

Meta-Entity Driven Triplet Mining for Aligning Medical Vision-Language Models

Melih B. Yilmaz*, Saban Ozturk*, Muti Kara, M. Talat Yavuz, Ekim Gumeler, Aykut Koç, *Senior Member* and Tolga Çukur, *Senior Member*

Abstract—Imaging-based diagnostics rely on evaluation of both medical images and radiology reports, but increasing data volumes strain medical experts, leading to errors and workflow delays. Medical vision-language models (med-VLMs) offer an efficient approach for processing multimodal imaging data, especially for chest X-rays (CXRs), though their success depends on effective image-text alignment. Existing alignment methods for med-VLMs, primarily based on contrastive learning, often focus on coarse-grained disease class separation and overlook fine-grained pathology attributes such as location, size, or severity, which results in suboptimal representations. We introduce MedTrim (Meta-entity-driven Triplet mining), a novel alignment method that improves precision via structured triplet learning guided by meta-entities extracted from radiology reports. Unlike conventional contrastive and triplet frameworks for representational learning that rely on global class labels or implicit similarity references, MedTrim explicitly models hierarchical relationships between pathology attributes to preserve clinically meaningful intra-class variation. To do this, MedTrim leverages a domain-specific ontology to identify adjectival qualifiers and directional descriptors of pathology, a novel entity-aware triplet mining score to capture hierarchical inter-sample similarity, and a multimodal alignment objective that enforces consistency across image-text pairs sharing detailed pathology attributes without compromising within-modality relationships. MedTrim improves performance in downstream retrieval, classification, and generation tasks compared to leading alignment methods. Code is available at: <https://github.com/icon-lab/MedTrim>.

Index Terms—multimodal alignment; vision-language model; triplet mining; radiography

I. INTRODUCTION

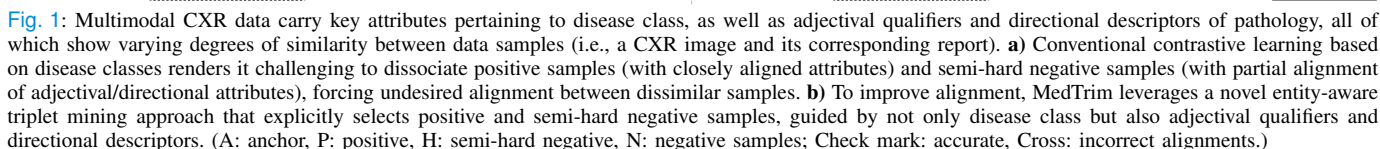
Chest X-ray (CXR) exams play a pivotal role in screening thoracic conditions such as pneumonia, lung cancer, and heart disease. These assessments rely on interpreting both CXR images and radiology reports [1]. Yet, the high volume of multimodal data can overwhelm clinicians and raise the risk of diagnostic errors, particularly for subtle or diffuse pathologies [2]. To alleviate this burden, medical vision-language models (med-VLMs) have emerged as powerful tools that integrate visual and textual information for improved clinical decision making [1], [3], enhancing interpretation accuracy and patient outcomes [4]–[7]. However, image-text misalignment in CXR

data can introduce spurious correlations, leading to erroneous predictions that undermine clinical reliability [8], [9]. Hence, robust alignment methods that preserve both within- and cross-modal fidelity are crucial for med-VLM-based assessments.

Recent work has explored various ways to strengthen CXR image-text alignment. Some approaches improve semantic conditioning on the text encoder [10], while others modify the image encoder through pathology-aware or anatomy-sensitive designs [11], [12]. Additional efforts adapt both encoders by coupling images with label prompts in a shared embedding space [13]. Despite differences in where the encoding pipeline is modified, these methods ultimately rely on contrastive learning as the core alignment framework for med-VLMs [14], [15]. In this framework, random pairs of multimodal samples (i.e., a sample being a CXR image and its corresponding report) are selected to push dissimilar pairs apart [16]–[19]. Dissimilarity is typically assessed at a relatively coarse level based on learned prototypes or predefined labels for disease classes [20], [21]. Consequently, conventional contrastive methods treat all negative pairs as uniformly dissimilar, enforcing inter-class separation without nuance (Fig. 1a), and overlooking subtle intra-class variations and shared features across disease classes [16], [17], [22]. Knowledge-guided contrastive learning has been considered as a potential remedy by constructing semantic similarity references from structured pathology labels in reports [23]. Although such guidance can provide more precise supervision than coarse disease labels, alignment objectives based on pairwise distances are inherently limited in their ability to capture graded or hierarchical similarity among multiple samples [24]. *As a result, knowledge-guided contrastive learning may still over-separate disease classes and suboptimally represent fine-grained intra-class variations.*

Contrastive objectives can be limited for CXR data because pairwise distances do not naturally encode relative similarity or graded variations in pathology attributes [25]. A promising framework in representational learning to address this limitation uses triplet-based formulations, which provide the minimal relational structure required to impose ordered similarity constraints [26], [27]. By explicitly structuring similarity through triplets of anchor, positive, and negative samples, this framework can enable finer control over both inter-class discrimination and intra-class variation [28]. *To our knowledge, however, triplet learning has not been previously explored for med-VLM alignment [29], [30].* While few prior med-VLM studies have referenced “triplets”, they use the term with an entirely different meaning: 3-tuples that enumerate pathology-related attributes within a single report, such as (entity, attribute, presence) or (suggestive of, modifier, location) [23], [31], [32]. This usage does not intend to encode any relational structure across samples, and is therefore fundamentally dis-

M.B. Yilmaz, S. Ozturk, M.T. Yavuz, A. Koç, and T. Çukur are with the Dept. of Electrical-Electronics Engineering, Bilkent University, Ankara 06800, Turkey. S. Ozturk is also with the Dept. of Management Information Systems, Ankara Hacı Bayram Veli University, Ankara 06570, Turkey. M. Kara is with the Dept. of Computer Engineering, Bilkent University, Ankara 06800, Turkey. E. Gumeler is with the Dept. of Radiology, Hacettepe University, Ankara 06230, Turkey. *S. Ozturk and M.B. Yilmaz contributed equally to this work. Corresponding author: Tolga Çukur (e-mail: cukur@ee.bilkent.edu.tr). This work was supported in part by TUBA GEBIP 2015 and BAGEP 2017 fellowships awarded to T. Çukur, in part by a BAGEP 2023 fellowship awarded to A. Koç, and in part by TUBITAK Grant no. 123E142 awarded to T. Çukur.



Although triplet learning offers a principled mechanism for modeling ordered similarity relationships, generic triplet mining is suboptimal for med-VLM alignment. This stems from the fact that conventional triplet mining strategies (whether based on random sampling, distance heuristics, or class labels [26]) rely on unidimensional selection criteria that neglect the hierarchy of fine-grained attributes present in radiological prose. As a result, they can struggle to capture subtle yet clinically critical distinctions in entities like adjectival qualifiers (mild vs. severe) or directional descriptors (upper lobe vs. peripheral). *Ignoring these entities during sample selection can blur important differences within a disease class and make it harder to separate closely related classes.*

The novelty of MedTrim stems from bringing triplet learning together with meta-entity-aware sample selection guided by pathology descriptions in radiology reports. Previous studies have used structured pathology descriptors to guide contrastive learning, but these descriptors have not been used to form triplets of CXR samples or to define ordered similarity relationships across data. Note that MedTrim moves beyond standard triplet learning, which typically relies only on disease labels, by introducing an entity-aware similarity score that considers overlap in disease classes as well as adjectival and directional entities. This score helps the model mine triplets that capture subtle clinical differences rather than broad category distinctions. As a result, MedTrim can represent variations in pathology that are often missed by class-based triplet mining and contrastive approaches.

