Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Selective data curation enables efficient pretraining of chest radiograph foundation models

Curtis Langlotz

langlotz@stanford.edu

Stanford University https://orcid.org/0000-0002-8972-8051

Chong Wang

Stanford University https://orcid.org/0000-0003-0022-0217

Yabin Zhang

Stanford University

Yunhe Gao

Stanford University

Maya Varma

Stanford University https://orcid.org/0000-0003-0693-7753

Clemence Mottez

Stanford University

Faidra Patsatzi

Stanford University

Fuying Wang

Stanford Center for Artificial Intelligence in Medicine and Imaging, Stanford University

Jiaming Liu

Stanford University

Jin Long

Stanford Medicine https://orcid.org/0000-0002-6225-727X

Jean-Benoit Delbrouck

Stanford University

Sergios Gatidis

Stanford University

Akshay Chaudhari

Stanford https://orcid.org/0000-0002-3667-6796

Article

Keywords: Foundation models, medical data, efficient learning, CXR interpretation

Posted Date: September 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10848130/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. Jean-Benoit Delbrouck is an employee of hoppr.ai. The other authors declare no competing interests.

Selective data curation enables efficient pretraining of chest radiograph foundation models

Chong Wang[1,2†], Yabin Zhang[1,2†], Yunhe Gao[1,2], Maya Varma[1,2,3], Clemence Mottez[1,2], Faidra Patsatzi[1,2], Fuying Wang[1,2], Jiaming Liu[1,2], Jin Long[1,4], Jean-Benoit Delbrouck[1,2], Sergios Gatidis[2], Akshay S. Chaudhari[1,2,5], Curtis P. Langlotz[1,2,5,6*]

[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.
[3]Department of Computer Science, Stanford University, Stanford, CA, USA.
[4]Department of Pediatrics, Stanford University, Stanford, CA, USA.
[5]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA.
[6]Department of Medicine, Stanford University, Stanford, CA, USA.

*Corresponding author(s). E-mail(s): langlotz@stanford.edu;
†These authors contributed equally to this work.

Abstract

Foundation models for medical imaging are pretrained on increasingly large datasets following a "scale-at-all-costs" paradigm, although medical corpora often contain substantial redundancy, class imbalance and heterogeneous data quality. Here we investigate whether selective data curation can reduce the data and computational requirements of chest X-ray (CXR) vision-language pretraining. We introduce CheXficient, which leverages representation-guided data curation to prioritize informative and important image-report pairs. CheXficient is pretrained using 22.7% of 1,235,004 paired CXR images and reports and under 27.3% of the total compute budget, yet maintains strong performance relative to full-data pretraining and remains competitive with other large-scale pretrained models. We assess CheXficient across 20 individual benchmarks spanning 5 task types, including non-adapted off-the-shelf evaluations (zero-shot

findings classification and cross-modal retrieval) and adapted downstream tasks (disease diagnosis, semantic segmentation, and radiology report generation). Further analyses show CheXficient systematically prioritizes under-represented training samples, improving generalizability on long-tailed or rare thoracic conditions. These findings support selective data curation as a practical strategy for resource-efficient development of medical foundation models.

Keywords: Foundation models, medical data, efficient learning, CXR interpretation.

1 Introduction

Medical foundation models have emerged as a powerful paradigm for building general-purpose artificial intelligence systems in healthcare, offering a unified framework for learning transferable representations from massive medical data [1, 2]. In medical imaging, such models have shown strong performance across diverse specialties, including radiology [3], pathology [4], dermatology [5], and ophthalmology [6], with the potential to support a wide range of clinically relevant downstream tasks. Typically, medical foundation models are pretrained on broadly collected large-scale datasets (e.g., chest X-rays and associated free-text radiology reports [7], figures with captions from scientific articles [8, 9], and social media posts [10]) to acquire highly generalizable representations, which can then be adapted to specific clinical applications using comparatively limited task-specific annotations [11].

Despite these promising achievements, the reliance on large-scale medical datasets presents significant practical challenges that limit the feasibility and sustainability of medical foundation models in real-world clinical settings. In practice, retrospective archives, data-sharing initiatives, and web-sourced medical content have made large-scale medical datasets increasingly accessible [12]. For example, the global volume of chest radiography continues to grow, with an estimated annual addition of over 1.4 billion chest X-ray examinations [13]. However, standard pretraining approaches that blindly train on all available data face two critical inefficiencies. First, medical datasets often exhibit high redundancy and severe class imbalance [14], e.g., normal findings and common pathologies dominate while rare but clinically important conditions are under-represented, biasing representation learning toward over-represented patterns. Second, exhaustive training on millions of samples without accounting for heterogeneity in data quality incurs prohibitive computational burden [2] (e.g., RadFM [15] trained on 16M radiology scans using 32 NVIDIA A100 GPUs for weeks; MedGemma [16] trained on 33M medical image-text pairs using industrial-scale computational resources accumulating thousands of TPU-hours), creating substantial barriers for academic and clinical institutions with limited resources.

These challenges motivate exploring data-centric alternatives to pursuing ever-larger datasets, where the quality, composition, and utilization of training data are considered alongside dataset scale. Recent advances in machine learning suggest that optimizing which data are utilized for training can be more influential than simply increasing how much data is used [17]. A growing body of work [17–19] shows that

training samples differ markedly in their contribution to representation learning: some examples are highly informative, while others provide little benefit or can even hinder optimization [20]. Nevertheless, standard foundation model pretraining approaches typically treat all data points as equally valuable, overlooking substantial differences in informativeness, representativeness, relevance, or diversity. This insight has driven a shift toward data curation or selection strategies, in which compact, high-value subsets are selectively emphasized and employed for training. Empirical evidence [18, 20, 21] outside the medical domain demonstrates that models trained on carefully curated datasets can match or even surpass the performance of those trained on much larger corpora, challenging the long-standing assumption that generalization ability scales predominantly with dataset size. This data-centric perspective is particularly relevant to medical imaging, where data distributions are highly imbalanced across normal findings, rare pathologies, and acquisition protocols. Treating all training samples uniformly can therefore be computationally inefficient and bias representation learning toward over-represented patterns. These observations raise a central question: can medical foundation models still maintain strong performance while substantially reducing data and computational requirements through strategic data curation?

In this work, we introduce CheXficient, which uses representation-guided data curation to improve data and computation efficiency in chest X-ray (CXR) vision-language pretraining. As shown in Figure 1(a), CheXficient incorporates a prototype-driven data curator during the early pretraining stage, built within a contrastive language-image pretraining (CLIP [23]) framework. A set of evolving prototypes (i.e., prototypical centroids) [24, 25] is leveraged to characterize the current multimodal representation space, enabling prioritization of informative CXR image–report data pairs for model optimization. To be concrete, training samples that lie farther from the prototypes (corresponding to under-represented but informative regions of the data distribution) are emphasized with higher priority, while training samples near the prototypes, which tend to contain greater local redundancy, are down-weighted and under-sampled (Figure 1(b); see Methods). This strategy yields a compact yet comprehensive coverage of the data manifold, leading to improved data scalability and computational efficiency compared with existing vision–language pretraining approaches (Figure 1(d)). We systematically evaluate CheXficient under 5 complementary evaluation protocols (Figure 1(c)), encompassing both non-adapted tasks (zero-shot findings classification and cross-modal retrieval) and adapted downstream tasks (multi-disease diagnosis, semantic segmentation of anatomical structures and abnormalities, and radiology report generation). Together, these tasks reflect clinically relevant use cases of medical foundation models, spanning diagnostic decision support, anatomical understanding, and clinical reporting. Using only 22.7% of 1,235,004 paired CXR training samples (sourced from 13 public datasets across multiple countries (Table 1)) and under 27.3% of the full compute budget, CheXficient retains strong performance relative to its full-data pretraining counterpart across 20 evaluation benchmarks (Figure 1(d, e)). In addition, CheXficient remains competitive with several published state-of-the-art vision-language foundation models across representative tasks, despite using substantially fewer pretraining data and compute resources.

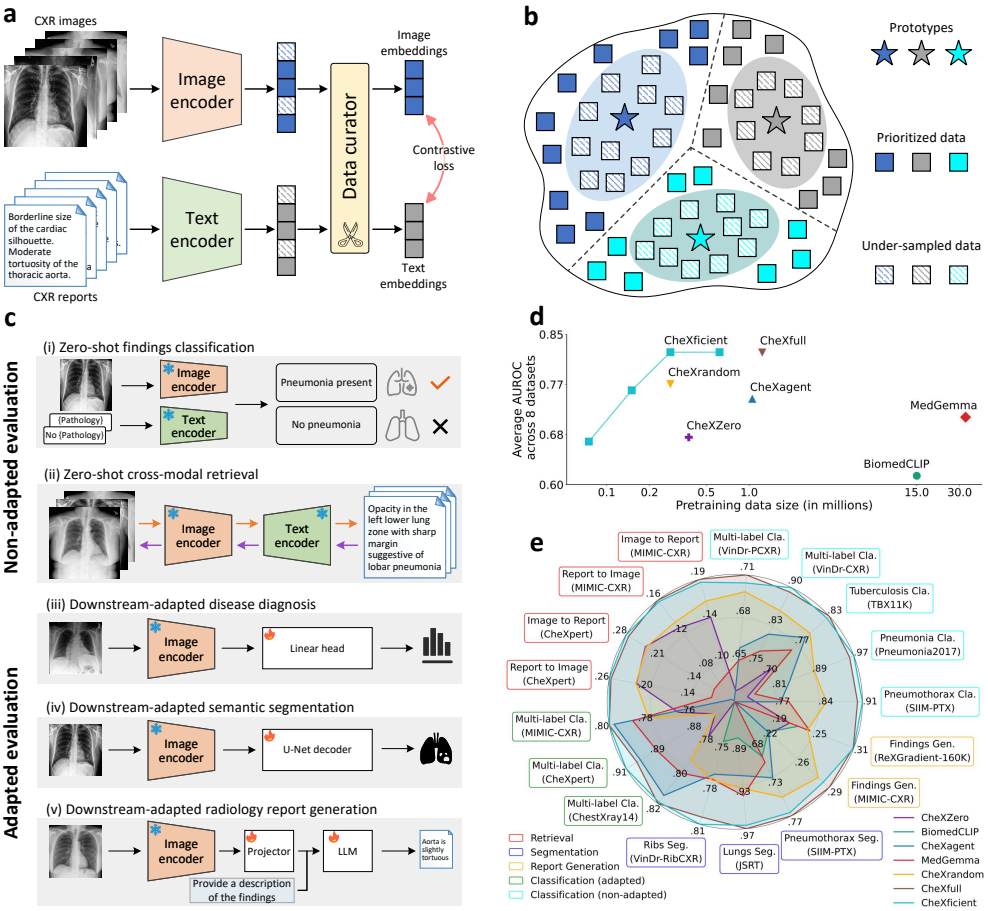


Fig. 1 Overview of this study. (a) **CheXficient pretraining strategy.** Paired chest X-ray (CXR) images and radiology reports are employed. In the early stage of pretraining, a data curator selectively determines which image–report pairs are included for training, followed by optimization with an InfoNCE [22] contrastive loss. (b) **Data curation mechanism.** The data curator prioritizes image–report pairs based on their proximity to a set of evolving prototypes. Samples farther from the prototypes are assigned higher priority, while more redundant data closer to the prototypes is down-sampled. (c) **Evaluation protocol.** CheXficient is evaluated on non-adapted tasks without any architectural or weight modifications, including (i) zero-shot finding classification and (ii) zero-shot cross-modal retrieval. CheXficient can also be adapted to downstream tasks via fine-tuning, including (iii) disease diagnosis, (iv) semantic segmentation, and (v) radiology report generation. (d) **Data efficiency analysis.** Zero-shot findings classification performance (average AUROC on 8 datasets) versus pretraining data size. (e) **Cross-task performance.** The radar plot summarizes results across 5 task types: retrieval (in Recall@1), segmentation (in Dice score), and radiology report generation (in RadGraph), adapted and non-adapted classification (in AUROC). Comparisons with published external models provide performance context but are not controlled for differences in architecture, pretraining objective, or training domain. The effect of data curation is evaluated through the controlled comparisons among CheXficient, CheXfull and CheXrandom.

