$	$50.3_{0.1}$
MI2 reports	$\mathbf{37.6_{0.2}}$	$\mathbf{23.6_{0.4}}$	$\mathbf{74.9_{0.0}}$	$\mathbf{18.3_{0.3}}$	$\mathbf{85.4_{0.0}}$	$\mathbf{81.6_{0.1}}$	$\mathbf{20.2_{0.1}}$	$\mathbf{32.1_{0.9}}$	$\mathbf{61.4_{0.0}}$	$\mathbf{59.7_{0.0}}$	$\mathbf{51.8_{0.3}}$	$12.6_{0.3}$	$68.6_{0.5}$	$\mathbf{80.5_{0.1}}$	$\mathbf{77.2_{0.0}}$	$\mathbf{27.5_{0.5}}$	$\mathbf{77.6_{0.0}}$	$\mathbf{45.0_{0.2}}$	$\mathbf{38.0_{0.2}}$	$\mathbf{82.9_{0.0}}$	$\mathbf{52.8_{0.1}}$
MI2 reports + tubes	$\mathbf{37.3_{0.2}}$	$23.3_{0.3}$	$\mathbf{74.9_{0.0}}$	$\mathbf{18.0_{0.2}}$	$85.3_{0.0}$	$\mathbf{81.6_{0.2}}$	$18.9_{0.5}$	$\mathbf{32.5_{0.7}}$	$\mathbf{61.3_{0.2}}$	$\mathbf{59.6_{0.1}}$	$51.4_{0.3}$	$\mathbf{13.0_{0.2}}$	$\mathbf{69.3_{0.3}}$	$\mathbf{80.8_{0.0}}$	$77.1_{0.0}$	$\mathbf{27.5_{0.3}}$	$77.4_{0.2}$	$\mathbf{45.0_{0.3}}$	$\mathbf{37.7_{0.1}}$	$\mathbf{82.8_{0.1}}$	$\mathbf{52.7_{0.1}}$
RAD-DINO OWM	$35.5_{0.4}$	$21.1_{0.2}$	$73.0_{0.2}$	$15.1_{0.3}$	$84.3_{0.2}$	$74.9_{0.3}$	$14.8_{1.1}$	$28.1_{1.2}$	$57.0_{0.1}$	$56.5_{0.1}$	$42.8_{0.8}$	$9.6_{0.3}$	$65.7_{0.2}$	$77.9_{0.1}$	$75.2_{0.1}$	$25.1_{0.2}$	$76.6_{0.1}$	$41.7_{0.4}$	$36.4_{0.2}$	$80.6_{0.0}$	$49.6_{0.1}$
RAD-DINO	$36.0_{0.3}$	$22.0_{0.2}$	$73.4_{0.1}$	$15.7_{0.4}$	$84.6_{0.1}$	$77.8_{0.2}$	$14.1_{0.2}$	$30.2_{0.6}$	$58.4_{0.2}$	$57.2_{0.2}$	$46.1_{0.6}$	$9.7_{0.2}$	$68.3_{0.2}$	$79.5_{0.3}$	$75.8_{0.0}$	$25.5_{0.4}$	$76.7_{0.1}$	$40.1_{0.4}$	$36.6_{0.3}$	$81.3_{0.1}$	$50.4_{0.1}$

HI: Hyperinflation, ILD: Interstitial Lung Disease Pattern, ATL: Atelectasis, CAB: Costophrenic Angle Blunting, PE: Pleural Effusion, PTX: Pneumothorax, AD: Adenopathy, EPA: Enlarged Pulmonary Artery, AC: Arterial Calcification, OA: Osseous Abnormalities, RF: Rib Fracture, BWT: Bronchial Wall Thickening, HRN: Hernia, SA: Subcutaneous Air/Emphysema, OP: Opacity, VC: Vascular Congestion, CM: Cardiomegaly, DE: Diaphragm Elevation, PDE: Pulmonary Edema, NF: No Finding, AVG: Mean of micro averaged AUPRC across 3 seeds.

A.3 Tube presence classification on INST-CXR-BENCH-TUBE-CLASS

In Table 3 we compare models trained with all of INST-CXR-BENCH (3.5M samples) vs the open-weights models.

