1 Supplementary material

1.1 Annotation pairs

Table 1: Distribution of the annotation pairs (image bounding-box and sentence) across different clinical findings. MS-CXR [1], VinDr [2]

Findings	**# of annotation pairs**	**Dataset origin**
Atelectasis	123	MS-CXR+VinDr-CXR
Cardiomegaly	2150	MS-CXR+VinDr-CXR
Consolidation	238	MS-CXR+VinDr-CXR
Edema	46	MS-CXR
Lung opacity	628	MS-CXR+VinDr-CXR
Pleural effusion	730	MS-CXR+VinDr-CXR
Pneumonia	182	MS-CXR
Pneumothorax	295	MS-CXR+VinDr-CXR
Aortic enlargement	2348	VinDr-CXR
Calcification	177	VinDr-CXR
Emphysema	14	VinDr-CXR
Enlarged PA	21	VinDr-CXR
Interstitial Lung Desease (ILD)	152	VinDr-CXR
Infiltration	245	VinDr-CXR
Lung cavity	21	VinDr-CXR
Mediastinal shift	85	VinDr-CXR
Nodule/mass	410	VinDr-CXR
Pulmonary Fibrosis	1017	VinDr-CXR
Pleural thickening	882	VinDr-CXR
Rib fracture	41	VinDr-CXR
Total	**9805**	

Table 1 presents the number of annotation pairs, specifically image bounding-boxes and corresponding sentences, across various clinical findings. The data utilized in this study is derived from the MS-CXR dataset [1], and the VinDr [2] dataset. The table categorizes the annotation pairs according to different clinical findings, providing a comprehensive overview of the dataset composition and representation of various conditions within the datasets. The plot on Figure 1 complements Table 1, offering a graphical perspective on the same statistical distribution of clinical annotations.

1.2 Qualitative Results

In Figure 2, we present qualitative examples of radiology report generation by our model. The model demonstrates the ability to generate precise reports and accurately identify the locations of specific entities within the images using bounding boxes.

