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

Hanbin Ko[1,2], Chang-Min Park[1,2,3*]
[1]Interdisciplinary Program in Bioengineering, Seoul National University Graduate School,
[2]Integrated Major in Innovative Medical Science, Seoul National University Graduate School,
[3]Department of Radiology, Seoul National University Hospital
{lucasko1994, morphius}@snu.ac.kr

Abstract

*The development of large-scale image-text pair datasets has significantly advanced self-supervised learning in Vision-Language Processing (VLP). However, directly applying general-domain architectures such as CLIP to medical data presents challenges, particularly in handling negations and addressing the inherent data imbalance of medical datasets. To address these issues, we propose a novel approach that integrates clinically-enhanced dynamic soft labels and medical graphical alignment, thereby improving clinical comprehension and improving the applicability of contrastive loss in medical contexts. Furthermore, we introduce negation-based hard negatives to deepen the model's understanding of the complexities of clinical language. Our approach is easily integrated into medical CLIP training pipeline and achieves state-of-the-art performance across multiple tasks, including zero-shot, fine-tuned classification and report retrieval. To comprehensively evaluate our model's capacity in understanding clinical language, we introduce **CXR-Align**, a benchmark uniquely designed to evaluate the understanding of negation and clinical information within chest X-ray (CXR) datasets. Experimental results demonstrate that our proposed methods are straightforward to implement and generalize effectively across contrastive learning frameworks, enhancing medical VLP capabilities and advancing clinical language understanding in medical imaging.*

1. Introduction

CLIP [26] has revolutionized Vision-Language Processing (VLP), with particularly promising applications in medical imaging analysis [41]. Medical imaging, especially in areas requiring specialized annotation expertise, greatly benefits from CLIP's ability to leverage image-text pairs without extensive labeled data, thus enabling efficient representation learning. Consequently, research has increasingly focused on adapting CLIP-like models for CXR data, which is rich in image-report pairs and well-suited to contrastive learning.

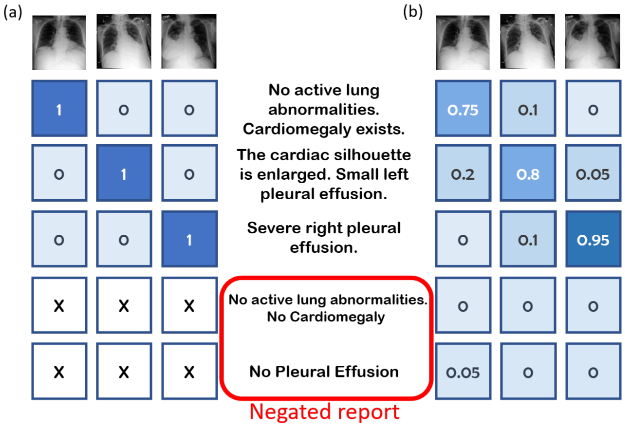
(a)
1
0
0
0
1
0
0
0
1
X
X
X
X
X
X
No active lung abnormalities. Cardiomegaly exists.
The cardiac silhouette is enlarged. Small left pleural effusion.
Severe right pleural effusion.
No active lung abnormalities. No Cardiomegaly
No Pleural Effusion
Negated report
(b)
0.75
0.1
0
0.2
0.8
0.05
0
0.1
0.95
0
0
0
0.05
0
0

Figure 1. (a) Standard visual-language pre-training approaches using contrastive learning (e.g., InfoNCE). (b) Our approach, leveraging unique medical domain characteristics (e.g., imbalance and negations), dynamically generates soft labels based on clinical, textual, and relational similarities while integrating negations as hard negatives.

However, directly applying CLIP's contrastive learning framework to medical data poses significant challenges due to unique characteristics of medical image-text data. For example, medical reports often contain negations and are subject to considerable data imbalance. While several adaptations, such as Xlip [34], CXR-CLIP [38], GLORIA [14], BioViL [6], MedKlip [35], MLIP [21], have sought to improve image-report alignment in the medical field, many overlook specific aspects of CXR reports, such as references

[*] Corresponding author

to interval changes requiring temporal context from prior images. BioViL-T [3] addresses this challenge by incorporating prior images to account for temporal considerations.

Report	Count
No acute cardiopulmonary process	15829
No acute intrathoracic process	4643
No acute cardiopulmonary abnormality	4488
No evidence of acute cardiopulmonary process	1255
No evidence of pneumonia	1014

Table 1. Top most frequent reports and their counts in the MIMIC training data. Due to the frequent use of predefined templates, duplicated reports amplify imbalance.

In addition, radiologists routinely document both the presence and absence of findings (e.g., "No pneumothorax"), making negation a critical feature for precise clinical communication. However, CLIP-based models, which often exhibit "bag-of-words" characteristics [39], can struggle to fully interpret negated terms [27]. For effective comprehension of CXR reports, it is essential for the model to correctly understand the purpose and implications of these negated entities.