Table 3: Tube presence classification on the INST-CXR-BENCH-TUBE-CLASS subset. Comparison of the open-weights models and encoders pretrained with the full INST-CXR-BENCH dataset (3.5M samples).

Tube category	ETT	TT	NGT	SGC	CT	MD	IABP	IJ	PICC	SC	NT	AVG
Prevalence	11.39%	3.23%	12.73%	5.17%	14.03%	5.56%	0.56%	12.38%	13.49%	5.49%	14.04%	
MI2 OWM	$95.6_{0.1}$	$94.8_{0.1}$	$92.4_{0.0}$	$94.8_{0.2}$	$96.4_{0.1}$	$83.1_{0.2}$	$81.5_{0.4}$	$81.8_{0.2}$	$95.9_{0.1}$	$81.7_{0.7}$	$97.6_{0.0}$	$90.5_{0.0}$
MI2 reports	$96.2_{0.1}$	$\mathbf{95.5_{0.0}}$	$\mathbf{93.3_{0.1}}$	$\mathbf{95.5_{0.1}}$	$97.2_{0.0}$	$85.9_{0.2}$	$\mathbf{83.6_{0.5}}$	$83.4_{0.2}$	$\mathbf{96.7_{0.0}}$	$85.6_{0.2}$	$98.0_{0.0}$	$\mathbf{91.9_{0.1}}$
MI2 reports + tubes	$\mathbf{96.3}_{0.1}$	$\mathbf{95.7}_{0.1}$	$\mathbf{93.4}_{0.0}$	$\mathbf{95.7}_{0.1}$	$\mathbf{97.5}_{0.0}$	$\mathbf{86.7}_{0.3}$	$\mathbf{84.5}_{0.3}$	$\mathbf{84.3}_{0.1}$	$\mathbf{96.9}_{0.0}$	$\mathbf{87.0}_{0.1}$	$\mathbf{98.1}_{0.0}$	$\mathbf{92.4}_{0.0}$
RAD-DINO OWM	$95.1_{0.0}$	$94.4_{0.2}$	$92.2_{0.0}$	$94.6_{0.1}$	$96.1_{0.0}$	$82.8_{0.1}$	$82.3_{0.4}$	$82.0_{0.1}$	$96.1_{0.0}$	$83.5_{0.3}$	$97.6_{0.0}$	$90.6_{0.0}$
RAD-DINO	$95.9_{0.0}$	$\mathbf{95.2_{0.1}}$	$\mathbf{93.3_{0.1}}$	$\mathbf{95.3_{0.1}}$	$97.0_{0.0}$	$86.2_{0.3}$	$\mathbf{85.6_{0.5}}$	$83.8_{0.1}$	$\mathbf{96.9_{0.0}}$	$85.9_{0.2}$	$98.0_{0.0}$	$\mathbf{92.1_{0.1}}$

ETT: Endotracheal Tube, TT: Tracheostomy Tube, NGT: Nasogastric Tube, SGC: Swan-Ganz Catheter, CT: Chest Tube, MD: Mediastinal Drain, IABP: Intra-Aortic Balloon Pump, IJ: Internal Jugular CVC, PICC: Peripherally Inserted Central Catheter, SC: Subclavian CVC / Port-a-Cath, NT: No Tubes, AVG: Mean of micro averaged AUPRC across 3 seeds.

A.4 Findings classification on holdout dataset VinDR

In Table 4, we compare models trained on the full INST-CXR-BENCH (3.5M samples) to open-weights models.

Table 4: Findings classification on holdout dataset VinDR: Classification accuracy of open-weights encoder checkpoints and encoders pretrained with the full INST-CXR-BENCH (3.5M samples)

