Bringing CLIP to the Clinic: Dynamic Soft Labels and Negation-Aware Learning for Medical Analysis

Supplementary Material

A. Motivation

In general-domain datasets, captions involve millions of unique objects, scenes, and entities interacting in a multitude of combinations. Due to the diverse nature of general-domain data, contrastive learning is highly effective, as diversity is guaranteed even with random sampling of data into a batch. However, in medical settings, there are far fewer entities, and their relationships are limited, which does not align well with the objectives of contrastive learning.

A.1. Imbalance

Clinical data is often highly skewed, containing many duplicate templated reports, as shown in Tab. 7. Even when reports differ slightly in wording, semantically identical information still limits the effectiveness of standard contrastive learning. This has led many medical researchers and companies to discard duplicates and train models in a more balanced setting. However, fully normal chest X-rays (CXRs) contain crucial information for triage in clinical practice, as identifying normal cases can significantly reduce radiologists' workload. Our goal, therefore, is to develop a method that leverages all data—including duplicates—without discarding valuable information.

Furthermore, many reports are semantically similar even if the textual expressions differ. This occurs because there is an imbalance in the entities themselves; similar symptoms are commonly found across medical reports. This can cause semantic overlaps within a batch, where larger batch sizes might introduce more complexity in a contrastive learning context. As illustrated in Fig. 4, when trained with clinical data, one must consider this imbalance of clinical findings within the dataset. Using a general hospital dataset—which has more long-tailed characteristics compared to public data—could introduce noise into the training process due to this imbalance.

Impression	
No acute cardiopulmonary process.	37,962
No acute cardiopulmonary abnormality.	10,806
No acute intrathoracic process.	10,744
Findings	
Heart size is normal. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.	2,209
PA and lateral views of the chest provided. There is no focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.	1,763
The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.	1,635

Table 7. Most frequent reports from MIMIC impressions and findings. Note that the counts differ from Tab. 1 since the reports used in training prioritize findings over impressions.

A.2. Similarity

Standard contrastive learning frameworks typically pull positive pairs together and push negative pairs apart. From a clinical perspective, it would be beneficial if similarity could be weighted according to clinical context. For instance, a report noting "Right large pleural effusion. No pneumothorax." should be considered closer to "Right small pleural effusion." than to "No pleural effusion. Cardiomegaly exists," reflecting the clinical relevance of both findings. This is why we incorporate soft labels using similarity measures rather than a uniform distribution of soft labels, which has already been shown to be beneficial in [12]. Notably, using this characteristic, we can also handle duplicates or overlaps of clinical semantics since this method shares labels with similar or identical data within the batch.

We explore three types of similarity—textual, clinical, and graph-based—to achieve this nuanced approach. Similarity measures play a crucial role in contrastive learning, particularly in the medical domain. The SOFTCLIP [12] method, which also uses soft labels, relies primarily on textual similarity and is not well-suited for medical data where textual and clinical meanings often diverge. As shown in Fig. 6, textual similarity alone does not align well with clinical importance. For example, for the report "Mild cardiomegaly. The lungs are clear," the textual similarity score is higher with "The cardiomediastinal silhouette is normal. The lungs are clear." than with "The cardiac silhouette is moderately enlarged. No pleural effusion." Although the latter is closer in clinical meaning, textual similarity alone fails to capture this. Therefore, using solely textual similar-

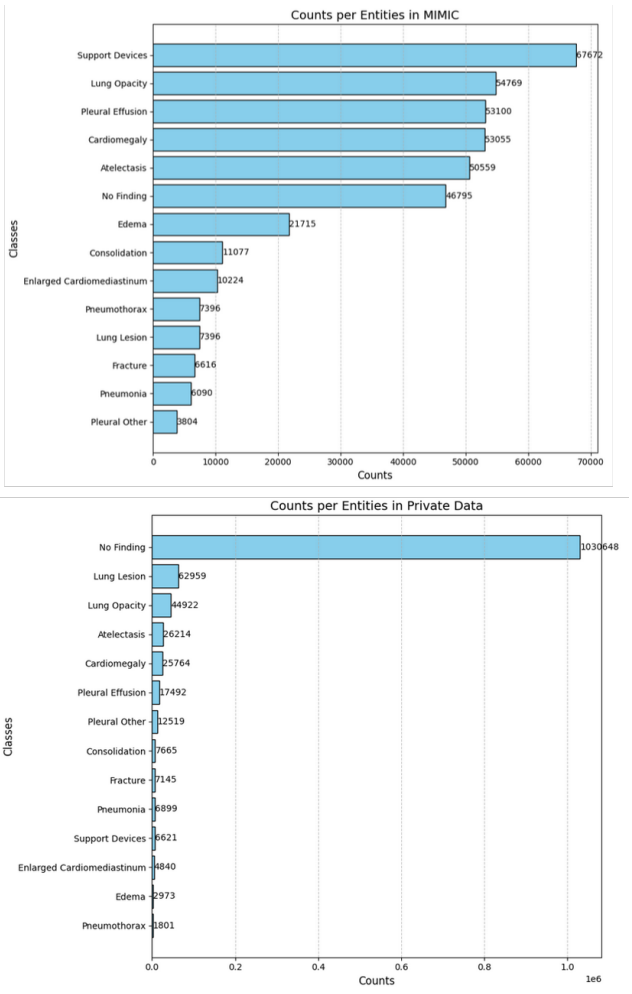


Figure 4. Counts of clinical entities in the whole MIMIC training set and a private dataset collected from a tertiary hospital. The private dataset comprises around 1.3 million records collected over 20 years, each from unique patients.

ity as soft labels can inadvertently bring unrelated reports closer rather than pushing them apart. This effect is demonstrated in Sec. 5.3 where SOFTCLIP performs worse than the baseline model.

While clinical similarity captures context better than text alone, it does not account for critical details like severity or location, such as "severe" or "mild." To address this, we introduce graph similarity, which can capture these nuanced attributes and improve alignment.

