

Learning Pathology-Localized Medical Vision-Language Representations with Atomic and Grounded Disease Supervision

Mong Yuan Sim^{1,2} Wei Emma Zhang¹ Xiang Dai² Byron Fang³ Miao Liu⁴

¹Adelaide University ²CSIRO ³Oracle ⁴Wuhan University

{mongyuan.sim;wei.e.zhang}@adelaide.edu.au

Abstract

Medical vision-language models are typically trained by aligning whole images with entire radiology reports, assuming that global correspondence is sufficient for learning clinically meaningful representations. We argue that effective medical representation learning requires pathology-localized supervision, which explicitly aligns specific disease observations with corresponding anatomical image regions. To enable this at scale, we propose a fully automated pipeline that decomposes radiology reports into atomic disease observations, identifies relevant regions, and pairs them with corresponding image regions without manual annotation, resulting in a fine-grained dataset, **LocalizedCXR**, with over 6 million region-caption pairs. Extensive experiments demonstrate that pretraining with our pathology-localized semantic supervision signal significantly improve model’s performance in cross-modal retrieval, image classification, phrase grounding, and report generation. Our results suggest that moving beyond global alignment to disease-aware region-level alignment substantially improve the model performance. ¹

1 Introduction

Medical images, such as X-Rays and CT scans play a critical role in clinical practice, as physicians rely on them for screening and diagnosis (Kahn et al., 2009; Gershnik et al., 2011). Effectively aligning these images with associated text reports is crucial for automated diagnosis, clinical decision support, and retrieval of relevant cases. Despite advances in general-domain vision-language representation learning, transferring methods such as CLIP (Radford et al., 2021) to the medical domain remains challenging. Medical images exhibit subtle, domain-specific visual patterns, while radiology reports contain specialized language, implicit

¹Our code is available at: <https://github.com/raymondsim/LocalizedCXR/>

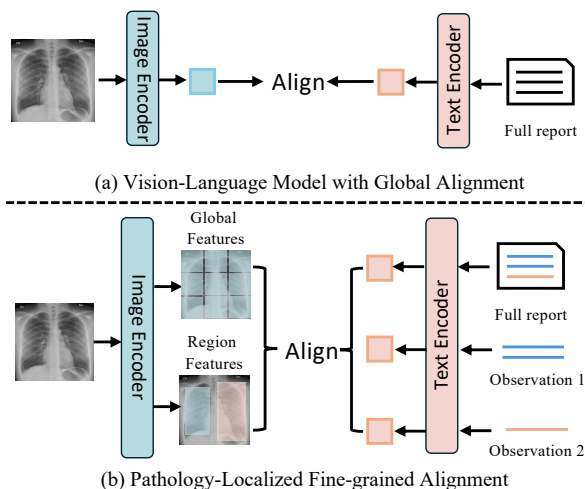


Figure 1: (a) Existing medical vision-language models align an entire image with full radiology reports. (b) We propose an automatic dataset construction pipeline and a dataset, **LocalizedCXR**, to support pathology-localized supervision, which decomposes reports into atomic disease observations and explicitly aligns each observation with its corresponding image region, in addition to global image-report alignment.

clinical reasoning, and compositional descriptions of multiple findings (Chen et al., 2024), complicating the supervision signal for pretraining.

As shown in Figure 1, most existing medical vision-language models align entire images with entire reports (Wang et al., 2022; Codella et al., 2024; Lu et al., 2025). While effective for learning global semantics, such supervision implicitly assumes that a report describes a single, homogeneous visual concept. In practice, radiology reports often describe multiple disease findings, each associated with distinct anatomical regions. This mismatch introduces systematic noise into representation learning, where textual content may be aligned with irrelevant image regions.

To address this issue, several works have proposed methods to capture fine-grained information in medical images using attention mechanisms

(Huang et al., 2021) or region-aware modeling (Chen et al., 2024; Shui et al., 2025; Gu et al., 2025). Others incorporate disease-related knowledge by capturing disease triplets from text reports (Wu et al., 2023) and mining study-level knowledge (You et al., 2023). However, these approaches still rely on implicit alignment or coarse supervision and do not provide explicit signals that localize disease-related atomic observations to their corresponding image regions. Achieving such region-level alignment has remained largely under-explored, primarily due to the limited availability of scalable region-caption pairs.

In this work, we argue that pathology-localized supervision, aligning atomic observations with their relevant anatomical regions, is a missing but crucial component for effective medical vision-language representation learning. To enable this supervision at scale, we propose a fully automated dataset curation framework that decomposes radiology reports into atomic observations, identifies relevant regions, and pairs them with corresponding image regions without manual annotation. A medical expert validation study demonstrates that the domain-specific LLM selects anatomically relevant regions with high agreement compared to radiologist annotations. We further introduce clinically informed hard negatives via minimally contrasting disease descriptions to encourage sensitivity to subtle visual differences.

Experimental results show that models trained with pathology-localized supervision achieve strong overall performance in few-shot and supervised settings, including the best fully supervised results across all three datasets. This demonstrates that grounding atomic observations to localize visual evidence leads to more semantically precise and transferable representations. To summarize, our contributions are threefold:

- We constructed **LocalizedCXR**, which contains over 6 million region-caption pairs, using a fully automated pipeline that derives atomic, grounded disease observations from radiology reports, enabling scalable construction of pathology-aware region-caption pairs without manual annotation.
- We train a medical vision-language model using our dataset, which captures pathology-aware fine-grained information from text reports and X-ray images.

- Extensive experiments demonstrate that pathology-localized supervision improves semantic grounding, retrieval, and classification performance across datasets and pretrained backbones.

2 Related Work

2.1 Medical Vision-Language Models

Existing medical vision-language models adopt contrastive learning by aligning text and images on radiology datasets (Zhang et al., 2020), medical domain articles (Eslami et al., 2023), and various downstream tasks datasets (Lu et al., 2025; Codella et al., 2024). BioViL (Boecking et al., 2022) and MedKLIP (Wu et al., 2023) focus on biomedical-related content from the text report, and enforce alignment on this text instead of the original text report. MedCLIP (Wang et al., 2022) trains encoders with unpaired data and uses medical semantic matching loss instead. CheXzero (Tiu et al., 2022) and RadDINO (Pérez-García et al., 2025) are radiology image encoders trained with a self-supervised method. GLoRIA (Huang et al., 2021) is the most similar to ours, where it focuses on aligning both global and local representations with attention. However, none of these methods provides an explicit fine-grained signal to align text and image representations for the medical domain.