Finding	AE	CM	LO	PE	PL-T	PF	TB	AVG
Prevalence	28.52%	23.63%	7.01%	8.28%	11.68%	13.71%	7.18%	
MI2 OWM	$36.3_{7.0}$	$66.0_{7.3}$	$14.1_{0.3}$	$\mathbf{79.1_{2.3}}$	$40.2_{2.6}$	$57.4_{4.5}$	$71.3_{5.5}$	$52.0_{2.3}$
MI2 reports	$\mathbf{49.0_{3.6}}$	$78.6_{1.4}$	$\mathbf{16.0_{1.7}}$	$75.8_{2.6}$	$47.7_{2.8}$	$\mathbf{65.6_{1.0}}$	$\mathbf{79.8_{0.8}}$	$\mathbf{58.9_{1.2}}$
MI2 reports + tubes	$\mathbf{50.6_{1.9}}$	$\mathbf{81.4_{1.0}}$	$\mathbf{17.1_{0.9}}$	$77.0_{4.7}$	$\mathbf{52.6_{0.7}}$	$\mathbf{65.9_{0.7}}$	$\mathbf{80.4_{0.5}}$	$\mathbf{60.7_{1.1}}$
RAD-DINO OWM	$\mathbf{48.1_{0.7}}$	$\mathbf{83.2_{0.4}}$	$\mathbf{17.5_{0.9}}$	$\mathbf{83.5_{0.1}}$	$47.3_{1.1}$	$\mathbf{65.2_{0.1}}$	$76.6_{0.9}$	$\mathbf{60.2_{0.5}}$
RAD-DINO	$43.3_{5.2}$	$77.9_{1.4}$	$\mathbf{17.8_{0.5}}$	$\mathbf{80.5_{0.6}}$	$46.3_{1.3}$	$\mathbf{63.5_{0.8}}$	$74.7_{0.7}$	$\mathbf{57.7_{0.8}}$

AE: Aortic Enlargement, CM: Cardiomegaly, LO: Lung Opacity, PE: Pleural Effusion, PL-T: Pleural Thickening, PF: Pulmonary Fibrosis, TB: Tuberculosis, AVG: Mean of micro averaged AUPRC across 3 seeds.

A.5 Lines and tubes segmentation on holdout dataset RANZCR-CLiP

In Table 5, we compare models trained on the full INST-CXR-BENCH dataset (3.5M samples) with open-weights models.

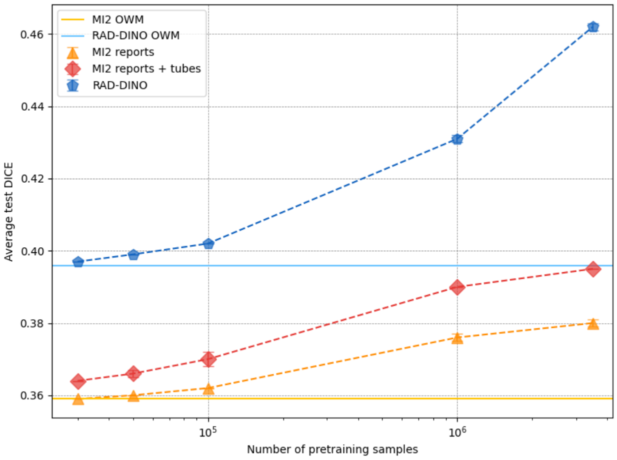


Figure 6: Lines and tubes segmentation performance on RANZCR-CLiP as a function of vision encoder pretraining with increasing sample sizes from INST-CXR-BENCH. DICE averaged across four l&t.

Table 5: Lines and tubes segmentation on holdout dataset RANZCR-CLiP: Hausdorff distance for segmentation with different open-weights models and encoders pretrained with the full INST-CXR-BENCH dataset (3.5M samples).

Tube category	SGC	CVC	ETT	NGT	AVG
Prevalence	0.91%	51.54%	17.43%	18.49%	
MI2 OWM	$39.6_{0.7}$	$73.5_{1.4}$	$23.1_{0.6}$	$80.0_{1.7}$	$23.1_{0.6}$
MI2 reports	$39.4_{0.2}$	$48.3_{0.6}$	$17.6_{0.6}$	$58.9_{0.8}$	$17.6_{0.6}$
MI2 reports + tubes	$40.5_{0.3}$	$40.0_{0.3}$	$15.6_{0.1}$	$49.8_{0.9}$	$15.6_{0.1}$
RAD-DINO OWM	$39.0_{0.2}$	$35.3_{0.2}$	$21.4_{0.3}$	$38.9_{0.6}$	$21.4_{0.3}$
RAD-DINO	**38.2**$_{0.1}$	**26.9**$_{0.3}$	**13.1**$_{0.1}$	**24.8**$_{0.2}$	**13.1**$_{0.1}$

SGC: Swan Ganz Catheter, CVC: Central Venous Catheter, ETT: Endotracheal Tube, NGT: Nasogastric Tube, AVG: Mean of micro averaged Hausdorff Distance across 3 seeds.