2 Results

2.1 Model design and experimental setup

To validate the data and compute efficiency of CheXficient in CXR vision-language pretraining, we construct a large-scale pretraining database containing over 1.235 million CXR image–report pairs collected from 13 publicly available sources, including CheXpert-Plus [26], MIMIC-CXR [27], ReXGradient-160K [28], PadChest [29], BIMCV COVID-19 [30], CANDID-PTX [31], CASIA-CXR [32], Open-I [33], ChestX-ray14 [34], BRAX [35], VinDr-CXR [36], VinDr-PCXR [37], and ChestDR [38]. The detailed composition of the full pretraining database is provided in Table 1.

We compare the performance of CheXficient against that of the following baseline foundation models: CheXrandom and CheXfull. To obtain a fair comparison, the same network architectures (DINOv2 [39] as the image encoder and BioClinicalBERT [40] as the text encoder), pretraining recipe (image-text contrastive learning by InfoNCE loss [22]), and evaluation protocols are employed in all models. CheXfull is pretrained on the entire 1.235 million paired CXR data. CheXficient is pretrained on a carefully curated subset from the full pretraining corpus. CheXrandom is pretrained on a random subset of the full set, with an identical size as the curated subset. Additional comparisons with representative data selection strategies are provided in Appendix B.1. We evaluate CheXficient on 20 benchmarks across two evaluation protocols: (1) non-adapted zero-shot evaluation: findings classification (47 thoracic findings) and cross-modal retrieval (image to reports, image to *Findings*, image to *Impression* and vice versa); (2) adapted downstream fine-tuning evaluation: multi-disease diagnosis, radiology report generation, as well as semantic segmentation of anatomic structures and abnormalities. Detailed information on downstream datasets is given in Table 2. The area under the receiver operating curve (AUROC), Recall@1, Dice score, and standard radiology report generation metrics (e.g., RadGraph) are utilized for evaluating the task performance of these models. In experiments, we train the models with different seeds to obtain multiple results and evaluate their performances for statistical analysis.

2.2 CheXficient prioritizes under-represented training data

To characterize the curated training data relative to the original pretraining distribution, we extract feature embeddings for both full-set and curated-set training samples using the pretrained CheXfull model. CheXfull is used as the reference encoder because it is trained on the complete pretraining corpus and therefore provides a natural representation of the data distribution prior to curation. Similar enrichment patterns are observed when the analysis is repeated using CheXficient embeddings (Extended Data Figure 1). For comparison, we additionally consider a random subset (with the same size as the curated set) used by CheXrandom. We apply principal component analysis (PCA) to project the high-dimensional features (unified multimodal embeddings, see Section 4.4) into a two-dimensional space for visualization. Figure 2(a) illustrates the resulting feature distributions using two-dimensional histograms. The curated subset

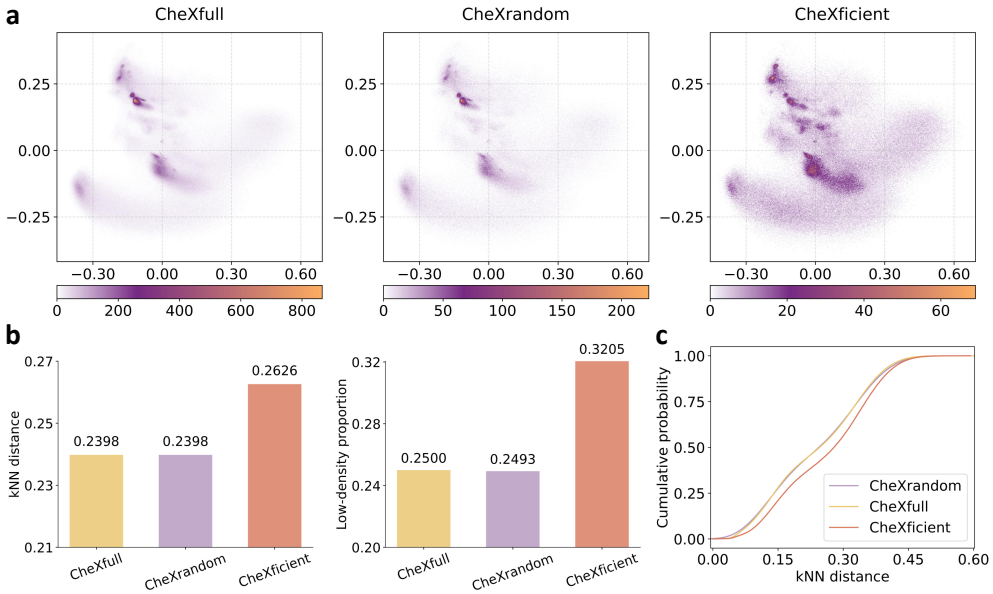


Fig. 2 (a) Feature distribution of training samples from the full set (CheXfull, n = 1,235K), random subset (CheXrandom, n = 280K), and curated subset (CheXficient, n = 280K). Qualitative histograms of feature embeddings projected onto a two-dimensional PCA space (Colorbars are independently normalized to reveal the structural organization of the PCA space, absolute density values are therefore not directly comparable across methods). The curated subset exhibits a distribution distinct from those of both the full corpus and the random subset, and enriches low-density regions. (b) Quantitative analysis of local feature density (left) for samples from the full set, random subset, and curated subset, measured by the average k-nearest neighbor (kNN) distance ($k = 20$) computed in the raw feature space. The low-density proportion (right) denotes the fraction of samples falling within the top 25% lowest-density regions of the full dataset. (c) CDF plot of the average kNN distance. Curated samples (red) are shifted toward higher kNN distances, indicating enrichment in low-density regions, whereas the random subset (purple) closely overlaps with the full set (yellow).

exhibits a distribution distinct from those of both the full corpus and the random subset. Particularly, curated samples extend into regions that are sparsely populated in the full corpus, likely suggesting an over-representation of uncommon imaging patterns, atypical report semantics, or long-tailed thoracic diseases.

To quantitatively characterize these differences, we also measure the local sample density using the average k-nearest neighbor (kNN) distance in the raw feature space. As reported in Figure 2(b), curated samples exhibit a much larger kNN distance than the full set (0.2626 vs. 0.2398), consistent with enrichment in sparsely populated regions. Moreover, more than 32% of curated samples fall within the top 25% lowest-density regions of the full set. In Figure 2(c), the CDF (cumulative distribution function) plot of average kNN distances shows that the curated subset (red) is systematically shifted to the right, further supporting preferential retention of samples from low-density regions. In contrast, the random subset (purple) closely follows the density distribution of the full corpus (yellow), demonstrating that it does not prioritize under-represented training data from low-density regions.

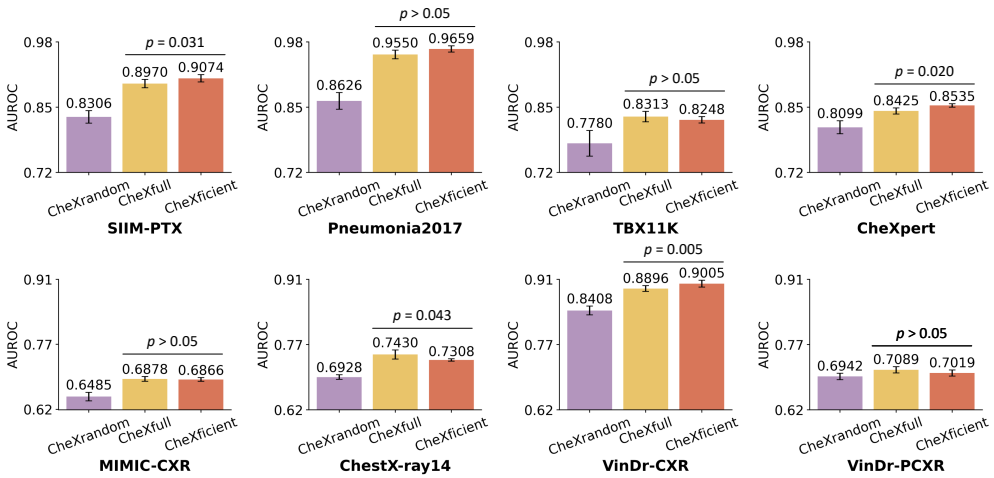


Fig. 3 Performance on zero-shot findings classification. We evaluate the pretrained models on 8 public datasets. Among them, SIIM-PTX, Pneumonia2017, and TBX11K are from unseen domains not included in pretraining, while the remaining 5 datasets are used for internal evaluation. Compared to CheXfull, CheXficient achieves higher AUROC on three datasets ($p < 0.05$), shows no statistically significant differences on four datasets ($p > 0.05$), and lower performance on one dataset ($p < 0.05$). Across all datasets, CheXficient achieves higher AUROC than CheXrandom. We present the mean ± 95% confidence interval (CI) of AUROC for each model. The listed p-values are computed using two-sided t-tests.

2.3 Performance in non-adapted zero-shot tasks

We first assess the performance of the three pretrained vision–language models (CheXficient, CheXrandom, and CheXfull) on two non-adapted zero-shot tasks, where these models are directly used without any architecture or weight modifications. For zero-shot findings classification, we follow prior work [7] and prompt the text encoder with class-specific descriptions. For zero-shot cross-modal retrieval, we employ radiology reports as queries to prompt the text encoder.

2.3.1 Zero-shot findings classification

We evaluate the pretrained CheXficient, CheXrandom, and CheXfull on the SIIM-PTX [41] dataset (n = 1,372), containing binary classification labels (pneumothorax vs. no pneumothorax). As shown in Figure 3, CheXficient significantly outperforms CheXfull (p = 0.031) and substantially exceeds the performance of CheXrandom (AUROC: 0.9074 (95% CI 0.9001, 0.9148) vs. 0.8306 (95% CI 0.8181, 0.8431)). We also evaluate the above three pretrained models on the Pneumonia2017 [42] dataset (n = 624), where each CXR image is labeled either pneumonia (bacterial/viral) or normal. On Pneumonia2017, no statistically significant difference in AUROC is detected between CheXficient and CheXfull ($p > 0.05$), whereas CheXficient achieves an absolute AUROC improvement of more than 0.1033 over CheXrandom. We conduct tuberculosis classification on the TBX11K [43] dataset (n = 1,800), where each image is labeled either tuberculosis or non-tuberculosis. CheXficient largely preserves the AUROC performance of CheXfull and outperforms CheXrandom.