2.2 Fine-grained Radiology Datasets

Several chest X-ray datasets provide fine-grained annotations of abnormalities, but most focus on report-level only, and are either manually or semi-automatically curated, which limits their scale and coverage (Filice et al., 2020; Jaeger et al., 2014; Shih et al., 2019). Bustos et al. (2020) manually annotated a subset of radiology reports and trained a model to label the remaining reports, while Datta and Roberts (2020) curated a small expert-annotated dataset covering a broader range of abnormalities. However, these efforts primarily provide report-level labels and do not establish explicit correspondences between individual disease statements and their associated anatomical image regions. To the best of our knowledge, the only large-scale dataset that contains localized abnormalities in X-ray images is Chest ImaGenome (Wu et al., 2021), which uses a small number of expert annotations to build a prediction model and capture localized annotations for MIMIC-CXR. Nevertheless, these datasets are not designed to support

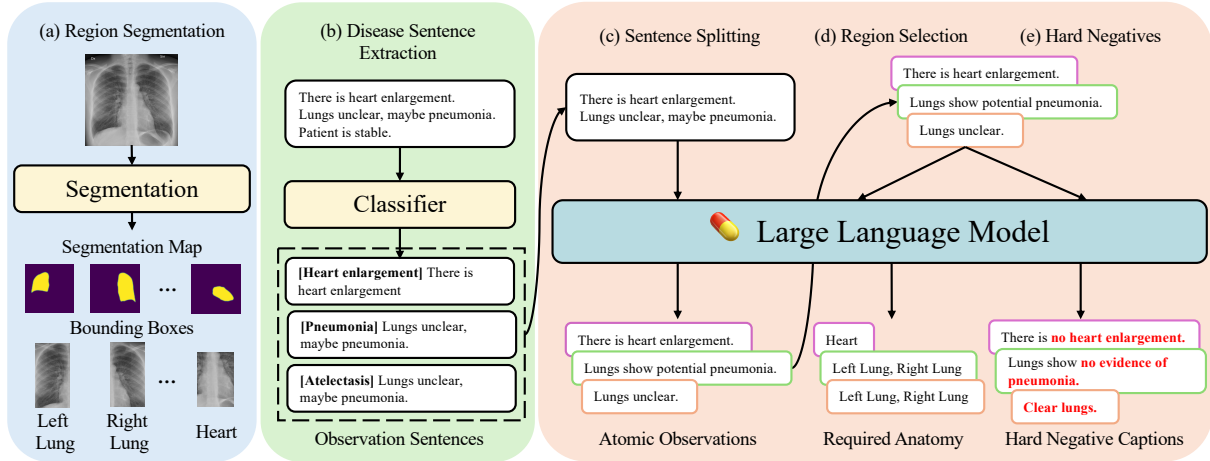


Figure 2: **Method overview.** We propose to automatically build a pathology-aware fine-grained dataset to enable region-level CLIP model training. (a) We obtain anatomical region’s coordinates through segmentation. (b) We extract disease-relevant sentences by forming it into a rule-based classification problem. Then, we use a domain-specific LLM to: (c) split sentences into atomic observations, each containing one disease only; (d) for each atomic observation, we use the LLM to select required anatomy parts from the regions for which coordinates are obtained using a segmentation tool; (e) generate hard negative captions that are semantically similar but clinically different.

large-scale pretraining, motivating our approach to automatically derive pathology-localized supervision from free-text radiology reports.

3 Method

3.1 LocalizedCXR Construction

Radiology reports often contain multiple clinical findings across different anatomical regions. Aligning the full report with the image may introduce supervision noise, preventing the model from gaining a fine-grained understanding of the image (Sim et al., 2025). We address this by decomposing reports into atomic clinical observations and grounding each observation to the relevant image region. Our pipeline consists of five stages: region segmentation, disease sentence extraction, atomic observation construction, anatomical region selection, and hard negative generation.

Region Segmentation To obtain image regions corresponding to anatomical structures, we use an existing segmentation tool (Cohen et al., 2022) to segment each frontal X-ray into anatomical regions. For each region, we extract spatial coordinates from the segmentation masks, enabling region-level image-text alignment during training.

Disease Sentence Extraction We isolate disease-related content from radiology reports, using CheXpert (Irvin et al., 2019), a rule-based classifier that assigns positive, negative, or uncertain labels to 14 disease categories. Since each label is linked to

a sentence span, we extract disease-relevant sentences and discard non-diagnostic content. This step ensures that subsequent supervision is derived from clinically meaningful content and reduces noise introduced by non-diagnostic text.

Atomic Observation Construction Disease-relevant sentences frequently mention multiple findings or anatomical regions within a single sentence, resulting in ambiguous supervision signals. To enforce a one-to-one mapping between language and visual evidence, we further decompose each sentence into atomic disease observations, where each sentence describes exactly one disease finding.

We perform this decomposition using a biomedical large language model (i.e., medgemma-27b (Sellersgren et al., 2025)). For example, the sentence “There is no pleural effusion and pneumothorax.” is decomposed into two atomic observations: “There is no pleural effusion.” and “There is no pneumothorax.”. This constraint reduces compositional ambiguity, enabling more precise alignment during representation learning.

Anatomical Region Selection To ground each atomic observation to the corresponding regions, we utilize the same domain-specific LLM (i.e., medgemma-27b (Sellersgren et al., 2025)) to predict related anatomical region(s). For example, the atomic observation “Observed heart enlargement” will be paired with the region *heart* only. This step provides explicit pathology-localized supervision

	MIMIC-CXR	ChestXray14	CheXpert-Plus
# exp.	240,197	69,718	190,954
# atom. sent.	8,878	1,828	347,611
# region	1.80m	299k	4.15m
# avg. region	7.48	4.29	21.73

Table 1: Dataset statistics for the constructed fine-grained dataset with region-caption pairs in each dataset. * #atom. sent. refers to the number of unique atomic observations; # region is the total number of region-caption pairs per dataset; # avg. region is the average number of regions per image.

signal by linking disease sentence to anatomically meaningful image regions.

Hard Negatives Generation To further strengthen the supervision signal, we generate clinically informed hard negatives for each region-caption pair. Using the same biomedical LLM, we modify a single clinical attribute of the original observation, such as disease presence/absence, severity, or anatomical location, producing captions that are semantically similar but clinically different. By contrasting true and counterfactual captions within the same anatomical region, the model is encouraged to attend to subtle but clinically important visual distinctions (Yang et al., 2023; Xie et al., 2025).

As a result, each image-report pair is associated with multiple atomic observation-region pairs. The resulting fine-grained dataset contains over 6 million region-caption pairs for MIMIC-CXR, ChestXray14, and CheXpert-Plus datasets in total, as summarized in Table 1. The prompts used are listed in Appendix D, and the tools employed are described in the implementation details.