A.6 REPORT GENERATION ON INST-CXR-BENCH-REPORT-GEN

In Table 6, we compare models trained on all INST-CXR-BENCH data (3.5M samples) to open weights.

Table 6: Findings generation on INST-CXR-BENCH-REPORT-GEN dataset. Comparison of the open-weights models and encoders pretrained with the full INST-CXR-BENCH dataset (3.5M samples).

	ROUGE-L	CheXbert Macro F1-14	Incorrect Placement F1	RadFact: logical F1
RAD-DINO OWM	38.5 [38.2, 38.7]	50.1 [49.6, 50.6]	30.2 [26.0, 35.3]	63.3 [63.1, 63.6]
RAD-DINO	39.0 [38.8, 39.2]	51.1 [50.6, 51.6]	**41.3 [36.1, 47.3]**	63.0 [62.8, 63.2]
MI2 OWM	38.7 [38.5, 38.9]	51.0 [50.5, 51.4]	29.8 [26.0, 33.3]	63.8 [63.6, 64.0]
MI2 reports	**39.5 [39.3, 39.7]**	**52.4 [51.9, 52.9]**	33.2 [30.2, 36.3]	**64.7 [64.5, 64.9]**
MI2 reports + tubes	**39.5 [39.3, 39.8]**	**52.3 [51.8, 52.8]**	37.2 [33.7, 40.3]	**64.8 [64.6, 65.0]**

A.7 LABEL EXTRACTION

We implemented an LLM-based pipeline to extract findings labels as well as l&t labels from CXR radiology reports. This involved the use of detailed prompts that were engineered with the help of radiologists and iteratively refined for accurate label extraction. We used GPT-4o (OpenAI, 2024) endpoints for extracting the findings and l&t labels. This process was designed to support structured data generation for downstream clinical applications, with an emphasis on consistency, reliability, and alignment with radiologist expectations. We extracted the presence or absence of 19 findings: hyperinflation, interstitial lung disease pattern, atelectasis, costophrenic angle blunting, pleural effusion, pneumothorax, adenopathy, enlarged pulmonary artery, arterial calcification, osseous abnormalities, rib fracture, bronchial wall thickening, hernia, subcutaneous air/emphysema, opacity, vascular congestion, cardiomegaly, diaphragm elevation, pulmonary edema. The findings were chosen to cover a wide range of appearances (some are more diffuse/texture like, others are more localized/shape like) and areas of a CXR (from the esophagus to the diaphragm), and have a good support (at least 10k in the test set) in the dataset. There are ten l&t types whose presence or absence we extracted: Internal Jugular Central Venous Catheter (CVC), Peripherally Inserted Central Catheter, Subclavian CVC or Port-A-Cath, Endotracheal Tube, Tracheostomy Tube, Nasogastric Tube, Swan-Ganz Catheter, Chest Tube, Mediastinal Drain, and Intra-Aortic Balloon Pump. These are the most common l&t devices observed in CXR reporting in practice, and radiologists considered them important for analysis. We evaluated the accuracy of the GPT-based l&t labels extraction using a manually annotated hold-out dataset of 115 samples and achieved an F1-score of 0.94. This manual evaluation provided key insights into error modes and prompted refinements to both the input formatting and the GPT prompt design. What follows is an example prompt for extracting chest tube labels:

```
You are an AI radiology assistant.  You are helping to process reports for
Chest X-rays by extracting information about lines and tubes visible in the
image, by looking at the reports.  In radiology reports, "left" corresponds to
the left side of the patient, which is the right side of the X-ray; similarly
"right" corresponds to the right side of the patient, which is the left side
of the X-ray; use the same terminology.

You will be given the report for the current study (marked by "Current Study")
which describes the findings from the chest X-ray(s) taken at the that time.
Each report will have the date of the report, the reason for exam, and the
impression, which contains the radiologist’s observations.

The goal is to use the reports to extract information about lines and tubes
which can be seen in the current X-ray.  Look at current report for the
specified line/tube and its side.  Check if the specified line/tube is
mentioned.  Check if the current report states if the line/tube is correctly
placed or indicates any malpositioning (for instance, doubled up, looped,
kinked, coiled), and should be repositioned or retracted.   Only extract lines
and tubes mentioned in the current report.  Only describe changes which are
described in the current report.

Extract information in JSON format as a list of each line/tube visible in the
current X-ray image.  Each line/tube should have a single entry.  There can be
multiple types of lines/tubes in the report, as well as multiple instances of
the same type or even the same subtype; in all cases, ensure that each one has
a separate entry in the JSON list.  If there are no lines/tubes then output an
empty list.

# JSON entry fields

- reference_sentence (this should contain the original sentence, sub-sentence,
or multiple sentences from the report describing all details about the
line/tube) - type:  the line/tube type exactly as written in the report -
type_categorical:  the line/tube type formatted to fall into one of a fixed
number of categories that will be defined later.  - placement:  if described
in the report, whether the line/tube is correctly placed or incorrectly
placed (correct or incorrect).  If it is not explicitly described, use the
tip location to infer the placement, that will be defined later.  If it is
described but it’s unclear what category it falls into, write \unclear".
Otherwise N/A.

# Lines and tubes to extract

In this pass, only extract information about chest tubes.  Chest tubes are
inserted through the chest wall into the pleural space and are used to
```