Contrastive learning in medical datasets also contends with significant imbalance. In general domains, increasing batch sizes during CLIP training improves gradient estimation by introducing a wider variety of negative samples [8, 26]. However, medical datasets are heavily skewed towards normal cases and exhibit template-based duplication, as illustrated in Tab. 1. Radiologists frequently use predefined templates for routine findings, resulting in numerous near-duplicate reports that amplify data imbalance. In this context, larger batch sizes increase the likelihood of semantically identical or duplicate reports being treated as negatives, introducing noise that conflicts with the objectives of contrastive learning. Although the ICU-focused MIMIC [17, 18] dataset provides some diversity, such imbalances can be even more pronounced in general hospital datasets where templated language and normal cases are prevalent, comprising over three-quarters of the data.

In this study, we address the challenges of data imbalance and negation specifically within the context of CLIP for medical datasets. Unlike traditional medical imbalance issues, this imbalance arises uniquely in contrastive settings, where solutions like report rewriting [10] are insufficient. To our knowledge, this is the first paper to directly tackle these medical-specific features common in clinical reports at a global-feature scale. We define data imbalance as primarily semantic overlap within the batch, often but not exclusively limited to normal CXR reports.

Our approach focuses on single-image scenarios, leaving temporal considerations for future work. To mitigate imbalance, we introduce a *clinically-enhanced dynamic soft-labeling* method that incorporates clinical and textual similarity into contrastive learning, allowing the model to better interpret the clinical relationship of reports in each batch. For handling negations, we generate negation-based hard negatives to enhance CLIP training, strengthening the model's capacity to create accurate clinical representations. Additionally, we integrate *graph embeddings* to enrich the image-text architecture, capturing domain-specific relationship that result in stable learning and improved performance across tasks such as zero-shot and fine-tuned classification, adversarial prediction, CXR-report alignment through negations and clinical entities, normal case detection, and report retrieval.

- We propose a contrastive learning method that leverages dynamic soft labels, incorporating clinical and textual similarity, to address data imbalance and improve training stability in medical settings.
- We introduce the *CXR-Align* benchmark, designed to evaluate models on negation handling and clinical alignment, advancing the assessment of medical VLP models.
- We create a negation generation pipeline to synthesize hard negatives, strengthening the model's understanding of negated findings, which works synergistically with dynamic soft labels.
- We integrate graph embeddings into the contrastive framework to capture the unique characteristics of medical data, refining soft-labels and enhancing negation comprehension.
- Our method demonstrates strong performance across tasks like classification, adversarial prediction, CXR-report alignment, normal case detection and retrieval, surpassing baseline and state-of-the-art CLIP-based models.

2. Related Works

2.1. Medical VLP for Chest X-Rays

Recently, contrastive learning approaches inspired by CLIP [26] have gained traction in medical applications, benefiting from the abundant paired data in CXR tasks [7, 18]. Notable models include CheXzero [30] and ConViRT [40], which align image and text representations trained on the MIMIC dataset, and GLORIA [14], which employs local representations for fine-grained alignment. CXR-CLIP [38] explores image-to-image alignment, while XLiP [34] and BioViL [6] adopt masked modeling to predict masked elements in both images and text. BioViL-T [3] uniquely incorporates temporal information using prior images to capture interval changes in CXR reports. MedKlip [35] and MLIP [21] incorporate clinical knowledge to enhance the models with domain-specific information. Notably, MLIP highlighted that semantic overlap within batches can cause problems in contrastive learning settings,

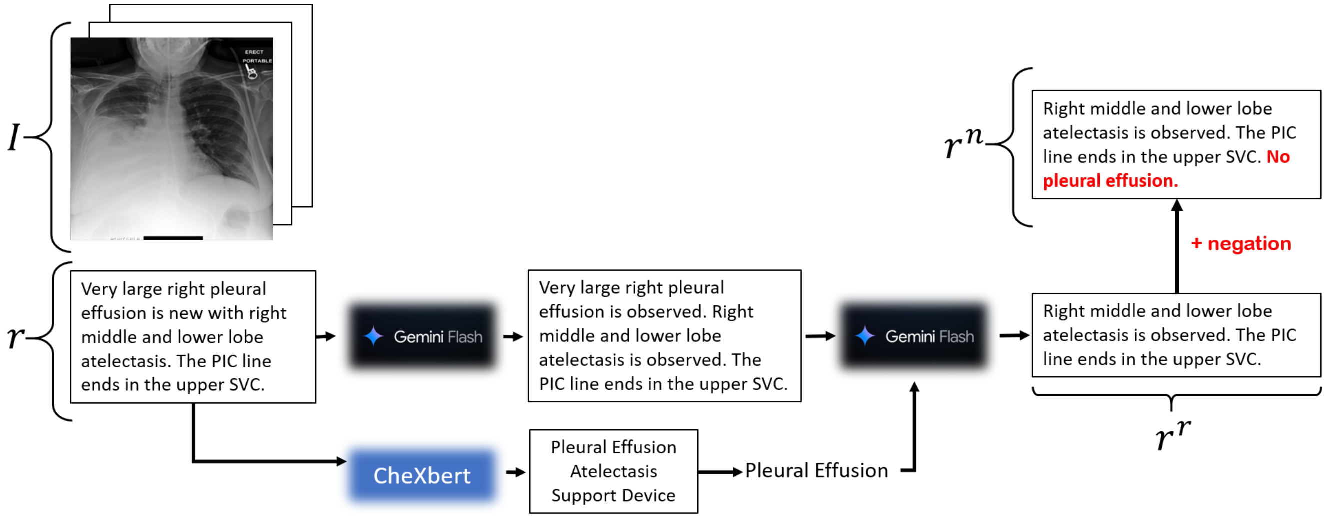
I
ERECT
PORTABLE
r
Very large right pleural effusion is new with right middle and lower lobe atelectasis. The PIC line ends in the upper SVC.
Gemini Flash
Very large right pleural effusion is observed. Right middle and lower lobe atelectasis is observed. The PIC line ends in the upper SVC.
Gemini Flash
Right middle and lower lobe atelectasis is observed. The PIC line ends in the upper SVC.
+ negation
r^n
Right middle and lower lobe atelectasis is observed. The PIC line ends in the upper SVC. No pleural effusion.
r^r
CheXbert
Pleural Effusion
Atelectasis
Support Device
Pleural Effusion