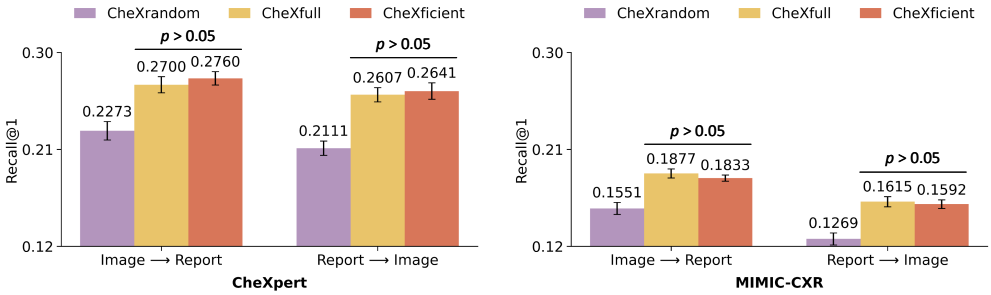


Fig. 4 Performance on zero-shot cross-modal retrieval. We evaluate the pretrained models on the public CheXpert and MIMIC-CXR benchmarks. CheXficient achieves performance without statistically significant differences from CheXfull on both datasets ($p > 0.05$), while outperforming CheXrandom. We report the mean ± 95% CI of Recall@1 (The recall of retrieving the exact paired CXR report (or image) within the top-1 result for a given CXR image (or report)).

Beyond single-disease tasks, recent research [44, 45] has focused on comprehensive detection of multiple thoracic diseases and radiological findings. We therefore evaluate CheXficient on five multi-disease chest X-ray benchmarks: CheXpert [26] (n = 500, 14 categories), MIMIC-CXR [27] (n = 3,082, 14 categories), ChestX-ray14 [34] (n = 25,596, 15 categories), VinDr-CXR [36] (n = 3,000, 28 categories), and VinDr-PCXR [37] (n = 1,397, 15 categories) datasets. For each dataset, we compute category-wise AUROC and report the macro-average. CheXficient achieves significantly higher macro-average AUROC than CheXfull on CheXpert ($p = 0.020$) and VinDr-CXR ($p = 0.005$), shows no statistically significant difference on MIMIC-CXR and VinDr-PCXR ($p > 0.05$), and yields lower performance on ChestX-ray14 ($p = 0.043$). Across all five benchmarks, CheXficient consistently achieves higher AUROC than CheXrandom.

Notably, three datasets (SIIM-PTX, Pneumonia2017, and TBX11K) originate from institutions and data distributions entirely disjoint from the pretraining corpus and are therefore treated as unseen-domain evaluation benchmarks (Table 2). As shown in Figure 3, CheXficient preserves strong performance relative to CheXfull across these unseen-domain evaluations, despite being pretrained on less data. AUPRC results on the 8 public datasets are reported in Extended Data Figure 2.

2.3.2 Zero-shot cross-modal retrieval

Zero-shot cross-modal retrieval aims to evaluate a model's ability to match a CXR image to its corresponding radiology report (*Findings* section, *Impressions* section, or their combination) and vice versa. As illustrated in Figure 4, no statistically significant differences are detected between CheXficient and CheXfull, on the task of retrieving the correct CXR report for a given CXR image (Image → Report) across both evaluation datasets (CheXpert (n = 1,000) and MIMIC-CXR (n = 3,082)), whereas CheXficient achieves higher Recall@1 results than CheXrandom. A similar trend is observed for retrieving the correct CXR image given a CXR report (Report → Image). In this evaluation, *Findings* and *Impressions* sections are combined together to constitute the textual report, retrieval results based solely on either section are provided in Extended Data Figure 3 and Figure 4.

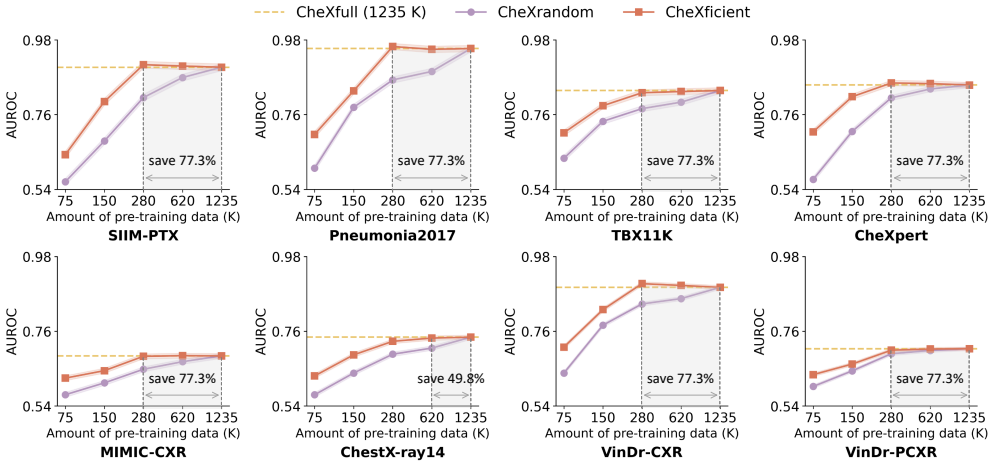


Fig. 5 Zero-shot findings classification performance under different pretraining data budgets. The performance of both CheXficient and CheXrandom progressively increases on all 8 benchmarks as the quantity of pretraining samples increases. CheXficient consistently outperforms CheXrandom at all data budgets and across all benchmarks, demonstrating higher data efficiency. When pretrained on only 280K curated image–report pairs (22.7% of the full data, except for ChestX-ray14), CheXficient maintains strong performance relative to the full-data pretrained model CheXfull.

2.3.3 Pretraining data and compute efficiency

We validate the advantage of data curation by assessing the pretraining data and compute efficiency, defined as the amount of paired image–report data and pretraining time (GPU-hours) required to build a high-performing CXR foundation model.

We compare CheXficient with the baseline CheXrandom model, which is trained under identical pretraining data budgets ranging from 75K to 1.235M image–report pairs. Figure 5 shows zero-shot findings classification performance across different pretraining set sizes, with the CheXfull model (trained on the entire dataset) included as a reference. The performance of both models consistently improves across the eight evaluation benchmarks as the size of the pretraining corpus increases. Across all data budgets and benchmarks, CheXficient achieves higher performance than CheXrandom. When pretrained on 280K curated image–report pairs (22.7% of the full data), CheXficient retains strong performance relative to CheXfull on most benchmarks, including SIIM-PTX, Pneumonia2017, CheXpert, MIMIC-CXR, and VinDr-CXR.

We further investigate the pretraining compute efficiency in the context of zero-shot findings classification. Computational cost is quantified in NVIDIA H100 GPU-hours. For a fair comparison, all models are pretrained for the same total number of epochs (20 epochs), which we found sufficient for convergence. CheXfull is pretrained on the entire dataset comprising 1.235 million image–report pairs, whereas CheXrandom and CheXficient are pretrained on the same subset size (280K). As demonstrated in Figure 6, CheXficient uses 72.7–81.8% fewer H100 GPU-hours than CheXfull to reach the reported operating points. Under matched pretraining compute budgets, CheXficient achieves higher AUROC classification performance than both CheXfull and CheXrandom across all eight evaluation benchmarks.

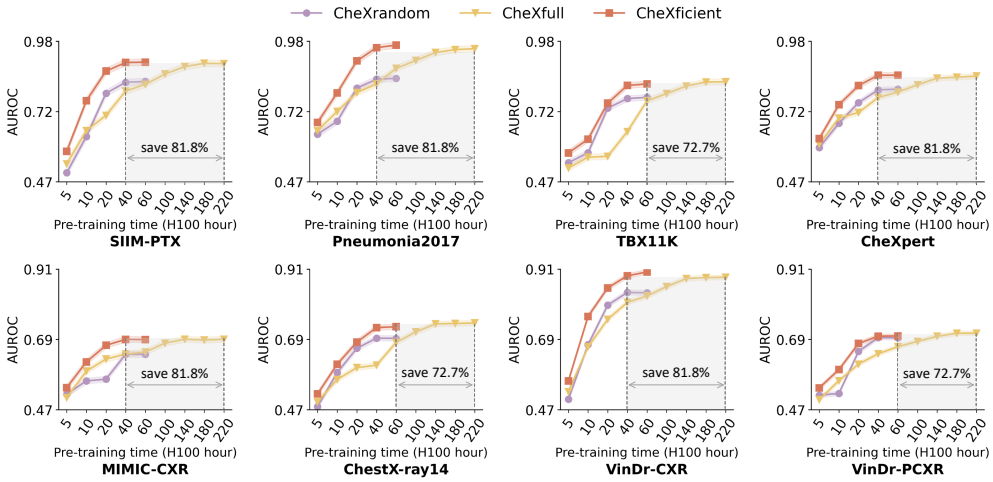


Fig. 6 Zero-shot findings classification performance under different pretraining compute budgets. The three models are pretrained for the same number of epochs. CheXficient (280K) outperforms CheXrandom (280K) and CheXfull (1235K) on all evaluation benchmarks under matched pretraining time, highlighting the compute efficiency of curated pretraining of CheXficient.

We additionally compare CheXficient with several representative data selection strategies under controlled pretraining settings (Appendix B.1), where CheXficient demonstrates an overall advantage over methods based on different selection principles.

2.4 Performance in adapted downstream tasks

We next evaluate the performance of pretrained vision–language models (CheXficient, CheXrandom, and CheXfull) on three adapted downstream tasks: disease diagnosis (image classification), semantic segmentation, and radiology report generation. In all tasks, the pretrained image encoder is leveraged as a fixed feature extractor, while task-specific heads or decoders are appended and fine-tuned on target labeled datasets. Detailed fine-tuning configurations are provided in Appendix C.

2.4.1 Classification label efficiency and fine-tuning performance

To assess the representation quality learned through efficient pretraining in CheXficient, we evaluate downstream image classification performance on three representative multi-disease CXR datasets: CheXpert, MIMIC-CXR, and ChestXray14, which cover a diverse set of pathological findings and diseases. We adopt a linear-probing protocol with a frozen image encoder backbone. Specifically, a randomly initialized linear classification head is attached to the pretrained image encoder, and only this linear classifier is trained using images and corresponding labels from the target dataset. For comparison purposes, we also include DINOv2 [39], which is pretrained solely on 142 million natural images.