A.3. Negation

Negations are prevalent in medical reports, as illustrated in Fig. 7, where negated terms dominate the dataset. Unlike general domains, medical reports use diverse negation forms, such as "resolved," "removed," or "rule out," in addition to common terms like "no" or "not." Understanding negation is critical for accurate model performance, but using negated terms as hard negatives in standard contrastive learning often introduces noise. This is why, even though negations are a serious concern, few studies attempt to tackle this issue.

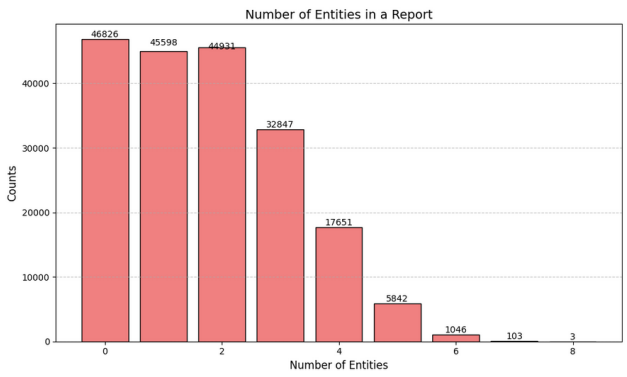


Figure 5. Counts of clinical entities in reports for the MIMIC training set.

For example, negating the report "Pneumothorax is present on the right upper lung zone" to "No pneumothorax" would yield a hard negative that overlaps semantically with other normal CXRs or cases without pneumothorax in the same batch, causing confusion. As shown in Fig. 5, using negation as a hard negative will introduce more overlaps as entity counts become smaller in the report. By incorporating dynamic soft labels, we can address this issue, allowing the model to handle clinical semantics effectively without adding noise from negated terms.

B. Dataset

B.1. Dataset Preprocessing

All CXR images undergo preprocessing through a pipeline that includes monochrome fixation, rotation correction, out-of-distribution (OOD) filtering, and view position selection. The monochrome fixation and rotation correction models were trained on Chexpert dataset using a MobileNetV3 CNN architecture, while view position and OOD detection utilize a DeepMCDD pipeline with a ResNet34 backbone. An example of image post-processing is shown in Fig. 8. All images are resized to 224×224 pixels and min-max normalized.

B.2. Dataset Split

Details of the training, validation, and test splits for our experiments (Sec. 5.3 and Sec. 5.5) are provided in Tab. 8. We use the same dataset splits as GLORIA [14] for CheXpert, VinDR, RSNA, and SIIM, while the CXR14 test set follows the split from ProbMed [36]. For MIMIC and OpenI, we exclude lateral and OOD images to ensure data consistency.

There is no pneumothorax or pleural effusion.		
Similarity	**Textual**	**Clinical**
No pleural effusion. No pneumothorax.	0.960	1
Lung volumes remain low. Small pleural effusion in the right middle fissure is present. There is no pneumothorax.	0.787	0
No acute cardiopulmonary process.	0.932	1
Right small pneumothorax. Left pleural effusion.	0.875	0

Mild Cardiomegaly. The lungs are clear.		
Similarity	**Textual**	**Clinical**
The cardio mediastinal silhouette is normal. The lungs are clear.	0.768	0
Severe Cardiomegaly. The lungs are clear.	0.998	1
The cardiac silhouette is moderately enlarged. No pleural effusion.	0.692	1
The cardiac silhouette is moderately enlarged. Mild pulmonary edema is present.	0.774	0.707

Figure 6. Comparison of textual, clinical similarity between reports.

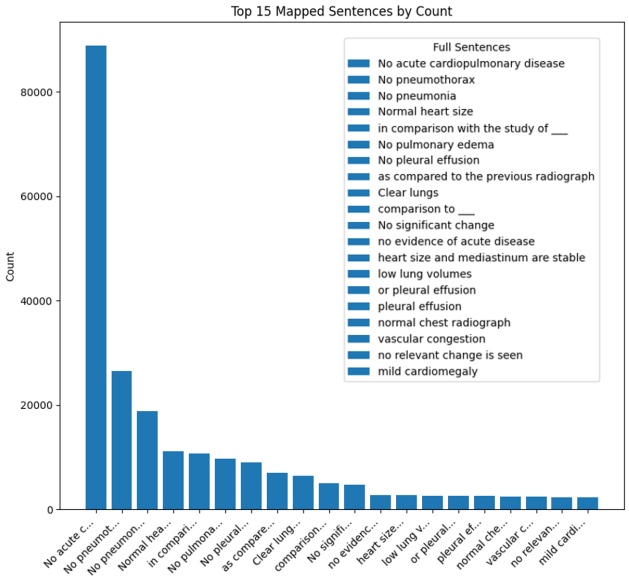


Figure 7. N gram frequent keyword extraction for MIMIC reports. The list is sorted by the top most frequently used phrases.

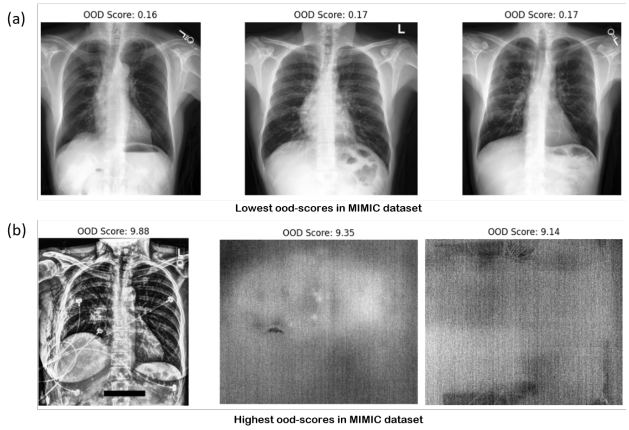


Figure 8. (a) Data with the lowest OOD score in the MIMIC dataset. (b) Data with the highest OOD score in the MIMIC dataset. The OOD detection model is implemented using the DeepMCDD pipeline.