Figure 2. Given a CXR report, CheXbert identifies all positive entities, and one is randomly selected. A language model then (i) splits the report so each sentence contains a single clinical entity without temporal statements and (ii) removes sentences related to the selected entity. Finally, a negation for the selected entity is added at a random position within the report (beginning, middle, or end).

proposing a solution that uses external knowledge to bind similar semantics in a local-scale environment. Despite these advancements, challenges such as data imbalance and frequent negations in medical text still remain largely unresolved, especially on a global scale, hindering the development of reliable clinical models.

2.2. CLIP for Compositional Understanding and Negations

Assessing the compositional capabilities of vision-language models like CLIP is essential for evaluating their generalization to new combinations of visual and textual information. The CREPE benchmark [23] introduced metrics revealing that large-scale pretraining often falls short in compositional reasoning. Yuksekgonul *et al.* [39] highlighted that CLIP often behaves like a ”bag-of-words” model, raising concerns about its textual comprehension. SUGAR-CREPE [13] addresses these biases by generating fluent and plausible hard negatives through language models with adversarial refinement. To enhance negation handling, CoN-CLIP [27] achieves strong results on the CC-Neg benchmark, underscoring the importance of handling negations in VLMs. However, incorporating negations in medical contrastive settings can exacerbate semantic overlap, introducing significant noise into the training pipeline.

2.3. Soft Alignment for CLIP

Although CLIP shows resilience to imbalanced and long-tailed data distributions [33], its performance deteriorates with highly imbalanced datasets. Re-weighting strategies and specialized loss functions have been explored [32], but often lack adaptability to the medical domain. PyramidCLIP [11] and SoftCLIP [12] relax the strict one-to-one constraint of CLIP's contrastive loss by implementing soft cross-modal alignment based on intra-modal self-similarity. However, this can introduce ambiguity in text embeddings within the clinical domain, leading to noisy outputs and unstable training. Medical datasets, characterized by extreme imbalance demand innovative solutions tailored to medical imaging to effectively address CLIP's limitations.

3. CXR-Align: A Benchmark for CXR-Report Alignment with Negations

Negations are rarely present in image-text pairs within general-domain datasets, limiting CLIP models' ability to accurately interpret negated information. In contrast, medical datasets frequently contain negations, which are critical for precise clinical interpretation. To address this gap, we introduce *CXR-Align*, the first benchmark specifically designed to evaluate models' comprehension of negations in CXR reports—an essential aspect for clinical applications. *CXR-Align* is synthesized from the test sets of MIMIC [18], CheXpert [7], and OpenI [9]. We begin by transforming the original reports using a large language model (LLM), specifically Gemini-Flash [29], to ensure that each sentence is limited to a single clinical entity [4]. To focus on diagnostically relevant cases, normal CXRs are excluded from this benchmark. We further standardize the reports by removing any temporal references, centering each on a single CXR instance. Using a CXR report labeler, specifically CheXbert [28], we identify positive findings, diseases, or medical devices within each report, then randomly select one entity for further processing. We create report varia-