- MedTrim is a novel med-VLM alignment method that moves beyond conventional contrastive alignment with a triplet learning framework driven by meta-entity-guided sample mining in multimodal CXR data.
- We introduce an OBER module to extract disease labels, adjectival qualifiers, and directional descriptors from radiology reports for entity-aware alignment supervision.
- We propose a novel entity-aware score function that guides triplet selection using disease, adjectival, and directional attributes to encode graded clinical similarity.
- We formulate a multimodal triplet alignment loss to jointly optimize within- and cross-modal relationships, enabling ordered similarity constraints.

Medical vision-language models. Med-VLMs encode data samples into a shared embedding space for multimodal analy-

sis [33], [34], yet they often struggle with naive combinations of image and text embeddings [35]–[37]. While radiological images capture the full range of normal anatomical structures and pathological abnormalities in a given subject, radiology reports provide condensed descriptions that emphasize abnormalities [38] and suffer from variability in prose [39]. These discrepancies can cause inaccuracies in med-VLM-based evaluations [40]–[43]. Therefore, ensuring proper image-text alignment is critical for training reliable med-VLMs.

Contrastive learning in med-VLMs. As the mainstream alignment framework, contrastive learning performs pairwise sample processing [24], [34], [44]. Given an anchor image-text sample from a batch, samples from the same disease class are pulled closer, while those from different classes are pushed away from the anchor [16], [17], [22], [45], [46]. Information regarding disease classes can be defined via external labels or prototype clustering [18], [19]. While some methods align image-text embeddings at a global level, others have sought local alignment between image patches and text fragments [47], [48]. Despite its ubiquity, contrastive learning primarily aims at segregating disease classes, rendering the aligned image-text embeddings relatively insensitive to fine-grained distinctions in pathology attributes.

Knowledge-guided contrastive learning in med-VLMs. To improve alignment precision, knowledge-guided contrastive learning methods have recently been proposed. These methods commonly represent knowledge of pathology attributes using structured labels [49], e.g., extracted via medical dictionaries and learned entity extractors, as well as general-purpose named-entity recognition tools and language models [42], [50], [51]. Structured labels are then compared across samples to derive a reference measure of semantic similarity, and image-text embeddings are aligned against this reference [41]. Although attribute-based sample relationships can implicitly influence the learning objective, contrastive learning lacks the ability to explicitly select samples reflecting fine-grained variations with respect to the anchor, as it flattens similarity relationships into pairwise distances [52]. As a result, gradual variations in pathology attributes are rendered implicit and difficult to capture. In addition, many previous knowledge-guided methods have neglected within-modal alignment objectives that can cause suboptimal image-image or text-text alignment and compromise performance in unimodal tasks [53].

Triplet learning. Although triplet learning can offer advantages over pairwise contrastive approaches in representational learning [26], its potential for med-VLM alignment remains largely unexplored. Some prior med-VLM studies have referred to “triplets” in a different meaning: 3-tuples that list pathology-related elements within a single report, such as (*entity*, *attribute*, *presence*) or (*suggestive of*, *modify*, *located at*) [23], [31], [32]. Note, however, that these studies still fundamentally rely on contrastive loss functions and pairwise sample processing for their core alignment objectives. A recent alignment method based on contrastive learning has augmented cross-entropy loss with a ranking loss, with positive samples taken as image/text pairs from the same subject and negative samples taken as image/text from different subjects [54]. While the ranking loss simultaneously

specifies positive and negative samples akin to a triplet loss, this previous method does not use information on disease classes or pathology attributes for guided sample selection. *To our knowledge, prior med-VLM frameworks have not adopted triplet learning as the primary alignment mechanism. Existing studies that use structured meta-entities apply them only within contrastive objectives, without using them to form triplets or to express ordered similarity relations across samples.* This gap motivates an approach that brings triplet learning together with meta-entity-aware sample selection, offering a more expressive and fine-grained supervision.

III. THEORY

A. Overview of MedTrim

MedTrim is devised to align image-text representations for multimodal CXR data. Let $\mathbb{D} = \{(I, T)_1, \dots, (I, T)_K\}$ denote a training dataset with CXR images and reports from K subjects, where $I \in \mathbb{R}^{H \times W}$ are single-view images with H and W specifying height and width, and T are reports comprising J tokens $\{t_1, t_2, \dots, t_J\}$ (where tokens are single words and J is variable across reports). For each training iteration, MedTrim draws a distinct mini-batch with k image-text samples from $\mathbb{D}$ denoted as $\mathbb{B} = \{(I, T)_1, \dots, (I, T)_k\}$, $k \ll K$. The four training stages per mini-batch are given below:

▷ In the first stage, MedTrim leverages its OBER module (Sec. III-B) to extract a set of meta-entities from each radiology report as $\mathbf{m} = \text{OBER}(T)$. The meta-entities for each report are cast as a structured list $\mathbf{m} = (\mathbf{d}, \text{adj}(\mathbf{d}), \text{dir}(\mathbf{d}))$, where $\mathbf{d}$ is a list of diseases classes, $\text{adj}(\cdot)$ is a disease-specific list of adjectival qualifiers, and $\text{dir}(\cdot)$ is a disease-specific list of directional descriptors. Note that each report can contain multiple disease classes, and there can be multiple adjectival qualifiers or directional descriptors for each disease class.

▷ In the second stage, MedTrim performs entity-driven triplet mining within the given mini-batch (Sec. III-C). A total of n_{tri} triplets are formed, where the n th triplet is composed as $\mathbb{T}ri_n = \{(I_A, T_A), (I_P, T_P), (I_N, T_N)\}_n$ with A denoting the anchor, P denoting the positive, and N denoting the negative sample. MedTrim leverages an entity similarity score, *score*, used to mine semi-hard negative samples for improved sensitivity to fine-grained variations in CXR data.

▷ In the third stage, the mined triplets are encoded onto a latent embedding space via transformer-based image and text encoders (Sec. III-D). This results in an embedding vector per image $E^I = \text{Trans}(I) \in \mathbb{R}^c$ that globally pools information across image patches, and an embedding vector per report $E^T = \text{Trans}(T) \in \mathbb{R}^c$ that pools information across tokens (c : embedding dimensionality). As such, each mined triplet is projected onto image representations $\{E_A^I, E_P^I, E_N^I\}$ and text representations $\{E_A^T, E_P^T, E_N^T\}$.