B.3. CXR-Align

B.3.1. Counts

The number of test samples for each dataset in CXR-Align is shown in Tab. 9 and the distribution of selected entities is illustrated in Fig. 9. Entities are randomly selected with weights following the original distribution across test sets. Note that we prioritize cardiomegaly, atelectasis, edema, pleural effusion, pneumothorax, and consolidation, since the generated negations occur more often compared to other entities.

Dataset	Train	Valid	Test
MIMIC-CXR	194,847	1,984	2,490
CheXpert	-	-	1000
VinDR	-	-	3,000
RSNA	18,678	4,003	4,003
RSNA-ab	-	-	3,165
SIIM	8,432	1,808	1,807
Open-I	-	-	3,318
CXR14	-	-	880

Table 8. Data Summary for training and evaluation.

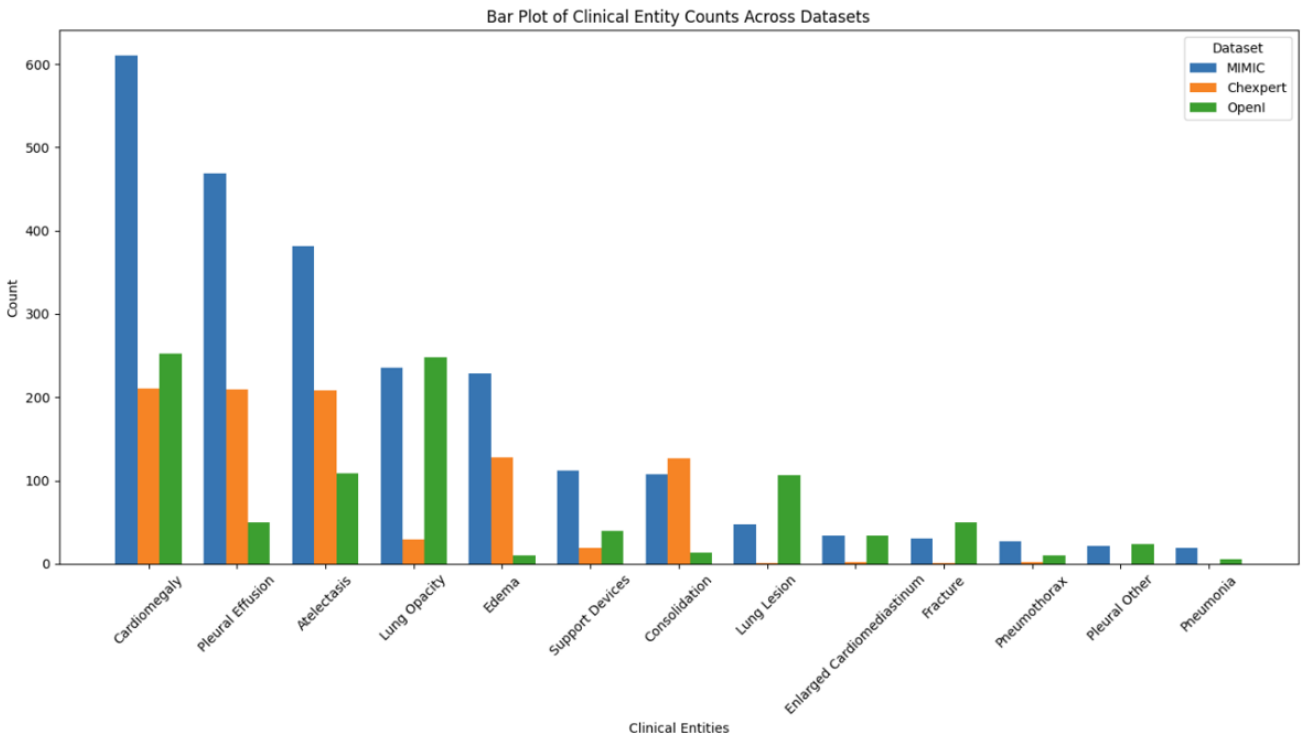


Figure 9. The number of selected entities in each dataset for CXR-Align.

Dataset	MIMIC	Chexpert	OpenI
Count	2323	937	953

Table 9. Count of datasets used in the *CXR-Align*.

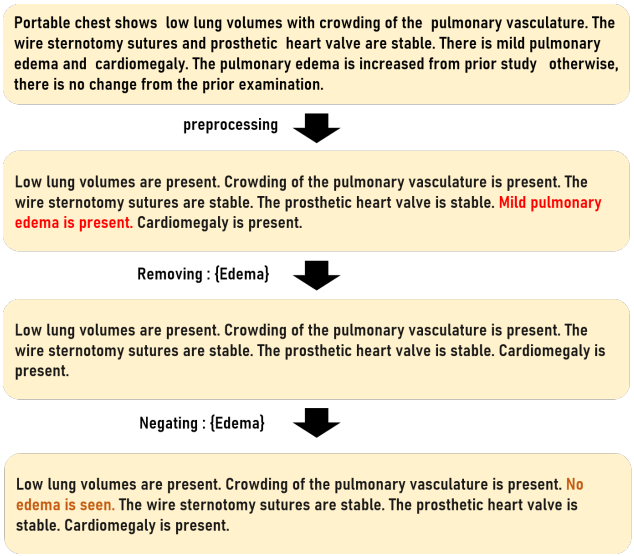


Figure 10. Example of the CXR-Align generation process.

B.3.2. Process

The process of CXR-Align generation is shown in Fig. 10. The removal of findings is a very important step to avoid contradictions or inconsistencies within the report. When mediastinal-related finding is chosen, we add one of the following sentences into the report: 'The cardiomediastinal silhouette is normal.', 'The cardiac silhouette is unremarkable.', 'The heart size is normal.', 'The cardiomediastinal silhouette is within normal limits.', or 'No cardiomegaly.'. If other findings are chosen, we add one of the following templates: "No (finding) is seen.", "No (finding) is observed.", "There is no (finding).", or "No evidence of (finding).". Note that the negated sentence is inserted randomly within the report, either at the beginning, middle, or end. If all the sentences related to the finding were removed, we simply insert the negated statement.