drain fluid, blood, or air. There are other ways to describe a chest tube including chest drain, pleural drain, pleural catheter, pigtail pleural drain, pigtail catheter, drainage catheter, drainage tube, thoracostomy tube, PleurX catheter, etc. Different terms may be used in different reports; use in such cases, if there are multiple chest tubes and it is ambiguous which one corresponds to which in previous reports, use information about insertion side and tip location to determine which are which. If chest tubes are described as bilateral or bibasilar etc., means that more than one chest tubes are present in both sides of the chest i.e. there is one on each side of the body, then output two entries, one for side_categorical left, and the other for side_categorical right.

Additional information Do not confuse chest tubes with mediastinal drains and pericardial drains, which are inserted in the mediastinum rather than the pleural space. Also do not confuse chest tubes with any other kinds of tubes such as feeding tubes, tracheostomy tubes, or endotracheal tubes.

It is common for there to be multiple chest tubes in place at one time. Remember that each each individual chest tube must have a separate entry in the output list.

Placement Information For the placement field: Write "incorrect" if that line/tube is described as misplaced or malpositioned (e.g. kinked, coiled, doubled up) and/or should be repositioned or withdrawn. Write "incorrect" if the report mentions a pleural effusion or pneumothorax on the same side as the chest tube that is at least moderate in size or larger/worsened than before. Write "incorrect" if the tube or side port is outside of the chest cavity. Write "correct" if the current report describes a "stable position" of that line/tube or that line/tube being "in place". If correct/incorrect placement is not explicitly described in the report, use the following mapping from the extracted tip location: 'upper': 'correct', 'lower': 'correct', 'middle': 'correct', 'below diaphragm': 'incorrect', 'side port outside rib cage': 'incorrect', 'outside chest': 'incorrect', 'adjacent to mediastinum/esp aorta': 'incorrect', 'unclear': 'unclear', 'N/A': 'N/A' If tip location is described but placement can't be inferred from the above mapping, write "unclear". Write "N/A" if there is no tip location or placement information about that line/tube in the report. Write "N/A" if the current report describes that line/tube as having been removed.

What follows is an example prompt for extracting findings from CXR reports:

You are an AI radiology assistant. You are helping process reports from chest X-rays. In radiology reports, \left" corresponds to the left side of the patient, which is the right side of the X-ray; similarly, \right" corresponds to the right side of the patient, which is the left side of the X-ray. Each radiology report contains several sections, such as the findings, impression, comparison, indication, and technique sections.

