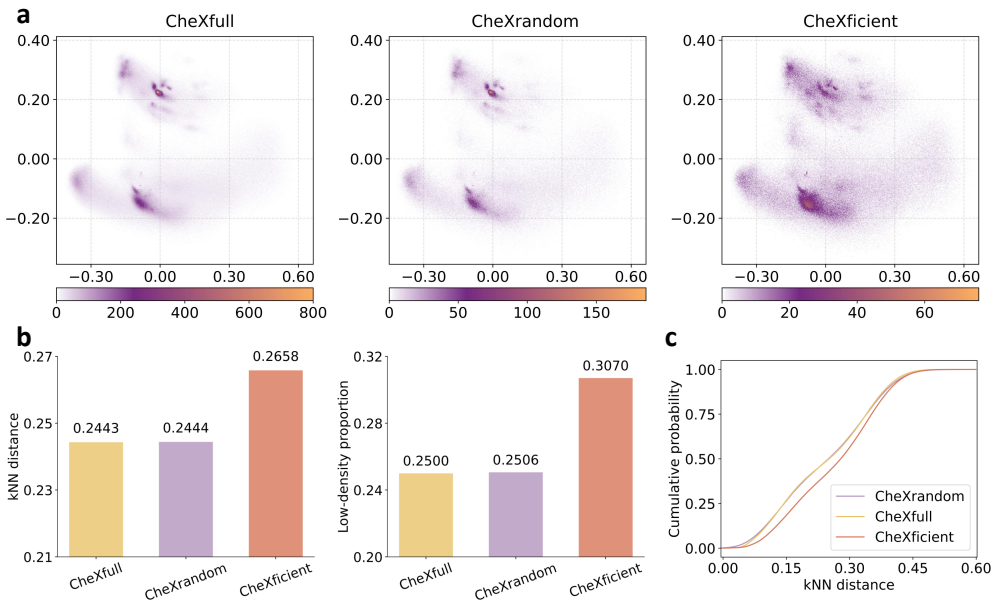
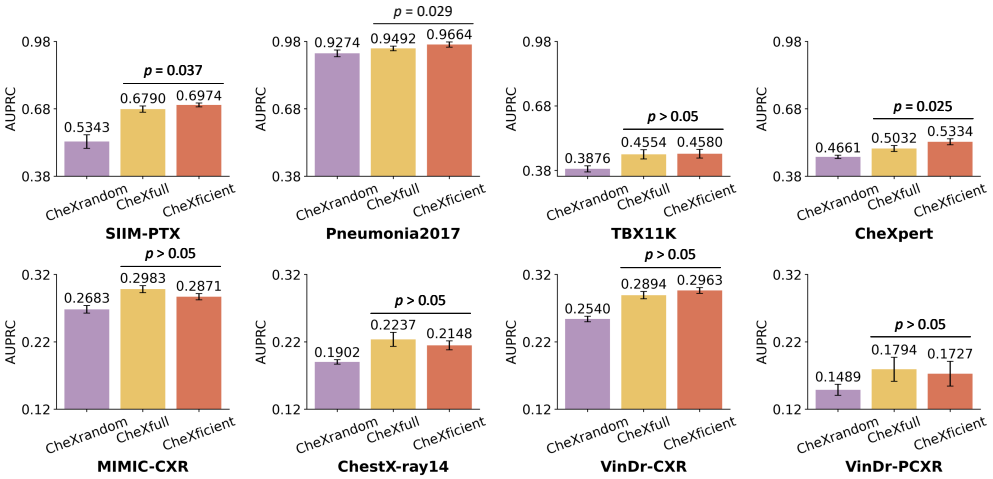
**Extended Data Table 1** Comparative overview of CheXficient with other large-scale pretrained medical foundation models.