B.3.3. Prompts

Below is the prompt for each step in LLM text preprocessing as in Fig. 2.

Splitting We use the prompt from MAIRA2 [4] for splitting reports so that each sentence represent and describe only one entity.

Removing Prior Reference "You are an expert chest X-ray (CXR) radiologist familiar with radiologic reports. Your task is to rewrite the given radiology reports by removing all references to prior reports or comparisons, while preserving the original structure as much as possible. Input: A radiology report for a chest X-ray (CXR). Output: A revised CXR report focusing solely on current medical findings, excluding references to prior reports, comparisons, and irrelevant details. Guidelines: Remove Comparisons: Eliminate any terms or phrases that suggest a comparison, such as "compared to," "in comparison with," "change", "cleared", "constant", "decrease", "elevate", "expand", "improve", "decrease", "increase", "persistent", "reduce", "remove", "resolve", "stable", "worse", "new", etc. Focus on Current Findings: Ensure the report only describes the current state of the patient's lungs and related structures. Preserve Medical Context: Maintain the original medical terminology and descriptions of abnormalities. Retain Negations: Keep any negative statements about the absence of abnormalities.

Example 1: Original: The left apex has not been included on this radiograph. The ET tube terminates 3.9 cm above the carina. The NG tube terminates in the stomach. Surgical clips and a faint metallic coil project over the chest. A left PICC terminates in the mid SVC. EKG leads overlie the chest wall. The lung volumes are low. There are persistent bilateral mid and lower zone hazy opacities. There are persistent bilateral hilar and perihilar linear opacities. No significant interval change is observed in the lung opacities. Bilateral pleural effusions are present. The right pleural effusion is greater than the left. No pneumothorax is observed on the right. No cardiomegaly is present. No interval change is observed in the mediastinal silhouette. No significant interval change is observed in the bony thorax. Revised: The left apex has not been included on this radiograph. The ET tube terminates 3.9 cm above the carina. The NG tube terminates in the stomach. Surgical clips and a faint metallic coil project over the chest. A left PICC terminates in the mid SVC. EKG leads overlie the chest wall. The lung volumes are low. There are persistent bilateral mid and lower zone hazy opacities. There are bilateral hilar and perihilar linear opacities. Bilateral pleural effusions are present. The right pleural effusion is greater than the left. No pneumothorax is observed on the right. No cardiomegaly is present. "

Omitting selected entity "Task: Given a specific finding or disease and a chest X-ray report, remove the sentences relevant to that finding or disease.

Context:

Lung lesion: Refers to nodule or mass. Pleural other: Refers to pleural thickening.

Example:

Finding: Lung Lesion Report: No pneumothorax is observed. No pleural effusion is observed. No evidence of hemorrhage is observed in the lung or mediastinum. Emphysema is severe. The heart size is normal. A complex of nodule and large bullae is present in the axillary region of the right upper lobe. Expected Output: No pneumothorax is observed. No pleural effusion is observed. No evidence of hemorrhage is observed in the lung or mediastinum. Emphysema is severe. The heart size is normal.

Finding: Cardiomegaly Report: The feeding tube, with the wire stylet in place, is in the mid stomach. Heterogeneous pulmonary opacification is most pronounced in the left mid and lower lung. Heterogeneous pulmonary opacification is also present on the right, sparing only the upper lobe. The heart is mildly enlarged. Expected Output: The feeding tube, with the wire stylet in place, is in the mid stomach. Heterogeneous pulmonary opacification is most pronounced in the left mid and lower lung. Heterogeneous pulmonary opacification is also present on the right, sparing only the upper lobe."

B.4. Normal Case Detection

As described in Sec. 5.5, we augmented the MIMIC dataset by adding 130,000 normal CXR images from a single tertiary hospital, each labeled with the report "No active lung lesion." This augmentation results in an imbalanced dataset with 176,726 normal CXRs and 148,121 abnormal CXRs in the training set. For the reports which is used for the test set of this task, we included 2,999 abnormal reports sampled from the MIMIC test set with one normal report "No active lung lesion.". Data counts for the normal case detection experiment are provided in Tab. 10.

Dataset	Train	Valid	Test
MIMIC-CXR	194,847	-	-
Private	130,000	-	1,026
Open-I	-	-	1,289

Table 10. Data counts for normal case detection experiment.

C. Model

This section details the model implementation, augmentations, details with clinical information and hyperparameters.

C.1. Implementation Details

The model is trained using the AdamW optimizer with a cosine learning rate schedule and linear warm-up. The learning rate is set to 4×10^{-6}, with a batch size of 64 over 10 epochs on a single A6000 GPU. For fine-tuning experiments, we set the learning rate to 1×10^{-4}, with a batch size of 128. We train for 200 epochs when fine-tuning with 10%

of the data, and for 20 epochs when fine-tuning with 100% of the data, all on a single A6000 GPU. Each graph node's word is embedded using ClinicalBERT, and a one-hot code for class 'ANAT-DP', 'OBS-DP', 'OBS-DA', and 'OBS-U' is concatenated. The Graph Convolutional Neural Network (GCNN) for graph embeddings consists of two GCN conv layers with an input dimension of 772, a hidden dimension of 256, and an output dimension of 512 which is same with the other modalities. The max token length is set to 300.

C.2. Augmentations

For image augmentation, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clip limit of 4, random resized cropping, and rotations of up to 10 degrees. Text augmentation consists of sentence shuffling only.

C.3. Clinical Information

For the clinical information, we use CheXbert to extract the presence of entities. We additionally add one more label, where the value is 1 if all other labels are 0, and 0 otherwise. This accounts for cases where no findings are present, including entities that CheXbert may not cover. The entities are: ["Cardiomegaly", "Lung Opacity", "Atelectasis", "Lung Lesion", "Pleural Effusion", "Fracture", "Support Devices", "Enlarged Cardiomediastinum", "Pleural Other", "Consolidation", "Edema", "Pneumothorax", "Pneumonia", "No Findings"].