Please extract information about all the findings and diseases from the radiology report that refer to findings visible in a chest X-ray or disease diagnosed from a chest X-ray, and categorize certain elements. Your task is to extract information about all findings and diseases from the current report and prior structured reports (if available) in JSON format as a list of dictionaries. **If a finding or disease is present in the prior structured report but not in the current report, ensure it is included in the current report output with all of its details from the prior report.** Each unique combination of finding/disease and region should have a single entry, carrying forward all prior information as needed.

Each entry should use the keys given below: "finding_type": The finding type, value should be either DISEASE or FINDING. FINDING represents an observation in the chest x-ray. DISEASE represents the interpretation or diagnosis from the observations in the chest x-ray. Return the finding_type value depending on which list the extracted "label" below belongs to. Return DISEASE if label is in DISEASE list or FINDING if label is in FINDING list.

We will use the word "finding" in the rest of the prompt, to represent a FINDING or a DISEASE.

"reference_phrase": The phrase associated with the finding. Make sure to provide the exact phrase from the report. Don’t change the phrase at all, extract it as it is. If the same finding is present in the current report, update it with the new phrase. Otherwise, retain the phrase from the prior report.

"label" : The finding label mentioned in the phrase. The value must come from the provided list of DISEASE or FINDING. Provide the value "No finding" when a phrase mentions anatomical structures with normal observations. For example: "The cardiomediastinal silhouette is normal", "The imaged upper abdomen is unremarkable", "Lungs are clear", "Pulmonary vasculature is normal", "The cardiomediastinal silhouette is within normal limits", "The cardiac, mediastinal and hilar contours are normal".

DISEASE: $DISEASE

FINDING: $FINDINGS

Please note: 1. ’Pulmonary vascular engorgement’ and ’Vascular engorgement’ are other ways of refering to Pulmonary venous hypertension. 2. ’Mediastinal widening’ and ’Enlarged cardiomedistinum’ are other ways of refering to Enlarged cardiomediastinum. 3. ’Hyperaeration’ and ’Overinflation’ are other ways of refering to Hyperinflation. 4. ’Negative chest, ’chest negative’ and ’no acute disease in the chest’ are other ways of refering to No finding. 5. ’Prosthetic valve’ is another way of refering to Valve prosthesis. 6. ’Enlarged cardiomegaly’ is another way of refering to Cardiomegaly. 7. ’Fibrosis’ is another way of refering to Pulmonary fibrosis. 8. ’Pulmonary opacity’ is another way of refering to lung opacity. 9. Only when ’linear’ when used with ’fibrosis’ i.e. ’linear fibrosis’ is another way of refering to scarring. 10.’Infiltration’ is another way of refering to infiltrate.

Instruction for handling out of list values **You should strictly stick to FINDING and DISEASE labels for "label" category.** **When you find a value of a category which is not from one of the given values for that category (except for "label" category), assign "Other" to it. If the category doesn’t exist in the phrase, assign "N/A" to it.**

Format of each output structured finding: [{ "finding_type" : "", "reference_phrase": "", "label": "", }]

Instructions for Handling All Findings 1. If the same finding is mentioned in multiple sections with different regions, comparison status, is_positive status, severity, anatomy, morphology, spatial distribution or spatial comparison, then extract each instance separately. 2. If the finding is present in multiple phrases, return multiple JSON items for each finding separately. 3. Include normal or negative findings as well. If a finding is negative give the label for that and mark is_positive as "No" for it. e.g: labels present in the phrase : "There is no atelectasis or lung opacity seen." are [’Atelectasis’, ’Lung Opacity’] 4. Match the finding sentences first from the current report with the prior structured report phrases, then create the final structured report.

Only if the input has prior report incorporate the below changes otherwise ignore.

```
**Instructions for Incorporating Prior Findings**:  1.  **Inclusion of
Previous Findings**:  All findings, including negative findings, from the
prior structured report should be included in the current report output,
even if they are not mentioned in the current report.  If no prior structured
report is provided, treat the current report as the first report for the
patient.  2.  **Current Report First**:  In the final structured report, if
there is a prior structured report, give the current report phrases first
and then the prior report phrases.  3.  **Finding Propagation**:  Propagate
all findings with their corresponding values from the prior report unless new
values are provided in the current report.  4.  **Unchanged Findings**:  If
the current report does not specify changes in a finding, it should retain
its values from the prior report in the output.  5.  **Updates to Prior
Findings**:  If the current report updates an existing finding replace the
previous values with the updated values for that finding.

Don’t provide any explanations.
```