As shown in Figure 7, AUROC performance of all models increases with the amount of labeled data used for fine-tuning. Models pretrained on CXR data achieve higher performance than DINOv2. Across different proportions of labeled training

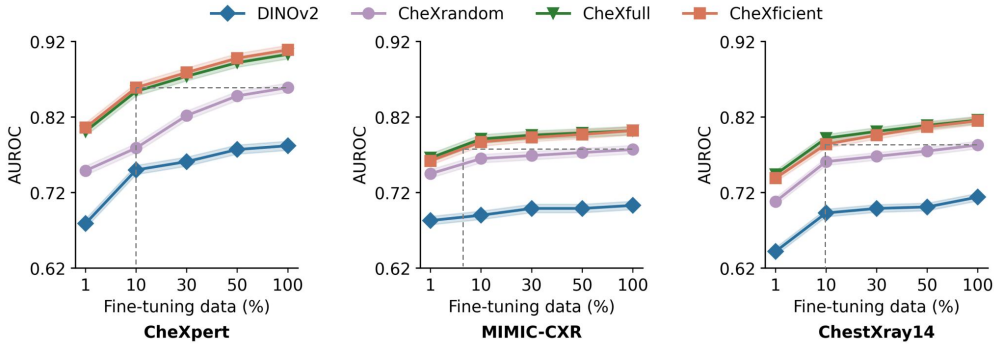


Fig. 7 Comparison of CheXficient with other pretrained models in label-efficient disease diagnosis across three representative datasets: CheXpert (n = 500, 14 categories), MIMIC-CXR (n = 3,082, 14 categories), and ChestXray14 (n = 25,596, 15 categories). Performance is reported at varying proportions of labeled training data. Vertical dashed lines indicate the amount of labeled data required for CheXficient to match the performance of its counterpart CheXrandom.

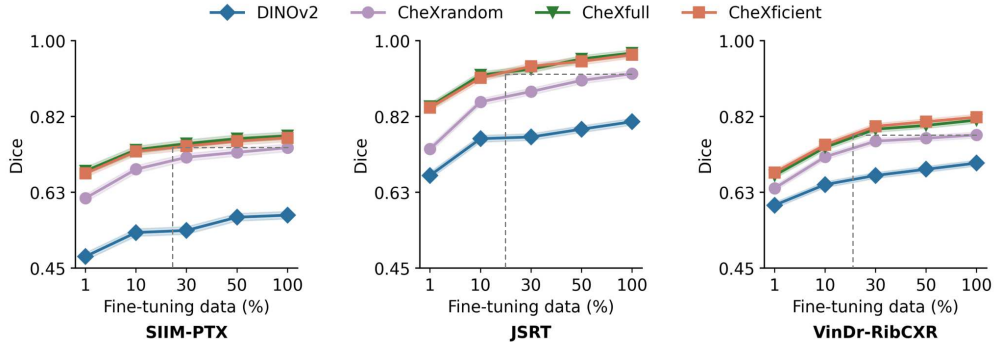


Fig. 8 Comparison of CheXficient with other pretrained models in annotation-efficient pathological and anatomical segmentation across three datasets: SIIM-PTX (n = 1,372), JSRT (n = 50), and VinDr-RibCXR (n = 49). Performance is reported at varying proportions of annotated training samples. Vertical dashed lines indicate the amount of annotated data required for CheXficient to match the performance of its counterpart CheXrandom.

data, CheXficient realizes AUROC performance similar to CheXfull. CheXficient achieves performance competitive relative to CheXrandom while using approximately 10% of the labeled training samples. A similar observation can be found in AUPRC results (Extended Data Figure 5). Furthermore, CheXficient also achieves competitive performance relative to several published large-scale pretrained models (e.g., BiomedCLIP [8] and MedGemma [16]) trained on datasets more than 50× larger than that used by CheXficient (Figure 1(e) and Extended Data Table 1). These comparisons contextualize CheXficient among existing medical foundation models but are not controlled for differences in architecture, pretraining objective, or training domain.

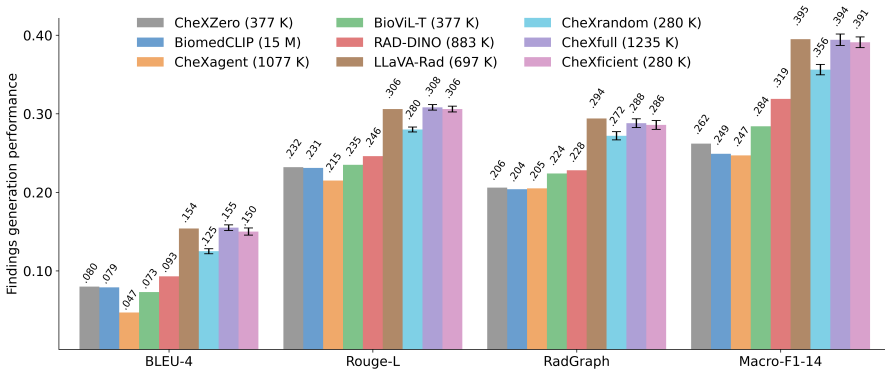


Fig. 9 Downstream radiology report (*Findings* section) generation performance (%) on official test split of MIMIC-CXR (n = 2,461). Text in parentheses denotes the pretraining data size of the model.

2.4.2 Segmentation annotation efficiency and fine-tuning performance

To further evaluate the quality of dense representations learned through curated pretraining, we focus on downstream tasks of semantic segmentation. Specifically, we consider three commonly adopted CXR benchmarks covering both pathological and anatomical structures: pneumothorax segmentation on SIIM-PTX [41], lung segmentation on JSRT [46], and rib segmentation on VinDr-RibCXR [47]. We employ a frozen image encoder backbone within an encoder–decoder segmentation framework. A U-Net [48] decoder is attached to the pretrained image encoder and fine-tuned using CXR images and their corresponding pixel-wise annotations from each target dataset, while the image encoder remains fixed. This setup allows us to assess the quality and transferability of learned representations for pixel-level dense prediction tasks.

As shown in Figure 8, CheXficient achieves similar segmentation performance relative to CheXfull across all three tasks. CheXficient exhibits strong annotation efficiency, matching the performance of CheXrandom while using 10–30% of the densely-annotated training samples. These results demonstrate that curated, data-efficient pretraining enables the learning of high-quality image representations that transfer effectively to challenging downstream segmentation tasks. CheXficient also achieves performance competitive relative to several published off-the-shelf image encoders (e.g., BiomedCLIP [8], MedGemma [16], and CheXagent [49]), in spite of being trained on a substantially smaller-scale pretraining corpus (Figure 1(e)). These results position CheXficient among existing medical foundation models, whereas the contribution of data curation is isolated through the controlled CheXficient–CheXfull comparison.

2.4.3 Radiology report generation

We also evaluate the quality of image representations learned through efficient pretraining on a generative vision–language downstream task, namely radiology report generation for the *Findings* section. For the model architecture, we adopt a LLaVA-style multimodal framework [50]. To generate the textual report output, patch

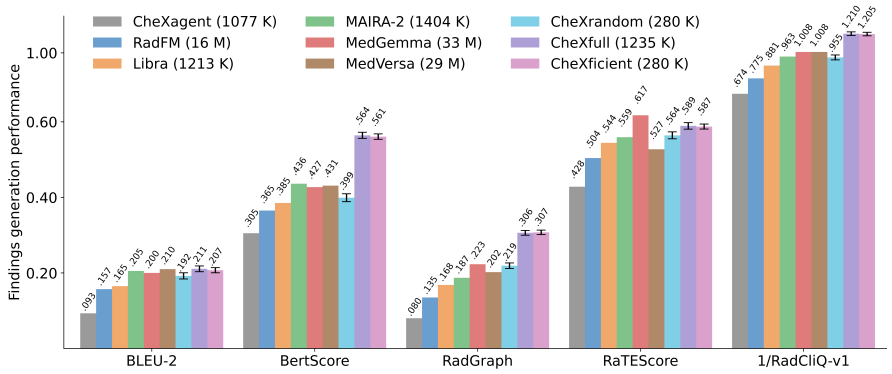


Fig. 10 Downstream radiology report (*Findings* section) generation performance (%) on official test split of ReXGradient-160K (n = 10,000). Text in parentheses denotes the pretraining data size of the model.

embeddings extracted from the frozen image encoder are projected and concatenated with a textual instruction: "Provide a description of the findings in the radiology image given the following Indication: ⟨ ⟩". If the Indication information is unavailable, the optional ⟨ ⟩ field is omitted. Consistent with LLaVA-Rad [51], we employ a multi-layer perceptron (MLP) projector and Vicuna-7B-v1.5 [52] as the language decoder model. The projector network is randomly initialized and jointly trained with the language decoder, while the image encoder remains frozen. Performance is quantified using standard lexical metrics (BLEU-4 [53], Rouge-L [54]) and radiology-specific metrics, including RadGraph [55] and CheXbert-based Macro-F1-14 [26] (mapping "uncertain" labels to negative). For ReXGradient-160K, we additionally report RadCliQ [56], RaTEScore [57], and BertScore [58].

CheXficient consistently outperforms CheXrandom across all lexical and clinical metrics on both MIMIC-CXR (Figure 9) and ReXGradient-160K (Figure 10). Despite being pretrained on substantially fewer image–report pairs, CheXficient largely maintains performance relative to CheXfull. On MIMIC-CXR, CheXficient slightly under-performs the recent leading report generation framework LLaVA-Rad, while demonstrating improvements over other specialized vision–language baselines, including BiomedCLIP, CheXagent, and RAD-DINO [59], which are pretrained on much larger-scale datasets (Extended Data Table 1). On ReXGradient-160K, CheXficient achieves a higher reported 1/RadCliQ score relative to several advanced medical foundation models, e.g., RadFM [15], Libra [60], MAIRA-2 [61], and MedVersa [62].

2.5 Generalization across long-tailed thoracic conditions

Chest radiograph interpretation is inherently a long-tailed problem [44, 45]. The distribution of thoracic findings is heavily skewed toward a small number of common abnormalities, as many clinically relevant conditions occur infrequently. This severe class imbalance poses substantial challenges for model learning, as training objectives are often dominated by majority classes. Consequently, AI models tend to exhibit

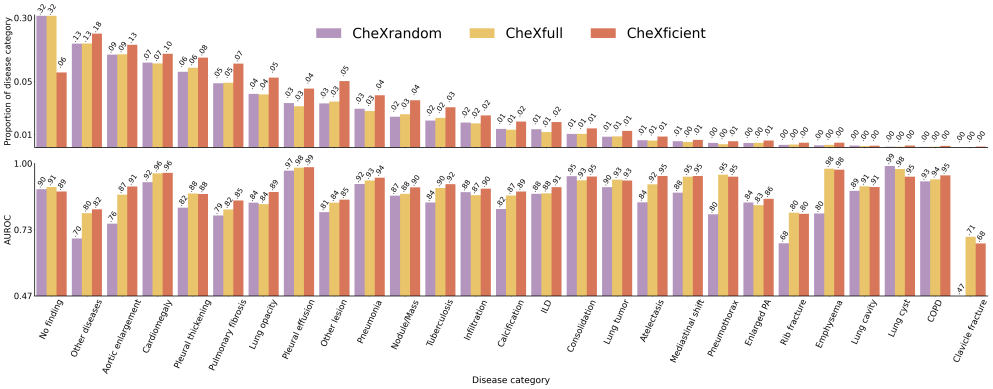


Fig. 11 Impact of curated pretraining on per-disease generalization in VinDr-CXR. Top: disease label distributions from the VinDr-CXR training set (the y-axis uses a symmetric logarithmic scale to visualize disease label frequencies spanning multiple orders of magnitude). Bottom: corresponding zero-shot per-disease AUROC results on the VinDr-CXR test set. Curated pretraining (CheXficient) re-balances label distributions and yields improved zero-shot generalization for most disease labels compared to non-curated pretraining (CheXfull and CheXrandom). Note that the disease label "edema" is not shown in the figure as there are no samples with this label in the official test data.

biased representations that favor common findings, potentially at the expense of underrepresented conditions. Robustness and generalizability under long-tailed distributions therefore constitute critical criteria for evaluating medical foundation models.