C.4. Hyperparameters

The temperature τ is set to 0.1, and similarity thresholds for textual τ_t, clinical τ_c, graph τ_g set at 0.9, 0.8, and 0.7, respectively. The weights for text w_T, clinical w_C, graph w_G weights in Eq. (10) are all set to 0.167.

D. Evaluation Settings

D.1. zero-shot prompt

Zero-shot prompt used for Sec. 5.3 is shown in Tab. 11. For CheXpert multi-class classification, we follow the prompt used in CXR-CLIP. For adversarial prediction, we used the same prompts as in the "Others" category.

	Positive	**Negative**
RSNA	Findings suggesting pneumonia.	No evidence of pneumonia.
SIIM	There is pneumothorax	There is no pneumothorax
Others	There is {findings}	There is no {findings}

Table 11. Positive and negative prompts for zero-shot evaluation.

D.2. Report retrieval

For report retrieval, we use the CheXbert F1 score rather than the standard BERTScore to evaluate how the retrieved or generated report clinically reflects the original report. The Macro F1 score is used since the Micro F1 score does not reflect the imbalance of the dataset. Furthermore, rather than focusing on top-k retrieval performance, we emphasized clinical metrics because the test set contains reports with similar clinical semantics, which could bias the performance evaluation if based solely on top-k retrieval metrics.

E. Additional Experiment

In this section, we provide a detailed discussion of our experiments. A notable finding from Sec. 5.3 is that our model's performance improves as we incorporate each similarity measure and hard negatives. Surprisingly, our baseline CLIP model's finetuned performance is comparable to or surpasses most of the SOTA CLIP models, implying that preprocessing steps like splitting reports and omitting prior references enhance the discriminability of CLIP models. Furthermore, adding similarity measures narrowed the gap between the RSNA and RSNA-ab results, indicating that our method helps the model to discriminate and correctly identify entities within abnormalities. In the following subsections, we present a sub-analysis of our methodology using different models, hyperparameters, and methods. We also provide a more detailed analysis of our benchmark *CXR-Align*, adversarial prediction, normal case detection, and report retrieval.

E.1. Different Models

We applied dynamic soft labels and negations as hard negatives with different model settings for both the image encoder and text encoder. We additionally compared a CNN-based encoder (ResNet50) and a BERT-based encoder (CXR-BERT [6]). As shown in Tab. 12, our method improved performance across all model combinations, demonstrating its applicability.

E.2. Different Methods

We conducted additional experiments using only the clinical similarity measure as soft labels (Experiment 1). Additionally, instead of using similarity measures for dynamic soft labels, we performed an experiment where we uniformly distributed the labels for indices that exceeded a predefined similarity threshold (Experiment 2). The experimental results in Tab. 13 reveal some surprising outcomes. Experiment 1 yielded better performance compared to our proposed model, which is unexpected since using only clinical similarities ignores other factors like "severity" and "location" relationships, focusing solely on the presence of an entity. This suggests that more comprehensive experiments and evaluation benchmarks, such as report retrieval, should be used when evaluating a CLIP model beyond zero-shot performance. In Experiment 2, although uniform soft labels improved results compared to the base model, our proposed model using similarity-based soft labels performed

Model	Image Encoder		Text Encoder		RSNA		RSNA-ab		SIIM		Chexpert
	Swin	Resnet	BioClinical	CXR	ZS	FT	ZS	FT	ZS	FT	Acc
Base	✓		✓		81.2	88.3	70.6	83.7	74.3	87.8	52.3
Ours	✓		✓		86.6	90.7	78.3	84.8	87.2	89.6	57.3
Base	✓			✓	82.0	88.4	70.5	83.8	74.1	87.4	53.1
Ours	✓			✓	84.7	90.3	75.7	84.9	86.5	89.0	55.0
Base		✓	✓		81.8	88.2	69.7	83.9	81.4	87.9	52.5
Ours		✓	✓		85.9	90.6	77.9	84.5	87.4	89.8	57.8
Base		✓		✓	80.6	88.1	70.8	83.7	82.3	87.6	54.3
Ours		✓		✓	85.2	90.5	75.6	84.2	85.6	89.3	56.9

Table 12. Zeroshot, and fine-tuned (10%) classification performance with each model settings comparing the base model and our proposed method.

significantly better. This implies that understanding clinical relationships within the batch and utilizing that information to create labels enhances the model's ability to comprehend clinical reports.

	RSNA (ZS)	RSNA-ab (ZS)	SIIM (ZS)
Ex 1	86.8	79.6	87.9
Ex 2	85.2	77.0	84.1

Table 13. Results across different settings. Experiment 1 uses only the clinical similarity measure as soft labels. Experiment 2 employs uniformly distributed soft labels for indices exceeding a predefined similarity threshold.

E.3. Different Hyperparameters

We explored the hyperparameters related to similarity thresholds to observe how changes affect the model's performance. Specifically, we varied the similarity thresholds for textual τ_t, clinical τ_c, graph τ_g from our default settings in Appendix C.4. As shown in Fig. 11, higher textual and clinical similarity thresholds yielded better results. For graph similarity, although a threshold of 0.9 provided the highest score for SIIM zero-shot AUC, a threshold of 0.7 resulted in more balanced performance overall.