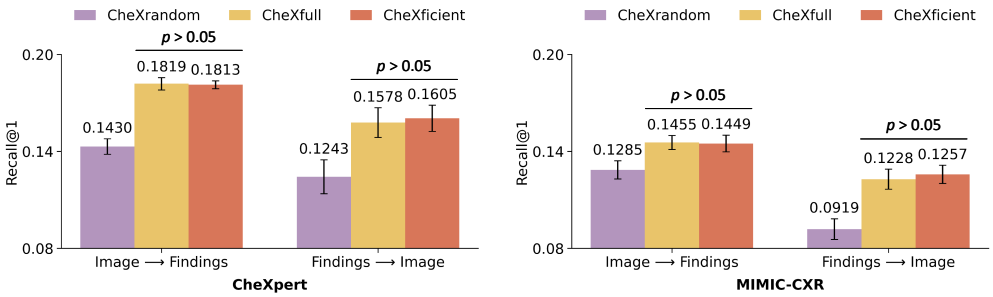
| Model | Image encoder | Parameter size | Image resolution | Pretraining data size | Pretraining type | Data accessibility |
|---|---|---|---|---|---|---|
| CheXficient | ViT-Base | 86 M | $378 \times 378$ | 280 K | image-text | public |
| CheXfull | ViT-Base | 86 M | $378 \times 378$ | 1,235 K | image-text | public |
| CheXZero [7] | ViT-Base | 86 M | $224 \times 224$ | 377 K | image-text | public |
| BioViL-T [74] | ResNet-50 | 25.6 M | $512 \times 512$ | 377 K | image-text | public |
| LLaVA-Rad [51] | ViT-Base | 86 M | $518 \times 518$ | 697 K | image-text | public |
| RAD-DINO [59] | ViT-Base | 86 M | $518 \times 518$ | 883 K | self-supervised | public + private |
| CheXagent [49] | ViT-Large | 307 M | $512 \times 512$ | 1,077 K | image-text | public |
| Libra [60] | ViT-Base | 86 M | $518 \times 518$ | 1,213 K | self-supervised | public |
| MAIRA-2 [61] | ViT-Base | 86 M | $518 \times 518$ | 1,404 K | self-supervised | public + private |
| BiomedCLIP [8] | ViT-Base | 86 M | $224 \times 224$ | 15 M | image-text | private |
| RadFM [15] | 12-layer 3D ViT | 90 M | $512 \times 512$ | 16 M | image-text | public + private |
| MedVersa [62] | Swin Transformer | 88 M | $224 \times 224$ | 29 M | image-text | public |
| MedGemma [16] | SoViT-400m | 400 M | $448 \times 448$ | 33 M | image-text | public + private |



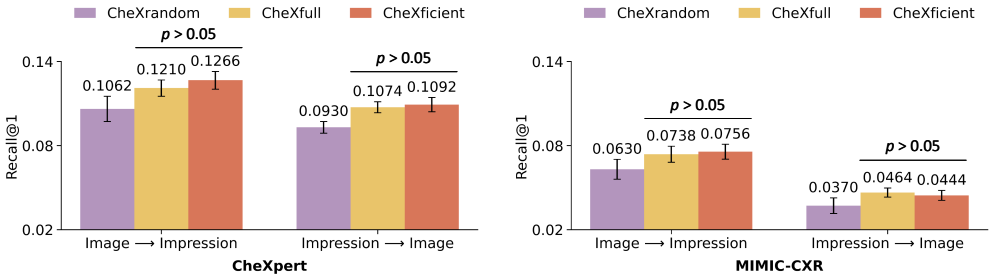

**Extended Data Fig. 1** (a) Feature distribution of training samples from the full set (CheXfull, $n$ = 1,235K), random subset (CheXrandom, $n$ = 280K), and curated subset (CheXficient, $n$ = 280K). Features are extracted by CheXficient and projected onto a two-dimensional PCA space for visualization (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).



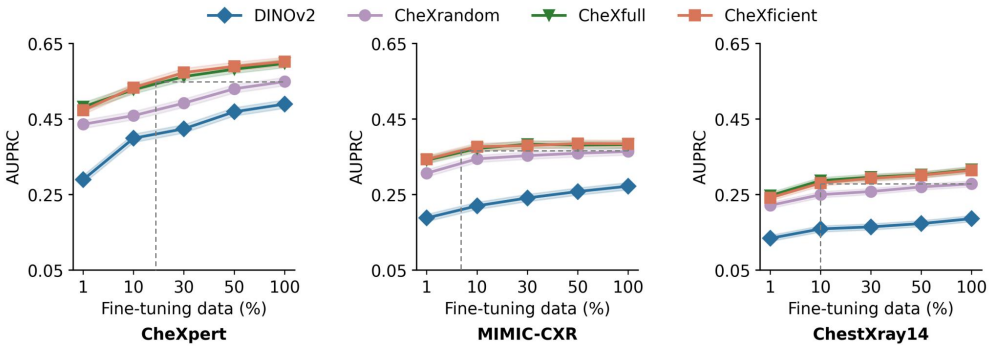

**Extended Data Fig. 2** Performance (AUPRC) on zero-shot findings classification. We evaluate the pretrained models on 8 public datasets. Among them, SIIM-PTX, Pneumonia2017, and TBX11K are from unseen domains not included in pretraining, while the remaining 5 datasets are used for internal evaluation. Compared to CheXfull, CheXficient achieves higher AUPRC on 3 datasets ($p < 0.05$) while no statistically significant differences are detected on the remaining 5 datasets ($p > 0.05$). For each dataset, CheXficient outperforms CheXrandom by large margins. We present the mean ± 95% confidence interval (CI) of AUPRC for each model. The listed $p$-values are calculated using two-sided $t$-tests.




**Extended Data Fig. 3** Performance on zero-shot cross-modal retrieval (Image → Findings and Findings → Image). 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 Findings section (or image) within the top-1 result for a given CXR image (or Findings section)).