To investigate how data curation influences long-tailed generalization, we analyze the relationship between disease label distributions in the pretraining data and zero-shot generalization ability on test data. We focus on VinDr-CXR [36], as it provides comprehensive disease-level annotations covering a total of 28 findings and diagnoses, with long-tailed distributions. Specifically, we first compute the disease label distributions in the VinDr-CXR training set (see Figure 11(Top)). Then, we compare the per-category zero-shot findings classification performance (AUROC) on the VinDr-CXR test data for different models (Figure 11(Bottom)).

We observe that *No finding* samples dominate the non-curated training sets used by CheXfull and CheXrandom. In contrast, CheXficient greatly reduces the proportion of *No finding* samples, while almost preserving the generalization performance for this category on the test set. Meanwhile, the curated pretraining set increases the relative prevalence of most disease labels, which is accompanied by improved zero-shot generalization across a broad range of pathologies (16 out of 26), such as *Other diseases*, *Aortic enlargement*, *Pulmonary fibrosis*, *Lung opacity*, *Pleural effusion*, *Other lesion*, and *Pneumonia*. However, this trend does not consistently hold for a small number of extremely-rare diseases, such as *Lung cyst* and *Clavicle fracture*, which each account for less than 0.1% of the training data. We hypothesize that, despite being relatively up-weighted by data curation, the absolute number of samples for these conditions remains insufficient to support robust representation learning. This observation suggests a potential limitation of selective data curation strategies alone in addressing long-tailed distributions and highlights that complementary approaches,

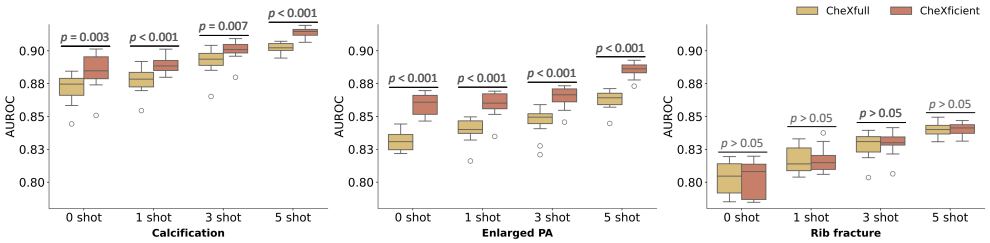


Fig. 12 Adaptability of CheXficient in detecting long-tailed conditions (Calcification, Enlarged PA (pulmonary artery), and Rib fracture) under a two-way (no finding versus each long-tailed condition) k-shot learning setting. The results demonstrate the ability of CheXficient to adapt to data-scarce conditions using only a few labeled training samples. Box plots show AUROC distributions over 20 experimental replicates, with the median (central line), interquartile range (box indicating the 25th and 75th percentiles), whiskers extending to 1.5 times the interquartile range, and outliers shown as individual points. p-values from two-sided t-tests are reported.

such as targeted data acquisition or task-specific data augmentation, may be necessary for extremely-rare disease modeling.

We further examine the adaptability of CheXficient to long-tailed conditions under limited supervision by finetuning the pretrained models with a linear-probing configuration, using only 1–5 labeled samples (k-shot learning) [44]. Figure 12 presents box plots of AUROC distributions for three representative VinDr-CXR conditions. CheXficient outperforms CheXfull, achieving higher median and maximum AUROC scores for *Calcification* and *Enlarged PA*, and competitive performance for *Rib fracture*. These results demonstrate that CheXficient can leverage scarce labeled data to detect long-tailed conditions, underscoring its practical value in data-scarce environments.

3 Discussion

In this study, we propose a data- and compute-efficient approach for pretraining medical foundation models that leverages representation-guided data curation. Rather than indiscriminately enlarging training datasets [63, 64] and incurring substantial computational costs, our approach optimizes the utilization of data and compute budgets, offering a cost-effective alternative for building high-performing medical foundation models. By curating 280K image–report data pairs (22.7%) from a pool of 1.235 million paired CXR images and radiology reports and consuming under 27.3% of the total compute budget, we pretrained a new CXR foundation model, CheXficient. Extensive experiments demonstrate that CheXficient achieves strong performance compared with a full-data pretrained counterpart across diverse evaluation benchmarks spanning five evaluation protocols. These evaluations cover both non-adapted tasks, including zero-shot findings classification and cross-modal retrieval, as well as adapted downstream tasks such as multi-disease diagnosis, semantic segmentation of anatomical structures and abnormalities, and radiology report generation. Together, these results provide promising evidence that curated pretraining on large-scale medical datasets can substantially improve data and compute efficiency as well as generalizability. We also analyze feature distributions of the curated training samples (Figure 2), where

the curated data exhibits distinctive local sample-density structures, reflecting its ability to preserve data diversity while emphasizing under-represented and informative samples. This structural property is closely associated with strong model performance (Figure 3 and Figure 4) and improved long-tailed generalization (Figure 11 and Figure 12) in the resulting CXR foundation model.

During the early-stage of pretraining, CheXficient integrates a data curator that employs a set of evolving prototypes (i.e., prototypical centroids) to approximate the global data manifold. The curator selectively prioritizes informative CXR image–report pairs for model optimization: samples that lie far from the learned prototypes are emphasized, while those close to the prototypes are down-weighted and under-sampled. Those unselected samples are excluded from the pretraining process. In comparison with CheXfull, which shares an identical network architecture but is pretrained on the entire training set of 1.235 million data pairs, CheXficient achieves substantially higher pretraining data efficiency (Figure 5). In addition, CheXficient requires significantly less pretraining computational cost than CheXfull (Figure 6). Comparisons with representative alternative data selection strategies further support the effectiveness of representation-guided curation in CheXficient (Appendix B.1).

Despite being pretrained on only 280K image–report pairs, CheXficient also demonstrates strong performance across a wide range of downstream applications, including both unimodal evaluations (multi-disease diagnosis and semantic segmentation) and multimodal tasks such as radiology report generation. Across eight datasets, CheXficient achieves performance similar to that of CheXfull (Figures 7, 8, 9, and 10). Another appealing capability of CheXficient is its ability to reduce the amount of expert-annotated training data required for downstream tasks. Figure 7 demonstrates the high labeling efficiency of CheXficient in terms of disease classification. On MIMIC-CXR dataset, it achieves the classification performance of CheXrandom (pretrained on a random subset with an identical size as the curated subset), using fewer than 10% of the labeled finetuning data. A similar trend is observed for downstream semantic segmentation tasks (Figure 8). Furthermore, CheXficient shows highly competitive reported performance in report generation compared with several existing large-scale pretrained models (Figure 9 and Figure 10).

We also conducted additional analyses of the curated subset to examine data-source composition and patient demographic characteristics, including age and race. First, we noticed that the data efficiency gains are not attributable to excluding specific data source. Instead, all 13 pretraining data sources remain represented in the curated subset (Appendix A.2). We also observed that the curated subset contains a relatively higher proportion of training samples from two age groups: pediatric patients (0–10 years-old), whose CXRs exhibit greater anatomical variability, and older adults (over 65 years old), whose CXRs often present a broader spectrum of age-related pathologies due to their higher prevalence in this population (Appendix A.3). Regarding patient race, the curated subset shows a racial distribution closely resembling that of the full training set, suggesting no obvious shift in demographic representation at the dataset level (Appendix A.4). In addition, the curation process exhibits a mild preference for frontal CXR scans over lateral CXR scans (Appendix A.5), as frontal CXRs often provide more informative visual cues for many radiological findings. It also demonstrates

a pronounced preference for samples with longer CXR reports (Appendix A.6), which may contain richer pathological findings and greater diagnostic complexity. Together with the disease-level analyses in Figure 11, these results inspire an important and broad implication: data acquisition or collection practices for developing medical foundation models may benefit from prioritizing under-represented data, diverse sources, long-tailed pathology conditions, and informative samples with rich visual and textual clinical content.

Although our experiments focus on chest X-rays, the proposed strategy is not inherently tied to chest radiography and may be extensible to other medical imaging modalities and multimodal datasets, such as CT. We acknowledge several limitations in our methodology and evaluations. First, due to the limited scope of our study, we have not studied alternative encoder architecture adaptations, e.g., ViT-Large architecture adopted by other CXR foundation models (Extended Data Table 1). We expect that incorporating more advanced encoder architectures into our framework could yield additional performance gains due to enhanced representational capacity, and we leave this exploration for future work. Second, our multimodal evaluation setting is currently limited to textual report generation. Other important radiological multimodal tasks, such as visual question answering [65] and visual grounding [66], remain unexplored and represent promising directions for future research. Third, although our validation covers 47 thoracic findings and diseases (e.g., cardiomegaly, pneumonia, and nodule), this constitutes only a fraction of known thoracic conditions (for example, over 170 radiographic findings reported in [29]), with limited coverage of rare diseases and complex systemic diseases. Future studies should aim for broader and more comprehensive evaluation across a wider spectrum of thoracic conditions. Our evaluations are retrospective and primarily based on publicly available benchmark datasets. Further evaluation across healthcare systems, acquisition settings and patient populations is needed before considering clinical deployment.

In summary, we present a data- and compute-efficient pretraining strategy to build high-performing CXR foundation models based on representation-guided data curation. Under non-adapted evaluation settings, CheXficient demonstrates strong zero-shot generalization performance while achieving high pretraining data and compute efficiency. When adapted to a range of downstream tasks, CheXficient consistently exhibits high labeling and fine-tuning efficiency while maintaining competitive performance. In addition, these findings underscore the potential of strategic dataset acquisition—particularly the prioritization of informative and long-tailed samples—to reduce development costs. They also offer practical insights for medical domains in which foundation-model development is limited by restricted access to large-scale training datasets and computational resources.

