A Additional Experiments

A.1 Zero-shot Text-prompt Sensitivity Analysis

Vision-language pretraining aligns image and text data in a joint representation space, which enables impressive zero-shot downstream image classification performance via input text prompts. However, some recent work [31,85] has shown that downstream task performance can heavily depend on the choice of text prompts. Constructing good text prompts (prompt engineering) may require expert domain knowledge and can be costly and time-consuming. In Table A.1, we study RSNA pneumonia zero-shot classification performance using different text prompt combinations. Compared to the baseline, BioViL demonstrates much lower sensitivity to prompt choices selected from the data distribution. BioViL maintains its high performance even when faced with relatively long queries, which is not the case for the baseline model. These observations suggest that our improved text encoder CXR-BERT is more robust to prompt variations, and makes prompt engineering easier and less of a requirement to achieve high zero-shot classification performance.

Table A.1: Text prompt sensitivity analysis on the RSNA pneumonia zero-shot classification task. Image-text models trained without the proposed text modelling improvements (Table 4) show higher sensitivity to different input text prompts as the latent text embeddings are inconsistent for synonym phrases. For this reason, baseline methods often require post-hoc text prompt engineering heuristics (e.g. [31]).

Method	Pos. Query	Neg. Query	F1 Score	ROC-AUC	$\lvert\Delta AUC\rvert$
BioViL	"Findings suggesting pneumonia"	"There is no evidence of acute pneumonia"	0.657	0.822	-
ClinicalBert	"Findings suggesting pneumonia"	"There is no evidence of acute pneumonia"	0.581	0.731	-
BioViL	"Findings suggesting pneumonia"	"No evidence of pneumonia"	0.665	0.831	-
BioViL	"Consistent with the diagnosis of pneumonia"	"There is no evidence of acute pneumonia"	0.669	0.839	0.008
ClinicalBert	"Findings suggesting pneumonia"	"No evidence of pneumonia"	0.614	0.815	-
ClinicalBert	"Consistent with the diagnosis of pneumonia"	"There is no evidence of acute pneumonia"	0.621	0.694	0.121
BioViL	"Findings consistent with pneumonia"	"No evidence of pneumonia"	0.672	0.838	-
BioViL	"Findings consistent with pneumonia"	"There is no pneumonia"	0.679	0.847	0.009
ClinicalBert	"Findings consistent with pneumonia"	"No evidence of pneumonia"	0.640	0.782	-
ClinicalBert	"Findings consistent with pneumonia"	"There is no pneumonia"	0.586	0.724	0.058

A.2 Qualitative Results – Phrase Grounding

In Fig. A.1, we show and describe some phrase grounding examples obtained with different models on the `MS-CXR` dataset. From left to right, the figure shows the ClinicalBERT baseline, ConVIRT, GLoRIA, and BioViL similarity maps. While the figure only illustrates a few examples, the results demonstrate that phrase grounding performance can be significantly enhanced by leveraging improved text modelling (BioViL). The examples include clinical findings that differ in size, type, and anatomical location.

Additionally, in Fig. A.2, we show and describe some failure cases of BioViL on the `MS-CXR` dataset to motivate any further research on this topic. In particular, the models show limitations in grounding the descriptions relating to smaller structures (e.g., rib fracture, pneumothorax), and in a few cases the location

Table A.2: An extension of Table 6 to include Sensitivity and Specificity for the RSNA Pneumonia zero-shot and fine-tuned classification. We compare to GLoRIA scores reported in [31] which outperforms ConVIRT [85] (see [31]). Training size: GLoRIA ($N = 186k$, private dataset), BioViL ($N = 146.7k$ of MIMIC-CXR).

Method	Type	Text Model	Loss	% of labels	Acc.	Sens.	Spec.	F1	AUROC
SimCLR [6]	Image only	-	Global	1%	0.545	0.776	0.436	0.522	0.701
				10%	0.760	0.663	0.806	0.639	0.802
				100%	0.788	0.685	0.837	0.675	0.849
GLoRIA [31]	Joint	ClinicalBERT	Global & local	Zero-shot	0.70	0.89	0.65	0.58	-
				1%	0.72	0.82	0.69	0.63	0.861
				10%	0.78	0.78	0.79	0.63	0.880
				100%	0.79	0.87	0.76	0.65	0.886
Baseline	Joint	ClinicalBERT	Global	Zero-shot	0.719	0.648	0.781	0.614	0.812
BioViL	Joint	CXR-BERT	Global	Zero-shot	0.732	0.831	0.685	0.665	0.831
				1%	0.805	0.791	0.812	0.723	0.881
				10%	0.812	0.781	0.826	0.727	0.884
				100%	0.822	0.755	0.856	0.733	0.891

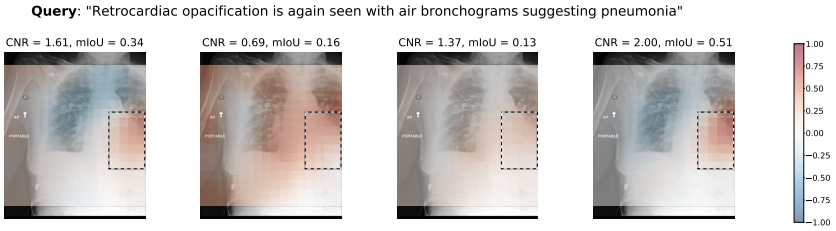
modifier is not disassociated from the entities corresponding to abnormalities, see (a) in Fig. A.2.