A.8 Significance test

We compare two binary classifiers across multiple tasks and multiple random seeds using a hierarchical paired bootstrap procedure to estimate the difference in performance and its uncertainty. For each task, we collect the ground-truth labels and the predicted probabilities from both models across all seeds. We begin performing stratified bootstrap resampling (500 bootstrap samples) of the test set for each task to preserve the original class balance. For every bootstrap replicate, we compute the AUPRC for each seed of both models using the resampled data. These seed-level metrics are then avaraged within each model. The difference between the aggregated metrics of the two models is saved for that replicate. Repeating this process across many bootstrap replicates produces a distribution of differences for each task. From this distribution, we report the average difference and construct a percentile-based confidence interval at the 95% level. To obtain an overall comparison across all tasks, we pool the bootstrap differences from every task into a single distribution (micro AUPRC) and compute the overall mean difference and its confidence interval.

A.9 Metadata stratification

To evaluate potential biases and subgroup performance disparities, we stratify model performance across five metadata variables: ethnicity, sex, age, scanner manufacturer, and patient type (inpatient vs. outpatient). For each variable, we compare the best-performing MI2 and RAD-DINO models (trained with 3.5M INST-CXR-BENCH samples) with their corresponding paper checkpoints. We observe that performance trends are largely consistent across MI2 and RAD-DINO; that is, subgroups where MI2 underperforms tend to also show lower performance for RAD-DINO. For all models we are focusing on the findings classification task from Section 5.1.1.

For the ethnicity metadata variable, we stratify the test set into two categories: White (87%) and Non-White (13%). A performance drop of 3% is observed for the Non-White group. This is expected given the reduced sample size in this group. In addition, we assess performance by patient care setting. We find that models perform 5% worse on outpatient scans, with a notably higher standard deviation across runs. We hypothesize that the drop in performance stems from a greater variability in outpatient imaging protocols and patient conditions.

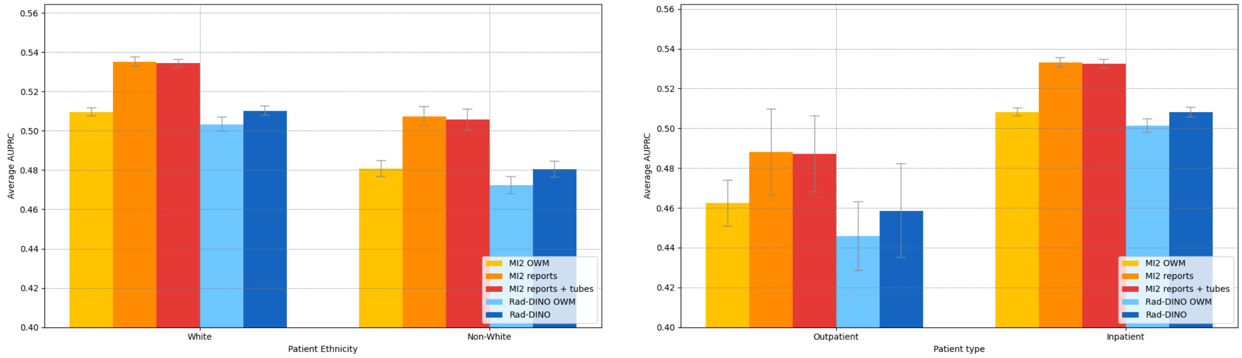


Figure 7: Metadata stratification findings classification on INST-CXR-BENCH. Left: Average AUPRC across 20 findings stratified by ethnicity. Right: Average AUPRC across 20 findings stratified by patient type.