3.2 Pretraining with LocalizedCXR

In order to use **LocalizedCXR** for pretraining, we adapt the same architecture as CLIP (Radford et al., 2021), which consists of a vision encoder and a language encoder, with multiple training objectives.

Global-level Contrastive Learning aims to enhance the model’s coarse-grained understanding by aligning text and image representations of the image-report pairs, which contain rich global information. The objective function is based on the InfoNCE loss (He et al., 2020), which maximizes the similarity between matching image-text pairs while minimizing the similarity between mismatched pairs. Specifically, the similarity between each pair is calculated using cosine similarity be-

tween embeddings:

$$s(v, t) = \frac{v \cdot t^T}{\|v\| \|t\|}. \quad (1)$$

To align image and text representations, we use InfoNCE loss for a batch of N image-text pairs:

$$L_{global} = -\frac{1}{2N} \sum_{i=1}^N \left(\log \frac{\exp(s(v_i, t_i)/\tau)}{\sum_{j=1}^N \exp(s(v_i, t_j)/\tau)} + \log \frac{\exp(s(t_i, v_i)/\tau)}{\sum_{j=1}^N \exp(s(t_i, v_j)/\tau)} \right), \quad (2)$$

where τ is a learnable temperature parameter.

Pathology-Aware Region Aggregation aligns related regions to the corresponding atomic observation. Formally, for each atomic observation, we aggregate all related region embeddings extracted using RoIAlign: where a_k is a learnable attention weight. The loss is computed as the contrastive loss between the atomic observation text embedding and the aggregated regions.

For each atomic observation o_i , suppose the LLM selects a set of relevant anatomical regions:

$$\mathcal{R}_i = \{r_1, r_2, \dots, r_m\}.$$

Each selected region is first encoded using the same RoIAlign-based region extraction module as in our current pipeline. Instead of treating the selected regions independently or averaging them uniformly, we introduce a pathology-aware region aggregation module to combine them into a single observation-specific visual representation.

Specifically, for atomic observation o_i :

$$z_i = \sum_{k=1}^m a_k r_k, \quad (3)$$

where r_k denotes the embedding of the k -th selected region, and a_k is a learned attention weight over the selected regions.

The attention weight is computed as:

$$a_k = \frac{\exp(q_{o_i}^\top W r_k)}{\sum_{j=1}^m \exp(q_{o_i}^\top W r_j)}, \quad (4)$$

where q_{o_i} is the text embedding of the atomic observation, and W is a learnable projection matrix.

The resulting aggregated region representation z_i is then aligned with the text embedding t_i of the corresponding atomic observation using a pathology-level contrastive objective. For a batch of M

atomic observation–region pairs, the pathology-level contrastive loss is defined as:

$$L_{path} = -\frac{1}{2M} \sum_{i=1}^M \left(\log \frac{\exp(s(z_i, t_i)/\tau)}{\sum_{j=1}^M \exp(s(z_i, t_j)/\tau)} + \log \frac{\exp(s(t_i, z_i)/\tau)}{\sum_{j=1}^M \exp(s(t_i, z_j)/\tau)} \right), \quad (5)$$

where $s(\cdot, \cdot)$ denotes cosine similarity and τ is a learnable temperature parameter. The first term aligns each aggregated pathology-aware visual representation with its corresponding atomic observation, while the second term performs the reverse text-to-region alignment.

Hard Fine-grained Negative Samples Learning

To enable our model to learn more discriminative differences between disease-related features from image and text, we introduce a hard negative data generation strategy. These hard negatives are constructed by modifying disease-related attributes in the atomic observations using a domain-specific LLM as introduced in Section 3.1.

To incorporate hard negative samples into the learning process, we follow Xie et al. (2025) to include a hard negative loss term. For each text-region pair, we compute the similarity between the regional feature and both the positive caption and the corresponding hard negative caption. The hard negative loss L_{hard} is defined as:

$$L_{hard} = -\frac{1}{K} \sum_{i=1}^K \log \frac{\exp(s(r_i, l_{i,1})/\tau)}{\sum_{j=1}^M \exp(s(r_i, l_{i,j})/\tau)}, \quad (6)$$

where M denotes the total number of captions per region, with $j = 1$ referring to the positive sample, and $j > 1$ referring to the negative samples, which is set to exactly 1 in all of our experiments.

The overall learning objective of our model is:

$$L = L_{global} + \alpha * L_{path} + \beta * L_{hard}. \quad (7)$$

where hyper-parameters α and β are both set to 0.3, to balance the global loss, pathology contrastive loss, and hard negative loss.

4 Experiments

4.1 Datasets for Constructing Localized CXR

Using our automatic dataset construction pipeline, we derive pathology-localized region-caption pairs from three large-scale radiology image-text

datasets. **MIMIC-CXR** (Johnson et al., 2019) contains 227k image-report pairs, covering both frontal and lateral X-ray images. **Chexpert-Plus** (Chambon et al., 2024) contains 223k image-report pairs, from 64k patients. Following You et al. (2023), we use frontal-view images and binary labels from **ChesX-ray14** (Wang et al., 2017), and convert these labels to reports by randomly sampling disease sentences. Detailed description for pretraining datasets is provided in Appendix G.

4.2 Datasets for Downstream Tasks

We evaluate medical vision-language representations across retrieval, classification, and grounding tasks. Cross-modal retrieval is evaluated on **MIMIC-CXR** (Johnson et al., 2019), **CheXpert-Plus** (Chambon et al., 2024), and **OpenI** (Demner-Fushman et al., 2015), which contain image-report pairs. Classification is evaluated on **RSNA Pneumonia** (Shih et al., 2019), **SIIM Pneumothorax** (Zawacki et al., 2019), and **VinDR-CXR** (Nguyen et al., 2020). Phrase grounding is evaluated on **MS-CXR** (Boecking et al., 2022). A more detailed description of each downstream dataset is provided in Appendix C.

4.3 Baseline Models

We compare against representative medical vision-language representation learning methods, including **CXR-CLIP** (You et al., 2023), **MedCLIP** (Wang et al., 2022), **PubMedCLIP** (Eslami et al., 2023), **RadCLIP** (Lu et al., 2025), **MedImageInsight** (Codella et al., 2024), which mainly differ in their supervision sources. We also compare with **RadDINO** (Pérez-García et al., 2025), a self-supervised model, which we evaluate only on image classification because it lacks a text encoder for retrieval, zero-shot classification, and grounding.

5 Results

5.1 Dataset Quality Validation

To assess the quality of our pathology-localized supervision, we conduct two validation studies. First, an experienced radiologist evaluates region selection by identifying anatomically relevant regions for each atomic disease observation. On 500 randomly sampled instances from MIMIC-CXR and CheXpert-Plus, LLM achieves an agreement score of 91.27%, and 93.14% recall, indicating strong alignment with expert judgment and barely misses relevant regions.

Method	MIMIC-CXR			OpenI			CheXpert-Plus		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
<i>Image-to-Text Retrieval (I2T)</i>									
GLoRIA	0.68	4.80	10.10	0.77	1.53	2.30	1.80	6.50	10.80
CXR-CLIP	23.15	52.22	65.64	7.35	16.08	21.75	9.30	22.70	31.80
MedCLIP	0.06	0.31	0.62	0.15	0.77	1.53	0.10	0.50	1.00
PubMedCLIP	0.12	0.80	1.35	0.00	1.53	2.30	0.20	0.70	1.10
RadCLIP	0.18	0.74	1.54	0.30	1.50	2.50	0.30	1.50	2.50
MedImageInsight	24.69	55.60	67.24	6.74	15.31	21.13	9.20	20.80	29.10
Ours	26.35	55.79	68.66	7.96	17.30	22.82	18.20	32.60	43.00
<i>Text-to-Image Retrieval (T2I)</i>									
GLoRIA	8.13	22.84	35.28	3.68	9.49	13.48	9.60	29.50	43.20
CXR-CLIP	24.20	53.82	66.87	7.66	17.15	22.36	8.90	22.60	32.40
MedCLIP	0.06	0.31	0.55	0.15	0.77	1.68	0.10	0.80	1.40
PubMedCLIP	0.25	0.68	1.11	0.00	0.61	1.84	0.20	0.80	1.20
RadCLIP	0.06	0.55	1.23	0.40	1.40	2.40	0.40	1.40	2.40
MedImageInsight	27.28	55.85	67.06	8.42	20.21	26.49	10.60	22.70	31.80
Ours	28.76	57.02	69.64	7.96	15.77	22.82	14.20	30.70	40.30

Table 2: Image-to-text (I2T) and text-to-image (T2I) retrieval performance across MIMIC-CXR, OpenI, and CheXpert-Plus.

We further manually validate sentence splitting and hard negative generation. The extracted observations achieve 99.8% disease atomicity, 93.67% completeness, and 99.42% faithfulness, indicating that the splitting step preserves clinical meaning while producing fine-grained observations. Hard negatives receive an average score of 4.17/5, suggesting challenging but appropriate contrastive supervision. Together, these validations show that our pipeline and LocalizedCXR provide reliable pathology-localized supervision.

5.2 Cross-Modal Retrieval

We evaluate image-to-text and text-to-image retrieval using Recall@K. As shown in Table 2, our model consistently outperforms existing methods across all three image-report datasets. It shows improvement over the existing best baseline models, MedImageInsight, by a great margin. There is an exception for this, where our model performs slightly worse than MedImageInsight on OpenI for the text-to-image retrieval task. We also noticed that our model’s performance on CheXpert-Plus is significantly better than other baseline models.

It is noticeable that models trained with disease-level data instances like CXR-CLIP and our model perform much better on retrieval tasks compared to other models like MedCLIP and PubMedCLIP. This highlights the importance of explicitly introducing disease-related features in medical vision-language representation learning.

The near-zero results for MedCLIP, PubMedCLIP, and RadCLIP likely reflect pretraining mismatch with exact CXR report-pair retrieval. These models use decoupled, article-based, or broad radiology supervision, respectively, whereas CXR-CLIP, MedImageInsight, and our model are trained on chest X-ray-report pairs, leading to substantially stronger retrieval performance.

5.3 Image Classification

To evaluate representation quality and transferability, we evaluate models under zero-shot, few-shot, and supervised classification settings on binary classification datasets (i.e., SIIM and RSNA) and multi-class classification datasets (i.e., VinDr-CXR) using Area Under the Curve (AUC).

Table 3 presents classification results on different datasets. Our model consistently outperforms existing methods in few-shot and supervised settings. Although pretrained on RSNA Pneumonia, RadCLIP underperforms CXR-CLIP and our model, suggesting that task-related pretraining alone is insufficient and fine-grained supervision plays a more critical role in learning transferable representations.

For zero-shot classification, we follow prior work (You et al., 2023), to use text prompts to represent positive and negative labels for each disease, and compute the similarity of the image and the text prompts. Our model achieves the strongest zero-shot performance on RSNA and competitive performance on SIIM, while GLoRIA and MedImageIn-

	VinDR-CXR			SIIM			RSNA		
	ZS	10%	100%	ZS	10%	100%	ZS	10%	100%
GLoRIA	87.02	54.72	55.91	87.02	91.99	92.85	71.18	87.91	89.26
CXR-CLIP	38.61	41.67	48.00	83.30	91.92	93.16	85.20	89.14	90.25
MedCLIP	31.79	42.80	43.05	71.73	71.73	88.67	81.76	87.84	88.99
Rad-DINO	-	44.76	53.00	-	52.53	55.32	-	62.08	68.74
PubMedCLIP	40.28	44.27	51.97	71.79	72.09	76.96	76.94	79.48	83.71
RadCLIP	40.53	49.73	50.63	70.12	73.29	84.48	78.09	83.20	85.67
MedImageInsight	58.27	43.62	50.33	83.25	95.31	95.94	86.72	84.36	91.12
Ours	43.66	45.31	56.33	85.63	95.72	96.33	88.61	89.95	91.54

Table 3: The performance of image encoders on the image classification task under both zero-shot (ZS) and fine-tuning settings evaluated using AUC. The percentage indicates the portion of training data used. * Rad-DINO only has an image encoder; therefore, it cannot be used in a zero-shot manner.

Model	Atelectasis	Cardiomegaly	Consolidation	Lung Opacity	Edema	Pneumonia	Pneumothorax	Pl. Effusion	Avg.
GLoRIA	0.98	0.53	1.38	1.05	0.66	1.18	0.47	1.20	0.93
CXR-CLIP	-0.15	-0.46	0.00	-0.27	0.01	-0.14	0.55	-0.54	-0.11
MedCLIP	-0.07	-0.85	0.09	-0.13	-0.06	0.02	0.24	-0.97	-0.26
PubMedCLIP	-0.46	-0.24	-0.89	-0.62	-0.92	-0.80	-0.80	-0.25	-0.60
RadCLIP	-1.28	-0.67	-1.16	-1.14	-1.34	-1.28	-0.58	-1.04	-0.96
MedImageInsight	1.55	1.72	1.80	1.82	1.32	1.93	1.74	2.18	1.78
Ours	0.59	1.90	1.90	2.10	1.22	2.28	2.04	2.14	1.96

Table 4: Grounding Task. Contrast-to-noise ratio (CNR) on MS-CXR dataset.

Model	R-L	B-4	C.Bert	R.Graph	GREEN
Gloria	19.12	4.60	17.49	40.16	28.97
CXRCLIP	24.81	5.51	22.64	39.47	27.53
MedCLIP	25.41	6.61	23.31	40.68	28.35
RadDINO	25.35	5.98	23.44	41.11	28.15
PubMedCLIP	23.20	4.46	21.11	38.48	24.66
RadCLIP	24.32	5.58	22.44	38.21	26.69
MedImageInsight	24.74	5.21	22.90	40.33	27.07
Ours	25.65	6.63	24.40	42.11	29.88

Table 5: Radiology report generation performance on the official MIMIC-CXR dataset. Higher scores indicate better performance for all metrics.

sight perform better on VinDR-CXR. We exclude Rad-DINO for zero-shot evaluation as it lacks a text encoder needed for image-text similarity. VinDR-CXR is a multi-label dataset with 28 radiologists-defined findings, and uses label names/prompts that does not always match the disease-observation language used during pretraining. Therefore, the result is very poor compared to SIIM and RSNA zero-shot settings. This explains why performance improves under fine-tuning while zero-shot setting remains disadvantaged.

5.4 Phrase Grounding

We evaluate phrase grounding on the MS-CXR dataset (Boecking et al., 2022), a radiologist-annotated dataset with bounding boxes paired with radiology text descriptions. We use contrast-to-noise ratio (CNR) to measure how well the model localizes disease-related image regions given tex-

tual input, by comparing scores inside and outside the box. More details about CNR calculation can be found at Appendix I. As in zero-shot classification, Rad-DINO is excluded as it lacks a text encoder. Table 4 shows the effectiveness of our model over other existing models on the phrase grounding task across diseases. Our model performs better than MedImageInsight in 5 disease types out of 8, and achieves a better overall result.

5.5 Report Generation

We evaluate the performance of different image encoders in radiology report generation. Experimental setup for report generation model training is provided in Section H.2. Table 5 presents the radiology report generation results. In addition to widely used generation evaluation metrics (i.e., ROUGE-L (Lin, 2004) and BLEU-4 (Papineni et al., 2002)), and radiology report generation metrics (i.e., CheXbert and RadGraph), we also use GREEN (Ostmeier et al., 2024), an LLM-based evaluation metric that assesses the clinical correctness of generated reports by identifying clinically significant errors. Results show superior results across all evaluation metrics, compared to other image encoders. This suggests that our pathology-localized pretraining provides more effective visual representations for downstream report generation, capturing more fine-grained and clinically meaningful information.

	R@1	R@5	R@10
MedImageInsight	24.69	55.60	67.24
+ global-only	25.49	54.50	66.69
+ pathology-loss	26.54	56.96	69.77
+ atomic observations	26.52	55.43	68.20
+ hard negative	26.97	57.27	69.33
FG-CLIP	0.12	0.37	1.05
+ global-only	15.54	38.49	50.71
+ pathology loss	16.19	39.47	53.08
+ atomic observations	16.54	40.11	52.90
+ hard-negative	17.49	42.18	56.16
CLIP	0.06	0.31	0.62
+ global-only	14.89	36.82	50.30
+ pathology loss	15.21	36.15	48.28
+ atomic observations	15.21	39.50	51.43
+ hard negative	15.76	38.42	49.57

Table 6: The effectiveness of our pipeline in different pretrained models, tested using MIMIC-CXR on image-to-text retrieval task.

5.6 Ablation Studies

To evaluate the contributions of different components in our dataset construction pipeline and training objectives, we conduct an extensive ablation study with results summarized in Table 6.

Region-level Supervisory Signal We compare models trained using only global image-text alignment with those incorporating both global and region-level information. As shown in Table 6, introducing region-level fine-grained supervision consistently improves image-to-text retrieval performance. Results also show that introducing region-level fine-grained further demonstrate that this improvement holds when initializing from general-domain CLIP models (i.e., FG-CLIP and CLIP). These findings indicate that region-level captions provide effective fine-grained supervisory signals for aligning visual and textual representations in the medical domain.

Atomic Observations To reduce dataset noise and improve alignment quality, we split disease descriptions into atomic observations such that each sentence refers to a single disease only, as described in Section 3.1. The motivation of this design is to enable more precise image-text correspondence during training. Compared to region-level alignment alone, this strategy yields substantial performance gains across different datasets, as shown in Table 6. These results highlight the importance of dataset granularity and cleanliness for effective vision-language alignment.

	R@1	R@5	R@10
MedImageInsight	24.69	55.60	67.24
w/ ImaGenome	25.87	54.40	67.53
w/ Ours	26.27	55.83	68.54
FG-CLIP	0.12	0.37	1.05
w/ ImaGenome	13.21	34.65	46.64
w/ Ours	14.69	36.20	50.11
CLIP	0.06	0.31	0.62
w/ ImaGenome	13.35	35.20	46.90
w/ Ours	12.82	35.58	48.00

Table 7: The effectiveness of **LocalizedCXR** compared to ImaGenome dataset, testing using MIMIC-CXR on image-to-text retrieval task

Hard Negative Samples Learning We further introduce hard negative samples learning, to encourage the model to distinguish semantically similar but clinically different patterns. As shown in Table 6, adding L_{hard} consistently improves performance, particularly for models initialized from FG-CLIP and CLIP weights. These results underscore the importance of L_{hard} in guiding the model to learn subtle disease-related semantic differences.

Table 6 compares models initialized from different pretrained weights and subsequently trained on our constructed dataset. Across both general-domain and domain-specific initializations, pathology-aware and fine-grained supervision consistently improve performance. In particular, all models benefit from the use of region-level supervision and hard negative losses, leading to substantial improvements over their corresponding baselines. This suggests that the proposed supervisory signals are complementary to pretraining and remain effective regardless of the initialization domain.

5.7 Comparison to Existing Datasets

ImaGenome (Wu et al., 2021), to the best of our knowledge, is the only large-scale dataset that contains localized abnormalities in X-ray images. To compare the effectiveness of existing resources and our method, we train both general and medical domain models on ImaGenome and LocalizedCXR². From the Table 7, we observe that models trained using our dataset outperforms models trained with ImaGenome, when evaluated on image-to-text retrieval. Importantly, our dataset construction framework is not specific to MIMIC-CXR and can be readily applied to other radiology

²For a fair comparison, we only use the subset of LocalizedCXR that are derived from MIMIC-CXR.

datasets, enabling scalable and flexible fine-grained supervision without requiring manual annotation.

6 Conclusion

In this work, we presented an automated framework for constructing a pathology-aware, fine-grained radiology image-text dataset, that enables fine-grained alignment between radiology text and image. By focusing on disease-related sentences and regions, our framework integrates global and region-level supervision for medical vision-language representation learning. Extensive experiments across three tasks, seven datasets, and three backbone models validate our framework’s superior performance. Future work could focus on extending this automatic fine-grained dataset construction framework to other radiology images and a broader range of disease types. In addition, further study could be conducted to understand the difference between radiologist and LLM annotations to reduce the error rate and improve the dataset quality for better alignment.

Limitations

Despite the strong empirical performance of our framework, we acknowledge several limitations of our work, and the results should be interpreted with caution.

- Our work focuses on chest X-ray images and primarily uses data from U.S. hospitals (except for one downstream task dataset). Extending this framework to other radiological modalities, such as MRI and CT scans, as well as to data from other regions, requires further investigation.
- Our focus is on constructing high-quality training data to enable the model to gain fine-grained understanding of images through CLIP training. Therefore, we selected our evaluation task accordingly. Whether such an improved model can enhance diagnostic accuracy, support writing radiology report, or aid clinical decision-making requires more rigorous investigation.
- We restrict the anatomical regions to 14 regions in this study. While practical, this fixed schema may miss finer anatomical details. Errors from the segmentation model can also propagate to the constructed region-caption

pairs. Nevertheless, we view this design as a first step toward fully fine-grained anatomical supervision, with future work extending it to more detailed and uncertainty-aware region annotations.

Ethical Considerations

This work focuses on medical vision-language representation learning and is intended to support, rather than replace, a real clinician for clinical decision-making. All experiments are conducted using de-identified radiology datasets (see dataset license in Section F, and no patient-identifiable information is used. Although our framework aims to improve fine-grained medical image-text alignment, model predictions may still contain errors and should not be used as a standalone diagnostic evidence. Deployment in real clinical settings would require careful validation across institutions, imaging devices, patient populations, and disease distributions, as well as oversight by qualified medical professionals.

Acknowledgment

Mong Yuan Sim is supported by CSIRO’s Data61 Top-up scholarship. Wei Emma Zhang is supported by the Australian Research Council Early Career Industry Fellowships (IE230100119).

References

- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie L. Hyland, Maria Wetscherek, Tristan Naumann, Aditya V. Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. [Making the most of text semantics to improve biomedical vision-language processing](#). In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 1–21. Springer.
- Aurelia Bustos, Antonio Pertusa, José María Salinas, and María de la Iglesia-Vayá. 2020. [Padchest: A large chest x-ray image dataset with multi-label annotated reports](#). *Medical Image Anal.*, 66:101797.
- Pierre J. Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven Q. H. Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. [Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats](#). *CoRR*, abs/2405.19538.

- Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. [Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 9494–9509. Association for Computational Linguistics.
- Noel C. F. Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, Hoi-fung Poon, Stephanie Hyland, Shruthi Bannur, Javier Alvarez-Valle, Xue Li, John Garrett, Alan McMillan, Gaurav Rajguru, Madhu Maddi, and 12 others. 2024. [Medimageinsight: An open-source embedding model for general domain medical imaging](#). *Preprint*, arXiv:2410.06542.
- Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarnera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. 2022. [Torchxrayvision: A library of chest x-ray datasets and models](#). In *International Conference on Medical Imaging with Deep Learning, MIDL 2022, 6-8 July 2022, Zurich, Switzerland*, volume 172 of *Proceedings of Machine Learning Research*, pages 231–249. PMLR.
- Surabhi Datta and Kirk Roberts. 2020. [A dataset of chest x-ray reports annotated with spatial role labeling annotations](#). *Data in Brief*, 32:106056.
- Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. [Preparing a collection of radiology examinations for distribution and retrieval](#). *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. [Pubmedclip: How much does CLIP benefit visual question answering in the medical domain?](#) In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1151–1163. Association for Computational Linguistics.
- Ross W. Filice, Anouk Stein, Carol C. Wu, Veronica A. Arteaga, Stephen Borstelmann, Ramya Gaddikeri, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C. Godoy, Stephen B. Hobbs, Jean Jeudy, Paras Lakhani, Archana Laroia, Sundeep M. Nayak, Maansi R. Parekh, Prasanth Prasanna, Palmi Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and George Shih. 2020. [Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest x-ray dataset](#). *J. Digit. Imaging*, 33(2):490–496.
- E. F. Gershanik, R. Lacson, and R. Khorasani. 2011. [Critical finding capture in the impression section of radiology reports](#). *AMIA Annu Symp Proc*, 2011:465–469.
- Difei Gu, Yunhe Gao, Mu Zhou, and Dimitris N. Metaxas. 2025. [Anatomy-vlm: A fine-grained vision-language model for medical interpretation](#). *CoRR*, abs/2511.08402.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. [Momentum contrast for unsupervised visual representation learning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE.
- Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. [Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition](#). In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3922–3931. IEEE.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghighi, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. [Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597. AAAI Press.
- Stefan Jaeger, Sema Candemir, Sameer Antani, Yi-Xiang J. Wang, Pu-Xuan Lu, and George Thoma. 2014. [Two public chest x-ray datasets for computer-aided screening of pulmonary diseases](#). *Quantitative Imaging in Medicine and Surgery*, 4(6).
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. [MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs](#). *CoRR*, abs/1901.07042.
- C. E. Kahn, C. P. Langlotz, E. S. Burnside, J. A. Carrino, D. S. Channin, D. M. Hovsepian, and D. L. Rubin. 2009. [Toward best practices in radiology reporting](#). *Radiology*, 252(3):852–856.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2021. [Grounded language-image pre-training](#). *CoRR*, abs/2112.03857.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summariza-*