A.3 Additional Evaluation Metrics

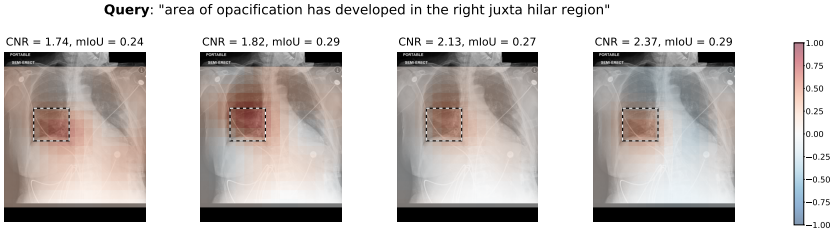
In Table A.2, an extension of Table 6 is provided to include the sensitivity and specificity metrics for the zero-shot and fine-tuned classification experiments presented in Section 4.4. The classification thresholds are set to maximise the F1 scores for each method. Further, in Table A.4 we provide mean IoU scores for the phrase grounding experiments presented in Section 4.3, which evaluates the pretrained BioViL model on the `MS-CXR` dataset. We observed that the distribution of similarity scores is different for GLoRIA and BioViL-L due to the different temperature parameter used in the local loss term in [31]. To provide a fair comparison, we adjust the similarity scores via min-max scaling to the full $[-1, 1]$ range. The same scaling strategy is utilised in the implementation of the baseline method [31]. Note that the CNR scores are not affected by this linear re-scaling.

A.4 Ablations on Training Dataset Size & Use of Raw Input Images

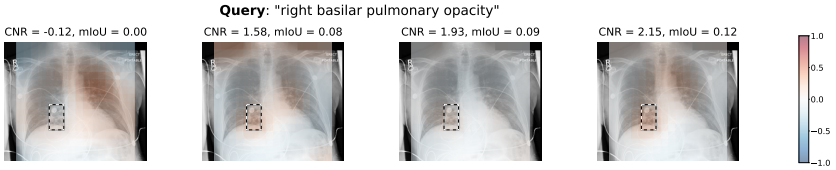
An additional set of experiments are conducted to test the impact of (I) training dataset size and (II) the use of raw DICOM images instead of JPEG images on phrase grounding performance. In the former case, the number of training pairs is increased from $146.7k$ to $176k$, where we used all available studies with IMPRESSION section and AP/PA scans after excluding the test set. In the latter ablation, the JPEG images are replaced with the raw DICOM images to reduce image artefacts due to compression. Table A.3 shows that further performance gains can be achieved by utilising the DICOM data and matching the training set size to related methods (e.g., GLoRIA [31]), where the raw data is empirically



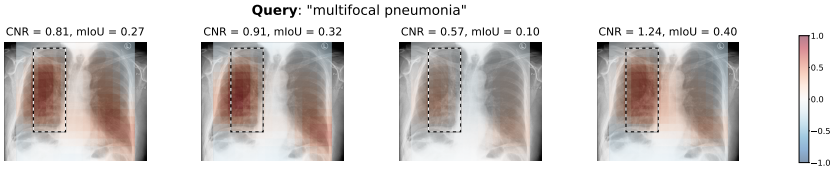
(a) Relatively long and complex query



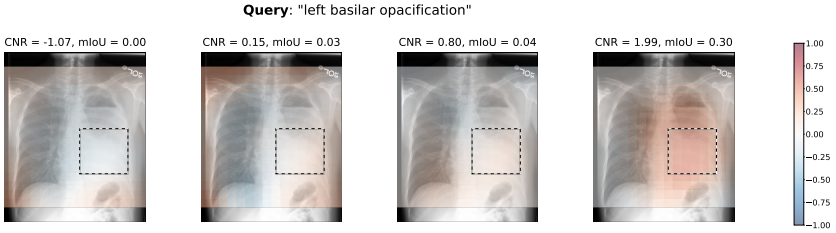
(b) Complex anatomical location specification



(c) Small ground-truth bounding box

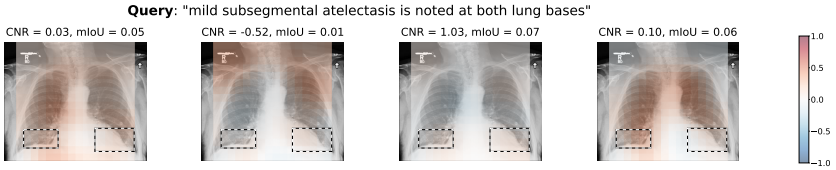


(d) Multifocal pneumonia example which is localised in the right lobe

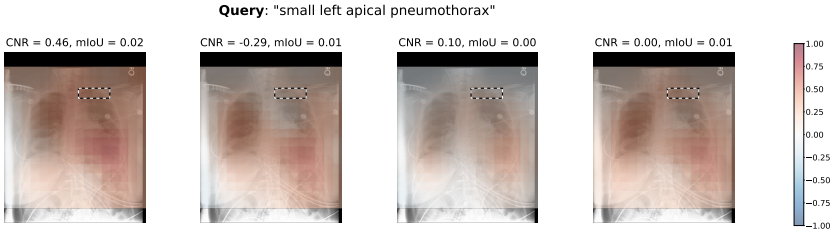


(e) Location modifier “left basilar”