4 Methods

4.1 Datasets for pretraining

We construct a large-scale pretraining corpus comprising over 1.235 million paired CXR image–report samples, collected and aggregated from 13 publicly available datasets covering diverse patient populations worldwide, including MIMIC-CXR [27],

CheXpert Plus [26], ReXGradient-160K [28], PadChest [29], BIMCV COVID-19 [30], CANDID-PTX [31], CASIA-CXR [32], Open-I [33], ChestX-ray14 [34], BRAX [35], VinDr-CXR [36], VinDr-PCXR [37], and ChestDR [38]. The detailed composition and distribution of the full dataset are summarized in Table 1. This diverse aggregation enables broad coverage of imaging protocols, disease spectra, and textual reporting styles, thereby mitigating potential data acquisition biases and supporting robust, generalizable representation learning. To ensure transparency and reproducibility, we strictly follow the official train/validation/test splits provided by each dataset when available, and incorporate only the train split into the pretraining corpus.

Among the 13 contributing datasets, 8 provide raw radiology reports, including MIMIC-CXR [27], CheXpert Plus [26], ReXGradient-160K [28], PadChest [29], BIMCV COVID-19 [30], CANDID-PTX [31], CASIA-CXR [32], and Open-I [33]. Notably, several of these datasets (PadChest, BIMCV COVID-19, and CASIA-CXR) contain non-English reports, which are automatically translated into English to ensure linguistic consistency across the corpus. For all datasets with reports, regex-based filtering is applied to handle common formatting inconsistencies such as repeated whitespace and malformed line breaks. Samples in which both the *Findings* and *Impression* sections are blank are excluded from the corpus.

The remaining 5 datasets do not include free-text reports but instead provide structured diagnostic labels (e.g., pleural effusion, cardiomegaly, atelectasis). They are ChestX-ray14 [34], BRAX [35], VinDr-CXR [36], VinDr-PCXR [37], and ChestDR [38]. For these datasets, we generate pseudo-reports using a template-based report synthesis strategy introduced in the prior work [51]. Inclusion of template-generated reports enabled us to leverage these well-known public CXR datasets that would otherwise be excluded from multimodal pretraining. The resulting synthetic reports exhibit a token-length distribution similar to that of real-world reports (Table 1) and account for only a small portion (12.6%) of the overall pretraining corpus.

4.2 Datasets for non-adapted zero-shot tasks

As shown in Table 2, non-adapted zero-shot findings classification is evaluated on the official test sets of 8 datasets: MIMIC-CXR [27], CheXpert [26], ChestX-ray14 [34], VinDr-CXR [36], VinDr-PCXR [37], TBX11K [43], SIIM-PTX [41], and Pneumonia2017 [42]. Since the pretraining pool assembles multiple public datasets, some of which (MIMIC-CXR, CheXpert, ChestX-ray14, VinDr-CXR, VinDr-PCXR) are also used for evaluation, we strictly follow the official test splits for these datasets, ensuring patient-level separation between pretraining and evaluation. This precaution addresses potential data leakage, which is the first consideration in our experiments. Notably, SIIM-PTX [41], Pneumonia2017 [42], and TBX11K [43] are unseen out-of-distribution datasets, with their test sets completely disjoint from the pretraining sources, and are therefore treated as unseen-domain evaluation. Following the zero-shot classification protocol in CheXzero [7], for each disease label (e.g., mass, edema, atelectasis), we leverage the prompts "⟨ disease ⟩" and "no ⟨ disease ⟩" as positive and negative prompts for the text encoder, obtaining softmax-based probabilistic predictions.

For the non-adapted zero-shot cross-modal retrieval, evaluation is performed on two datasets: MIMIC-CXR and CheXpert. On MIMIC-CXR, the standard test split

Table 1 An extensive corpus of 1.235 million CXR image-report pairs employed for model pretraining in our experiments. We exclude images that do not belong to standard views (i.e., frontal or lateral). Real-world *Findings* are extracted from the free-text radiology reports. Synthetic *Findings* are generated using template sentences derived from clinical diagnostic labels.

Source dataset	Country	Patients	Labels	Text tokens			Image-text pairs
				5%-ile	Average	95%-ile	
Real-world Findings (87.4%)							
MIMIC-CXR	U.S.	65,379	–	42	94	176	354,619
ReXGradient-160K	U.S.	95,716	–	35	76	152	238,965
CheXpert-Plus	U.S.	64,725	–	38	98	209	222,116
PadChest	Spain	67,625	–	6	30	76	160,672
BIMCV-COVID19	Spain	25,082	–	8	46	102	65,421
CANDID-PTX	New Zealand	13,744	–	8	42	97	19,609
CASIA-CXR	France	11,111	–	57	96	145	11,111
Open-I	U.S.	3,996	–	29	72	139	7,424
Synthetic Findings (12.6%)							
NIH ChestX-ray14	U.S.	28,008	15	17	72	159	86,524
BRAX	Brazil	18,442	14	11	29	65	40,967
VinDr-CXR	Vietnam	15,000	28	17	107	325	15,000
VinDr-PCXR	Vietnam	7,728	15	17	72	189	7,728
ChestDR	China	4,848	19	20	122	239	4,848
						Total:	1,235,004

is utilized. Since CXR reports are unavailable for the CheXpert test split, we instead use CheXpert5x200 [67, 68] for the retrieval evaluation. To prevent data leakage, all the 1000 paired CXR samples in CheXpert5x200 are excluded from the pretraining pool in advance, ensuring a fair and disjoint evaluation set.

4.3 Datasets for adapted downstream tasks

For downstream disease diagnosis tasks, evaluations are performed on the official test sets of MIMIC-CXR, CheXpert, and ChestX-ray14 (Table 2), as in the zero-shot setting. The key difference is that varying percentages (1–100%) of data from the training split are used from each dataset to fine-tune the linear classification head.

For downstream image segmentation tasks, the SIIM-PTX [41], JSRT [46], and VinDr-RibCXR [47] datasets are employed for fine-tuning and evaluation, with the official data splits summarized in Table 2.

For downstream radiology report generation tasks, fine-tuning and evaluation are performed on the official splits of MIMIC-CXR and ReXGradient-160K. Following the protocols of LLaVA-Rad [51] and RAD-DINO [59], non-frontal images and samples without a *Findings* section are excluded, forming 146,909 training, 7,250 validation, and 2,461 testing image–text pairs for MIMIC-CXR. A similar pre-processing pipeline is applied to ReXGradient-160K, yielding 140,000, 10,000, and 10,000 image–text pairs for training, validation, and testing, respectively.

Table 2 Summary of the evaluation datasets used in this study. We evaluate on a total of 11 datasets (6 seen datasets for internal evaluation and 5 datasets for unseen-domain evaluation), comprising diverse evaluation benchmarks across 5 task types. We follow the official train/test splits provided by each dataset, except for CheXpert, where 1,000 paired CXR samples were excluded from the training set for retrieval evaluation.

Dataset	Learning setup	Task	Clinical conditions	# Train	# Test
Internal evaluation					
MIMIC-CXR	Non-adapted	Classification	14 thoracic findings	0	3,082
	Non-adapted	Retrieval	–	0	3,082
	Adapted	Classification	14 thoracic findings	227,459 (1–100%)	3,082
	Adapted	Report generation	–	146,909	2,461
CheXpert	Non-adapted	Classification	14 thoracic findings	0	500
	Non-adapted	Retrieval	–	0	1,000
	Adapted	Classification	14 thoracic findings	223,414 (1–100%)	500
NIH ChestX-ray14	Non-adapted	Classification	15 thoracic findings	0	25,596
	Adapted	Classification	15 thoracic findings	86,524 (1–100%)	25,596
VinDr-CXR	Non-adapted	Classification	28 thoracic findings	0, k-shot	3,000
VinDr-PCXR	Non-adapted	Classification	15 thoracic findings	0	1,397
ReXGradient-160K	Adapted	Report generation	–	140,000	10,000
Unseen-domain evaluation					
SIIM-PTX	Non-adapted	Classification	Pneumothorax	0	1,372
	Adapted	Segmentation	Pneumothorax	10,675 (1–100%)	1,372
Pneumonia2017	Non-adapted	Classification	Pediatric pneumonia	0	624
TBX11K	Non-adapted	Classification	Tuberculosis	0	1,800
JSRT	Adapted	Segmentation	Lungs	172 (1–100%)	50
VinDr-RibCXR	Adapted	Segmentation	Ribs	196 (1–100%)	49

4.4 Representation-guided data curation

Our CheXficient adopts a contrastive language–image pretraining framework based on CLIP [23] and actively curates informative and diverse training samples from a large-scale corpus containing over 1.235M paired CXR image–report samples. Unlike conventional pretraining pipelines that treat all samples equally, CheXficient performs data curation to identify samples that are most beneficial for the representation learning. This strategy enables efficient utilization of large-scale medical datasets, which are often redundant, imbalanced, and noisy, while substantially reducing computational cost. Formally, we consider a large image–text dataset $\mathcal{D} = \{(x_i^{\text{img}}, x_i^{\text{txt}})\}_{i=1}^{|\mathcal{D}|}$, where each data pair consists of a CXR image x_i^{img} and its corresponding textual radiology report x_i^{txt}. Our objective is to construct a compact yet representative subset $\mathcal{D}_c \subset \mathcal{D}$ during the early stage of pretraining, such that a vision–language model subsequently trained on $\mathcal{D}_c$ achieves performance comparable to training on the full dataset $\mathcal{D}$, but with considerably fewer samples and lower training cost.

The detailed data curation and training procedure consists of the following steps:

(1) **Multimodal embedding.** We randomly sample a large super-batch of paired CXR data $\mathcal{S} = \{(x_i^{\text{img}}, x_i^{\text{txt}})\}_{i=1}^{|\mathcal{S}|}$ from the full pretraining set $\mathcal{D}$. Each radiology report x_i^{txt} is tokenized and encoded by the text encoder θ_{txt} to obtain a text embedding $\mathbf{z}_i^{\text{txt}}$. Correspondingly, each CXR image x_i^{img} is resized and processed by the image encoder θ_{img} to extract an image embedding $\mathbf{z}_i^{\text{img}}$ of the same dimensionality. Both embeddings are ℓ_2-normalized to ensure scale consistency and then concatenated to form a unified multimodal representation $\mathbf{z}_i = \mathsf{concat}(\mathbf{z}_i^{\text{img}}, \mathbf{z}_i^{\text{txt}})$, capturing complementary semantic information from both visual and textual modalities (see ablations in Appendix B.2).

(2) **Prototype-driven selection.** The model maintains a set of evolving prototypes $\{\mathbf{p}_k\}_{k=1}^{K}$ that characterize representative multimodal patterns in the global data manifold (Figure 1(b)). Given the multimodal embedding $\mathbf{z}_i$ of each image-report pair in the super-batch $\mathcal{S}$, we assign it to its nearest prototype by:

$$d_i = \min_{k \in 1,\ldots,K} \mathsf{dist}(\mathbf{z}_i, \mathbf{p}_k), \tag{1}$$

where d_i denotes the distance to the assigned prototype.