**Extended Data Fig. 4** Performance on zero-shot cross-modal retrieval (Image → Impression and Impression → Image). We evaluate the pretrained models on the public CheXpert and MIMIC-CXR benchmarks. CheXficient shows no 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 Impression section (or image) within the top-1 result for a given CXR image (or Impression section)).




**Extended Data Fig. 5** Performance (AUPRC) 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.

**Algorithm 1:** Data-curated Pretraining Stage of CheXficient

**Data:** Pretraining set $\mathcal{D} = \{(x_i^{\text{img}}, x_i^{\text{txt}})\}_{i=1}^{|\mathcal{D}|}$, number of training iterations $T$ for curation, super-batch size $|\mathcal{S}|$, mini-batch size $|\mathcal{M}|$.

**Result:** Image encoder $\theta_{\text{img}}$, text encoder $\theta_{\text{txt}}$, curation set $\mathcal{D}_c$.

Initialize $\mathcal{D}_c = \emptyset$, warm-up model, initialize prototypes $\{\mathbf{p}_k\}_{k=1}^{K}$ by k-means;

**for** *iteration* $t = 1$ **to** $T$ **do**

- Given a large super-batch $\mathcal{S}$ sampled from $\mathcal{D}$;
- Extract unified embeddings $\mathbf{z}_i = \mathsf{concat}(\mathbf{z}_i^{\text{img}}, \mathbf{z}_i^{\text{txt}})$ for each sample in $\mathcal{S}$;
- Compute the distance of each $\mathbf{z}_i$ to its nearest prototype: $\min_k \mathsf{dist}(\mathbf{z}_i, \mathbf{p}_k)$;
- Select samples based on prototype distances and form the mini-batch $\mathcal{M}$;
- Contrastive pretraining on $\mathcal{M}$ with InfoNCE loss;
- Calculate prototypes based on embeddings of the selected samples in $\mathcal{M}$;
- Update prototypes $\{\mathbf{p}_k\}_{k=1}^{K}$ via exponential moving average;
- Add selected samples into the curation set: $\mathcal{D}_c = \mathcal{D}_c \cup \mathcal{M}$.

# Appendix A Additional analysis

To further investigate the effect of data curation in medical foundation model pre-training, we additionally analyze curation stability, data source composition, patient demographic information (e.g., age and race), distribution of frontal and lateral CXR scans, and distribution of token length of CXR reports.

## A.1 Stability of early-stage curation

We first investigate the stability of the early-stage curated subset by examining whether the curated subset changes substantially during subsequent optimization. To do so, we repeat the curation procedure at multiple training epochs and quantify the overlap degree between the subset curated in the first round and those curated in subsequent rounds: $\text{overlap}(n) = \frac{|\mathcal{D}_c^{(1)} \cap \mathcal{D}_c^{(n)}|}{|\mathcal{D}_c^{(1)}|}$, where $\mathcal{D}_c^{(n)}$ denotes the subset curated at the $n$-th round. As shown in Figure A1, the curated subsets remain highly stable across curation rounds, with overlap degree remaining above 0.85 even after 5 rounds of re-curation and plateauing thereafter. This suggests that samples prioritized by CheXficient's curation emerge early and remain largely stable throughout subsequent optimization. Repeated curation produces only marginal improvements in zero-shot findings classification performance, measured by the average AUROC across eight datasets. In contrast, repeated curation substantially increases both pretraining time and cumulative data usage. This observation is consistent with prior studies showing that sample importance can often be estimated reliably from early training dynamics and tend to stabilize rapidly during optimization [75, 76]. Taken together, these findings support the use of a single early-stage curation pass, which captures most of the useful training signal while preserving the data and compute efficiency of CheXficient.

## A.2 Data source composition

We examine the source composition of the curated subset to determine whether the observed data reduction is driven by the preferential exclusion of particular datasets. As shown in Figure A2, the curated subset exhibits a moderately increased contribution from MIMIC-CXR and CheXpert-Plus, two large-scale datasets containing high-quality CXR images and free-text radiology reports. In particular, all

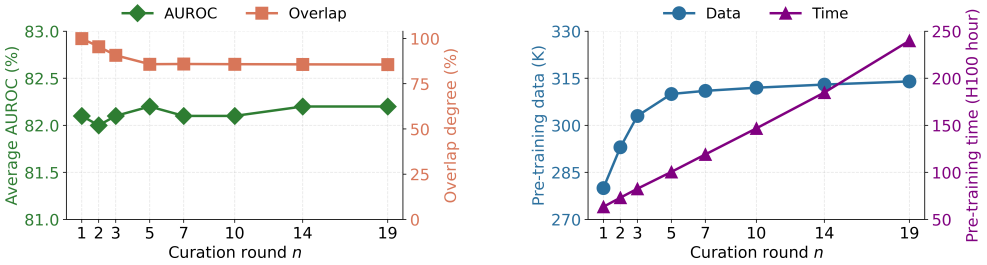



**Figure A1** Stability analysis of the curated subset in CheXficient.

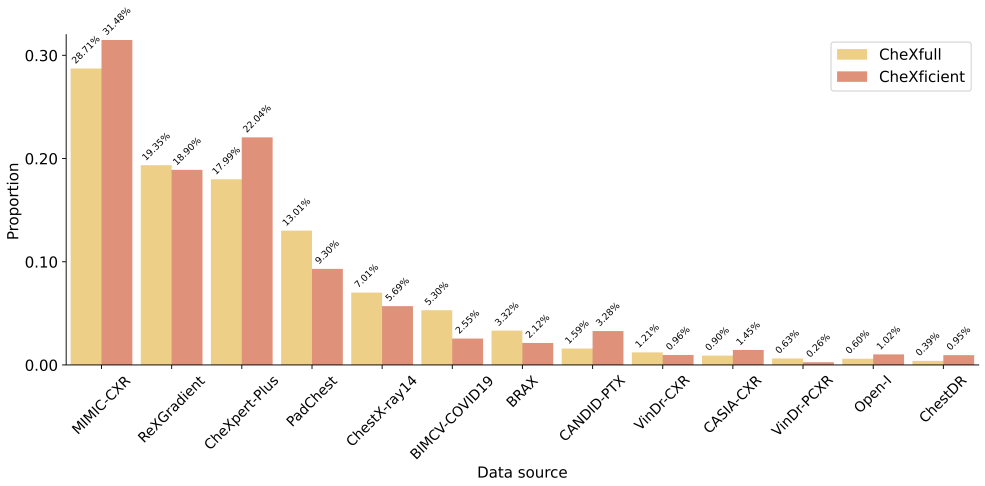



**Figure A2** Distribution of data sources in the full pretraining corpus (CheXfull) and the curated subset (CheXficient).

five template-generated report datasets (ChestX-ray14, BRAX, VinDr-CXR, VinDr-PCXR, and ChestDR) remain represented in the curated subset, indicating that the proposed curation strategy does not simply remove synthetic template-report data.

## A.3 Patient age

Patient demographic information is an important factor to consider when developing medical AI models. We examine how training data curation affects the distribution of patient age by comparing the full training set and the curated subset. Figure A3 shows that the curated subset tends to include relatively more samples from two age groups: pediatric patients (0–10 years old) and older adults (over 65 years old). This observation may be relevant because pediatric CXRs exhibit greater anatomical variability, and CXRs of older adults often show a wider range of age-related pathologies due to their higher prevalence in this population. Hence, the curated pretraining data emphasizes samples from these two age groups.

## A.4 Patient race

Understanding the distribution of patient race in the training data is important for assessing whether data curation may inadvertently alter demographic representation. We therefore analyze the race distribution in the CheXpert dataset, which provides self-reported race information, including White, Asian, Black, Pacific Islander, Native American, and Other. As shown in Figure A4, the racial composition of the curated subset closely mirrors that of the full training set, suggesting that the curation process does not substantially alter the demographic distribution of the pretraining data. These results indicate that the proposed curation strategy does not introduce obvious race imbalance at the dataset level.

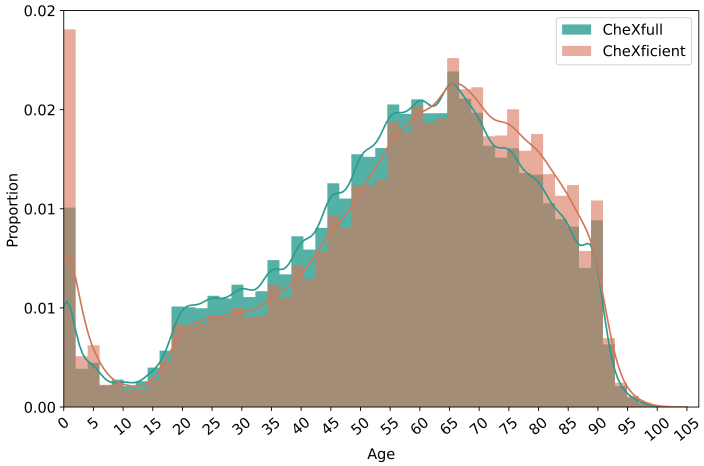



**Figure A3** Distribution of patient age in the full training set (CheXfull) and the curated subset (CheXficient). The curated subset includes relatively more samples from two age groups: pediatric patients (0–10 years old) and older adults (over 65 years old).

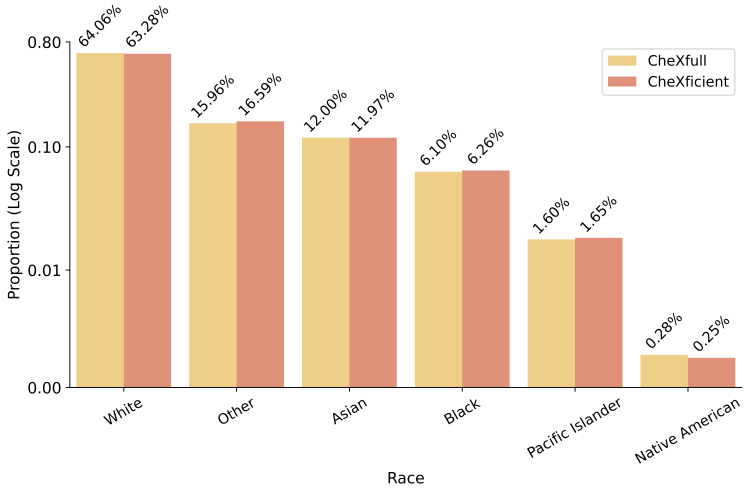



**Figure A4** Distribution of patient race in the full training set (CheXfull) and the curated subset (CheXficient). The two sets show similar race distributions.

We additionally conduct an exploratory race-stratified analysis on the CheXpert test set to examine whether data-efficient pretraining is associated with subgroup-specific changes in performance. We evaluate three representative thoracic findings: Enlarged Cardiomediastinum, Cardiomegaly, and Lung Opacity. For each subgroup, we calculate the AUROC performance difference between CheXfull and CheXficient, with 95% confidence intervals (CI) calculated from 1,000 patient-level paired bootstrap replicates. As shown in Table A1, across the nine finding–subgroup combinations, all confidence intervals for the AUROC differences include zero, and the analysis does not

**Table A1** Exploratory race-stratified subgroup performance analysis on the CheXpert test set. AUROC is reported for CheXfull and CheXficient across three thoracic findings. Total, positive, and negative sample counts are provided for each finding–subgroup combination. The AUROC difference is defined as $\Delta$ AUROC = $\text{AUROC}_{\text{CheXfull}} - \text{AUROC}_{\text{CheXficient}}$.

| Thoracic findings | Race | Total | Positive | Negative | $\text{AUROC}_{\text{CheXfull}}$ | $\text{AUROC}_{\text{CheXficient}}$ | $\Delta$ AUROC | 95% CI |
|---|---|---|---|---|---|---|---|---|
| Enlarged Cardiomediastinum | White | 277 | 140 | 137 | 0.8912 | 0.8971 | -0.0059 | [-0.0643, 0.0539] |
| | Asian | 58 | 27 | 31 | 0.9254 | 0.9295 | -0.0041 | [-0.0533, 0.0432] |
| | Black | 21 | 13 | 8 | 0.8321 | 0.8269 | +0.0052 | [-0.0498, 0.0628] |
| Cardiomegaly | White | 277 | 83 | 194 | 0.8919 | 0.8902 | +0.0017 | [-0.0221, 0.0234] |
| | Asian | 58 | 16 | 42 | 0.8852 | 0.8810 | +0.0042 | [-0.0501, 0.0594] |
| | Black | 21 | 8 | 13 | 0.8902 | 0.8942 | -0.0040 | [-0.0558, 0.0455] |
| Lung Opacity | White | 277 | 143 | 134 | 0.9088 | 0.9054 | +0.0034 | [-0.0489, 0.0581] |
| | Asian | 58 | 29 | 29 | 0.9572 | 0.9584 | -0.0012 | [-0.0328, 0.0289] |
| | Black | 21 | 10 | 11 | 0.8073 | 0.8091 | -0.0018 | [-0.0246, 0.0230] |

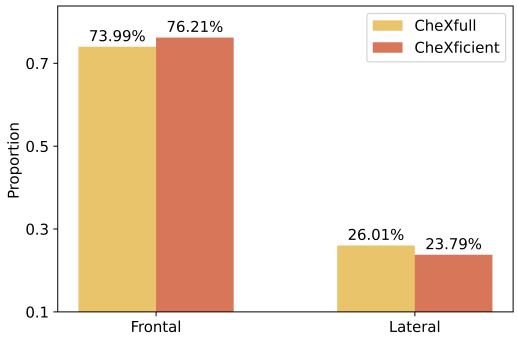



**Figure A5** Distribution of frontal and lateral scans in the full training set (CheXfull) and the curated subset (CheXficient). The curated subset contains a higher proportion of frontal scans in comparison with the full training set.

identify a consistent subgroup-specific reduction in performance associated with data curation. However, the Black subgroup includes only 21 examinations, and several finding–subgroup strata contain limited numbers of positive or negative cases, resulting in considerable statistical uncertainty. These exploratory findings should therefore not be interpreted as a comprehensive assessment of algorithmic fairness.