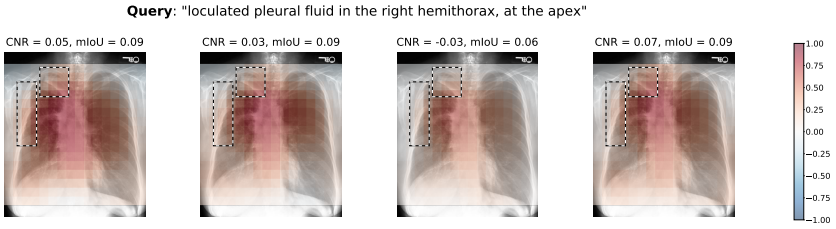
Fig. A.1: Qualitative examples from `MS-CXR` phrase grounding benchmark. Model outputs (latent vector similarity) are compared (from left, ClinicalBERT baseline, ConVIRT, GLoRIA, and BioViL)



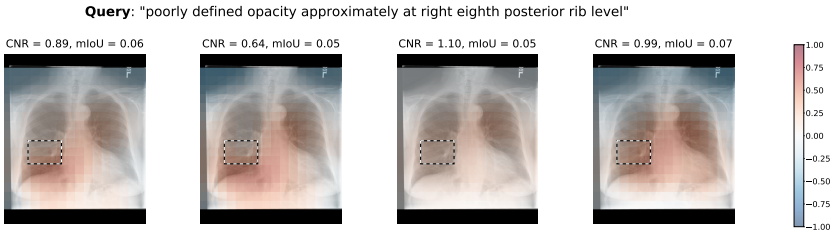
(a) Failed to recognise atelectasis despite having "both lung bases" location specification



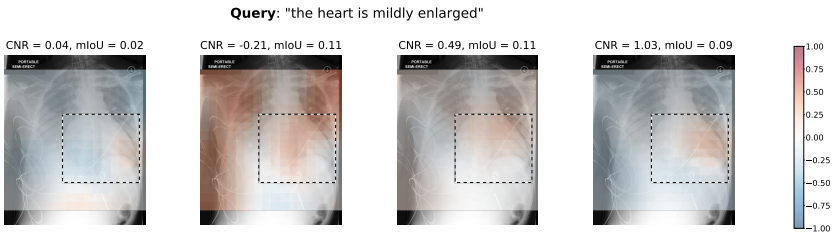
(b) Failed to recognise small pneumothorax despite having "apical" location modifier



(c) Failed to recognise loculated pleural fluid despite having "apical" and "right hemithorax"



(d) Failed to recognise the rib position



(e) Mismatch between bounding box and salient region: Models attend to the salient region (enlarged area) to identify the abnormality instead of the entire heart.

Fig. A.2: Failure cases from `MS-CXR` phrase grounding benchmark. Model outputs (latent vector similarity) are compared (from left, ClinicalBERT baseline, ConVIRT, GLoRIA, and BioViL)

Table A.3: Ablations on BioViL – Increasing training set size and use of raw DICOM images instead of compressed JPEG images. The approaches are compared in terms of contrast-to-noise ratio (CNR) obtained on the newly released `MS-CXR` dataset, averaged over four runs with different seeds.

Method	Training	Atelectasis	Cardiomegaly	Consolidation	Lung opacity	Edema	Pneumonia	Pneumothorax	Pl. effusion	Avg.
BioViL	146.7k	1.02±.06	0.63±.08	1.42±.02	1.05±.06	0.93±.03	1.27±.04	0.48±.06	1.40±.06	1.03±.02
+ More data	176.0k	1.01±.07	0.70±.03	1.45±.01	1.04±.04	0.94±.01	1.27±.05	0.54±.05	1.43±.04	1.05±.02
+ Raw images	176.0k	1.03±.06	0.64±.09	1.51±.02	1.12±.06	1.00±.07	1.39±.04	0.56±.05	1.46±.05	1.09±.02

Table A.4: Mean IoU scores obtained on the newly released `MS-CXR` dataset, averaged over four runs with different seeds. The results are collected using different text encoder and training objectives (e.g., G&L: Global and local loss).

Method	Objective	Text encoder	Atelectasis	Cardiomegaly	Consolidation	Lung opacity	Edema	Pneumonia	Pneumothorax	Pl. effusion	Avg.
Baseline	Global	ClinicalBERT	0.228	0.269	0.293	0.173	0.268	0.249	0.084	0.232	0.224
Baseline	Global	PubMedBERT	0.225	0.293	0.297	0.167	0.266	0.286	0.077	0.222	0.225
ConVIRT [85]	Global	ClinicalBERT	0.257	0.281	0.313	0.177	0.272	0.238	0.091	0.227	0.238
GLoRIA [31]	G&L	ClinicalBERT	0.261	0.273	0.324	0.198	0.251	0.246	0.100	0.254	0.246
BioViL	Global	CXR-BERT	0.296	0.292	0.338	0.202	0.281	0.323	0.109	0.290	0.266
BioViL-L	G&L	CXR-BERT	0.302	0.375	0.346	0.209	0.275	0.315	0.135	0.315	0.284

observed to contribute more. These improved results and pre-training models are neither reported nor used in the experiments presented in the main body of this paper. We hope that these findings can provide useful insights for future research on this topic.