For each prototype $\mathbf{p}_k$, let $\mathcal{S}_k$ be the set of samples assigned to it. We compute the mean and standard deviation of the within-prototype distances as:

$$\mu_k = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} d_i, \quad \sigma_k = \sqrt{\frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} (d_i - \mu_k)^2}. \tag{2}$$

Rather than using fixed global thresholding, we perform a Gaussian-inspired adaptive stratification within each prototype cluster, partitioning the samples assigned to prototype $\mathbf{p}_k$ into central, peripheral, tail, and outlier regions, denoted by $\mathcal{C}_k$, $\mathcal{P}_k$, $\mathcal{T}_k$, and $\mathcal{O}_k$, respectively:

$$\begin{aligned}
\mathcal{C}_k &= \{i \in \mathcal{S}_k \mid 0 \le d_i \le \mu_k + \sigma_k\}, \\
\mathcal{P}_k &= \{i \in \mathcal{S}_k \mid \mu_k + \sigma_k < d_i \le \mu_k + 2\sigma_k\}, \\
\mathcal{T}_k &= \{i \in \mathcal{S}_k \mid \mu_k + 2\sigma_k < d_i \le \mu_k + 3\sigma_k\}, \\
\mathcal{O}_k &= \{i \in \mathcal{S}_k \mid d_i > \mu_k + 3\sigma_k\}.
\end{aligned} \tag{3}$$

This adaptive stratification allows the selection criterion to adjust to the local density and scale of each prototype cluster. The central region contains samples close to the prototype and is typically dominated by common and redundant patterns. The peripheral region contains moderately distant samples that provide harder and more informative training signals. The tail region contains samples far from the prototype but not extreme; these samples are likely to correspond to rare pathologies, atypical imaging appearances, or underrepresented report semantics. Finally, samples in the outlier region are treated as potential noise, e.g., corrupted images, severe artifacts, report translation errors, image–report mismatches, or template-generated reports with weak visual grounding. We retain all samples in $\mathcal{T}_k$ to preserve long-tailed knowledge

and discard all samples in $\mathcal{O}_k$. For the central region $\mathcal{C}_k$ and peripheral region $\mathcal{P}_k$, we allocate a compact sampling budget B_k controlled by a curation ratio r:

$$B_k = \lfloor r \cdot (|\mathcal{C}_k| + |\mathcal{P}_k|) \rfloor . \tag{4}$$

We then design a generalized power allocation rule to distribute the budget B_k between the two regions:

$$B_k^c = \left\lfloor B_k \frac{|\mathcal{C}_k|^\alpha}{|\mathcal{C}_k|^\alpha + |\mathcal{P}_k|^\alpha} \right\rfloor, \quad B_k^p = B_k - B_k^c, \tag{5}$$

where $\alpha \in [0, 1]$ controls the allocation sharpness. Within both $\mathcal{C}_k$ and $\mathcal{P}_k$, we apply a farthest point sampling [69] in the multimodal embedding space to select B_k^c and B_k^p samples, respectively. This encourages geometric coverage within each region and avoids repeatedly selecting near-duplicate samples.

(3) **Contrastive learning.** The selected samples from the central, peripheral, and tail regions constitute a mini-batch $\mathcal{M} \subset \mathcal{S}$, which is added to the curated subset $\mathcal{D}_c$. Meanwhile, the mini-batch $\mathcal{M}$ is used to optimize via the InfoNCE contrastive loss [22], updating the image and text encoders jointly. This objective encourages matched CXR image–report pairs to have aligned representations, while pushing apart mismatched pairs, thereby improving cross-modal alignment and semantic consistency.

(4) **Prototype update.** The prototypes are updated concurrently with model training to reflect the evolving data distribution. Specifically, new prototype representations are computed as weighted averages of the concatenated embeddings from the selected samples, where the weight is determined by each sample's embedding proximity to the corresponding prototype. We achieve this via an optimal transport based clustering algorithm [70] that can enforce unique assignment and equipartition constraints across prototypes. To further stabilize prototype estimation and incorporate historical information, we apply an exponential moving average (EMA) [71] update scheme.

These above steps are iteratively executed throughout the early data-curated pretraining stage to form the curated subset $\mathcal{D}_c$, enabling the model to progressively identify and emphasize informative, diverse, and under-represented samples. The data-curated pretraining procedure is summarized in Extended Data Algorithm 1. During the later standard pretraining stages, $\mathcal{D}_c$ is fixed and used for improving efficiency.

4.5 Implementation details

CheXficient, CheXfull, and CheXrandom share the same model architecture, training configuration, and hardware setup, and all models are trained for a total of 20 epochs. The architecture follows a specific instantiation of the CLIP framework, comprising an image encoder and a text encoder. We adopt DINOv2 [39] as the image encoder, using the [CLS] token representation, and BioClinicalBERT [40] as the text encoder. Both encoders are followed by a linear layer that maps image and text features into a shared

embedding space of 512 dimensions. DINOv2 is implemented as ViT-Base [72] with 12 transformer blocks (approximately 86 million parameters), while BioClinicalBERT consists of 12 transformer layers with about 110 million parameters.

CXR images are resized to 378×378 before being fed into the image encoder. No data augmentation is applied to either image or text inputs. For the textual modality, the *Findings* and *Impression* sections of each radiology report are combined (when both are available) and used as input to the text encoder, with a maximum context length of 256 tokens. Model training is driven by an InfoNCE contrastive loss with a learnable temperature parameter, which is initialized to 0.01. The InfoNCE loss encourages alignment between paired CXR images and reports while separating mismatched pairs. The number of prototypes $K = 32$, with a sensitivity analysis provided in Appendix B.3. Before starting the curation process, we use 1% of pretraining data for model warm-up, which already contains sufficient samples at the million-scale, and initialize the prototypes using k-means centroids [73] computed from these warm-up samples. In the EMA update of prototypes, a smoothing factor of 0.1 is applied. During the first training epoch, the early data-curated pretraining stage is performed to build the curated subset, after which the model's training uses the resulting fixed curated subset. An analysis of repeated curation at later training stages shows high subset stability and limited additional performance benefit, while substantially increasing cumulative data usage and training time (Appendix A.1). Optimization is conducted using the AdamW optimizer with an initial learning rate of 5×10^{-5}, a weight decay of 1×10^{-4}, and a cosine annealing learning rate schedule. Curation ratio $r = 0.2$, and $\alpha = 0.5$ in Eq. (5). Each super-batch contains 1920 paired samples (8 GPUs $\times$ 240 data pairs per GPU). Notably, we find that CheXficient can also be trained on a single GPU by leveraging an embedding accumulation strategy, in which samples are repeatedly drawn from the pretraining corpus, their feature embeddings are extracted sequentially, and then accumulated to construct a large super-batch. This strategy can effectively enlarge the super-batch size without increasing the GPU count, enabling memory-efficient training while preserving the benefits of large-batch curation.

In fact, CheXficient introduces negligible additional pretraining overhead in comparison to each base model. The potentially increased overhead involves two major steps: 1) sample-prototype distance calculation: candidate training samples are drawn from in-batch data, which typically consists of only a few thousand samples, enabling efficient distance calculations to prototypes in the embedding space; 2) prototype update: new prototypes are obtained as weighted averages of embeddings from the selected samples and updated using an EMA scheme. This operation incurs minimal computational cost, as confirmed by our empirical observations.

4.6 Evaluation metrics

Classification performance is evaluated using well-established metrics widely adopted in prior studies [7, 49, 59], including the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC). AUROC characterizes the trade-off between the true positive rate (recall) and the false positive rate across varying decision thresholds, while AUPRC captures the trade-off between precision and recall. For multi-disease classification, AUROC and AUPRC

are computed independently for each disease category and then averaged to obtain the macro AUROC and macro AUPRC. Evaluation of retrieval tasks leverages the metric of Recall@1, which measures whether the exact paired radiology report is retrieved within the top-1 result for a given CXR image, and vice versa. Segmentation performance is evaluated using the Dice score, which quantifies pixel-wise overlap between the predicted segmentation maps and the corresponding ground-truth annotations. Performance of radiology report generation is evaluated using standard lexical metrics (BLEU-2, BLEU-4, and Rouge-L) which measure word-level overlap between the generated and reference reports, general semantic similarity metrics (BERTScore) that measure contextual embedding-level semantic similarity between generated and reference reports, and radiology-specific metrics (RadGraph, Macro-F1-14, RaTEScore, and RadCliQ) which assess the clinical correctness and factual consistency of the generated reports.

4.7 Statistical analysis

To guarantee reliability and robustness of this study, we trained the model with five different random seeds and calculated the mean and standard deviation of the performance metrics, from which the standard error is computed as the standard deviation divided by $\sqrt{5}$. The corresponding normal-approximation 95% confidence interval (CI) is estimated as 1.96 $\times$ the standard error. We perform two-sided paired t-tests to evaluate the statistical significance of performance differences between CheXficient and CheXfull, where metrics were computed across repeated experiments. A p-value less than 0.05 indicates a statistically significant difference, whereas a p-value greater than or equal to 0.05 suggests no statistically significant differences between them.

4.8 Computation resources

All experiments were conducted on NVIDIA H100 GPUs with 80 GB memory. CheXficient required approximately 60 H100 GPU-hours for curation and pretraining on 280K paired CXR data, whereas CheXfull required about 220 H100 GPU-hours to train on 1,235K paired CXR data. For downstream fine-tuning, CheXficient used one H100 GPU for classification and segmentation tasks, and four H100 GPUs for radiology report generation. All pretraining and fine-tuning procedures were carefully monitored to ensure adequate convergence.

5 Data availability

All data used in this study are publicly available: MIMIC-CXR (https://physionet.org/content/mimic-cxr-jpg/2.0.0/); CheXpert (https://stanfordmlgroup.github.io/competitions/chexpert/); CheXpert-Plus (https://aimi.stanford.edu/datasets/chexpert-plus); ReXGradient-160K (https://huggingface.co/datasets/rajpurkarlab/ReXGradient-160K); PadChest (https://bimcv.cipf.es/bimcv-projects/padchest/); BIMCV-COVID19 (https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/); CANDID-PTX (https://figshare.com/articles/dataset/CANDID-PTX/14173982); CASIA-CXR (https://www.casia-cxr.net/); Open-I (https://openi.nlm.nih.gov/faq);

NIH ChestX-ray14 (https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345); BRAX (https://physionet.org/content/brax/1.1.0/); VinDr-CXR (https://vindr.ai/cxr); VinDr-PCXR (https://vindr.ai/datasets/pediatric-chest-x-ray); ChestDR (https://springernature.figshare.com/articles/dataset/ChestDR_Thoracic_Diseases_Screening_in_Chest_Radiography/22302775); SIIM-PTX (www.kaggle.com/c/siim-acr-pneumothorax-segmentation); Pneumonia2017 (https://data.mendeley.com/datasets/rscbjbr9sj/2); TBX11K (https://datasetninja.com/tbx-11k); JSRT (http://db.jsrt.or.jp/eng.php); VinDr-RibCXR (https://vindr.ai/ribcxr).

6 Code availability

We have publicly released the data curation, pretraining, and evaluation code, along with the pretrained model weights, on GitHub at https://github.com/cwangrun/CheXficient. All experimental procedures are described in detail in the Methods section to support independent replication. To facilitate broader adoption, we also provide standardized downstream modeling tasks that are accessible to a wide scientific audience. Together, these resources are made available to ensure transparency and to promote further research in this area.