- tion Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Zhixiu Lu, Hailong Li, Nehal A. Parikh, Jonathan R. Dillman, and Lili He. 2025. [Radclip: Enhancing radiologic image analysis through contrastive language-image pretraining](#). *IEEE Trans. Neural Networks Learn. Syst.*, 36(10):17613–17622.
- Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Han Tong, Diep H. Dinh, Cuong Duy Do, Luu Tien Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, A. D. Hoang, Hien Ngoc Phan, Anh Thu Nguyen, Phuong Ho, and 5 others. 2020. [Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations](#). *Scientific Data*, 9.
- Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael E. Moseley, Curtis P. Langlotz, Akshay Chaudhari, and Jean-Benoit Delbrouck. 2024. [GREEN: generative radiology report evaluation and error notation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, Findings of ACL*, pages 374–390. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.
- Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. 2025. [Exploring scalable medical image encoders beyond text supervision](#). *Nat. Mac. Intell.*, 7(1):119–130.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla P. Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany L. Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. [Medgemma technical report](#). *CoRR*, abs/2507.05201.
- George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan Vummidi, Kavitha Yadanapudi, and Anouk Stein. 2019. [Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia](#). *Radiology: Artificial Intelligence*, 1(1):e180041. PMID: 33937785.
- Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. 2025. [Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoan Fang. 2025. [Can vlms actually see and read? A survey on modality collapse in vision-language models](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24452–24470. Association for Computational Linguistics.
- Ekin Tiu, Ellie Talius, Pujan Patel, Curt P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. 2022. [Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning](#). *Nature Biomedical Engineering*, 6:1399 – 1406.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. [Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3462–3471. IEEE Computer Society.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. [Medclip: Contrastive learning from unpaired medical images and text](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3876–3887. Association for Computational Linguistics.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. [Medclip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 21315–21326. IEEE.
- Joy T. Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and

Mehdi Moradi. 2021. [Chest imagenome dataset for clinical reasoning](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. 2025. [FG-CLIP: fine-grained visual and textual alignment](#). In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Haoran Yang, Hongxu Chen, Sixiao Zhang, Xiangguo Sun, Qian Li, Xiangyu Zhao, and Guandong Xu. 2023. [Generating counterfactual hard negative samples for graph contrastive learning](#). In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 621–629. ACM.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun Kyoung Hong, Woonhyuk Baek, and Byungseok Roh. 2023. [CXR-CLIP: toward large scale chest x-ray language-image pre-training](#). In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part II*, volume 14221 of *Lecture Notes in Computer Science*, pages 101–111. Springer.

Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. 2019. [Siim-acr pneumothorax segmentation](#). Kaggle.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. [Long-clip: Unlocking the long-text capability of CLIP](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LI*, volume 15109 of *Lecture Notes in Computer Science*, pages 310–325. Springer.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. [Contrastive learning of medical visual representations from paired images and text](#). *CoRR*, abs/2010.00747.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. 2021. [Regionclip: Region-based language-image pretraining](#). *CoRR*, abs/2112.09106.

A Extended Related Work

A.1 Fine-grained Visual Representation Learning

Contrastive learning for vision-language representation modeling is able to align text and image, which enables superior performance in zero-shot

classification and multimodal retrieval. While effective, CLIP (Radford et al., 2021) usually aligns the entire image with its corresponding caption, missing region-level fine-grained information. Several works have addressed this issue by introducing a region-level training objective to guide the model in learning more fine-grained information from the image and text pairs. LongCLIP (Zhang et al., 2024) extends the CLIP text encoder to receive longer text, in order to learn more details from the image caption. GLIP (Li et al., 2021) utilizes a grounding dataset for pretraining. RegionCLIP (Zhong et al., 2021) and FG-CLIP (Xie et al., 2025) automatically obtain region-specific information from existing datasets and incorporate it into pretraining. Nevertheless, these works focus on general-domain visual-text alignment only, while such a design is harder for medical domain due to data constraints.

B Implementation Details

We use pretrained weights from MedImageInsight (Codella et al., 2024) as the backbone text and image encoders. We use MedGemma-27b (Søllergren et al., 2025) as a domain-specific LLM for sentence splitting, region selection, and hard negative caption generation. To obtain coordinates for the anatomy parts, we utilize PSPNet from TorchXRayVision (Cohen et al., 2022) for segmentation and compute coordinates from the segmentation mask. For the classification task, we freeze the trained image encoder, add a linear layer as a classification head, and train for 10 epochs. For the retrieval task, we obtain the average pooled embeddings (i.e., [CLS] token) from both the image and text encoders, and compute an exact match. Our model is trained for 20 epochs, and it takes approximately 30 hours on 4 NVIDIA H100 GPUs.

C Detailed Description for Downstream Task Datasets

OpenI (Demner-Fushman et al., 2015) is a small-scale radiology report dataset collected by Indiana University. We use the official 653 image-report pairs in the official test set for the retrieval task.

RSNA Pneumonia (Shih et al., 2019) is a binary-labeled dataset where the task is to classify images into pneumonia or normal, collected by RSNA (Radiological Society of North America). We follow Huang et al. (2021) to split the train/valid/test set as 70%/15%/15% respectively.

SIIM-ACR Pneumothorax (Zawacki et al., 2019) is also a binary-labeled dataset, where images are classified as pneumothorax or normal. It contains more than 260k frontal-view chest X-rays. We split the dataset into 70%/15%/15% for train/valid/test set, following (Huang et al., 2021).

VinDR-CXR (Nguyen et al., 2020) contains 28 disease labels for 18k images, annotated by radiologists. Following (You et al., 2023), we obtain a validation set from the original training set, resulting in 12k/3k/3k for train/val/test set, respectively.

MS-CXR (Boecking et al., 2022) is a small-scale phrase grounding dataset that contains 1162 radiologist annotated image regions along with the corresponding phrase, obtained from the text report.

D Prompt Used for Dataset Construction

D.1 Prompts for Splitting Sentences

You are a radiology text normalization assistant. Your task is to split disease/finding sentences into multiple sentences, each with exactly one disease. Rules: Each sentence must be full, grammatically correct, and self-contained. Preserve negation accurately (“no”, “absent”, “without”). Do not create redundant or overlapping sentences. Do not invent new diseases or anatomy. Output always as a JSON list of strings.

Example 1:

Input: “heart size is normal, no mediastinal vein dilatation or vascular engorgement.”

Output: [

“Heart size is normal.”,

“There is no mediastinal vein dilatation.”,

“There is no vascular engorgement.”

]

Example 2:

Input: “Cardiomegaly is present.”

Output: [

“Cardiomegaly is present.”

]

Example 3:

Input: “no edema/effusion, atelectasis or infiltrate”

Output: [

“No edema.”,
“No effusion.”,
“No atelectasis.”,
“No infiltrate.”
]

Example 4:

Input: “there are patchy basilar opacities in the retrocardiac region of the left lower lobe.”

Output: [

“There are patchy basilar opacities in the retrocardiac region of the left lower lobe.”

]

Output the result ONLY in the following JSON format: [

sentence 1,

sentence 2,

...

]

D.2 Prompts for Region Selection

You are a radiology reasoning assistant. Your task is to analyze a disease description or clinical finding, and determine which anatomical structures must be evaluated to confirm the presence or absence of the condition. Only select anatomy parts from the provided list. Do not infer additional anatomy or add parts that are not explicitly listed. If the disease is too vague to map to anatomy, return an empty list. Be concise and deterministic. If an anatomy part is not in the list, do not include it under any circumstances.

Disease description: {DISEASE_SENTENCE}

Available anatomy parts: {ANATOMY_LIST}

Output the result ONLY in the following JSON format:

{{“disease”: [{DISEASE_SENTENCE}] ,

“required_anatomy_parts”: [...],

“reasoning”: “Short explanation of why these parts are needed”

}}

D.3 Prompts for Hard Negatives Generation

You are generating ONE hard negative medical sentence for training a vision-language model.

Given a disease description sentence, produce exactly ONE hard negative that:

- Is medically plausible
- Differs from the original by ONLY ONE clinically meaningful detail
- Preserves wording, structure, and length as much as possible
- Does NOT introduce new findings or organs
- Does NOT change more than one attribute

Allowed single modifications (choose exactly one):

- Presence or absence
- Severity (mild / moderate / severe)
- Laterality (left / right / bilateral)
- Temporal status (acute / chronic)
- Closely related disease subtype within the same disease family
- Location within the same organ

IMPORTANT:

- Change at most 5 words
- Do NOT rewrite the sentence
- Output ONLY the modified sentence

Example 1:

Input: "There is no pleural effusion."

Output: "There is a small pleural effusion."

Example 2:

Input: "Mild cardiomegaly is present."

Output: "Severe cardiomegaly is present."

Example 3:

Input: "Bilateral pulmonary infiltrates are noted."

Output: "Right-sided pulmonary infiltrates are noted."

E Medical Expert Validation

E.1 Annotation Guideline

To validate the quality of LLM annotation, we conduct an expert annotation to select the required anatomy parts, given a disease sentence. We provide an annotator with a simple interface, with the following short guidelines:

Given a disease sentence, fill in "1" in the column if the anatomy part is related to the sentence.

E.2 Annotation Interface

We use Microsoft Excel for the annotation study. We include a screenshot of the annotation interface of our task in Figure 3.

F Dataset License

MIMIC-CXR, MS-CXR, and VinDR-CXR are under the PhysioNet Credentialed Health Data License 1.5.0 and PhysioNet Credentialed Health Data Use Agreement 1.5.0. OpenI is publicly available, and no license terms are stated. CheXpert-Plus is also publicly available under the Stanford University Dataset Research Use Agreement. SIIM Pneumothorax³ and RSNA Pneumonia⁴ datasets are publicly available on Kaggle as competition data.

G Detailed Description for Pretraining Datasets

MIMIC-CXR (Johnson et al., 2019) contains 227k image-report pairs, collected from 65,379 patients, covering frontal and lateral X-Ray images. Each report is associated with one or more images, resulting in a total of 377,110 images.

Chexpert-Plus (Chambon et al., 2024) is a large-scale image-report dataset which contains 223,228 unique pairs of reports and X-rays from 187,711 studies and 64,725 patients.

ChestX-ray14 (Wang et al., 2017) has 112,120 frontal-view images with binary labels for 14 diseases. We convert these labels to reports by randomly sampling disease sentences, following (You et al., 2023).

³<https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation>

⁴<https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge>

disease	Left Clavicle	Right Clavicle	Left Scapula	Right Scapula	Left Lung	Right Lung	Left Hilus Pulmonis	Right Hilus	Heart	Aorta	Facies Diaphragmatica	Mediastinum	Weasand	Spine
EXAMPLE: Heart enlargement, obvious infection in lungs					1	1			1					
There is mild pulmonary edema.														
The position of the right internal jugular catheter is unchanged.														
The positioning of the right upper extremity PICC line is unchanged.														
Bilateral pleural effusions are present.														
There is a thoracic spine compression fracture of approximately % vertebral body height.														

Figure 3: Annotation interface. For simplicity, we provide the annotator with an Excel spreadsheet. Each row contains a disease sentence, and 14 columns of anatomy parts. Annotator is asked to label which anatomical parts are related to the given disease sentence. This is then compared to the LLM’s annotation to verify how well an LLM’s annotation aligns with the radiologists’.

H Radiology Report Generation

H.1 Task Definition

Radiology report generation aims to generate a clinical report from a chest X-ray image. A standard report consists of two main sections: *Findings*, which describe the observed radiographic evidence, and *Impression*, which summarizes the key clinical conclusions. In this experiment, we focus on image-to-impression generation, where the model generates the *Impression* section directly from the input image.

Input Image:



Findings: The lungs are clear without focal consolidation, pleural effusion, or pneumothorax. The cardiomedial silhouette is within normal size limits.

Impression: No acute cardiopulmonary abnormality.

H.2 Experiment Setup

Following MedImageInsight (Codella et al., 2024), we connect the image encoders to a text decoder via a projection layer. Specifically, the text decoder is a randomly initialized, 6-layer Transformer decoder, with dimension size of 768. The projection layer is a simple linear projection layer, projecting image representation from image encoders to the text decoder. The report generation model is

trained in two stages. We first pretrain the linear projection layer and text decoder using MIMIC-CXR from LLaVA-Rad pretraining data⁵. Then, we finetune the full model on the official MIMIC-CXR dataset. The prompt used for zero-shot classification is There is DISEASE for positive cases, and There is no DISEASE for negative cases. We use the mean pooled patches embeddings for all images. Cross-modal retrieval is performed on the test split from the original dataset.

I Contrast-to-Noise (CNR)

For each MS-CXR image-phrase pair, we compute cosine similarity between the phrase embedding and the spatial visual embeddings before global pooling, producing a phrase-conditioned similarity map. The map is reshaped to its spatial feature-grid resolution and upsampled to the input-image resolution. Let B denote the pixels inside the radiologist-annotated bounding box and $\bar{B}$ those outside it. We compute the signed contrast-to-noise ratio as

$$\text{CNR} = \frac{\mu_B - \mu_{\bar{B}}}{\sqrt{\sigma_B^2 + \sigma_{\bar{B}}^2 + \epsilon}},$$

where μ_B, σ_B^2 and $\mu_{\bar{B}}, \sigma_{\bar{B}}^2$ are the mean and variance of the similarity scores inside and outside the annotated region, respectively. A larger positive CNR therefore indicates that the phrase-conditioned activation is more concentrated within the ground-truth region, whereas a negative value indicates greater activation outside it. We will also explicitly report how multiple boxes are handled, the feature-map resolution and upsampling procedure, and how instance-level scores are aggregated across the eight pathology categories.

⁵MedImageInsight and LLaVA-Rad use both MIMIC-CXR and translated PadChest; however, the translated PadChest is not publicly available, we pretrain the projection layer and Transformer decoder using only MIMIC-CXR.