B Background in Chest Radiology

Chest X-rays are the most commonly performed diagnostic X-ray examination, and a typical text report for such an exam consists of three sections: a "Background" section describing the reason for examination and the exam type, a "Findings" section describing abnormalities as well as normal clinical findings in the scan, and an "Impression" section which summarises the findings and offers interpretation with possible recommendations. Multiple large Chest X-ray datasets have been released to the public (see [71] for an overview of CXR image datasets), including multi-modal ones of images and text such as MIMIC-CXR [34], some also accompanied by small sets of expert-verified ground-truth annotations of various nature, making the application a popular candidate for exploring self-supervised VLP on biomedical data.

The application area also possesses a strong clinical motivation. Globally, there is a shortage of qualified trained radiologists and a constantly increasing number of examinations in healthcare systems, workflows are hampered by issues such as a lack of standardisation in report writing, and fatigue-based errors occur too frequently. Thus, decision-support systems that can analyse incoming images or image-report pairs in order to provide real-time feedback to radiologists are a promising avenue towards improving workflow efficiency and the quality of medical image readings. In practice, the existing radiology workflow can for example be augmented via machine learning models by providing feedback on

any incorrect or missing information in reports, and by standardising the reports' structure and terminology.

B.1 Key NLP and Dataset Challenges in Radiology

In this work, we focus on developing text and image models to enable clinical decision-support systems for biomedical applications via self-supervised VLP, without ground-truth annotations, and we conduct experiments in CXR applications. Image and text understanding in the biomedical domain is distinct from general-domain applications and requires careful consideration. Medical images are elaborately structured, which is reflected in the corresponding notes. To be able to harness the dense information captured in text notes for free-text natural language supervision, it becomes imperative to obtain finely tuned text models.

Complex Sentence Structure. Linguistic characteristics in radiology reports, many shared with related clinical text settings, decidedly differ from general domain text and thus require carefully tuned text models to acquire the best possible free-text natural language supervision in self-supervised VLP. For one, negations are frequently used to indicate the absence of findings, in particular to make references as to how a patient's health has evolved, e.g. "there are no new areas of consolidation to suggest the presence of pneumonia". This sentence is for example falsely captured as positive for pneumonia by the automated CheXpert labeller [33]. Furthermore, as exemplified in this example, long-range dependencies are common, which makes understanding of relations within sentences challenging.

Use of Modifiers. Another characteristic is the use of highly specialised spatial language in radiology, which is crucial for correct diagnosis, often describing the positioning of radiographic findings or medical devices with respect to anatomical structures, see e.g. [13,14]. The use of words like "medial", "apical", "bilateral" or "basilar" as spatial modifiers is unlikely to appear in the general domain but very common in CXR radiology reports. In addition to spatial modifiers, severity modifiers such as "mild", "moderate" or "severe" are also commonly attached to an identified disorder or abnormality [18].

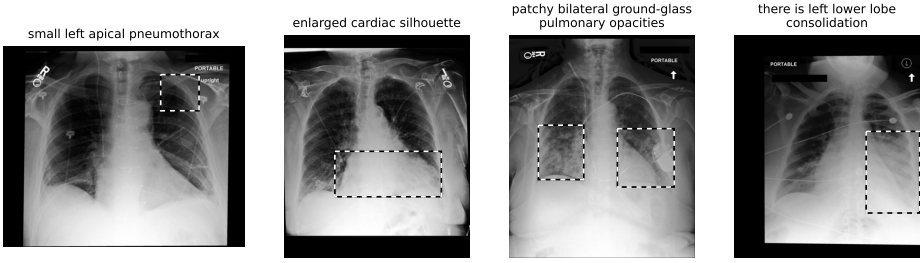
Expressions of Uncertainty. Another interesting difference to most general domain VLP applications and datasets such as Internet image captions, are expressions of uncertainty that one frequently encounters in radiology reports. We rarely expect to find an image caption to read "We see a person petting an animal, it is likely a dog but it could also be a cat". In contrast, consider the following real radiology example: "New abnormality in the right lower chest could be either consolidation in the lower lobe due to rapid pneumonia or collapse, and/or moderate right pleural effusion, more likely abnormality in the lung because of absent contralateral mediastinal shift." It is an extremely long description expressing uncertainty and containing long range dependencies.

Class Imbalance. Finally, a challenge for many domain-specific VLP applications that is far less pronounced in the general domain setting is that of imbalanced latent entities. An example of such entities are the normal and anomalous findings in radiology images that doctors will describe in their report. In the CXR application, reports can roughly be divided into normal and abnormal scans, where abnormal ones reveal signs or findings observed during the exam [11]. Normal scans that do not show any signs of disease are far more common than any other findings, which leads to a larger number of false negatives in contrastive objectives compared to the general domain. An important detail is that normal scans tend to be expressed in specific forms and doctors frequently use templates to produce reports with no abnormalities.