Acknowledgements

This work was supported in part by the Medical Imaging and Data Resource Center (MIDRC), which is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under contract 75N92020C00021 and through the Advanced Research Projects Agency for Health (ARPA-H).

Author contributions

Conceptualization: C.W., Y.Z., A.S.C., C.P.L.; Methodology: C.W., Y.Z. Y.G., M.V., F.W., J.Liu. A.S.C., C.P.L.; AI model and code development: C.W., Y.Z.; Dataset development: C.W., Y.Z., Y.G.; Data analysis: C.W., Y.Z., Y.G., C.M., J.Long.; Writing - original draft, review, and editing: C.W., Y.Z. Y.G., M.V., C.M., F.W., J.Liu., F.P, J.Long., J.B.D, S.G., A.S.C., C.P.L.; Supervision: S.G., A.S.C., C.P.L.; Project administration: C.W., Y.Z., A.S.C., C.P.L.; Funding acquisition: A.S.C., C.P.L..

Competing interests

J.B.D. is an employee of hoppr.ai. The remaining authors declare no competing interests.

References

[1] Bommasani, R.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)

[2] Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)

[3] Wiggins, W.F., Tejani, A.S.: On the opportunities and risks of foundation models for natural language processing in radiology. Radiology: Artificial Intelligence **4**(4), 220119 (2022)

[4] Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., *et al.*: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)

[5] Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z., Yang, L., Tschandl, P., Hu, M., Ju, L., Tan, G., et al.: A multimodal vision foundation model for clinical dermatology. Nature Medicine, 1–12 (2025)

[6] Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., *et al.*: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)

[7] 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**(12), 1399–1406 (2022)

[8] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)

[9] Lozano, A., Sun, M.W., Burgess, J., Chen, L., Nirschl, J.J., Gu, J., Lopez, I., Aklilu, J., Rau, A., Katzer, A.W., *et al.*: Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 19724–19735 (2025)

[10] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

[11] He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H.: Foundation model for advancing healthcare: Challenges, opportunities and future directions. IEEE Reviews in Biomedical Engineering (2024)

[12] Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nature medicine **28**(9), 1773–1784 (2022)

[13] Cid, Y.D., Macpherson, M., Gervais-Andre, L., Zhu, Y., Franco, G., Santeramo, R., Lim, C., Selby, I., Muthuswamy, K., Amlani, A., *et al.*: Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. The Lancet Digital Health **6**(1), 44–57 (2024)

[14] Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021)

[15] Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)

[16] Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)

[17] Joshi, S., Jain, A., Payani, A., Mirzasoleiman, B.: Data-efficient contrastive language-image pretraining: Prioritizing data quality over quantity. In: International Conference on Artificial Intelligence and Statistics, pp. 1000–1008 (2024). PMLR

[18] Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., *et al.*: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36**, 27092–27112 (2023)

[19] Mindermann, S., Brauner, J.M., Razzak, M.T., Sharma, M., Kirsch, A., Xu, W., Höltgen, B., Gomez, A.N., Morisot, A., Farquhar, S., *et al.*: Prioritized training on points that are learnable, worth learning, and not yet learnt. In: International Conference on Machine Learning, pp. 15630–15649 (2022). PMLR

[20] Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., Morcos, A.: Beyond neural scaling laws: beating power law scaling via data pruning. Advances in Neural Information Processing Systems **35**, 19523–19536 (2022)

[21] Goyal, S., Maini, P., Lipton, Z.C., Raghunathan, A., Kolter, J.Z.: Scaling laws for data filtering–data curation cannot be compute agnostic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22702–22711 (2024)

[22] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

[23] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G.,

Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PmLR

[24] Zhou, T., Wang, W., Konukoglu, E., Van Gool, L.: Rethinking semantic segmentation: A prototype view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2582–2593 (2022)

[25] Wang, C., Chen, Y., Liu, F., Liu, Y., McCarthy, D.J., Frazer, H., Carneiro, G.: Mixture of gaussian-distributed prototypes with generative modelling for interpretable and trustworthy image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

[26] Chambon, P., Delbrouck, J.-B., Sounack, T., Huang, S.-C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024)

[27] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

[28] Zhang, X., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. arXiv preprint arXiv:2505.00228 (2025)

[29] Bustos, A., Pertusa, A., Salinas, J.-M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)

[30] Vayá, M.D.L.I., Saborit, J.M., Montell, J.A., Pertusa, A., Bustos, A., Cazorla, M., Galant, J., Barber, X., Orozco-Beltrán, D., García-García, F., et al.: Bimcv covid-19+: A large annotated dataset of rx and ct images from covid-19 patients. arXiv preprint arXiv:2006.01174 (2020)

[31] Feng, S., Azzollini, D., Kim, J.S., Jin, C.-K., Gordon, S.P., Yeoh, J., Kim, E., Han, M., Lee, A., Patel, A., *et al.*: Curation of the candid-ptx dataset with free-text reports. Radiology: Artificial Intelligence **3**(6), 210136 (2021)

[32] Metmer, H., Yang, X.: An open chest x-ray dataset with benchmarks for automatic radiology report generation in french. Neurocomputing **609**, 128478 (2024)

[33] Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical

Informatics Association **23**(2), 304–310 (2015)

[34] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)

[35] Reis, E.P., De Paiva, J.P., Da Silva, M.C., Ribeiro, G.A., Paiva, V.F., Bulgarelli, L., Lee, H.M., Santos, P.V., Brito, V.M., Amaral, L.T., *et al.*: Brax, brazilian labeled chest x-ray dataset. Scientific Data **9**(1), 487 (2022)

[36] Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., *et al.*: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)

[37] Pham, H.H., Nguyen, N.H., Tran, T.T., Nguyen, T.N., Nguyen, H.Q.: Pedicxr: an open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. Scientific Data **10**(1), 240 (2023)

[38] Wang, D., Wang, X., Wang, L., Li, M., Da, Q., Liu, X., Gao, X., Shen, J., He, J., Shen, T., *et al.*: A real-world dataset and benchmark for foundation model adaptation in medical image classification. Scientific Data **10**(1), 574 (2023)

[39] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[40] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)

[41] Zawacki, A., Wu, C., Shih, G., Elliott, J., Fomitchev, M., Hussain, M., Lakhani, P., Culliton, P., Bao, S.: SIIM-ACR Pneumothorax Segmentation 2019. https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation/

[42] Kermany, D.: Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley data (2018)

[43] Liu, Y., Wu, Y.-H., Ban, Y., Wang, H., Cheng, M.-M.: Rethinking computer-aided tuberculosis diagnosis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2646–2655 (2020)

[44] Ma, D., Pang, J., Gotway, M.B., Liang, J.: A fully open ai foundation model applied to chest radiography. Nature, 1–11 (2025)

[45] Wang, C., Liu, F., Chen, Y., Frazer, H., Carneiro, G.: Cross-and intra-image prototypical learning for multi-label disease diagnosis and interpretation. IEEE

Transactions on Medical Imaging (2025)

[46] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y., Doi, K.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American journal of roentgenology **174**(1), 71–74 (2000)

[47] 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 arXiv:2107.01327 (2021)

[48] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[49] Chen, Z., Varma, M., Delbrouck, J.-B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: Chexagent: Towards a foundation model for chest x-ray interpretation. arXiv preprint arXiv:2401.12208 (2024)

[50] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

[51] Zambrano Chaves, J.M., Huang, S.-C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., *et al.*: A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. Nature Communications **16**(1), 3108 (2025)

[52] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., *et al.*: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) **2**(3), 6 (2023)

[53] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

[54] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[55] Delbrouck, J.-B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., Langlotz, C.: Improving the factual correctness of radiology report generation with semantic rewards. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 4348–4360 (2022)

[56] Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N.,

Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023)

[57] Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Ratescore: A metric for radiology report generation. arXiv preprint arXiv:2406.16845 (2024)

[58] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)

[59] Perez-Garcia, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., *et al.*: Exploring scalable medical image encoders beyond text supervision. Nature Machine Intelligence **7**(1), 119–130 (2025)

[60] Zhang, X., Meng, Z., Lever, J., Ho, E.S.: Libra: Leveraging temporal images for biomedical radiology analysis. In: Findings of the Association for Computational Linguistics: ACL 2025, pp. 17275–17303 (2025)

[61] Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)

[62] Zhou, H.-Y., Acosta, J.N., Adithan, S., Datta, S., Topol, E.J., Rajpurkar, P.: Medversa: A generalist foundation model for medical image interpretation. arXiv preprint arXiv:2405.07988 (2024)

[63] Bluethgen, C., Chambon, P., Delbrouck, J.-B., Van Der Sluijs, R., Połacin, M., Zambrano Chaves, J.M., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.S.: A vision–language foundation model for the generation of realistic chest x-ray images. Nature Biomedical Engineering **9**(4), 494–506 (2025)

[64] Sun, Y., Tan, W., Gu, Z., He, R., Chen, S., Pang, M., Yan, B.: A data-efficient strategy for building high-performing medical foundation models. Nature Biomedical Engineering, 1–13 (2025)

[65] Liu, B., Zou, K., Zhan, L.-M., Lu, Z., Dong, X., Chen, Y., Xie, C., Cao, J., Wu, X.-M., Fu, H.: Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21310–21320 (2025)

[66] Huy, T.D., Huynh, D.A., Xie, Y., Qi, Y., Chen, Q., Le Nguyen, P., Tran, S.K., Phung, S.L., Hengel, A., Liao, Z., *et al.*: Seeing the trees for the forest: rethinking weakly-supervised medical visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 24445–24455 (2025)

[67] Huang, S.-C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951 (2021)

[68] You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 101–111 (2023). Springer

[69] Eldar, Y., Lindenbaum, M., Porat, M., Zeevi, Y.Y.: The farthest point strategy for progressive image sampling. IEEE transactions on image processing **6**(9), 1305–1315 (1997)

[70] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)

[71] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)

[72] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[73] Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics) **28**(1), 100–108 (1979)

[74] Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., *et al.*: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15016–15027 (2023)

[75] Paul, M., Ganguli, S., Dziugaite, G.K.: Deep learning on a data diet: Finding important examples early in training. Advances in neural information processing systems **34**, 20596–20607 (2021)

[76] Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems **33**, 20331–20342 (2020)

[77] Hashir, M., Bertrand, H., Cohen, J.P.: Quantifying the value of lateral views in deep learning for chest x-rays. In: Medical Imaging with Deep Learning, pp. 288–303 (2020). PMLR

[78] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations, pp. 1–13 (2018)

[79] Mirzasoleiman, B., Bilmes, J., Leskovec, J.: Coresets for data-efficient training of machine learning models. In: International Conference on Machine Learning, pp. 6950–6960 (2020). PMLR

[80] Xu, H., Xie, S., Huang, P.-Y., Yu, L., Howes, R., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Cit: Curation in training for effective vision-language data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15180–15189 (2023)

[81] Kim, C.D., Moon, J., Moon, S., Yun, H., Lee, S., Kembhavi, A., Lee, S., Kim, G., Lee, S., Clark, C.: Respec: Relevance and specificity grounded online filtering for learning on video-text data streams. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 29040–29049 (2025)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf