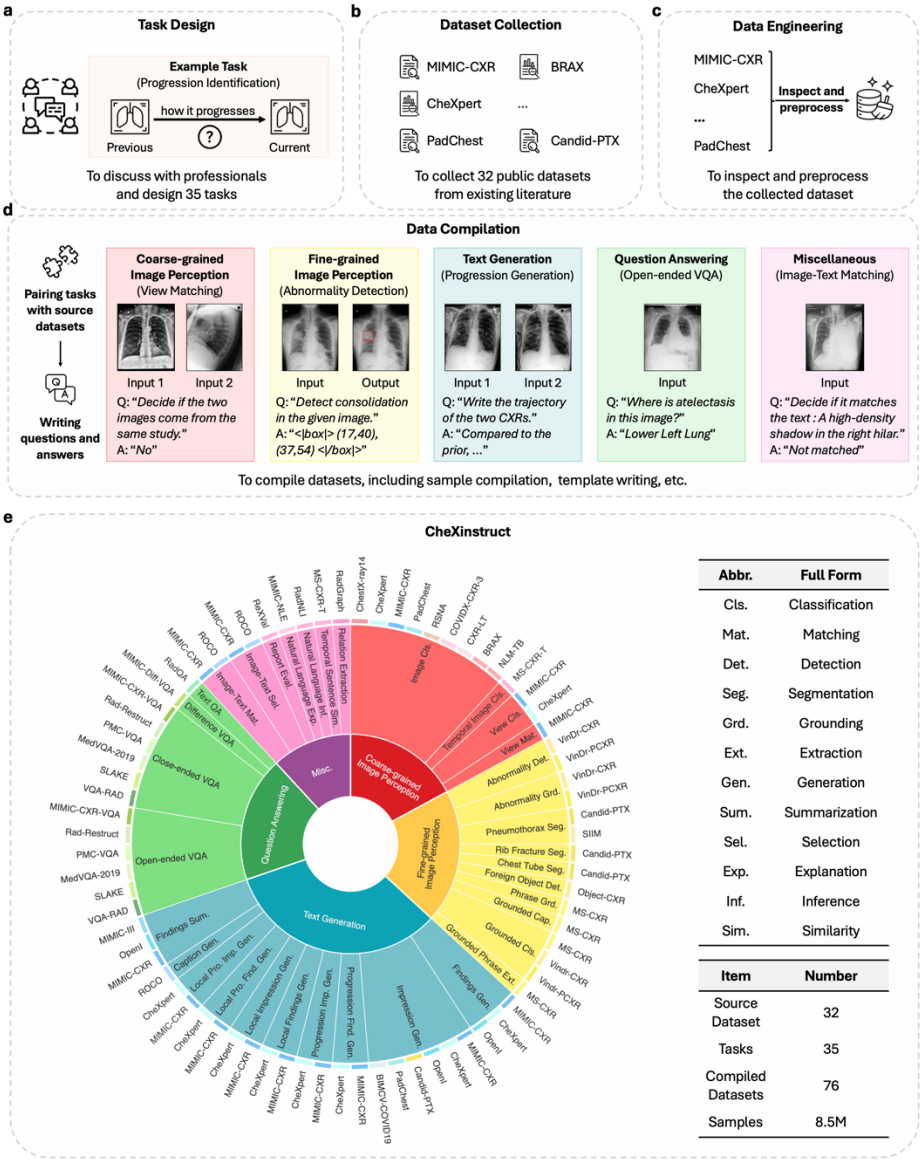
A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation

Zhihong Chen[1,2,*], Maya Varma[1,2,3,*], Justin Xu[1,2,4], Magdalini Paschali[1,2], Dave Van Veen[1,5], Andrew Johnston[2], Alaa Youssef[1,2], Louis Blankemeier[1,5], Christian Bluethgen[1,6], Stephan Altmayer[2], Jeya Maria Jose Valanarasu[1,3], Mohamed Siddig Eltayeb Muneer[2], Eduardo Pontes Reis[1,2], Joseph Paul Cohen[1], Cameron Olsen[2], Tanishq Mathew Abraham[7], Emily B. Tsai[2], Christopher F. Beaulieu[2], Jenia Jitsev[8,9], Sergios Gatidis[1,2], Jean-Benoit Delbrouck[1,2], Akshay S. Chaudhari[1,2,10], Curtis P. Langlotz[1,2,10,11]