▷ In the fourth stage, MedTrim expresses a unified triplet objective $\mathcal{L}_{MedTrim}$ to align image-text representations (Sec. III-E). $\mathcal{L}_{MedTrim}$ contains separate triplet terms for within-modal alignment (image-to-image, text-to-text) and cross-modal alignment (image-to-text, text-to-image). This unified objective enables MedTrim to precisely align CXR images

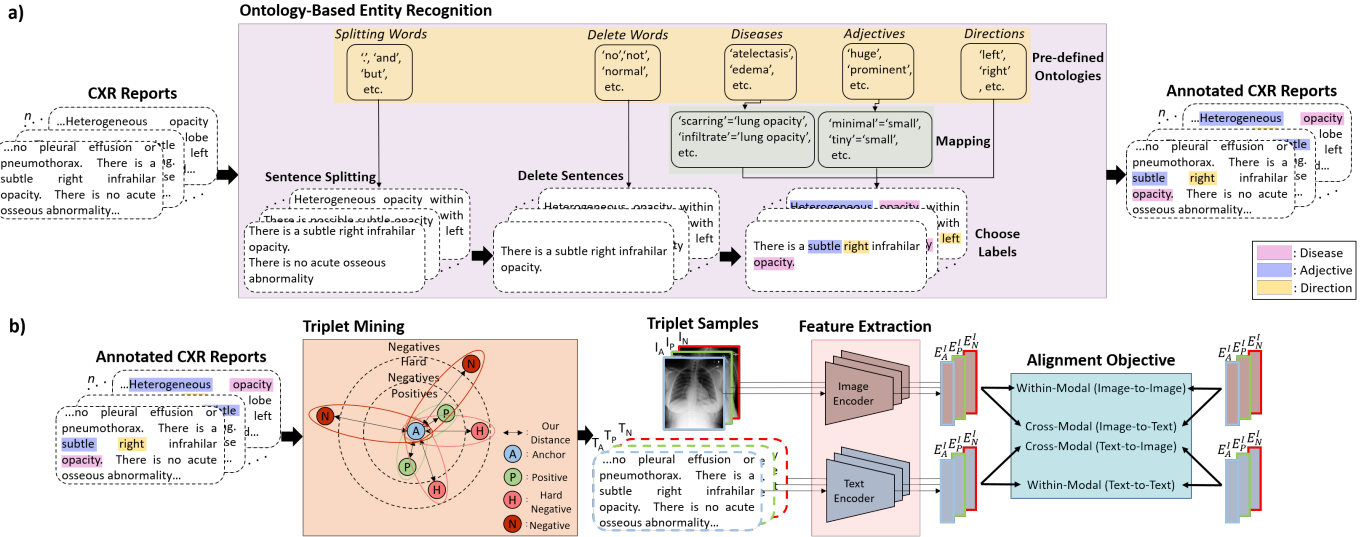


Fig. 2: **a)** Given a batch of multimodal CXR samples, MedTrim first deploys its OBER module to identify disease labels along with adjectival qualifiers and directional descriptors of corresponding pathology from radiology reports. **b)** Using these pathology attributes as meta entities, MedTrim performs triplet mining to select positive and semi-hard negative samples via entity-based weighting with respect to the anchor sample. The embeddings of the selected triplets are extracted with transformer-based image and text encoders. For image-text alignment, the encoders are fine-tuned via a multimodal triplet alignment objective that synergizes within-modal and cross-modal alignment expressed over embedding vectors.

against respective radiology reports, without compromising representations of within-modal relationships.

B. Ontology-Based Entity Recognition

Public CXR databases primarily provide raw radiology reports and disease labels, but they often lack isolated labels describing pathology attributes. As a result, attribute extraction must rely on indirect supervision from free-text reports, which can introduce variability and uncertainty in the extracted entities, with learning-based generative methods potentially more susceptible to this effect [55]. Meanwhile, many ontology-based extractors are either built for general language processing [56], potentially compromising sensitivity on medical data, or they are not publicly accessible. To address this gap, we introduce a new ontology-based recognition module (OBER) for high-fidelity extraction of meta-entities from CXR reports as outlined in Algorithm 1. We devised OBER as an ontology with disease words (O_{dis}), adjective words (O_{adj}), and direction words (O_{dir}) to detect meta-entities, as well as sentence splitting words (O_{spl}) to control text fragmentation, and delete words (O_{del}) to discard irrelevant text fragments. These words were curated from a medical dictionary, manually extended to include a comprehensive list of synonyms [57], and used in lemmatized form to eliminate the effects of various inflections such as singular-plural distinctions [58]. The resultant ontology comprised synsets for 12 disease, 98 adjective, 4 direction, 6 splitting, and 16 delete words.

Given a radiology report T , OBER first lemmatizes it by processing its individual tokens as $\bar{T} = f_{lem}(T)$. It then splits the report into sentences using splitting words as $\{s_1, \dots, s_V\} = f_{spl}(\bar{T}, O_{spl})$, where V is the number of sentences in $\bar{T}$. Sentences that contain irrelevant text are discarded using delete words to obtain a condensed list of v sentences: $\{s_1, \dots, s_v\} = f_{del}(\{s_1, \dots, s_V\}, O_{del})$. Following this clean-up, remaining sentences are processed to extract meta-entities starting from empty sets, $\mathbf{d} = \emptyset$, $adj(\mathbf{d}) = \emptyset$, $dir(\mathbf{d}) = \emptyset$. For the i th sentence, a search for disease

words is conducted to recognize the disease class: $\mathbf{d}[i] = \{f_{dis}(s_i, O_{dis})\}$, where f_{dis} returns an empty set when no disease words are present in the sentence. If a disease class is recognized, then further search for adjective and direction words is conducted as $adj(\mathbf{d}[i]) = f_{adj}(s_i, O_{adj})$ and $dir(\mathbf{d}[i]) = f_{dir}(s_i, O_{dir})$. Note that adjectival qualifiers and directional searches can return empty sets, single words or multiple words describing the pathology for the given disease class $\mathbf{d}[i]$. Processing of all sentences produces a list for disease classes $\mathbf{d} = \{\mathbf{d}[1], \mathbf{d}[2], \dots\}$, a list for adjectival qualifiers $adj(\mathbf{d}) = \{adj(\mathbf{d}[1]), adj(\mathbf{d}[2]), \dots\}$, and a list for directional descriptors $dir(\mathbf{d}) = \{dir(\mathbf{d}[1]), dir(\mathbf{d}[2]), \dots\}$. These lists are aggregated to form a structured list of meta-entities for each report, $\mathbf{m} = (\mathbf{d}, adj(\mathbf{d}), dir(\mathbf{d}))$.