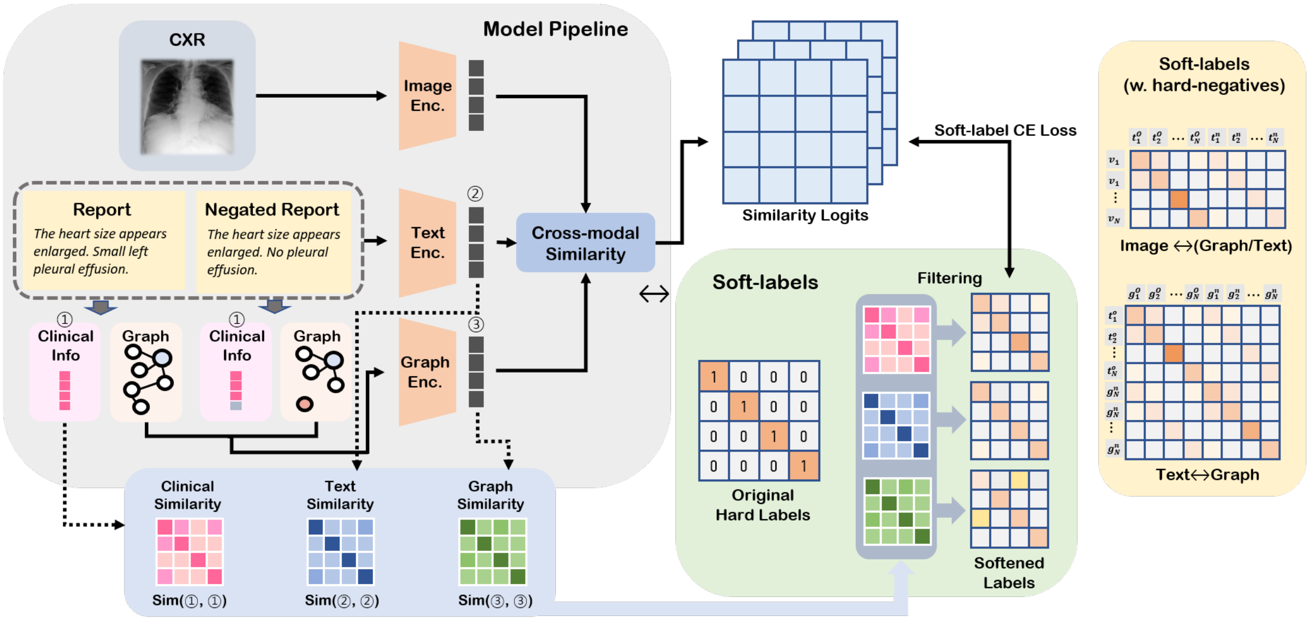

Figure 3. Overview of the proposed pipeline. Hard negative reports are created that differ from the original by only one clinical entity. Embeddings of each modality (CXR, report, graph) are extracted by their encoders, along with clinical labels from the report. Intra-modal self-similarities are computed for clinical labels, text embeddings, and graph embeddings, used as soft labels for each stream. The conventional InfoNCE loss is replaced by KL-Divergence when incorporating softened targets, ensuring labels reflect the textual, clinical, and graphical meanings correctly.

tions as follows:

- **Removing positive entity** (r^r): The LLM removes the selected positive entity from the report, resulting in r^r.
- **Adding negations** (r^n): From r^r, we generate r^n by inserting a predefined negative statement about the selected entity.

The final dataset contains two triplet structures: (I, r, r^n) and (I, r, r^r). The first triplet assesses the model's ability to understand and correctly reject an incorrectly negated statement, while the second evaluates its grasp of full clinical semantics by identifying the most complete report. This setup enables us to benchmark our model alongside other state-of-the-art vision-language models, focusing on negation comprehension and CXR-report alignment. More details for this dataset can be found on Appendix B.3.

4. Method

Our approach to training CLIP models on medical datasets addresses semantic overlap and negation by introducing a clinically-enhanced dynamic soft-label strategy combined with negation-based hard negatives. This section outlines our method for generating negated data as hard negatives, creating dynamic soft labels, and formulating the training loss, enhanced with graph embeddings.

4.1. Generating Data with Negations for Hard Negative Training

To improve the model's understanding of negations, we generate hard-negative samples. Following the process from Sec. 3, we create hard negatives r^n for abnormal CXRs by introducing negations into the report. For normal CXRs, we randomly select reports containing only a single positive entity, using these as hard negatives rather than generating negated reports. This approach ensures that the hard negative reports differ from the original reports by only one entity, making them challenging for the model to distinguish.

4.2. Dynamic Soft Contrastive Loss

To address imbalance and semantic overlap within a batch, we implement dynamic soft labels that reflect clinical similarities between text embeddings and clinical labels. Let $T_1 \in \mathbb{R}^{B \times D_t}$ denote the L2-normalized text embeddings and $T_2 \in \mathbb{R}^{B \times D_t}$ the hard negative text embeddings, where B is batch size, and D_t is the dimension of the text embeddings. Similarly, $C_1 \in \mathbb{R}^{B \times D_c}$ and $C_2 \in \mathbb{R}^{B \times D_c}$ represent the L2-normalized clinical labels and hard negative clinical labels, respectively, with D_c being the dimension of the clinical labels. We concatenate T_1 and T_2 to form combined text embeddings $T = [T_1; T_2] \in \mathbb{R}^{2B \times D_t}$ and likewise for the clinical embeddings $C = [C_1; C_2] \in \mathbb{R}^{2B \times D_c}$.

Since embeddings alone may not fully capture clinical semantics [2], we use 14 labels extracted from reports by CheXbert [28] as an alternative source of clinical information.

As in InfoNCE [25] for cross-modal alignment, the normalized cross-modal logits are calculated as:

$$p_{ij}(I,T) = \frac{\exp(\text{sim}(v_i, t_j)/\tau)}{\sum_{j=1}^{2B} \exp(\text{sim}(v_i, t_j)/\tau)}, \tag{1}$$

$$p_{ij}(T_1,I) = \frac{\exp(\text{sim}(t_i, v_j)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(t_i, v_j)/\tau)}, \tag{2}$$

where $v \in \mathbb{R}^{B \times D_{img}}$ are the image embeddings, τ is the temperature parameter, and sim$(\cdot)$ denotes dot-product similarity.

Next, we compute similarity matrices for text and clinical labels:

$$S_t = T \cdot T^T, \quad S_c = C \cdot C^T, \tag{3}$$

where $S_t \in \mathbb{R}^{2B \times 2B}$ and $S_c \in \mathbb{R}^{2B \times 2B}$ represent intra-modal similarities for text and clinical embeddings, respectively.