E.4. Detailed analysis on CXR-Align

Fig. 13, Fig. 14 and Fig. 15 provide a detailed sub-analysis for the *CXR-Align* benchmark on the MIMIC, CheXpert, and OpenI datasets, respectively. We analyze the following aspects:

1. **Entity Type**: For all datasets, negated entities related to 'pneumothorax', 'effusion', 'consolidation', 'enlarged cardiomediastinum', and 'pneumonia' performed below average, while the model best discriminated 'pleural other', 'support devices', and 'fracture'. This may be due to the prompts used to negate the latter entities being less frequent in the training set compared to the former.
2. **Location**: The insertion location of the negation did not significantly affect performance, as accuracy was similar across all positions.
3. **Mediastinal Prompt**: For prompts regarding mediastinal findings, Prompt 2 ('The heart size is normal') con-

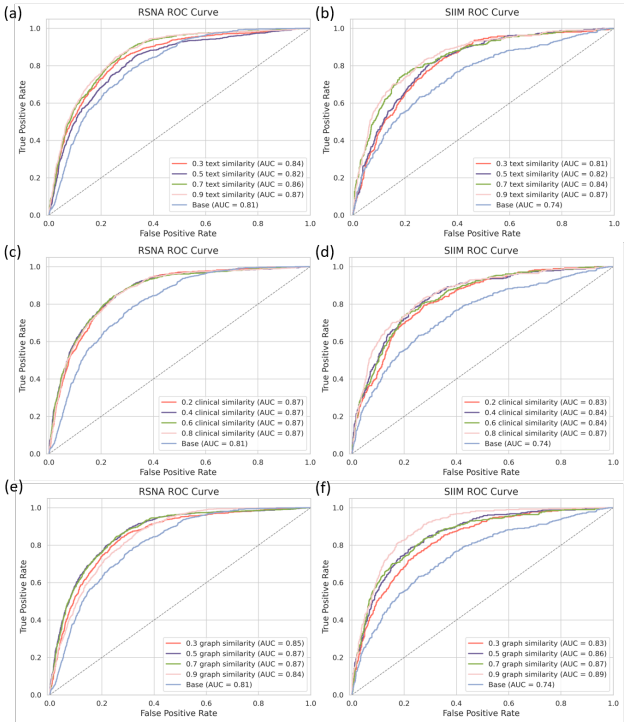


Figure 11. Threshold search for RSNA and SIIM zero-shot results. Starting from the settings in Appendix C.4, we varied the text similarity in (a) and (b), clinical similarity in (c) and (d), and graph similarity in (e) and (f).

sistently resulted in below-average accuracy when inserted as a negated statement across all datasets.

4. **Other Prompts**: For prompts related to lung entities, Prompt 2 ("There is no finding") performed the worst, falling below average. However, all prompts exhibited similar accuracy overall.

We hypothesize that the frequency of negated terms for each entity or prompt affects the model's performance and its comprehension of negations.

E.5. Detailed analysis on Adversarial Prediction

In this section, we perform a detailed analysis of adversarial prediction. We investigate how different models behave when subjected to this task compared to our model. As described in Sec. 5.4, this complex zero-shot task requires the model to determine whether one entity is present and another is absent. We conducted a total of 1,915 adversarial classification tasks. As shown in Tab. 14 most SOTA models tend to predict an entity as positive when given an abnormal CXR, indicating that they do not effectively discern which entities are present or absent. This raises concerns about the zero-shot classification task discussed in Sec. 5.3 suggesting that models may focus on the overall abnormality of the CXR rather than understanding the full context and associating positivity with specific entities. While CXR-CLIP mitigated this issue to some extent, our model demonstrated better clinical understanding regarding the presence and absence of clinical findings.

GT	Present		Absent	
Model	Positive	Negative	Positive	Negative
GLORIA	1671	244	1696	219
BioViL	1539	376	1281	634
BioViL-T	1625	290	1455	460
CXR-CLIP	754	1161	341	1574
OURS	720	1195	195	1720

Table 14. Positive/negative prediction counts in the adversarial prediction task for each model.

E.6. Detailed Analysis on Normal Case Detection

We conducted a detailed analysis of normal case detection, where the model must retrieve one normal report from 2,999 abnormal reports. As shown in Tab. 15, training with long-tailed data containing more than 50% normal CXR reports enables the model to effectively retrieve the normal report among all other abnormal reports. For the model trained only on the MIMIC dataset, the rank of the normal report was 68th. When using our internal test set as in Tab. 10, the model successfully retrieved the normal report with 99.7% accuracy. This suggests that further training with internal data containing normal CXRs can achieve higher performance for internal tasks, allowing hospitals to build their own specialized models.

OURS$_{mimic}$	
There is a right lower lobe airspace consolidation. The lungs are otherwise clear. The hilar and cardiomediastinal contours are normal. There is no pneumothorax. There is no pleural effusion. Pulmonary vascularity is normal.	12
A small residual area of linear atelectasis is present in the retrocardiac area. No pneumothorax is observed. No pleural effusion is observed. The heart size is normal. The hilar contours are normal. The mediastinal contours are normal. The visualized osseous structures are unremarkable.	12
The heart is normal in size. The mediastinal and hilar contours appear within normal limits. There is an inferolateral consolidation in the right upper lobe, consistent with pneumonia. The lungs appear clear elsewhere. No pleural effusions are present. No pneumothorax is present. The osseous structures are unremarkable.	11
OURS$_{mimic+private}$	
No active lung lesion.	1105
No focal consolidation is seen. No pleural effusion is seen. No pneumothorax is seen. No pulmonary edema is seen. Minimal bronchial wall thickening is present. The heart size is normal. Mediastinal contours are normal. No bony abnormality is detected.	48
No lung consolidation. The left lower lung atelectatic band has resolved. Mediastinal and cardiac contours are normal. No pneumothorax. No pleural effusion.	14

Table 15. Most frequent reports and their counts retrieved from the normal case detection task for the OpenI test images. The upper table shows results for our model trained only on MIMIC, while the lower table shows results for our model trained on MIMIC and private data.

E.7. Report Retrieval

We provide examples of report retrieval performance in Fig. 12. Compared to other SOTA models and the baseline model, our model successfully retrieves reports that share similar semantics with the original report, even if they are not identical. Notably, in the third example, our model linked the textual semantics of "There is infrahilar

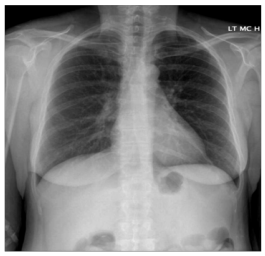
interstitial prominence which may represent bronchovascular crowding lung" to the original report's "The lungs are hypoinflated," demonstrating high correlation.