C. Entity-Driven Triplet Mining

Score function: MedTrim leverages a novel entity-driven triplet mining procedure to select samples that accurately capture fine-grained variations within individual disease classes. In this procedure, a random image-text sample from a given mini-batch is taken as the anchor sample. Afterwards, positive and negative samples are selected from the mini-batch to create a triplet based on similarity of meta-entities. Note that each sample can carry labels for multiple diseases, and there may be multiple adjectival qualifiers and directional descriptors within each disease class. These hierarchically-structured entities can show varying levels of match between CXR samples, and evaluating similarity of meta-entities hence requires a carefully constructed measure that respects the entity hierarchy. To address this issue, we introduce a novel score function that measures the aggregate similarity of disease classes, adjectival qualifiers and directional descriptors between two CXR samples as follows:

$$score(\mathbf{m}_i, \mathbf{m}_j) = \frac{\delta_{\mathbf{d}}(\mathbf{d}_i \cap \mathbf{d}_j)}{|\mathbf{d}_i \cup \mathbf{d}_j|} \times \sum_{\mathbf{d}[q] \in \mathbf{d}_i \cap \mathbf{d}_j} \left(\frac{\gamma_0 + \gamma_1 \mathbf{J}_{adj}(\mathbf{d}[q]) + \gamma_2 \mathbf{J}_{dir}(\mathbf{d}[q])}{\gamma_0 + \gamma_1 \delta_{adj}(\mathbf{d}[q]) + \gamma_2 \delta_{dir}(\mathbf{d}[q])} \right), \quad (1)$$

Algorithm 1: OBER Module

Input: I : CXR image, T : radiology report
 O_{dis} , O_{adj} , O_{dir} , O_{spl} , O_{del} : disease, adjective, direction, sentence splitting, and delete words.
 $f(\cdot)$: denotes various mappings for word extraction

Output: $\mathbf{m} = (\mathbf{d}, \mathbf{adj}(\mathbf{d}), \mathbf{dir}(\mathbf{d}))$

Do:
Lemmatize report $\rightarrow \bar{T} = f_{lem}(T)$
Split into sentences $\rightarrow \{s_1, \dots, s_V\} = f_{spl}(\bar{T}, O_{spl})$
Delete irrelevant $\rightarrow \{s_1, \dots, s_v\} = f_{del}(\{s_1, \dots, s_V\}, O_{del})$
for $i = 1:v$ **do**
Disease class for sentence $i \rightarrow \mathbf{d}[i] = f_{dis}(s_i, O_{dis})$
if $\mathbf{d}[i]$ exists:
Adjectival entity list $\rightarrow \mathbf{adj}(\mathbf{d}[i]) = f_{adj}(s_i, O_{adj})$
Directional entity list $\rightarrow \mathbf{dir}(\mathbf{d}[i]) = f_{dir}(s_i, O_{dir})$
end
return $(\mathbf{d}, \mathbf{adj}(\mathbf{d}), \mathbf{dir}(\mathbf{d}))$

where $\mathbf{m}_i$, $\mathbf{m}_j$ are meta-entities and $\mathbf{d}_i$, $\mathbf{d}_j$ are the label sets of disease classes for the i th and j th samples, $\cup$ is the union operator that pools labels across sets, and $\cap$ is the intersection operator that returns the common labels across sets.

The proposed score function identifies disease class labels shared between two CXR samples, and in case of a matching disease class, it proceeds onto evaluating potential matches between the labels for adjectival qualifiers and directional descriptors. Accordingly, the summation in Eq. 1 is taken over disease classes, where $\mathbf{d}[q]$ denotes the q th disease class in the intersection set of $\mathbf{d}_i$ and $\mathbf{d}_j$. For a given $\mathbf{d}[q]$ shared between i th and j th samples, the score is a combination of the degree of label matches for disease class (taken as unity), adjectival qualifiers and directional descriptors, with $\gamma_{0,1,2}$ denoting the respective weighting parameters such that $\gamma_0 + \gamma_1 + \gamma_2 = 1$. Meanwhile, $\text{JI}_{adj,dir}$ denote Jaccard indices that reflect the label matches for adjectival qualifiers and directional descriptors [59]:

$$\text{JI}_{adj}(\mathbf{d}[q]) = \frac{|\mathbf{adj}_i(\mathbf{d}[q]) \cap \mathbf{adj}_j(\mathbf{d}[q])|}{|\mathbf{adj}_i(\mathbf{d}[q]) \cup \mathbf{adj}_j(\mathbf{d}[q])|}, \quad (2)$$

$$\text{JI}_{dir}(\mathbf{d}[q]) = \frac{|\mathbf{dir}_i(\mathbf{d}[q]) \cap \mathbf{dir}_j(\mathbf{d}[q])|}{|\mathbf{dir}_i(\mathbf{d}[q]) \cup \mathbf{dir}_j(\mathbf{d}[q])|}. \quad (3)$$

Note that Jaccard indices have a range of $0 \leq \text{JI} \leq 1$. To ensure that the weighted score for $\mathbf{d}[q]$ is also normalized to the same range, the summation denominator in Eq. 1 contains δ_{adj} to indicate whether the union set of adjectival qualifiers for the disease class $\mathbf{d}[q]$ is non-empty, and δ_{dir} to indicate whether the union set of directional descriptors for the disease class $\mathbf{dir}[q]$ is non-empty. An indicator function for disease class, $\delta_{\mathbf{d}}$, is also included to assign a zero score to cases when disease class labels do not show any overlap:

$$\delta_{\mathbf{d}}(\mathbf{d}_i \cap \mathbf{d}_j) = \text{sgn}(|\mathbf{d}_i \cap \mathbf{d}_j|), \quad (4)$$

$$\delta_{adj}(\mathbf{d}[q]) = \text{sgn}(|\mathbf{adj}_i(\mathbf{d}[q]) \cap \mathbf{adj}_j(\mathbf{d}[q])|), \quad (5)$$

$$\delta_{dir}(\mathbf{d}[q]) = \text{sgn}(|\mathbf{dir}_i(\mathbf{d}[q]) \cap \mathbf{dir}_j(\mathbf{d}[q])|), \quad (6)$$

where sgn is the signum operator.

Sample selection: Given a mini-batch $\mathbb{B}$ with k samples of CXR data, MedTrim mines a total of n_{tri} triplets of the form $\text{Tri}_n = \{(I_A, T_A), (I_P, T_P), (I_N, T_N)\}_n$ (n : triplet index). Each triplet is devised to contain a unique anchor sample

A , randomly drawn from $\mathbb{B}$. Assume that the n th triplet is initiated by drawing the q th mini-batch sample as A , such that $(I_A, T_A) = \mathbb{B}[q] = (I, T)_q$ and $\mathbf{m}_A = \mathbf{m}_q$. MedTrim employs Eq. 1 to compute the similarity scores between A and remaining samples within the mini-batch. The image-text for the positive sample P are drawn from $\mathbb{B}$ by identifying the sample with maximum similarity score to A :

$$(I_P, T_P) = \mathbb{B} \left[\underset{j \in [1 \dots k], j \neq q}{\text{argmax}} |score(\mathbf{m}_A, \mathbf{m}_j)| \right]. \quad (7)$$

A crucial factor that influences the ability of triplet mining to learn discriminative representations is the selection of negative samples. While easy negatives that are substantially dissimilar from the anchor present trivial cases for representational learning, hard negatives that closely resemble positive samples can elicit underfitting when distinctions become excessively subtle. To improve learning of intra-class variations, MedTrim instead employs semi-hard negative samples that present challenging cases while still bearing moderate similarity to positive samples. Accordingly, the image-text for the semi-hard negative sample N are drawn from $\mathbb{B}$ by identifying the sample with minimum score in the range $[\tau_{min} \dots \tau_{max}]$:

$$(I_N, T_N) = \mathbb{B} \left[\underset{j \in [1 \dots k], j \neq q}{\text{argmin}} |score(\mathbf{m}_A, \mathbf{m}_j)| \right. \\ \left. \text{s.t. } \tau_{min} \leq score \leq \tau_{max} \right]. \quad (8)$$

The range constraint ensures that N shows a match to A in terms of disease classes, albeit it has mismatched adjectival qualifiers and/or directional descriptors. This semi-hard negative selection enforces the model to better capture fine-grained intra-class variations due to varying pathology attributes.