Dynamic soft labels are generated by applying thresholds τ_t and τ_c to retain values above each threshold:

$$y_t[i,j] = \begin{cases} \frac{S_t[i,j]-\tau_t}{1-\tau_t}, & \text{if } S_t[i,j] > \tau_t, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

$$y_c[i,j] = \begin{cases} \frac{S_c[i,j]-\tau_c}{1-\tau_c}, & \text{if } S_c[i,j] > \tau_c, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $y_t \in \mathbb{R}^{2B \times 2B}$ and $y_c \in \mathbb{R}^{2B \times 2B}$ are then normalized across each row to $\hat{y}_t$ and $\hat{y}_c$ ensuring that the sum of each row equals 1. Note that thresholding is crucial since sharing labels for data with minimal similarity introduces noise.

The image-to-text loss for each similarity measure incorporates KL-Divergence with the generated soft labels and is defined as follows:

$$\mathcal{L}_t(I,T) = \frac{1}{B} \sum_{i=1}^{B} \text{KL}(\hat{y}_t[i] \parallel p_i(I,T)), \tag{6}$$

$$\mathcal{L}_c(I,T) = \frac{1}{B} \sum_{i=1}^{B} \text{KL}(\hat{y}_c[i] \parallel p_i(I,T)), \tag{7}$$

where $\mathcal{L}_t(T_1, I)$, $\mathcal{L}_c(T_1, I)$ is also computed in a similar manner.

4.3. Dynamic Contrastive Loss with Graph Embeddings

To capture additional clinical relationships, we integrate graph embeddings. Clinical embeddings may lack specific attributes such as location, severity, or size of entities, as they only encode presence. We use RadGraph [16] to extract graphs from each report, embedding each node with ClinicalBERT [1]. A two-layer Graph Convolutional Network [37] then produces graph embeddings $G_1 \in \mathbb{R}^{B \times D_t}$ and hard negative graph embeddings $G_2 \in \mathbb{R}^{B \times D_t}$, which are concatenated as $G = [G_1; G_2] \in \mathbb{R}^{2B \times D_G}$. Using graph embeddings G, we compute pairwise similarity:

$$S_g = G \cdot G^T, \tag{8}$$

where S_g represents graph similarity within the batch. Graph-based soft labels y_g are generated with a threshold τ_g:

$$y_g[i,j] = \begin{cases} \frac{S_g[i,j]-\tau_g}{1-\tau_g}, & \text{if } S_g[i,j] > \tau_g, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The normalized $\hat{y}_g$ serves as soft labels, and KL-Divergence loss terms are computed with the graph-related logits extracted as like in Eq. (2) with Eq. (7) across text, clinical, and graph similarity stream.

The final training loss integrates all cross-modal components as follows:

$$\mathcal{L}_{\text{total}} = \sum_{i=1, i \neq j}^{3} \sum_{j=1}^{3} \Big(w_T \cdot \mathcal{L}_t(M_i, M_j) + w_C \cdot \mathcal{L}_c(M_i, M_j) + w_G \cdot \mathcal{L}_g(M_i, M_j) \Big). \tag{10}$$

Here, M_1, M_2, M_3 correspond to the image, text, and graph modalities, respectively, where we use T_1 and G_1 instead of T and G when compared with I. w_T, w_C, and w_G are weighting coefficients for each loss component.

5. Experiments & Results

Beyond traditional CXR evaluation tasks, including zero-shot, fine-tuned classification and report retrieval, we introduce new novel tasks such as the *CXR-Align* benchmark, RSNA-Abnormal (RSNA-*ab*) classification, adversarial prediction, and normal case detection to further assess our model's clinical understanding and robustness.

5.1. Dataset

For training, we use the MIMIC dataset, where original reports are split and prior references are omitted as described in Sec. 3. All datasets undergo our preprocessing pipeline detailed in Appendix B.1. Additionally, we utilize a private tertiary hospital dataset spanning 20 years with the last year as test set for our novel normal case detection task.

Evaluation is conducted on multiple datasets: zero-shot and fine-tuned classification on RSNA-Pneumonia, RSNA-*ab* (a subset of RSNA where the model is required to distinguish pneumonia cases from all other abnormal cases),