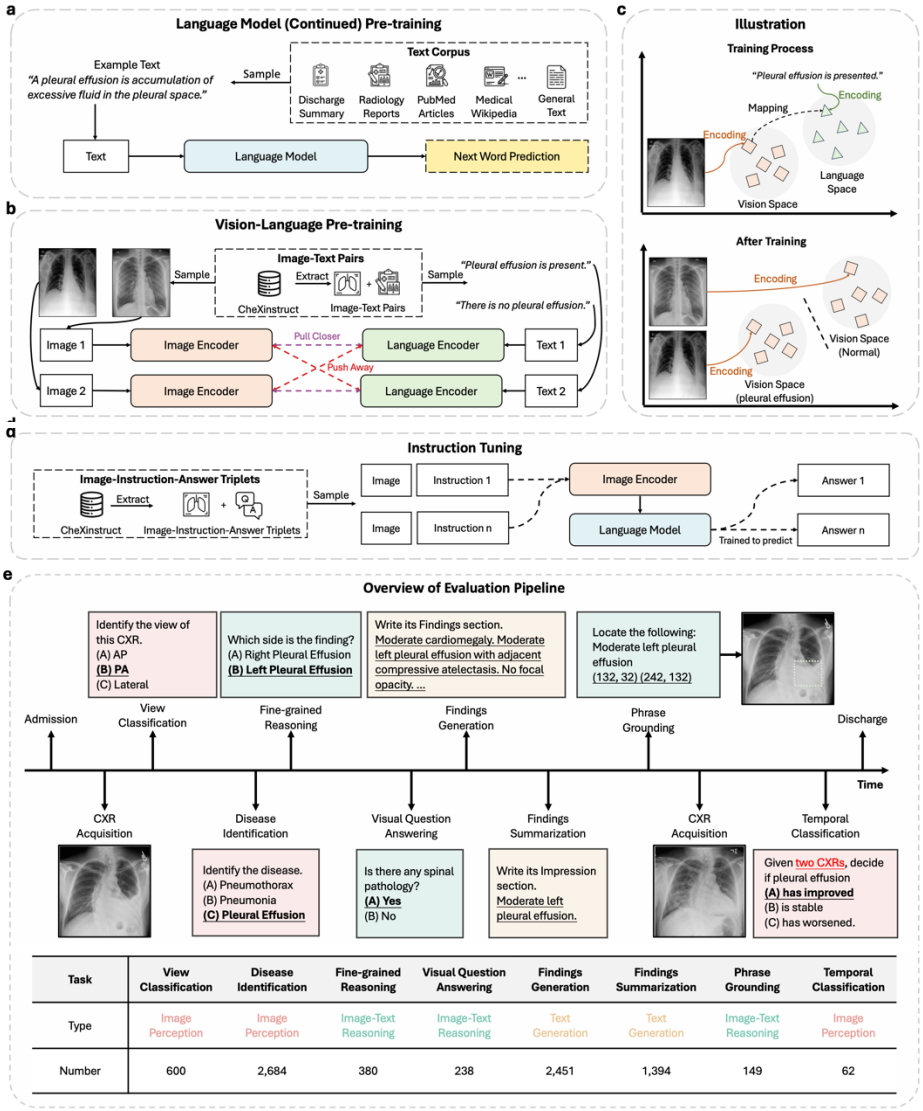
[1]Stanford Center for Artificial Intelligence in Medicine and Imaging, Stanford University, Palo Alto, CA, USA. [2]Department of Radiology, Stanford University, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. [4]Big Data Institute, University of Oxford, Oxford, UK. [5]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [6]Department of Radiology, University Hospital Zurich, Zürich, Switzerland. [7]Stability AI, London, UK. [8]Jülich Supercomputing Centre, Jülich, Germany. [9]LAION, Germany. [10]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [11]Department of Medicine, Stanford University, Stanford, CA, USA. Corresponding to: {zhihongc,mvarma2,jbdel,akshaysc,langlotz}@stanford.edu



Abbr.	Full Form
Cls.	Classification
Mat.	Matching
Det.	Detection
Seg.	Segmentation
Grd.	Grounding
Ext.	Extraction
Gen.	Generation
Sum.	Summarization
Sel.	Selection
Exp.	Explanation
Inf.	Inference
Sim.	Similarity

Item	Number
Source Dataset	32
Tasks	35
Compiled Datasets	76
Samples	8.5M

Figure 1 | Curation of CheXinstruct. a, Identification of CXR interpretation tasks. We defined 35 tasks that users are likely to perform with CXR FMs. b, Source dataset collection. To create training data samples for each of our defined tasks, we collected 32 public datasets. c, Data engineering. We performed both manual quality control and automated data engineering to preprocess collected source data. d, CheXinstruct compilation. We used the preprocessed datasets to generate training samples for each of our 35 defined tasks. e, Overview of CheXinstruct with data statistics.



Task	View Classification	Disease Identification	Fine-grained Reasoning	Visual Question Answering	Findings Generation	Findings Summarization	Phrase Grounding	Temporal Classification
Type	Image Perception	Image Perception	Image-Text Reasoning	Image-Text Reasoning	Text Generation	Text Generation	Image-Text Reasoning	Image Perception
Number	600	2,684	380	238	2,451	1,394	149	62

Figure 2 | Training and evaluating CheXagent. a, To develop CheXagent, we first trained a language model on clinical text. b, We then trained an image encoder to learn useful visual representations of imaging findings by leveraging paired text. c, This procedure enabled the visual encoder to capture semantic meaning with respect to key findings within its latent representation space. d, Finally, we jointly trained the image encoder and language model on data triplets from CheXinstruct, providing CheXagent with the capability to respond to user instructions. e, We constructed eight evaluation tasks to assess image perception, reasoning, and text generation capabilities.