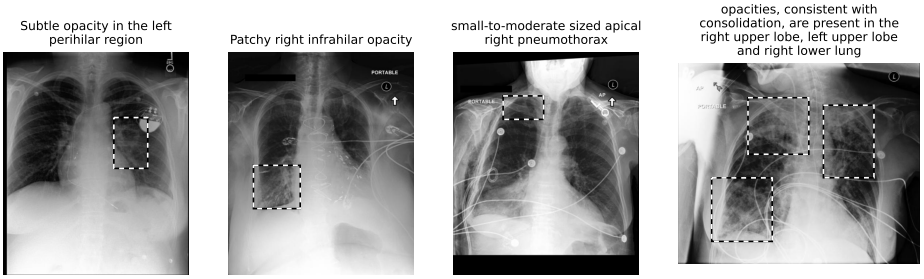
C MS-CXR Dataset Details

General Overview. With this new benchmark dataset, we provide bounding box and sentence pair annotations describing clinical findings visible in a given chest X-ray image. `MS-CXR` consists of 1047 images, with a total of 1153 bounding box and sentence pairs. Each sentence describes a single pathology present in the image, and there could be multiple manually annotated bounding boxes corresponding to the description of the single radiological finding. Additionally, an image may have more than one pathology present, and we provide separate sets of bounding boxes for each phrase describing a unique pathology associated with an image. The annotations were collected on a subset of MIMIC-CXR images, which additionally contains labels across eight different pathologies: atelectasis, cardiomegaly, consolidation, edema, lung opacity, pleural effusion, pneumonia and pneumothorax. These pathologies were chosen based on the overlap between pathology classes present in the existing datasets and the CheXbert classifier [69]. In Fig. C.1 and Table C.2, we show some representative image and text examples from `MS-CXR`. Additionally, the distribution of samples across the pathology classes is shown in Table C.1 together with demographics across subjects in `MS-CXR`.

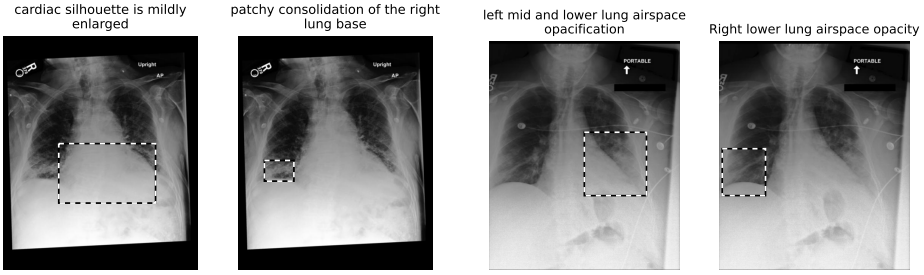
Differences to Existing Annotations. The proposed benchmark builds on top of publicly available bounding-box/ellipse annotations in REFLACX [37] and MIMIC-CXR-Annotations [71], where the latter also contains simplified text phrases for pneumonia and pneumothorax. `MS-CXR` extends and curates these annotation sets by (I) adding a new set of studies to cover a wider range of clinical findings and pathologies, (II) reviewing the clinical correctness and suitability of the existing annotations for the grounding task (see Section 3.1), (III) creating, verifying, and correcting bounding boxes where necessary, and (IV) pairing them up with real clinical descriptions extracted from MIMIC-CXR reports if none were present. Most importantly, the textual descriptions paired with dense image region annotations are sampled from the original distribution of word tokens, which capture dense text semantics and are better aligned with real-world clinical applications that build on good local alignment.



(a) Spatial extent of abnormalities ranging from highly localised to large and diffuse



(b) Complex spatial modifiers commonly seen in radiology reports



(c) Multiple pathologies reported for the same study

(d) Findings with multiple spatial locations reported separately

Fig. C.1: We here provide some examples illustrating important axes of variability present in the `MS-CXR` dataset. Text descriptions include clinical findings of varying spatial extent (a) and a range of different spatial modifiers (b). Additionally, a subset of studies contain multiple bounding-box and sentence annotations per image (c–d).

Table C.1: Distribution of the annotation pairs (image bounding-box and sentence) across different clinical findings. The demographic statistics (e.g., gender, age) of the subjects are collected from MIMIC-IV dataset for `MS-CXR` and all MIMIC-CXR.

Findings	# of annotation pairs	# of subjects	Gender - F (%)	Avg Age (std)
Atelectasis	61	61	28 (45.90%)	64.52 (15.95)
Cardiomegaly	333	282	135 (47.87%)	68.10 (14.81)
Consolidation	117	109	40 (36.70%)	60.08 (17.67)
Edema	46	42	18 (42.86%)	68.79 (14.04)
Lung opacity	81	81	33 (40.24%)	62.07 (17.20)
Pleural effusion	96	95	41 (43.16%)	66.36 (15.29)
Pneumonia	182	146	65 (44.52%)	64.32 (17.17)
Pneumothorax	237	151	66 (43.71%)	60.71 (18.04)
Total	1153	851	382 (44.89%)	64.37 (16.61)
Background (all MIMIC-CXR)	-	65379	34134.0 (52.39%)	56.85 (19.47)

C.1 Label Collection and Review

We first parse original MIMIC reports and REFLACX [37] radiology transcripts by extracting sentences to form a large pool of text descriptions of pathologies. These candidates are later filtered by deploying the CheXbert [69] text classifier, in order to keep only the phrases associated with the target pathologies whilst ensuring the following two criteria: (I) For a given study, there is only one sentence describing the target pathology, and (II) the sentence does not mention more than one findings that are irrelevant to each other. After extracting the text descriptions, they are paired with image annotations on a study level. At the final stage, a review process is conducted with two board certified radiologists mainly to verify the match between the text and bounding box candidates. Moreover, in this review process, we also assessed the suitability of the annotation pairs for the grounding task whilst ensuring clinical accuracy. In detail, the phrase-image samples are filtered out if at least one of following conditions is met:

1. Text describing a finding not present in the image.
2. Phrase/sentence does not describe a clinical finding or describes multiple unrelated abnormalities that appear in different lung regions.
3. There is a mismatch between the bounding box and phrase, such as image annotations are placed incorrectly or do not capture the true extent of the abnormality.
4. High uncertainty is expressed regarding reported findings, e.g. “there is questionable right lower lobe opacity”.
5. Chest X-ray is not suitable for assessment of the finding or has poor image quality.
6. Text contains differential diagnosis or longitudinal information that prohibits correct grounding via the single paired image.
7. Sentences longer than 30 tokens, which often contain patient meta-information that is not shared between the two modalities (e.g., de-identified tokens).