	RSNA			RSNA-*ab*			SIIM			VinDR	Chexpert	CXR14
Model	ZS	FT10	FT100	ZS	FT10	FT100	ZS	FT10	FT100	ZS	ZS	ZS
ConVIRT [40]	75.6	78.1	80.3	68.2	75.2	76.7	68.3	78.9	81.4	68.6	39.4	51.2
BIOVIL [6]	84.3	87.6	89.1	73.8	82.0	83.0	78.6	86.0	87.3	77.2	41.5	55.4
BIOVIL-T [3]	**87.8**	88.2	89.2	**79.4**	82.1	83.3	74.9	86.8	87.9	77.4	44.2	53.7
CXRCLIP [38]	81.4	88.1	89.3	72.0	83.6	84.0	85.4	87.2	88.7	78.3	53.0	55.9
CLIP	81.2	88.3	89.1	70.6	83.7	84.0	74.3	87.8	88.0	76.1	52.3	56.7
SOFTCLIP [12]	76.6	79.1	81.1	67.8	73.1	76.2	70.1	78.3	80.3	73.1	47.1	54.2
CLIP-D$_t$	81.6	88.5	89.5	71.8	83.8	84.2	80.1	88.0	88.3	78.1	54.1	59.2
CLIP-D$_{t+c}$	84.4	89.2	90.0	74.5	84.2	84.8	84.8	88.5	88.6	78.2	54.1	61.2
CLIP^N	82.0	88.2	89.0	72.5	83.3	83.9	74.1	88.4	88.9	76.1	53.7	57.2
CLIP^N-D$_{t+c}$	86.4	**90.7**	**91.2**	78.1	84.5	85.3	85.6	89.2	89.8	**79.1**	54.4	62.8
CLIP^G	81.8	88.6	89.2	70.5	83.3	84.1	70.8	87.3	87.8	77.0	52.6	57.1
CLIP^G-D$_{t+c+g}$	85.1	89.9	90.5	74.9	84.2	85.1	84.5	88.7	89.0	78.3	56.1	62.3
CLIP^{N,G}-D$_{t+c+g}$	86.6	**90.7**	91.1	78.3	**84.8**	**85.4**	**87.2**	**89.6**	**90.2**	78.8	**57.3**	**63.0**

Table 2. Performance comparison across datasets for zero-shot (ZS) and fine-tuned (FT) entity classification for models trained on MIMIC. FT10 and FT100 denote fine-tuning with 10% and 100% of the data, respectively. The upper part shows SOTA models' performance; the lower part shows performance improvements as features are added from the baseline. Here, N denotes training with hard negatives, G denotes training with graph embeddings and D indicates training with dynamic soft labels based on t for textual similarity, c for clinical similarity, and g for graph similarity. AUC is measured for RSNA and SIIM dataset while accuracy is measured for the others.

SIIM Pneumothorax, VinDr [24], CXR14 [31], and Chexpert; adversarial prediction on CXR14 following the approach in Probmed [36]; and normal case detection on OPEN-I dataset. The CXR-Align benchmark and report retrieval task is evaluated with OpenI, MIMIC, and Chexpert datasets. Note that RSNA, VinDr, SIIM, and Chexpert test sets align with those used in GLORIA [14]. For CXR14, we use the data selected by Probmed.

5.2. Model Settings

We utilize a Swin-Tiny [22] model as the image encoder, BioClinical-BERT [1] as the text encoder, and a 2-layer GCNN for graph encoding. Input resolution is set to 224 × 224 pixels. Additional details are provided in Appendix C.

5.3. Classification

In this section, we demonstrate that entity classification tasks benefit significantly from all proposed methodologies. Tab. 2 shows the zero-shot and fine-tuned performance across tasks, methods, and training datasets.

Dynamic Soft Labeling: Utilizing soft labels based on text similarity alone provides performance improvements over the baseline CLIP. Incorporating clinical and relational similarities further enhances performance, leading to even better results. Our approach outperforms other SOTA methods on most benchmark datasets without requiring lateral images, external knowledge, masked modeling, or MVS [20] methods. The performance enhancement is particularly evident on the SIIM dataset as more similarity measures are incorporated into the dynamic soft label approach. Although our final model's zero-shot performance is slightly below that of BIOVIL-T, our fine-tuned performance is superior, suggesting that our method extracts richer representations reflecting clinical information. Notably, our soft labels lead to more stable and improved performance compared to SOFTCLIP. Instead of relying solely on text similarity, we employ thresholding and distribute role among text, clinical, and relational graph similarities, helping the model allocate its outputs appropriately across categories and enhancing clinical comprehension. A hypothesis of this effect is discussed on Appendix A.2.

Hard Negatives: Using negations as hard negatives does not improve zero-shot and fine-tuned performance when used alone, possibly due to increased overlap of clinical semantics within the batch. However, employing the dynamic soft label approach effectively addresses this issue, stabilizing the effect of hard negatives and boosting both zero-shot and fine-tuned performances. As we incorporate each soft-label approach, performance improves across benchmarks, especially in fine-tuned classification.

Graph Embeddings: Incorporating graph contrastive loss using graph embeddings with dynamic soft labels further enhances performance, and combining this with hard negatives leads to notable gains. Overall, this method synergies well with both dynamic soft labels and hard negatives.

5.4. Adversarial Prediction

Following the approach of ProbMed [36], we constructed a zero-shot task comprising a positive first query and a negative second query, where the model should correctly identify the positive entity in the first query and the negative

CXR Model	RSNA		RSNA-*ab*		SIIM		NCD
	ZS	FT10	ZS	FT10	ZS	FT10	ACC
CLIP	77.2 (-4.0)	87.6 (-0.7)	65.0 (-5.6)	82.9 (-0.8)	72.4 (-1.9)	87.5 (-0.3)	84.3
CLIP-D$_t$	78.8 (-2.8)	88.5 (-0.0)	68.8 (-3.0)	83.7 (-0.1)	79.3 (-0.8)	87.9 (-0.1)	81.4
CLIP-D$_{t+c}$	82.1 (-2.3)	89.4 (+0.2)	72.4 (-2.1)	84.2 (+0.0)	84.1 (-0.7)	88.9 (+0.4)	86.8
CLIP^G-D$_{t+c+g}$	83.2 (-1.9)	90.1 (+0.2)	72.8 (-2.1)	84.5 (+0.3)	83.8 (-0.7)	88.8 (+0.1)	85.6