## A.5 Frontal and lateral CXR scans

Frontal and lateral chest X-rays (CXRs) are routinely acquired in clinical practice and are often jointly used to disambiguate radiological findings. We examine whether CheXficient exhibits a preference for either view during data curation. As shown in Figure A5, the curated subset contains a slightly higher proportion of frontal scans than the original dataset. One possible explanation is that frontal CXRs may provide more informative visual cues for many radiological findings, whereas certain findings are less clearly visible on lateral scans [77]. This observation suggests that the curation process tends to prioritize the informative frontal scans.

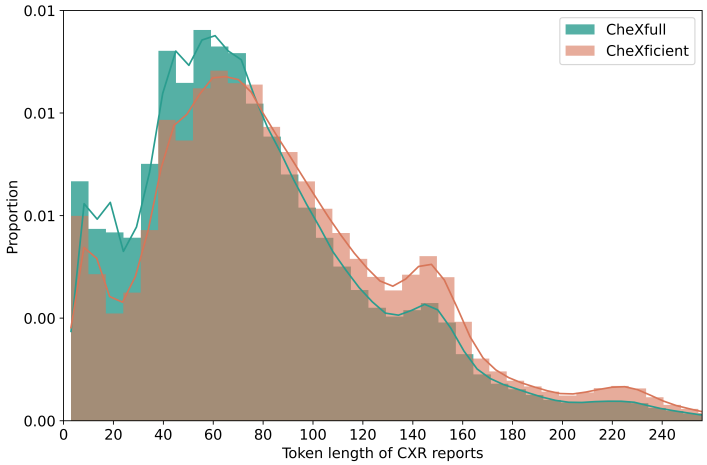



**Figure A6** Distribution of CXR report token lengths in the full training set (CheXfull) and the curated subset (CheXficient). The curated subset shows a higher proportion of longer CXR reports.

## A.6 Token length of CXR reports

We further analyze the token length of CXR reports to characterize the textual properties of the curated training samples. In general, CXR reports describing normal findings tend to be shorter, whereas those containing rich pathological findings, great diagnostic complexity or high reporting uncertainty require more detailed and longer descriptions. As shown in Figure A6, the curated subset exhibits a relatively higher proportion of longer CXR reports, suggesting that the data curation process may favor samples with more informative and complex textual descriptions.

# Appendix B Ablation study

To further evaluate CheXficient, we conduct additional analyses covering comparison with alternative data selection strategies (Section B.1), the contribution of its multimodal embedding design (Section B.2), and sensitivity to the number of prototypes (Section B.3).

## B.1 Comparison with alternative data selection strategies

We compare CheXficient with several representative data selection strategies spanning distinct curation principles: semantic-alignment-based filtering (CLIPScore [18]), geometric coverage (K-center [78]), submodular coreset selection (CRAIG [79]), downstream-relevance-based selection (CiT [80]), and hybrid filtering (ReSpec [81]). These methods were selected to cover complementary families of data curation approaches, ranging from sample-level quality or relevance scoring to representation-space coverage and coreset construction. Unless otherwise specified, all alternative strategies select 280K image–report pairs from the same 1.235-million-pair pretraining corpus and use the same model architecture, contrastive objective, optimization

**Table B1** Comparison of CheXficient with alternative data selection strategies under controlled pretraining settings. Unless otherwise specified, all methods select 280K image–report pairs from the same 1.235-million-pair pretraining corpus and use identical model architectures, pretraining objectives, optimization settings, and training epochs. Zero-shot findings classification performance is reported as mean AUROC across five independent runs on eight evaluation datasets.

| Method | Selection principle | Data | Compute | SIIM-PTX | Pneumonia2017 | TBX11K | CheXpert | MIMIC-CXR | ChestX-ray14 | VinDr-CXR | VinDr-PCXR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CheXfull | n.a. | 1,235K | 220.3h | 0.8970 | 0.9550 | **0.8313** | 0.8425 | **0.6878** | **0.7430** | 0.8896 | 0.7089 |
| CheXrandom | Random sampling | 280K | 60.2h | 0.8306 | 0.8626 | 0.7780 | 0.8099 | 0.6485 | 0.6928 | 0.8408 | 0.6942 |
| CLIPScore [18] | Semantic alignment | 280K | 60.2h | 0.7611 | 0.8221 | 0.6763 | 0.7844 | 0.6401 | 0.6543 | 0.8036 | 0.7044 |
| K-center [78] | Geometric coverage | 280K | 60.4h | 0.7655 | 0.8355 | 0.6941 | 0.7757 | 0.6478 | 0.6735 | 0.8065 | 0.6979 |
| CRAIG [79] | Submodular coreset | 280K | 62.6h | 0.7862 | 0.8453 | 0.7250 | 0.7635 | 0.6503 | 0.6826 | 0.8179 | 0.7027 |
| CiT [80] | Downstream relevance | 280K | 66.5h | 0.8408 | 0.8574 | 0.7826 | 0.8103 | 0.6626 | 0.7051 | 0.8417 | 0.7014 |
| ReSpec [81] | Hybrid filtering | 395K | 85.9h | 0.8715 | 0.8735 | 0.7978 | 0.8397 | 0.6654 | 0.7376 | 0.8807 | **0.7112** |
| CheXficient | Representation-guided | 280K | 60.2h | **0.9074** | **0.9659** | 0.8248 | **0.8535** | 0.6866 | 0.7308 | **0.9005** | 0.7019 |