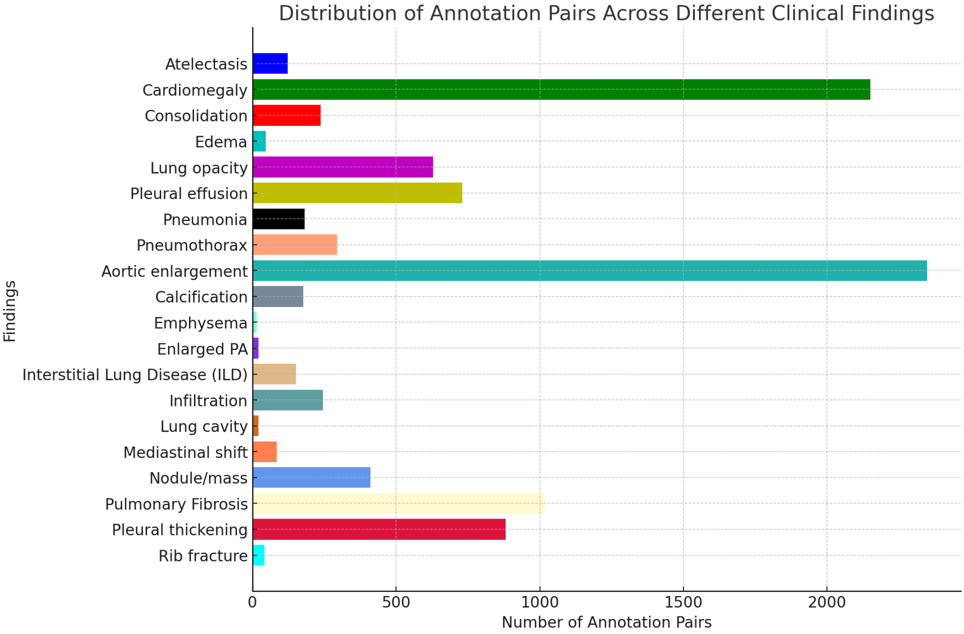
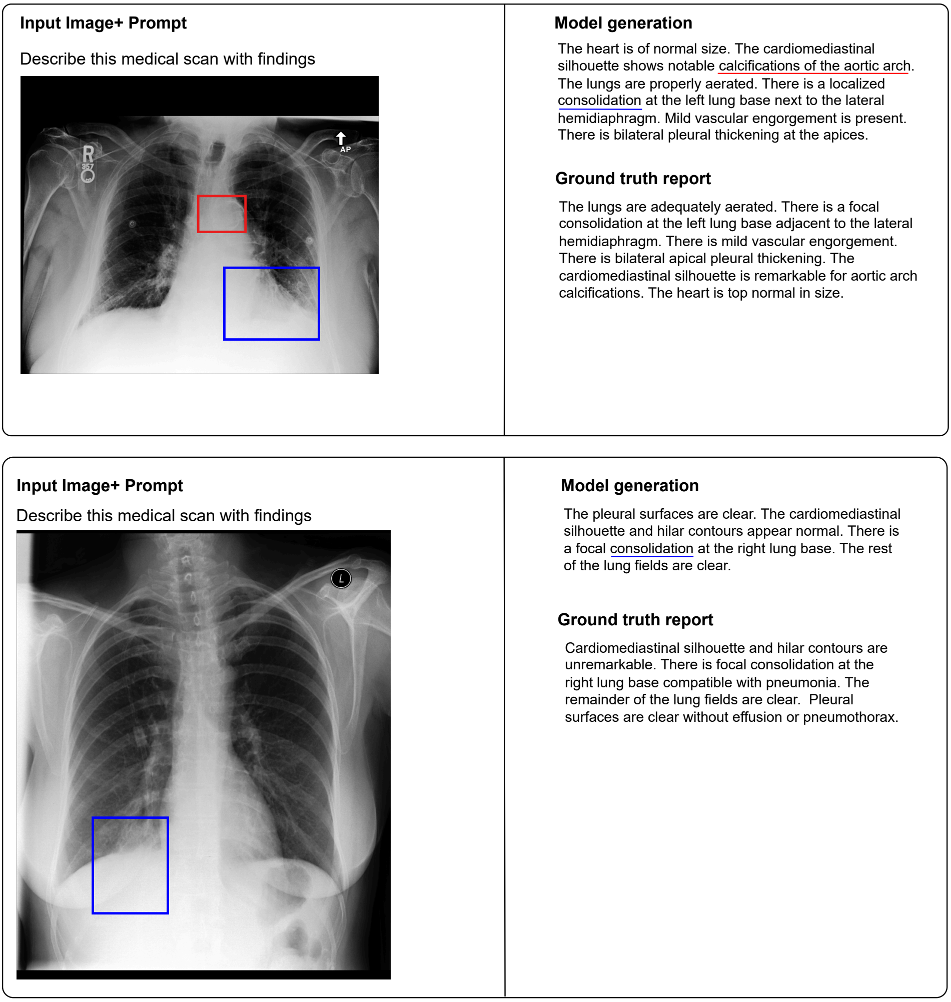


Figure 1: Proportional distribution of annotation pairs by clinical findings in the small dataset of grounded image-text pairs.



Input Image+ Prompt

Describe this medical scan with findings

Model generation

The heart is of normal size. The cardiomediastinal silhouette shows notable calcifications of the aortic arch. The lungs are properly aerated. There is a localized consolidation at the left lung base next to the lateral hemidiaphragm. Mild vascular engorgement is present. There is bilateral pleural thickening at the apices.

Ground truth report

The lungs are adequately aerated. There is a focal consolidation at the left lung base adjacent to the lateral hemidiaphragm. There is mild vascular engorgement. There is bilateral apical pleural thickening. The cardiomediastinal silhouette is remarkable for aortic arch calcifications. The heart is top normal in size.

Input Image+ Prompt

Describe this medical scan with findings

Model generation

The pleural surfaces are clear. The cardiomediastinal silhouette and hilar contours appear normal. There is a focal consolidation at the right lung base. The rest of the lung fields are clear.

Ground truth report

Cardiomediastinal silhouette and hilar contours are unremarkable. There is focal consolidation at the right lung base compatible with pneumonia. The remainder of the lung fields are clear. Pleural surfaces are clear without effusion or pneumothorax.

Figure 2: Qualitative Examples of Report Generation. This figure showcases several examples of reports generated by our model alongside reference reports. The underline color in the text corresponds to the specific bounding box colors on the images.

References

[1] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[2] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.