Table 3. Ablation study for the dynamic soft labels in a manually set up imbalanced dataset. Performance differences are measured compared to using only the MIMIC dataset for training, as in Tab. 2. Normal Case Detection (NCD) is performed with OpenI normal CXRs, where the model is required to retrieve one normal report from 2,999 abnormal reports.

entity in the second query in sequence. Although the original paper evaluated Large Language Models (LLMs), we adapted this task to evaluate CLIP. This is a complex task requiring the model to fully understand the image and recognize which entities are present and which are absent. As shown in Tab. 4, while all of the SOTA models performed below chance level, our model achieved significantly better results compared to both chance level and the best performing model by a significant margin.

Model	Adversarial CLS
CXRCLIP	21.4
BIOVIL	23.3
BIOVIL-T	14.0
MedCLIP	11.5
GLORIA	12.0
OURS	**34.4**

Table 4. Adversarial prediction accuracy where the model is required to guess both positive and negative entities correctly in a zero-shot setting.

5.5. Ablation with Imbalanced Dataset and Normal Case Detection

To simulate the effects of training in a general hospital setting, where class imbalance is significant, we added 130,000 normal CXR cases (all labeled as ”No active lung lesion”) from our private dataset to the MIMIC training data. This addition introduces substantial imbalance, with normal CXRs constituting over half of the dataset and a large number of duplicate reports. As shown in Tab. 3, this imbalance reduces zero-shot accuracy. However, our dynamic soft label approach mitigates this performance drop, and when clinical similarities are applied, fine tuned performance surpasses that of the original model.

This raises a key question: Why add more normal data and increase imbalance rather than remove excess normal data for a balanced dataset?” The answer lies in the normal case detection task, which assesses the model’s ability to identify normal cases. This challenging task requires the model to retrieve the one normal CXR report from among 2,999 abnormal reports in the test set—a *needle-in-a-haystack* scenario. Accuracy is measured by whether the model successfully retrieves the normal report.

In evaluating these measures with CXR-CLIP and our CLIP^{N,G}-D$_{t+c+g}$ in Tab. 2, we observe low accuracy (0.7, 3.1) percent, respectively. However, including normal data in the training set raises these measures to over 80%, demonstrating the importance that inclusion of normal data is crucial for enhancing the model’s comprehension of normal cases.

5.6. CXR-Align Benchmark Evaluation

The evaluation on *CXR-Align* demonstrates that introducing hard negatives enhances our model’s understanding of negation. As shown in Tab. 6, our model significantly outperforms other SOTA models. Notably, the performance of our model trained with hard-negatives on the negation-related task is unexpectedly high, leading us to hypothesize that the model may be learning to avoid unnatural negations by exploiting shortcuts.

To address this issue, we conducted a second task where the generated negated sentence was omitted(r^r) without being replaced with negations. Our final model also showed improved performance on this task, suggesting that introducing negations can enhance the full alignment between CXR images and reports. It is important to note that using negation-based hard negatives alone does improve performance on task 1, but the performance drops on task 2 compared to the baseline, indicating that semantic overlap may have introduced noise that hinders the model’s ability to learn clinical concepts.

5.7. Report Retrieval

As shown in Tab. 5, our model achieves competitive retrieval performance compared to other state-of-the-art models, with a particularly strong showing on the CheXbert F1 score. This implies that our model captures more clinically meaningful features from the reports and forms a better alignment compared to other methods.

	MIMIC				Chexpert				Open-I			
Model	@5	F1	Recall	Precision	@5	F1	Recall	Precision	@5	F1	Recall	Precision
MedCLIP	1.3	15.9	10.7	24.2	2.8	3.8	2.8	4.4	0.4	2.0	1.4	3.2
BIOVIL	10.9	36.4	35.6	38.9	10.5	24.6	24.9	25.6	3.2	20.6	22.9	22.3
BIOVIL-T	13.3	36.7	36.4	39.1	11.1	24.4	23.9	28.5	3.7	21.1	25.1	22.7
CLIP	35.5	44.8	45.0	44.6	24.5	35.4	33.3	39.9	6.4	28.7	30.6	28.5
CLIP^{N}-D$_{t+c}$	33.9	46.0	49.7	43.9	27.9	38.0	39.3	30.0	6.2	28.8	**33.5**	24.6
CLIP^{N,G}-D$_{t+c+g}$	**38.8**	**50.6**	**51.5**	**50.3**	**29.3**	**42.0**	**43.2**	**36.6**	**7.8**	**29.0**	31.4	**29.2**

Table 5. Retrieval performance on MIMIC, CheXpert, and OpenI, evaluated using Top-5 accuracy (@5) and CheXbert-based F1, Recall, and Precision scores.