configuration, and number of training epochs as CheXficient. For CLIPScore, image–report pairs are ranked according to their cross-modal semantic alignment measured by cosine similarity, and the highest-scoring samples are retained until the target data budget is reached. K-center iteratively selects samples (in the image-text concatenated embedding space) that maximize the minimum distance to the currently selected set, thereby encouraging broad geometric coverage of the training distribution. For CRAIG, we use its gradient-based coreset formulation, where per-sample gradients of the contrastive loss are used to select samples via a submodular facility-location objective. CiT estimates the relevance of candidate image–report pairs to downstream tasks and retains samples with the highest relevance scores. ReSpec combines alignment, relevance, and specificity criteria to filter training pairs and is implemented using the authors' recommended thresholding strategy. Under this threshold, ReSpec naturally retains 395K image–report pairs from our pretraining corpus; we therefore report its performance at this native subset size rather than forcing it to the 280K budget.

As shown in Table B1, under the matched 280K data budget, CheXficient achieves higher average AUROC across the eight evaluation benchmarks than random sampling, CLIPScore, K-center, CRAIG, and CiT, with absolute improvements of 0.0518, 0.0906, 0.0844, 0.0747, and 0.0462, respectively. Compared with ReSpec, which selects a larger subset of 395K image–report pairs and requires 85.9 GPU-hours, CheXficient achieves a higher average AUROC (0.8214 vs. 0.7972) while using fewer training samples and less compute. These controlled comparisons indicate that the gains of CheXficient cannot be attributed to data reduction alone and support the effectiveness of representation-guided curation relative to established data selection strategies.

## B.2 Multimodal embedding vs. unimodal embedding

In CheXficient, data curation can be performed using either unified multimodal embeddings obtained by concatenating image and text features, i.e., $\mathsf{concat}(\mathbf{z}_i^{\text{img}}, \mathbf{z}_i^{\text{txt}})$, or unimodal embeddings derived from images ($\mathbf{z}_i^{\text{img}}$) or reports ($\mathbf{z}_i^{\text{txt}}$) alone. Table B2 presents an ablation study comparing these choices. While no consistent trend is observed indicating the superiority of a single modality, combining image and text embeddings generally yields better performance than using either modality alone. This suggests that paired image-report data provide richer, complementary semantic cues

**Table B2** Effect of different feature embeddings used in CheXficient pretraining, e.g., image-only $\mathbf{z}_i^{\text{img}}$, text-only $\mathbf{z}_i^{\text{txt}}$, and image-text concatenation $\mathsf{concat}(\mathbf{z}_i^{\text{img}}, \mathbf{z}_i^{\text{txt}})$, on 280K curated CXR data pairs. Zero-shot findings classification performance (AUROC) is reported across eight datasets.

| Dataset | $\mathbf{z}_i^{\text{img}}$ | $\mathbf{z}_i^{\text{txt}}$ | $\mathsf{concat}(\mathbf{z}_i^{\text{img}}, \mathbf{z}_i^{\text{txt}})$ |
|---|---|---|---|
| SIIM-PTX | 0.8920 [0.8858, 0.8981] | 0.8937 [0.8900, 0.8975] | **0.9074 [0.9001, 0.9148]** |
| Pneumonia2017 | 0.9481 [0.9416, 0.9546] | 0.9490 [0.9447, 0.9532] | **0.9659 [0.9597, 0.9722]** |
| TBX11K | 0.8209 [0.8145, 0.8274] | 0.8174 [0.8110, 0.8238] | **0.8248 [0.8183, 0.8313]** |
| CheXpert | 0.8407 [0.8361, 0.8453] | 0.8335 [0.8290, 0.8380] | **0.8535 [0.8501, 0.8569]** |
| MIMIC-CXR | 0.6613 [0.6558, 0.6669] | 0.6758 [0.6708, 0.6807] | **0.6866 [0.6824, 0.6908]** |
| ChestX-ray14 | 0.7191 [0.7166, 0.7216] | **0.7312 [0.7255, 0.7369]** | 0.7308 [0.7281, 0.7336] |
| VinDr-CXR | 0.8966 [0.8929, 0.9003] | 0.8927 [0.8866, 0.8989] | **0.9005 [0.8929, 0.9081]** |
| VinDr-PCXR | 0.6992 [0.6915, 0.7070] | 0.6956 [0.6882, 0.7031] | **0.7019 [0.6951, 0.7087]** |