Note that we only filter out phrases containing multiple findings, not images with multiple findings. For instance, if an image contains both pneumonia and

atelectasis, with separate descriptions for each in the report, then we create two instances of phrase-bounding box pairs. Among those candidate annotations automatically extracted from radiology reports [34] or dictated transcripts [37], 222 of out 817 were rejected and not included in MS-CXR. Here the raw text data were first processed with an algorithm to extract caption candidates for the review process. The same review process is applied to adjudicate the annotation pairs released in [71], and 53 out of 367 pairs were rejected and not included in MS-CXR.

To further increase the size of our dataset, and to balance samples across classes, additional CXR studies are sampled at random, conditioned on the underrepresented pathologies. The following procedure is applied to create the pairs of image and text annotations for these selected studies: Text descriptions are extracted using the same methodology outlined above, using MIMIC-CXR and ImaGenome datasets [79], where the latter provides sentence extracts from a subset of MIMIC-CXR dataset for clinical findings. However, differently from the initial step, the corresponding bounding box annotations (either one or more per sentence) are created from scratch by radiologists for the finding described in the text, and the same filtering as above is applied by the annotator to discard candidates if the image and/or sentence is found unsuitable for the grounding task.

Analysis of Average Phrase Length. The average number of tokens (inc. full-stop) across all phrases is calculated for each benchmark dataset to better understand the characteristics of the dataset and domain. In that regard, the phrases released in [71] has an average of 6.76 tokens per sample and MS-CXR has an average of 7.49 of tokens per sample. The auto-extracted radiology sentences from transcriptions [37] whereas has an average of 8.49 tokens per sample. We observe that relatively long sentences auto-extracted from transcripts [37] were rejected more often in the review process as they often describe multiple clinical findings located in different image regions. This observation further emphasises the importance of review process of annotation pairs by the domain experts.

Patient Demographics. As shown in Table C.1, the average age of subjects in MS-CXR is higher than the average for all subjects in MIMIC-CXR. We explain this observation with the fact that we do not sample studies from healthy subjects that do not display any anomalous findings and who are statistically likely to be younger. Similarly, we do not expect gender bias to be present due to our sampling as none of the pathologies we sample are gender-specific. Overall MS-CXR does not deviate far from the MIMIC-CXR distribution.

D Related Work

Here we provide a more detailed overview of related work to complement the brief review provided in the main article.

Table C.2: Example findings in MS-CXR with complex syntactic structures. Please note how radiological sentences are most often not just a simple statement of the form "[class1, class2, ...]" that can be parsed with a simple bag-of-words approach, as in typical natural image captioning benchmarks (e.g., "A couple getting married" retrieved from Flickr30k [59]).

Sentence	Difficulty	Class
"Abnormal opacity in the basilar right hemithorax is likely atelectasis involving the right lower and middle lobes"	Complex syntactic structure	Atelectasis
"Multisegmental lower lobe opacities are present, consistent with areas of consolidated and atelectatic lung"	Complex syntactic structure	Atelectasis
"Parenchymal opacification in the mid and lower lung"	Less common expression	Pneumonia
"Air bronchograms extending from the left hilum throughout the left lung which has the appearance of infection"	Complex location description	Pneumonia
"Persistent focal bibasilar opacities, most consistent with infection"	Domain-specific modifier	Pneumonia
"Widespread infection, less severe on the left"	Location partially specified	Pneumonia
"Airspace consolidation in the right upper, right middle and lower lobes"	Multiple locations	Pneumonia
"Subsegmental-sized opacities are present in the bilateral infrahilar lungs"	Domain specific modifiers	Lung opacity
"There continues to be a diffuse bilateral predominantly interstitial abnormality in the lungs with more focal vague opacity in the left upper peripheral lung"	Complex syntactic structure	Lung opacity
"Left apical pneumothorax"	Domain-specific modifier	Pneumothorax
"Fluid level posteriorly, which represents a loculated hydropneumothorax"	Domain-specific language	Pneumothorax
"Mild-to-moderate left pneumothorax"	Severity modifier	Pneumothorax
"There is no pulmonary edema or pneumothorax, but small pleural effusions are still present"	Negated disease entities	Pleural effusion
"Pleural effusions are presumed but impossible to quantify, except say they are not large"	Complex sentence structure	Pleural effusion

Joint Image-Text Representation Learning. A variety of self-supervised VLP approaches have been proposed towards jointly learning visual and textual representations of paired data without supervision, such as frameworks using contrastive objectives [27,43,61], approaches based on joint transformer architectures [41,42,52,70], self-supervised VLP with word-region alignment and language grounding [7], and text prediction tasks to learn image features [16]. For example, [61] use a contrastive loss over embeddings of text and image pairs to train a model on large data collected from the internet (∼400M pairs) enabling zero-shot transfer of the model to downstream tasks. Some of the proposed approaches utilise a single architecture, usually a transformer, to learn a representation, following encoders for the individual modalities [7,42,70]. Another common theme is the use of use cross-modal attention mechanisms to improve the aggregation of image regions in convolutional architectures [1,12,27].

Table C.3: Example findings in ImaGenome which would make grounding of phrases difficult.

Sentence	Difficulty	Annotated Finding
"Even though Mediastinal veins are more distended, previous pulmonary vascular congestion has improved slightly, but there is more peribronchial opacification and consolidation in both lower lobes which could be atelectasis or alternatively results of recent aspiration, possibly progressing to pneumonia."	Multiple findings, uncertainty, different subparts of lung	Pneumonia
"Moderate right pleural effusion and bilateral heterogenous airplace opacities, concerning for pneumonia."	Multiple findings, differing laterality	Pneumonia
"It could be an early infection"	Region unclear	Pneumonia
"There is also a new small left-sided pleural effusion."	Differential diagnosis, there could be another effusion	Effusion

A number of different objectives have been explored for representation learning in VLP, including the prediction of words in image captions [36], predicting phrase n-grams [40], predicting of entire captions [16], *global* contrastive objectives defined on the embeddings of the entire image and text instances [85], and combinations of global and *local* contrastive terms [31,56], where local means that objectives are defined over text fragments (words or phrases) and image regions.

A task closely related to instance representation learning in VLP is *phrase grounding*, also known as visual grounding, phrase localisation, local alignment, or word–region alignment. The goal here is to connect natural language descriptions to local *image regions*. In a supervised learning setting such as in [53,55], this problem requires expensive manual annotation for region–phrase correspondence. Thus, settings for visual grounding have been explored in which cross-modal pairs are the only form of supervision that is available [7,12,22,27,49,75], i.e. the supervision signal is the knowledge of which caption belongs to which image. This setting of paired images and text has also been referred to as weakly supervision. Much of the general domain prior work on phrase grounding relies on off-the-shelf object-detection networks [7,12,27,75,83,86] such as Faster R-CNN [64] which are pretrained on large labelled datasets to extract region candidates from images. This considerably simplifies the problem of matching regions to phrases as the set of possible regions to match can be assumed to be known, a luxury that is often unavailable in domain specific contexts.

Biomedical VLP Representation Learning. Several studies [30,31,45,56,85] have explored joint representation learning for paired image and text data in the medical domain. Contrastive VIsual Representation Learning from Text (ConVIRT) [85] uses a contrastive learning formulation for instance-level representation learning from paired medical images and text. The authors uniformly sample sentences and maximise their similarity to true augmented paired images

via the InfoNCE contrastive loss [58], while reducing similarity between negative pairs in the same batch. [31,56] both introduce approaches that combine instance-level image–report contrastive learning with local contrastive learning for medical data. In contrast, [45] use a local-only objective in an approach that approximates the mutual information between grid-like local features of images and sentence-level text features of medical data. The formulation learns image and text encoders as well as a discriminator trained to distinguish positive and negative pairs. While most related approaches use no ground truth, [5] study a semi-supervised edema severity classification setting, and [28] assume sets of seen and unseen labels towards zero-shot classification on CXR data. [44] evaluate pretrained joint embedding models—general domain VLP representation learning models that use a transformer to learn a joint embedding—by fine-tuning the models on CXR data.

Multiple CXR datasets exist that enable a partial evaluation of phrase grounding, but all come with some limitations we hope to mitigate with our `MS-CXR` dataset (see Section 3.1). VinDr [57], RSNA Pneumonia [66], and the NIH Chest X-ray Dataset [76] are datasets that provide bounding-box image annotations, but lack accompanying free-text descriptions. REFLACX [37] provides gaze locations captured with an eye tracker, dictated reports and some ground truth annotations for gaze locations, but no full phrase matches to image regions. Phrase annotations for MIMIC-CXR data released in [71] are of small size (350 studies), only contain two abnormalities, and for some samples have shortened phrases that were adapted to simplify the task. ImaGenome [79] provides a large number of weak local labels for CXR images and reports, with a focus on anatomical regions. However, its ground-truth set is smaller (500 studies), bounding-box regions annotate anatomical regions rather than radiological findings. Furthermore, ImaGenome sentence annotations are not curated, see Table C.3 for some examples. Sentences often contain multiple diseases as well as uncertain findings, making an accurate, largely noiseless grounding evaluation difficult. Some sentences also contain differential diagnosis and temporal change information, which cannot be grounded without access to prior scans.