Acknowledgments

We would like to thank H.Y. Cho, I.H. Baek and Y.G. Kim for their valuable advice on this paper. This study was supported by the National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (MSIT) (Grant No. RS-2024-00354666).

There is mild cardiomegaly. The aorta is tortuous. There is XXXX opacities noted in the right lower lobe, XXXX subsegmental atelectasis. There is no pneumothorax or effusion. No displaced rib fractures. If there is high clinical concern, consider dedicated rib views for further evaluation.

MedCLIP	The heart is normal in size. The mediastinum is unremarkable. Emphysematous changes are identified. The lungs are otherwise grossly clear.
BIOVIL	The cardiomediastinal silhouette is normal in size and contour. No focal consolidation, pneumothorax or large pleural effusion. Negative for acute displaced rib fracture.
BIOVIL-T	Frontal and lateral views of the chest show normal size cardiac silhouette, allowing for an AP projection. Normal contour of the mediastinum and aorta. Grossly clear lungs. No obvious pneumothorax or hemothorax. No acute displaced clavicle or rib fractures.
CLIP	Normal heart size is prominent left ventricular contour. Unfolding of the thoracic aorta. No focal airspace consolidation. No pleural effusion or pneumothorax. Visualized osseous structures are unremarkable appearance.
OURS	There is mild cardiomegaly. The aorta is tortuous. There is XXXX opacities noted in the right lower lobe, XXXX subsegmental atelectasis. There is no pneumothorax or effusion. No displaced rib fractures. If there is high clinical concern, consider dedicated rib views for further evaluation.



No pneumothorax. Heart size is normal. Granulomas are seen within the right lung. No large pleural effusions. No focal airspace consolidation.

MedCLIP	The lungs are clear. There is no pleural effusion. The heart is normal. The mediastinum is normal. The skeletal structures show arthritic changes.
BIOVIL	Small 3.3 mm right-sided pneumothorax only visible on the left lateral decubitus film. Left lung is clear. Normal cardiac contour. No evidence of pleural effusion.
BIOVIL-T	Minimal right-to-left cardiomediastinal shift. The cardiomediastinal silhouette is otherwise normal size and configuration. Pulmonary vasculature within normal limits. There is a moderate sized right pneumothorax. This measures 3.2 cm at the level the right apex.
CLIP	Heart size normal. No focal airspace disease. No pneumothorax or effusions. No bony abnormalities.
OURS	Heart size and mediastinal contour normal. Lungs are clear except for residuals of prior granulomatous infection. No pleural effusions or pneumothoraces.

The heart is normal in size. The mediastinum is unremarkable. The lungs are hypoinflated. Small bilateral pleural effusions are seen.

MedCLIP	The lungs are grossly clear without focal pneumonic consolidation, large effusion or pneumothorax. Heart size is within normal limits.
BIOVIL	The heart is normal in size with normal appearance of the cardiomediastinal silhouette. There is a hiatal hernia with soft tissue projecting behind the mediastinum. The lungs are clear without focal airspace opacity, pleural effusion, pneumothorax. The osseous structures are intact.
BIOVIL-T	Frontal and lateral views of the chest show normal size and configuration of the cardiac silhouette. Normal mediastinal contour, pulmonary XXXX and vasculature, central airways and lung volumes. And scattered calcified granulomas. Left greater than right basilar opacity, probable atelectasis and/or scarring. No pleural effusion.
CLIP	Cardiomediastinal silhouette appears normal in size and contour. Right lung is clear. Stable blunting of costophrenic XXXX with improved aeration of the left base compared to prior exam. No visualized pneumothorax or focal consolidation. XXXX unremarkable.
OURS	Heart size and pulmonary vascularity normal. There is a small right pleural effusion. There is infrahilar interstitial prominence which may represent bronchovascular crowding lung. Small left pleural effusion. No pneumothorax.

Figure 12. Examples of retrieved reports. Blue text represents important entities that should be included in the report. Red text indicates hallucinations or falsely interpreted entities. Purple represents clinically similar entities.

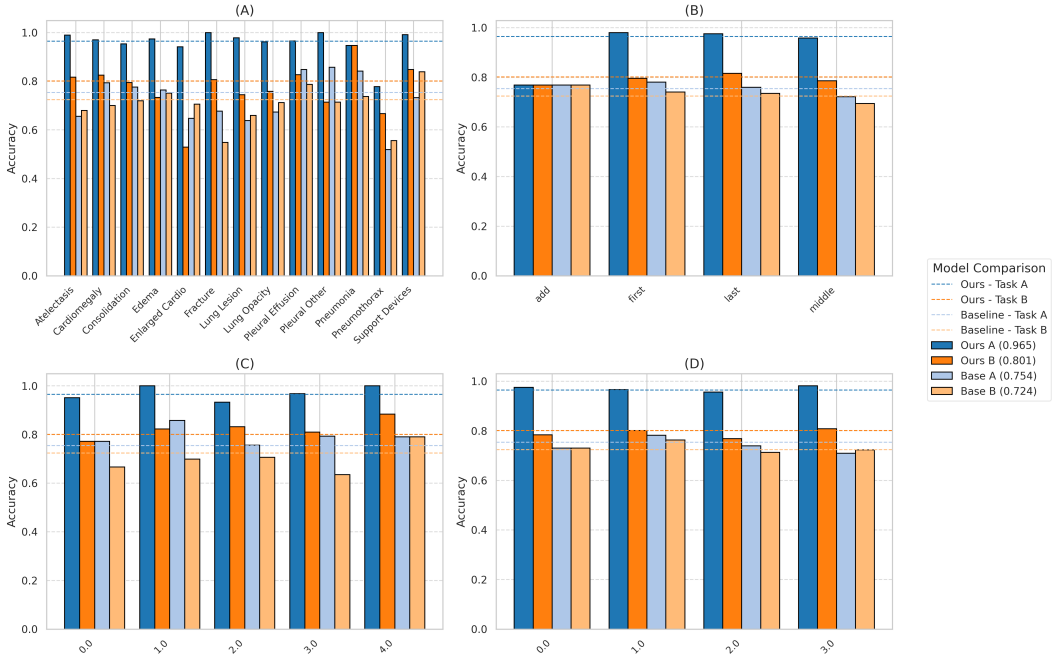


Figure 13. Detailed sub-analysis for CXR-Align on MIMIC dataset. (A) Task accuracy for entities that were either negated or removed. (B) Performance based on the location where the negated sentence was inserted. (C) Accuracy corresponding to the prompt used when the selected entity was related to mediastinal findings. (D) Performance corresponding to the prompt used when the selected entity was related to lung findings. For (C) and (D), refer to Appendix B.3.2