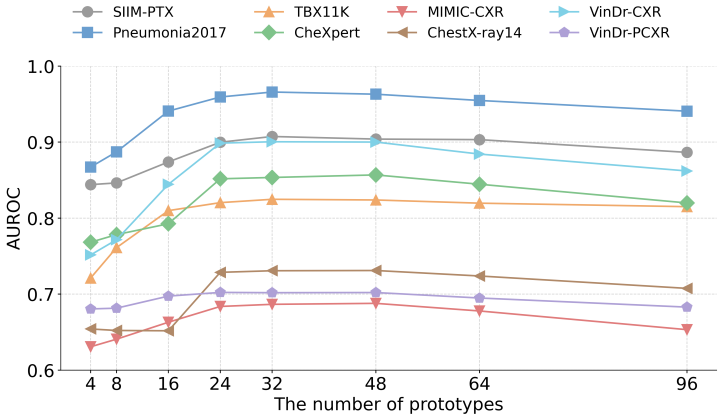



**Figure B1** Effect of the number of prototypes used in CheXficient pretraining on zero-shot findings classification performance across eight datasets.

that more accurately reflect data importance for effective data selection, ultimately contributing to improved representation learning.

## B.3 Effect of the number of prototypes

We further analyze the sensitivity of CheXficient to the number of prototypes $K$, as shown in Figure B1, by varying $K$ under the same pretraining data budget (i.e., 280K data pairs). Overall, the zero-shot findings classification performance is relatively robust when $K$ ranges from 24 to 48. Using too few prototypes fails to adequately capture the complexity and diversity of the training data, leading to sub-optimal performance. In contrast, employing an excessive number of prototypes introduces surplus or even harmful information, resulting in performance degradation and increased model complexity. Based on these observations, we set the number of prototypes to $K = 32$ for all remaining experiments.

# Appendix C Downstream implementation details

## C.1 Image classification

All downstream image classification experiments are conducted on a single NVIDIA H100 GPU. We use a batch size of 512 and optimize the models with AdamW, using a base learning rate of $3 \times 10^{-2}$, a weight decay of $1 \times 10^{-4}$, and a step-based learning rate decay scheduler. For image pre-processing, we apply center cropping and no additional image augmentation is applied during downstream training. CXR image intensities are normalized using a mean of [0.4814, 0.4578, 0.4082] and a standard deviation of [0.2686, 0.2613, 0.2757]. All classification models are trained for 15 epochs. The final checkpoint is used for inference on the test set, as no overfitting is observed based on the validation loss. We conduct experiments over five runs with different random seeds. Under this GPU setup, fine-tuning a linear classification head on top of the visual encoder takes approximately 0.41 seconds per iteration.

## C.2 Semantic segmentation

We evaluate the segmentation tasks on 1 NVIDIA H100 GPU. We use a training batch size of 256, AdamW optimizer, a base learning rate of $3 \times 10^{-4}$, a weight decay of $1 \times 10^{-4}$, and a step-based learning rate decay scheduler. For image pre-processing, we apply center cropping and no additional image augmentation is applied during downstream training. CXR image intensities are normalized using a mean of [0.4814, 0.4578, 0.4082] and a standard deviation of [0.2686, 0.2613, 0.2757]. The segmentation models are trained for a total of 120 epochs, using a combined Dice loss and cross-entropy loss as the training objective. For lung segmentation on the JSRT dataset, we use a data split of 172/25/50 for the training, validation, and test sets, respectively, and report Dice scores on the test set. For pneumothorax segmentation on SIIM-PTX and rib segmentation on VinDr-RibCXR, we follow the official data splits provided by the datasets. The model with the lowest validation loss is selected for evaluation on the test set. Under the given GPU setup, training a U-Net decoder as the segmentation head on top of the visual encoder takes approximately 0.45 seconds per iteration.

## C.3 Radiology report generation

All radiology report generation experiments are conducted on 1 compute node equipped with 4 NVIDIA H100 GPUs (80GB each). We follow the same training hyper-parameters as LLaVA-Rad [51]. Specifically, we use a total batch size of 128 (32 per GPU), a cosine learning rate scheduler with a warmup phase over 3% of the training steps, and a base learning rate of $3 \times 10^{-5}$. We adopt a two-stage fine-tuning strategy over three epochs, where the image encoder is kept frozen throughout both stages. In the first stage (feature alignment), both the image encoder and the large language model (LLM) are frozen, and only the MLP projection layer is trained. In the second stage (end-to-end fine-tuning), the MLP projection layer and the LLM are jointly updated, while the image encoder remains frozen. Training on MIMIC-CXR and ReXGradient-160K requires approximately 0.5 and 0.2 days, respectively.