We divide age into six groups: ’20-30’ (7%), ’30-40’ (9%), ’40-50’ (13%), ‘50-60’ (22%), ’60-70’ (22%), ’70-80’ (17%). A decrease in performance (5%) is observed in the two youngest age groups, which also represents the smallest proportion of the dataset. When stratifying by sex (Female (48%) and Male (49%)) we find that both MI2 and RAD-DINO models perform slightly better for female patients, with an average performance increase of 2%. We focus on the six most prevalent scanner manufacturers in the dataset. FUJIFILM Corporation (38%), Carestream Health (24%), GE Healthcare (15%), SIEMENS (8%), Philips (6%). Among these, we observe a 6% performance drop for scans from FUJIFILM Corporation compared to the best-performing group, Carestream Health. Performance on FUJIFILM scanners is likely worse because some original images are mislabeled as ‘derived’ in the DICOM tags. Since our subsetting prioritizes the latest original image (or derived if no original exists), this mislabeling may have caused FUJIFILM cases to rely on suboptimal images.

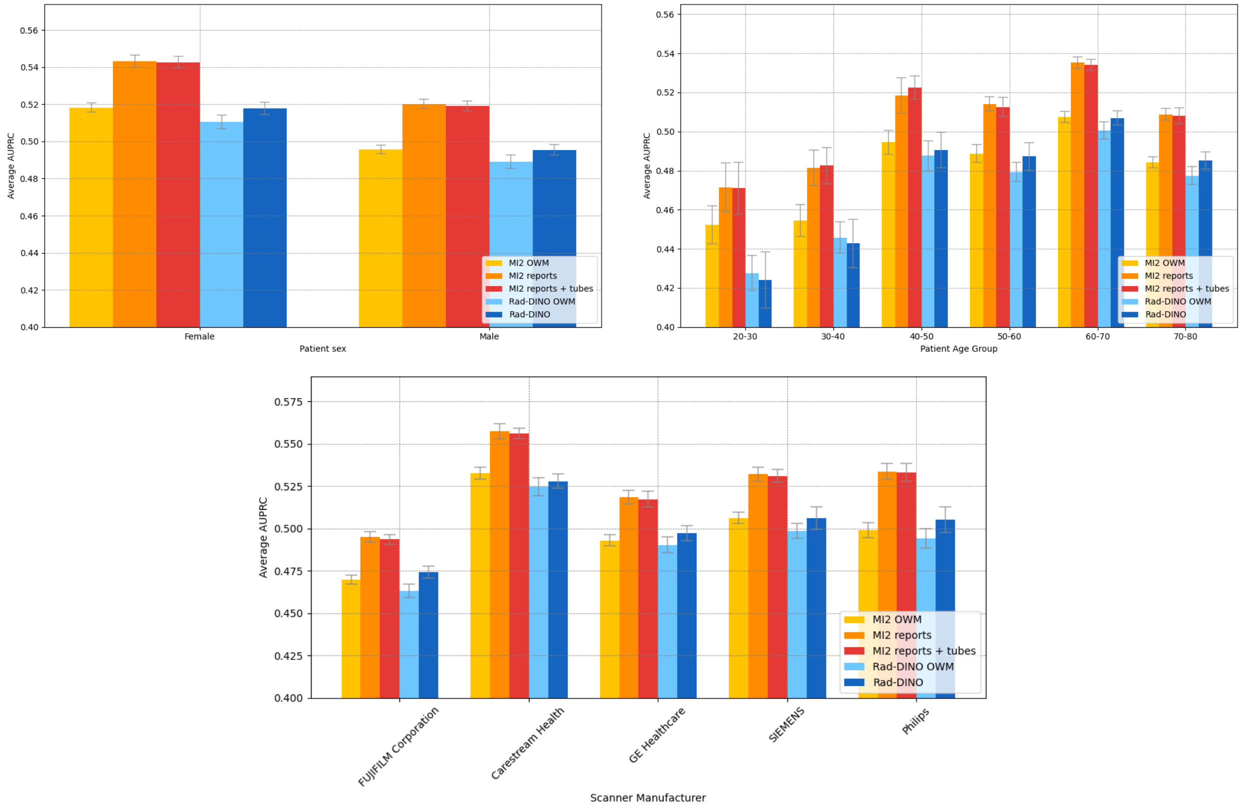


Figure 8: Metadata stratification findings classification on INST-CXR-BENCH. Top left: Average AUPRC across 20 findings stratified by sex. Top right: Average AUPRC across 20 findings stratified by age. Bottom: Average AUPRC across 20 findings stratified by manufacturer.