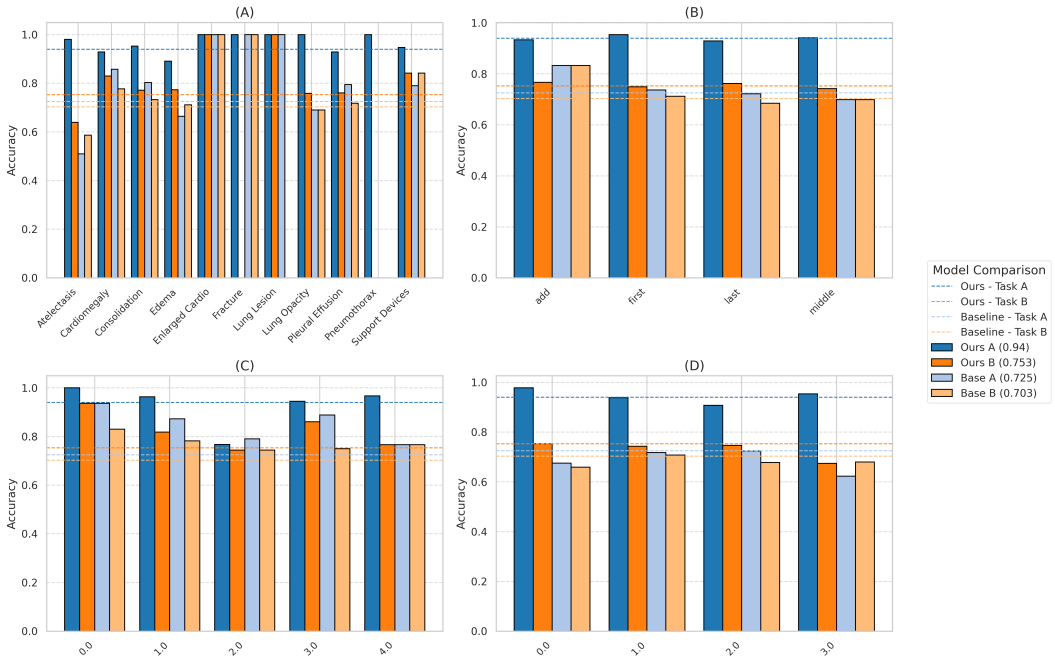


Figure 14. Detailed sub-analysis for CXR-Align on Chexpert dataset. (A) Task accuracy for entities that were either negated or removed. (B) Performance based on the location where the negated sentence was inserted. (C) Accuracy corresponding to the prompt used when the selected entity was related to mediastinal findings. (D) Performance corresponding to the prompt used when the selected entity was related to lung findings. For (C) and (D), refer to Appendix B.3.2

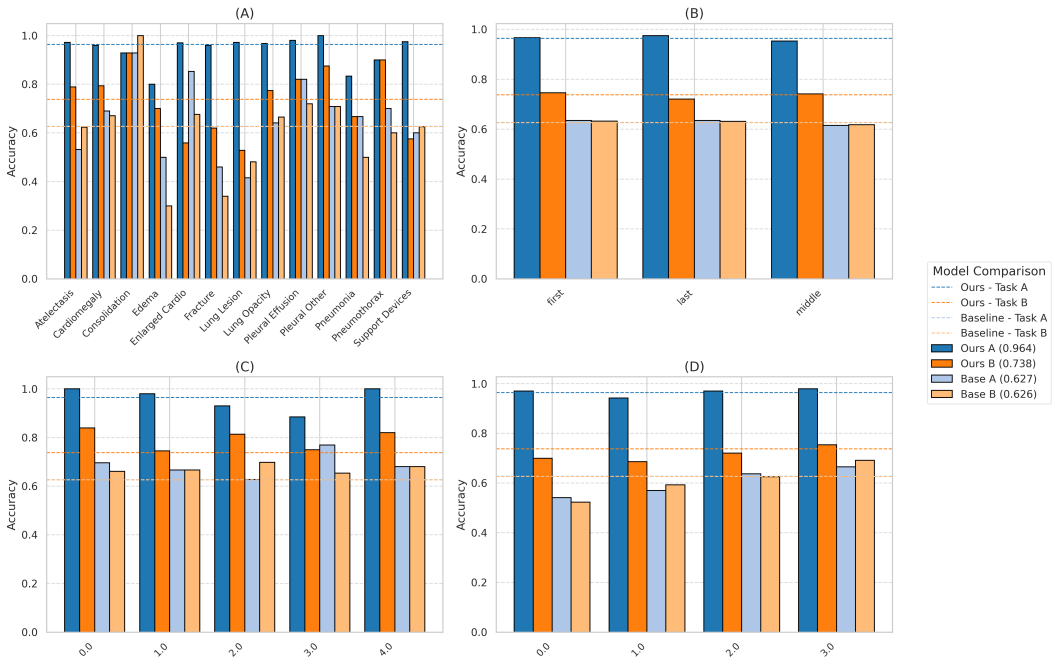


Figure 15. Detailed sub-analysis for CXR-Align on OPENI dataset. (A) Task accuracy for entities that were either negated or removed. (B) Performance based on the location where the negated sentence was inserted. (C) Accuracy corresponding to the prompt used when the selected entity was related to mediastinal findings. (D) Performance corresponding to the prompt used when the selected entity was related to lung findings. For (C) and (D), refer to Appendix B.3.2