	MIMIC		Chexpert		Open-I	
Model	A	B	A	B	A	B
GLORIA	50.0	34.4	54.5	36.8	59.6	35.0
BIOVIL	60.5	61.0	58.5	59.2	60.5	58.3
BIOVIL-T	64.3	65.1	65.7	63.9	60.9	62.6
CXRCLIP	78.3	73.6	74.6	70.4	72.2	68.7
CLIP	75.4	72.4	72.5	70.3	62.7	62.6
CLIP^{N}	**97.3**	71.7	**97.2**	69.1	**96.7**	62.3
OURS	96.5	**80.1**	94.0	**75.3**	96.4	**73.8**

Table 6. CXR-Align performance across different datasets. The model is required to perform two tasks: (A) selecting between the original report r and a report r^{n} where an entity present in the CXR has been negated; (B) select between the original report r and a report r^{r} where a sentence related to a specific entity has been removed.

6. Discussion

Our work presents a method for extracting medical-focused representations that addresses key challenges in adapting general-domain models to the medical domain, specifically semantic overlap (mostly caused by data imbalance), and negation handling. Our approach consistently outperforms both baseline and state-of-the-art models across classical tasks and novel benchmarks, demonstrating robustness and effectiveness. Key insights from our findings include: (1) While using negations as hard negatives alone provided limited benefits, combining them with dynamic soft labels significantly improved performance (Sec. 5.3); (2) our methods enhanced comprehensive clinical understanding, improving performance in adversarial tasks and alignment benchmarks (Sec. 5.6, Sec. 5.4); and (3) the inclusion of normal and duplicate reports contributed positively to training on imbalanced datasets, proving valuable insights for general hospital data where data imbalances are common (Sec. 5.5).

A potential area for refinement lies in refining the RadGraph [16]-based graph representation by focusing on structured elements such as *location*, *severity/size*, and *entity*, which could yield more precise clinical representations. Additionally, refining the text encoders [5, 15], with a better understanding of compositional context, remains important.

For clinical similarity, we leveraged CheXbert [28] outputs rather than embeddings, bypassing some limitations associated with embedding-based similarity measures. However, CheXbert does not encompass all clinical entities, and using more comprehensive labels could further enhance model performance. Future improvements might include embeddings that better capture the unique semantics of chest X-rays, thereby deepening the model's understanding of clinical relationships. Additionally, exploring alternative similarity measures beyond cosine similarity [19] could yield further improvements.

Our method was designed to integrate text, clinical, and graph similarities to capture the complexity of medical image interpretation in global scale. While results show notable improvements, there remains room to refine this approach for even greater impact in clinical applications.

7. Conclusion

In this work, we addressed two pivotal challenges in medical vision-language processing—data imbalance and negation handling—by introducing a specialized method that bridges the gap between general and medical domains. Our approach employs clinically-enhanced dynamic soft labels to mitigate semantic overlaps, incorporates negation-based hard negatives to improve the model's comprehension of complex clinical semantics, and integrates graph embeddings while leveraging clinical, relational, and textual similarities. This synergy yields substantial improvements over baseline and state-of-the-art models across various benchmarks. The *CXR-Align* benchmark also highlights our model's superior ability to process negations—an often overlooked yet crucial component in medical reporting. Overall, this study paves the way for more effective medical vision-language models that address the unique challenges of clinical environments, advancing the development of reliable AI tools in healthcare.

References

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019. 5, 6

[2] Oishi Banerjee, Agustina Saenz, Kay Wu, Warren Clements, Adil Zia, Dominic Buensalido, Helen Kavnoudias, Alain S Abi-Ghanem, Nour El Ghawi, Cibele Luna, et al. Rexamine-global: A framework for uncovering inconsistencies in radiology report generation metrics. *arXiv preprint arXiv:2408.16208*, 2024. 5

[3] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023. 2, 6

[4] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. 3, 4

[5] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024. 8

[6] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 1, 2, 6

[7] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*, 2024. 2, 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[9] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 3

[10] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022. 3

[12] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1860–1868, 2024. 3, 6, 1

[13] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024. 3

[14] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 1, 2, 6

[15] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, et al. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024. 8

[16] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021. 5, 8

[17] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55, 2020. 2

[18] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 2, 3

[19] Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11137–11146, 2024. 8

[20] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 6

[21] Zhe Li, Laurence T Yang, Bocheng Ren, Xin Nie, Zhangyang Gao, Cheng Tan, and Stan Z Li. Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11704–11714, 2024. 1, 2

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[23] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 3

[24] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022. 6

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[27] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024. 2, 3

[28] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020. 3, 5, 8

[29] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[30] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12): 1399–1406, 2022. 2

[31] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 6

[32] Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1):224–237, 2024. 3

[33] Xin Wen, Bingchen Zhao, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. Generalization beyond data imbalance: A controlled study on clip for transferable insights. *arXiv preprint arXiv:2405.21070*, 2024. 3

[34] Biao Wu, Yutong Xie, Zeyu Zhang, Minh Hieu Phan, Qi Chen, Ling Chen, and Qi Wu. Xlip: Cross-modal attention masked modelling for medical language-image pre-training. *arXiv preprint arXiv:2407.19546*, 2024. 1, 2

[35] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383, 2023. 1, 2

[36] Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *arXiv preprint arXiv:2405.20421*, 2024. 6, 2

[37] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7370–7377, 2019. 5

[38] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023. 1, 2, 6

[39] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 2, 3

[40] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 2, 6

[41] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023. 1