Language Modelling in Radiology. Most recent general domain VLP work relies on transformer based contextual word embedding models, in particular BERT [17], pretrained on general domain data from newswire and web domains such as Wikipedia. But specific domains often exhibit differences in linguistic characteristics from general text and even related domains, such as between clinical and non-clinical biomedical text as noted in [2], motivating the use of more specialised language models in most related work with a focus on the medical domain. Here, related multi-modal work commonly uses publicly available models including BioBERT [39], ClinicalBERT [2], BioClinicalBERT [2], or PubMedBERT [26], which are either trained from scratch or fine-tuned via continual pretraining using a Masked Language Modelling (MLM) objective. Sometimes additional objectives are added such as adversarial losses [47] or Next Sentence Prediction. [26] provide evidence that training language models from scratch for specialised domains with abundant amounts of unlabelled text can result in

substantial gains over continual pretraining of models first fit to general domain text. The specialised corpora these biomedical and clinical domain models use include PubMed abstracts and PubMed Central full texts, and de-identified clinical notes from MIMIC-III [35]. All the aforementioned language models have a pre-specified vocabulary size consisting of words and subwords, usually 30,000 words in standard BERT. The in-domain vocabulary plays a particularly important role in representative power for a specialised domain. A vocabulary that is not adapted will break up more words into subwords and additionally contain word pieces that have no specific relevance in the specialised domain, hindering downstream learning (see e.g. [26]). As [26] highlight, BERT models that use continual pretraining are stuck with the original vocabulary from the general-domain corpora.

Other closely related tasks in the CXR domain that share similar NLP challenges include report summarisation [11,84], automatic report generation [8,46,54], and natural language inference for radiology reports [54]. Finally, while the name implies close similarity to our CXR-BERT, CheXbert [69] is a BERT based sentence classification model developed for improving the CheXpert [33] labeller, and the model does not have a domain-specific vocabulary like ours or PubMedBERT.

We note that most related work on self-supervised multi-modal learning on CXR data neither explores text augmentation nor maintains text losses such as MLM during multi-modal training. An exception is found in [56], who use the Findings and Impression/Assessment sections of radiology reports, and randomly change the sentence order by swapping pairs of them.

E Model Details

E.1 CXR-BERT Pretraining Details

Our CXR-BERT text encoder is based on the BERT (base size) architecture [73]. We adopt an implementation available via the Huggingface transformers library [78]. The model weights are randomly initialised and pretrained from scratch. As described in Section 2.1, CXR-BERT is pretrained in three phases before the joint pretraining phase. For Phase (I), we use the Huggingface tokeniser library[5] to generate our custom WordPiece vocabulary of 30k tokens. For Phase (II), we use the AdamW [51] optimiser with a batch size of 2048 sequences and a linear learning rate schedule over 250k training steps with a 5% warm up period. We set a base learning rate of 4e-4. Following RoBERTa [48], we pack multiple sentences into one input sequence of up to 512 tokens and use dynamic whole-word masking. In Phase (III), we continue pretraining the model using only MIMIC-CXR text reports. In addition to the MLM loss, we add our RSM loss to pretrain the projection layer. The projection layer P_{txt} is used to project the 768-dimensional feature vector $\tilde{\mathbf{t}}$ to a 128-dimensional report representation $\mathbf{t}$. We use the AdamW optimiser with a batch size of 256 sequences

[5] https://github.com/huggingface/tokenizers

Table E.1: Hyper-parameter values used for image data augmentations.

	Image-Text Pretraining	Image-only Pretraining	Fine-tuning for Downstream Tasks
Affine transform – shear	15°	40°	25°
Affine transform – angle	30°	180°	45°
Colour jitter – brightness	0.2	0.2	0.2
Colour jitter – contrast	0.2	0.2	0.2
Horizontal flip probability	-	0.5	0.5
Random crop scale	-	(0.75, 1.0)	-
Occlusion scale	-	(0.15, 0.4)	-
Occlusion ratio	-	(0.33, 0.3)	-
Elastic transform (σ, α) [68]	-	(4, 34)	-
Elastic transform probability	-	0.4	-
Gaussian noise	-	0.05	-

and a linear learning rate schedule over 100 epochs with a 3% warm up period. We set the base learning rate to 2e-5.

E.2 Image Encoder

Pretraining Details. For the image encoder, we adopt the ResNet50 [29] architecture. The 2048-dimensional feature maps $\tilde{\mathbf{V}}$ of the ResNet50 are projected to 128-dimensional feature maps $\mathbf{V}$ using a two-layer perceptron P_{img} implemented with 1×1 convolutional layers and batch-normalisation [32]. The global image representation $\mathbf{v}$ is obtained by average-pooling the projected local features $\mathbf{V}$. Prior to image-text joint training, the model weights are randomly initialised and pretrained on MIMIC-CXR images using SimCLR [6] — an image-only self-supervised learning approach. We use a large-batch optimisation (LARS) technique [81] on top of ADAM with a batch size of 256 and a linear learning rate scheduler over 100 epochs with a 3% warm up period. We set the base learning rate to 1e-3.

Augmentations. For each training stage, we apply a different set of image augmentations to have a better control over the learnt feature invariances (e.g., laterality). During the image-text joint pretraining stage, we use affine transformations (random rotation and shearing) and contrast and brightness colour jitter. Unlike ConVIRT [85] and GLoRIA [31], we do not apply horizontal flips during the joint training to preserve location information (e.g. "pneumonia in the left lung"). During the image-only SSL (SimCLR) pretraining phase, we use additional image augmentations including random occlusion, additive Gaussian noise, and elastic spatial transforms [68]. We use the implementations available in the torchvision library[6]. The image augmentation parameters and their corresponding values are listed in Table E.1. Before applying these transformations, we normalise the input image intensities by re-scaling each colour channel values to the $[0, 255]$ range. During inference, we only apply centre cropping and resizing.

[6] https://pytorch.org/vision/stable/transforms.html