D. Encoding onto Latent Embedding Space

Following entity-driven triplet mining, multimodal CXR data are projected onto a latent embedding space, where the alignment objective is expressed. MedTrim leverages transformer encoders for both CXR images and reports, given their high representational power. To construct the encoder inputs, a given CXR image I is split into p patches to form the sequence $e^I = \{e_1^I, e_2^I, \dots, e_p^I\}$, and a given CXR report T is split into p tokens (i.e., words) to form the sequence $e^T = \{e_1^T, e_2^T, \dots, e_p^T\}$. For brevity, here we will present a common description of the encoder processing for the two modalities. Accordingly, the input sequence e is first embedded via a learnable linear layer (Lin) and supplemented with learnable position encodings (PE) [60]:

$$h^{(0)} = \text{Lin}(e) + \text{PE}(e). \quad (9)$$

The embedded sequence is then projected through L transformer blocks, comprising normalization (LN), multi-head self-attention (MHSA), and multi-layer perceptron (MLP) layers along with residual connections [60]. The mappings across the l th block can be expressed as:

$$\tilde{h}^{(l)} = \text{MHSA}(\text{LN}(h^{(l-1)})) + h^{(l-1)}, \quad (10)$$

$$h^{(l)} = \text{MLP}(\text{LN}(\tilde{h}^{(l)})) + \tilde{h}^{(l)}, \quad (11)$$

TABLE I: Ablation study on meta-entity guidance. MedTrim was compared against variants that selectively ablated guidance from disease classes, adjectival qualifiers and directional descriptors. Consistency of retrieved items was examined in terms of disease-class (Dis.), adjectival-qualifier (Adj.), and directional-descriptor (Dir.). Average performances across I2I, I2T, T2I, T2T tasks are reported as P@R (%) where R is the number of retrieved samples. Boldface marks the top-performing variant.

Meta Entity			P@10			P@20			P@50		
Dis.	Adj.	Dir.	Dis.	Adj.	Dir.	Dis.	Adj.	Dir.	Dis.	Adj.	Dir.
✓			95.6	81.5	85.3	96.4	82.6	85.4	97.3	83.5	88.3
✓	✓		95.7	84.2	84.8	95.8	86.1	84.9	96.9	86.7	88.5
✓		✓	95.4	81.3	88.5	95.7	83.5	88.9	96.8	83.4	91.6
✓	✓	✓	95.9	84.5	89.3	96.4	86.0	90.0	97.2	87.8	92.2

where $h^{(l)}$ is the block output. At the output of the L th block, the hidden representations are pooled [60]:

$$E = \text{GAP}(h^{(L)}), \quad (12)$$

where GAP denotes global average pooling across sequence elements. The overall mapping that encoders perform can be summarized as: $E^I = \text{Trans}(I)$ and $E^T = \text{Trans}(T)$.

E. Multimodal Triplet Alignment Objective

Conventional triplet learning performs unimodal representational learning by pulling together similar samples (i.e., anchor and positive), while pushing away dissimilar samples (i.e., anchor and negative) in the latent embedding space. For unimodal embedding vectors $\{E_A, E_P, E_N\}$, the standard triplet learning objective is given as [61]:

$$f_{tri}(E_A, E_P, E_N) = \max(0, \cos(E_A, E_P) - \cos(E_A, E_N) + \alpha), \quad (13)$$

where $\cos(\cdot)$ denotes the cosine similarity between embedding vectors, and α is a margin parameter. Adopting representational learning frameworks for image-text alignment, many previous alignment methods in the literature have expressed cross-modal learning objectives. Thus, a straightforward generalization of triplet learning to align multimodal CXR data can be implemented via the following terms:

$$\mathcal{L}_{I2T} = f_{tri}(E_A^I, E_P^T, E_N^T), \quad (14)$$

$$\mathcal{L}_{T2I} = f_{tri}(E_A^T, E_P^I, E_N^I), \quad (15)$$

where $I2T$ denotes image-to-text and $T2I$ denotes text-to-image alignment. Note, however, that solely using the cross-modal terms will attempt to align image against text embeddings, while neglecting within-modal alignment among images or among text. To seek image-text alignment without compromising an accurate capture of modality-specific relationships in CXR data, MedTrim leverages a new multimodal triplet learning objective that also includes within-modal terms:

$$\mathcal{L}_{I2I} = f_{tri}(E_A^I, E_P^I, E_N^I), \quad (16)$$

$$\mathcal{L}_{T2T} = f_{tri}(E_A^T, E_P^T, E_N^T), \quad (17)$$

where $I2I$ denotes image-to-image, and $T2T$ denotes text-to-text alignment. Finally, the overall objective for MedTrim can then be expressed as follows:

$$\mathcal{L}_{MedTrim} = \eta(\mathcal{L}_{I2T} + \mathcal{L}_{T2I}) + (1 - \eta)(\mathcal{L}_{I2I} + \mathcal{L}_{T2T}), \quad (18)$$

where η is a weighting parameter that balances the contributions of cross-modal versus within-modal alignment terms.

IV. EXPERIMENTAL METHODS

A. Datasets

In our experiments, we used three public datasets: MIMIC-CXR [62], CheXpert [57], and RSNA Pneumonia [63]. All alignment methods were originally trained on MIMIC-CXR. In MIMIC-CXR, 50,000 paired samples of CXR images and radiology reports were analyzed [62], covering 12 diseases as listed in Supp. Table 1 (Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Card., Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax). Data were split randomly into training (70%), validation (10%), test (20%) sets without any subject overlap. Models were learned on the training set, while hyperparameter selection was conducted on the validation set.

Trained models were subsequently evaluated on retrieval tasks using 14,000 samples from the test set of MIMIC-CXR. Meanwhile, zero-shot and few-shot disease classification tasks were performed on the test sets of CheXpert (1,000 samples) and RSNA Pneumonia (4,000 samples). For CheXpert, a subset referred to as CheXpert 5x200 was constructed following the protocol in [24], consisting of 200 images from each of five disease categories: Atelectasis, Cardiomegaly, Edema, Pleural Effusion, and Consolidation. For RSNA Pneumonia, 4,000 CXR images were sampled from the NIH CXR dataset [63], preserving the original class distribution with 20% pneumonia cases and 80% healthy controls. In zero-shot experiments, models trained on MIMIC-CXR were directly evaluated without further adaptation. In few-shot experiments, an independent fine-tuning set comprising 10% of the labeled training data from each dataset was used to adapt the models.

B. Implementation Details for MedTrim

MedTrim employed transformer-based encoders for projecting multimodal CXR data onto the latent embedding space. A ViT module pre-trained on natural images was taken as the image encoder [60], and a BERT module pre-trained on natural text was taken as the text encoder [64]. 1,250,000 unique triplets were mined from the training set. For consistency, image samples in each triplet were curated from a single CXR view (anteroposterior, posteroanterior or lateral), and images were resized to 224×224 pixels. Radiology reports that contained a findings or impressions section were curated. Key hyperparameters were selected via stratified cross-validation by minimizing validation loss. Accordingly, $\gamma_0 = 0.85$, $\gamma_1 = 0.1$, $\gamma_2 = 0.05$ were used for the score function; $\tau_{min} = 0.25$, $\tau_{max} = 0.6$ were used for triplet mining; $\eta = 0.5$ was used for the triplet loss (the margin parameter $\alpha = 0.3$ followed common practice in the triplet-learning literature [26]); and the Base variant of transformer backbones were used for image/text encoders [65].

C. Competing Methods

MedTrim was comparatively demonstrated against seven state-of-the-art alignment methods. For all methods, training was performed on NVidia RTX 4090 GPUs via the Adam

TABLE II: Ablation study on multi-modal triplet alignment objective. Variants were formed by selectively including the following components in the objective: triplet learning (f_{tri} , $\times$: contrastive, $\checkmark$: triplet), within-modal alignment losses ($\mathcal{L}_{I2I+T2T}$), image-to-text alignment loss ($\mathcal{L}_{I2T}$), and text-to-image alignment loss ($\mathcal{L}_{T2I}$). Average retrieval performances across the three meta-entities are reported as P@R.

Loss components				I2I			I2T			T2I			T2T		
f_{tri}	$\mathcal{L}_{I2I+T2T}$	$\mathcal{L}_{I2T}$	$\mathcal{L}_{T2I}$	P@10	P@20	P@50	P@10	P@20	P@50	P@10	P@20	P@50	P@10	P@20	P@50
$\times$	$\checkmark$			81.6	82.1	82.4	73.7	74.3	75.1	76.6	77.0	77.1	82.0	82.3	82.6
$\times$	$\checkmark$	$\checkmark$		88.6	89.1	89.5	84.8	85.0	85.3	83.6	84.2	84.8	89.3	89.8	90.1
$\times$	$\checkmark$		$\checkmark$	87.5	88.1	88.7	82.3	82.8	83.2	87.2	87.8	88.5	90.0	90.5	91.0
$\times$		$\checkmark$	$\checkmark$	86.6	87.0	87.2	85.5	86.0	86.3	87.0	87.5	88.2	88.3	89.0	89.6
$\times$	$\checkmark$	$\checkmark$	$\checkmark$	90.7	90.8	91.2	89.9	90.4	90.7	90.1	90.5	91.0	93.2	94.5	95.1
$\checkmark$	$\checkmark$			86.6	87.1	87.4	80.7	81.3	81.4	82.6	83.3	83.9	90.3	91.1	92.6
$\checkmark$	$\checkmark$	$\checkmark$		94.7	95.0	95.5	93.1	93.5	93.8	91.8	92.0	92.1	93.1	94.1	94.8
$\checkmark$	$\checkmark$		$\checkmark$	94.9	95.1	95.4	92.5	92.9	93.2	93.0	93.4	93.5	93.6	94.4	95.1
$\checkmark$		$\checkmark$	$\checkmark$	93.6	93.8	94.0	93.0	93.4	93.5	93.2	93.5	93.9	92.8	93.6	94.5
$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	96.7	96.8	97.3	96.2	96.5	97.5	96.1	96.3	96.8	94.5	95.8	97.0

TABLE III: Reliability in entity extraction. Accuracy in entity extraction was assessed via the intersection-over-union (IoU) metric between predicted entity tuples (i.e., disease, adjective, direction entities) and reference tuples. NegBio and CheXpert-Labeler only offer disease entities.

	Dis.	Adj.	Dir.
NegBio	72.6 $\pm$ 4.0	–	–
CheXpert-Labeler	74.1 $\pm$ 5.6	–	–
Radgraph-XL	82.2 $\pm$ 1.8	73.3 $\pm$ 3.6	78.1 $\pm$ 4.5
OBER	87.3$\pm$2.1	76.6$\pm$3.4	79.3$\pm$3.6

TABLE IV: Reliability in entity extraction in challenging report subsets. Accuracy in entity extraction was assessed in short reports and reports with entity tuples spread across multiple sentences via IoU.

	Dis.	Adj.	Dir.
NegBio	67.3 $\pm$ 6.2	–	–
CheXpert-Labeler	67.7 $\pm$ 5.1	–	–
Radgraph-XL	77.2 $\pm$ 3.1	65.8 $\pm$ 3.6	70.6 $\pm$ 4.5
OBER	80.3$\pm$3.5	68.8$\pm$4.2	71.5$\pm$4.6

optimizer, using a learning rate of 4×10^{-5} , a mini-batch size of 64, a weight decay factor of 1×10^{-4} , and 20 epochs.

ConVIRT [24]: uses contrastive learning to align medical images and random sentences within radiology reports at the global level.

JoImTeRNet [54]: uses contrastive learning for global image-sentence and local region-phrase alignment. Cross-entropy, ranking, masked modeling losses are combined.

GLoRIA [44]: uses contrastive learning for global image-text and local region-word alignment. Local and global contrastive losses are combined.

CheXzero [66]: uses contrastive learning to align medical images and impressions in radiology reports.

LIMITR [2]: uses knowledge-guided contrastive learning for image-report alignment at local and global levels. Location information and visual cues from multiple views (frontal and lateral) are employed for guidance.

MedFILIP [67]: uses knowledge-guided contrastive learning where an LLM extracts structured labels on disease class, severity and location from radiology reports. Image and text embeddings are aligned against a reference semantic similarity matrix derived from these labels.

MedKLIP [41]: uses knowledge-guided contrastive learning where a medical NER module extracts disease-related entities from radiology reports. Image and text embeddings are aligned via entity-level contrastive and cross-entropy losses.

D. Performance Evaluation

Models were first demonstrated for within-modal (image-to-image, text-to-text) and cross-modal (image-to-text, text-to-image) retrieval tasks. Consistency between query and retrieved items was evaluated in terms of disease class, adjectival qualifier and directional descriptor. In each case, performance was measured via Precision@R (P@R) at $R \in \{1, 10, 20, 50\}$ reflecting the percentage of consistent items among the top- R retrieved items, average Recall and mean average precision (mAP) metrics. Retrieved items (i.e., image or text) were ranked according to the cosine similarity between their embedding vectors. Consistency between the meta-entities of query and retrieved items was measured via Jaccard index. Models were also demonstrated for disease classification tasks from CXR images. Performance was measured via accuracy (ACC), area under the curve (AUC), and F1 metrics. For a given CXR image, the image embedding was compared against the embeddings of candidate text prompts in the form of “This is an X-Ray image of disease.”. Disease class was predicted based on the prompt with the maximum similarity between image and text embeddings. For robust evaluations, the test set was split into 20 non-overlapping subsets in MIMIC-CXR and 5 non-overlapping subsets in CheXpert and RSNA Pneumonia. Metrics were then reported as mean $\pm$ std across subsets. Statistical significance of differences in model performance were assessed via non-parametric Wilcoxon signed-rank tests.

V. RESULTS

A. Ablation Studies

Meta-entity guidance: To assess the impact of meta-entity guidance, we conducted an ablation study by training variant models that selectively enabled or disabled guidance from disease classes, adjectival qualifiers, and directional descriptors. Models were evaluated for downstream image-to-image (I2I), image-to-text (I2T), text-to-image (T2I) and text-to-text (T2T) retrieval tasks. Table I lists average retrieval precision (P@R with R denoting the number of retrieved items) of variants across tasks, while separately evaluating consistency between query and retrieved items in terms of disease-class, adjectival-qualifier and directional-descriptor entities. On average, including adjectival guidance when disease and directional guidance are already present improves precision for adjectival qualifiers by 1.7%, including directional guidance when disease and adjectival guidance are present improves precision for

TABLE V: Performance in multimodal retrieval. Retrieval performance was evaluated using precision at rank (P@R), along with average recall (averaged over R=1-50) and mean Average Precision (mAP), capturing both ranking accuracy and retrieval coverage. Results are reported as mean±std across test subsets in MIMIC-CXR.

		P@1			P@10			P@20			P@50			Avg Recall			mAP		
		Dis.	Adj.	Dir.	Dis.	Adj.	Dir.	Dis.	Adj.	Dir.	Dis.	Adj.	Dir.	Dis.	Adj.	Dir.	Dis.	Adj.	Dir.
I2I	ConVIRT	90.3±2.4	77.8±4.4	80.2±3.9	90.6±2.3	78.1±4.2	80.6±3.5	90.8±2.3	78.0±4.0	81.0±3.5	90.8±2.4	78.5±4.1	81.2±3.3	91.2±1.7	78.1±2.4	79.5±2.4	89.1±1.7	74.6±3.4	76.4±3.1
	JoImTeRNet	90.0±2.6	65.3±6.5	72.4±5.8	90.5±2.7	65.3±6.3	72.0±5.6	89.3±2.7	66.3±6.2	73.0±5.7	89.0±2.7	66.8±5.9	73.5±5.6	86.5±2.4	63.1±2.9	69.2±3.4	81.0±2.2	56.9±5.5	65.0±4.1
	GLORIA	90.0±2.6	75.0±5.1	80.0±4.1	89.0±2.8	72.5±4.8	78.0±4.5	88.0±2.6	74.0±4.5	80.0±3.9	90.0±2.3	80.0±3.8	80.0±3.7	87.1±2.3	73.1±2.7	82.1±1.9	86.4±1.2	69.6±2.7	79.3±2.8
	CheXzero	92.0±2.4	77.0±4.7	82.5±2.4	77.5±4.5	82.5±3.8	92.3±2.6	77.3±4.9	82.3±3.6	92.2±2.5	77.2±4.5	82.2±3.5	91.3±1.7	79.1±2.3	81.5±2.1	89.3±1.7	74.1±2.7	80.0±2.6	
	LIMITR	93.0±2.3	80.4±4.0	83.1±3.5	93.2±2.3	81.0±3.5	83.2±3.6	94.0±2.1	82.6±3.3	84.9±3.3	94.2±2.0	83.0±3.4	85.0±3.3	92.1±1.7	80.8±2.3	81.3±2.1	90.2±1.5	77.8±2.5	79.0±3.0
	MedFILIP	95.5±1.8	82.5±3.5	87.8±3.3	96.1±1.8	84.2±3.3	88.5±3.0	96.4±1.6	84.7±3.0	89.0±2.8	96.9±1.5	86.0±2.4	90.8±2.4	94.8±1.5	80.6±2.3	85.7±1.7	88.1±2.1	78.0±2.5	85.6±1.8
	MedKLIP	95.5±2.1	83.0±3.3	88.2±2.7	96.0±1.9	85.0±3.3	87.3±2.9	96.5±1.5	85.0±3.2	88.7±2.5	97.0±1.2	86.9±2.3	91.0±2.1	96.0±1.3	82.5±2.1	88.2±1.6	92.6±1.4	79.6±2.7	87.0±1.6
	MedTrim	96.4±1.8	83.9±4.1	90.0±2.3	96.7±1.6	85.1±3.5	91.0±2.1	96.8±1.6	85.6±3.1	91.2±2.0	97.3±1.4	87.7±2.8	92.6±2.0	95.8±1.6	83.5±1.9	89.1±1.4	93.1±1.4	80.2±2.4	90.6±1.4
I2T	ConVIRT	89.8±2.7	76.8±4.4	78.8±3.6	90.2±2.5	77.5±4.1	80.0±3.8	90.0±2.6	78.0±3.9	80.5±3.9	89.8±2.7	78.2±4.0	80.3±3.7	89.6±2.2	79.3±2.3	80.2±2.2	87.9±1.9	75.8±3.1	77.8±2.9
	JoImTeRNet	90.1±2.4	65.3±6.5	71.8±5.4	90.0±2.7	65.3±5.9	72.7±4.9	89.5±2.8	66.5±6.1	73.0±4.6	89.0±2.5	66.2±5.8	72.9±4.7	87.2±2.6	61.9±3.1	71.2±3.3	82.6±1.9	58.1±6.1	65.9±4.4
	GLORIA	90.5±2.4	76.0±4.8	81.0±3.9	90.0±2.4	73.0±4.6	80.0±4.1	90.3±2.3	76.0±4.9	83.0±3.9	90.5±2.5	76.0±4.8	83.0±3.7	89.2±2.0	73.5±2.6	79.8±2.2	87.5±1.5	70.3±2.5	76.4±2.2
	CheXzero	93.0±2.3	79.0±4.4	84.0±3.6	93.5±2.6	79.5±4.1	84.5±3.2	93.3±2.3	79.3±4.5	84.3±3.5	93.2±2.2	79.2±4.3	84.2±3.3	91.6±1.9	76.7±2.5	81.3±2.2	89.8±1.9	75.3±2.6	80.8±2.6
	LIMITR	94.0±2.3	82.0±3.5	85.6±3.5	94.3±2.3	82.4±3.6	85.5±3.4	95.0±2.1	84.0±3.4	87.0±3.3	95.3±1.9	84.1±3.4	85.0±3.4	93.6±1.6	81.5±2.2	84.2±1.9	90.0±1.6	79.2±2.2	80.5±2.7
	MedFILIP	92.6±2.5	73.5±4.8	82.5±3.6	95.2±2.5	79.8±4.7	86.1±3.4	95.4±2.4	80.3±4.4	86.6±3.0	95.8±2.2	81.0±3.8	87.4±2.7	92.2±1.7	76.0±3.3	82.8±2.1	90.9±1.7	77.2±2.6	80.2±2.6
	MedKLIP	94.7±2.1	83.3±3.3	86.2±2.8	95.0±1.9	84.1±3.3	86.5±2.5	94.7±2.0	85.9±2.9	88.0±2.2	96.2±1.8	87.7±2.3	91.4±1.5	94.8±1.6	82.7±2.2	85.7±1.9	92.2±1.5	80.2±2.3	84.6±1.9
	MedTrim	96.2±1.8	82.9±3.2	89.2±2.1	96.2±1.9	85.5±2.7	89.5±2.2	96.5±1.8	88.3±2.3	90.2±2.0	97.5±1.6	90.7±2.0	93.3±1.9	95.2±1.3	83.3±1.9	88.5±1.5	92.4±1.6	80.8±2.1	86.7±2.1
T2I	ConVIRT	91.6±2.6	81.3±3.3	81.0±3.4	92.0±2.6	81.5±3.3	81.7±3.5	91.7±2.8	82.0±3.0	82.3±3.1	92.2±2.6	82.3±3.2	83.5±3.3	90.3±1.7	78.8±2.6	81.0±1.9	87.5±1.9	77.3±3.0	78.3±2.7
	JoImTeRNet	88.3±2.9	63.0±5.5	75.0±5.1	88.0±2.6	63.5±5.2	76.0±4.8	87.4±2.7	62.0±5.3	74.0±4.7	87.6±2.7	63.0±5.2	77.2±4.6	87.5±2.1	64.2±2.8	73.1±2.8	79.9±2.2	57.4±5.6	67.1±3.6
	GLORIA	92.0±2.8	78.0±4.7	85.0±4.1	92.0±2.8	78.5±4.8	85.0±4.0	92.3±2.6	78.0±4.9	85.0±3.8	92.0±2.6	78.0±4.6	85.0±3.5	90.3±2.1	77.7±2.2	81.3±1.8	85.2±1.4	72.5±2.5	75.9±2.3
	CheXzero	93.5±2.3	80.0±4.0	86.0±3.2	93.5±2.1	80.5±3.8	86.0±3.0	93.8±2.1	80.3±3.5	86.0±3.1	93.5±1.9	80.2±3.4	86.0±3.0	91.8±2.0	77.5±2.2	83.2±2.0	90.0±1.4	76.9±2.4	81.7±2.3
	LIMITR	94.0±2.1	80.0±3.4	86.1±3.1	94.3±2.2	82.0±3.0	85.5±3.0	94.9±1.9	83.3±2.8	87.0±2.5	95.2±1.9	84.8±2.6	88.5±2.4	94.0±1.9	79.7±2.6	84.3±2.0	89.6±1.7	79.0±2.3	83.1±2.5
	MedFILIP	92.8±2.4	69.0±4.7	79.8±3.8	93.1±2.3	70.3±4.5	81.0±3.6	93.2±2.5	72.3±4.3	82.0±3.2	93.6±2.3	73.4±4.0	84.2±3.2	92.5±1.6	76.2±2.7	82.4±2.2	91.5±1.6	72.1±3.3	78.3±2.9
	MedKLIP	94.5±1.9	81.2±2.8	86.6±2.3	94.9±2.0	83.2±2.7	88.0±2.2	95.3±1.9	84.5±2.2	89.3±2.1	95.0±2.0	85.1±2.3	90.2±1.8	93.4±1.6	80.9±2.4	86.3±1.9	91.7±1.6	81.5±2.3	83.2±2.1
	MedTrim	95.9±2.0	80.3±3.0	87.3±2.4	96.1±1.6	83.8±2.6	89.4±2.1	96.3±1.6	85.4±2.4	90.0±2.4	96.8±1.4	86.0±2.5	91.7±2.1	94.5±1.4	82.5±2.1	87.6±1.6	92.0±1.7	80.9±2.5	85.5±1.8
T2T	ConVIRT	92.2±2.4	80.3±3.1	80.5±3.2	92.8±2.3	80.9±3.0	82.2±3.0	93.0±2.3	81.4±3.2	83.1±3.1	93.0±2.2	82.0±3.0	83.5±3.1	90.3±1.6	77.6±1.9	80.6±2.1	89.3±1.6	77.5±2.9	78.6±3.0
	JoImTeRNet	91.0±2.5	64.3±5.3	73.0±4.8	91.2±2.8	63.0±5.1	74.6±4.6	91.0±2.6	63.5±5.4	74.0±4.8	90.7±2.5	63.0±5.3	73.8±4.6	87.7±1.9	63.7±3.0	73.0±3.3	81.5±2.1	57.4±5.3	68.0±3.0
	GLORIA	92.0±2.5	78.0±4.8	84.0±3.6	92.0±2.7	78.5±4.8	84.5±3.5	92.3±2.6	78.3±4.6	84.7±3.1	92.0±3.0	78.2±4.5	84.9±2.9	90.7±1.8	78.2±2.0	82.2±1.8	86.2±1.7	73.0±2.3	78.3±2.2
	CheXzero	93.0±2.2	79.0±4.3	85.0±3.6	93.0±2.5	79.5±4.1	85.0±2.3	93.5±1.9	79.3±4.1	85.0±2.2	93.0±2.0	79.2±4.1	85.0±2.4	92.1±1.7	77.6±2.1	84.0±2.1	91.0±1.4	74.2±2.9	82.0±2.4
	LIMITR	93.7±2.4	81.5±3.2	85.9±3.3	94.1±2.8	82.0±3.8	87.0±2.4	94.4±2.0	83.0±3.6	87.6±3.4	95.1±1.6	84.0±3.1	88.0±3.3	91.6±1.3	81.8±2.4	85.1±2.1	90.8±1.6	78.3±2.6	83.6±2.6
	MedFILIP	94.0±1.6	82.4±3.2	86.2±2.9	95.2±1.6	83.4±3.2	87.0±2.9	95.6±2.0	84.3±2.8	87.9±2.6	96.5±1.8	85.6±2.3	90.4±3.0	94.7±1.7	79.5±2.8	87.3±1.8	92.1±1.5	80.1±2.4	84.8±2.3
	MedKLIP	94.0±1.8	82.2±2.9	85.8±2.3	94.6±1.7	83.0±2.7	86.9±2.3	95.3±1.8	84.1±2.7	87.2±2.2	95.9±1.7	84.0±2.5	88.1±2.9	92.0±1.8	82.7±2.0	86.5±1.6	91.4±1.5	82.1±2.2	83.1±1.9
	MedTrim	94.3±1.8	83.1±3.0	85.6±2.0	94.5±1.4	83.6±2.1	87.4±2.1	95.8±1.3	84.7±3.1	88.8±2.2	97.0±1.5	86.7±3.1	91.1±2.7	93.2±1.6	83.6±1.8	87.9±1.4	92.4±1.6	82.6±2.0	83.5±2.2

directional descriptors by 1.6%. Including both simultaneously when disease guidance is present results in improvements of 2.5% for adjectival qualifiers and 2.6% for directional descriptors. Supp. Table 2 lists average performance of variants for disease classification tasks. The full MedTrim model yields the strongest overall performance averaged across datasets, with improvements of 1.0% ACC, 0.8% F1, and 0.6% AUC relative to the average of the ablated variants, while intermediate variants that add adjectival or directional guidance alone also consistently improve over disease-only supervision. These results suggest that incorporating meta-entity guidance into triplet learning elicits attribute-aware alignment that in turn improves downstream task performance.

Multimodal triplet objective: To assess the impact of the multimodal triplet alignment objective, we trained variant models that selectively ablated different objective components, including the base triplet-type loss (f_{tri}), within-modal loss terms ($\mathcal{L}_{I2I+T2T}$) and cross-modal loss terms ($\mathcal{L}_{I2T}, \mathcal{L}_{T2I}$). Variants ablating triplet-type loss replaced f_{tri} with a contrastive-type loss between anchor-positive and anchor-negative pairs. Models were evaluated for multimodal retrieval tasks, and precision was measured separately for disease class, adjectival qualifiers and directional descriptors. Table II lists average retrieval precision across meta-entities, separately for I2I, I2T, T2I, T2T tasks. In general, triplet-type variants outperform contrastive-type variants when the two types are implemented with matching sets of within- and cross-modal loss terms. On average, triplet variants offer 6.2% higher precision than contrastive variants. Among triplet

variants, inclusion of either $\mathcal{L}_{I2T}$ or $\mathcal{L}_{T2I}$ enhances precision over the variant that only uses within-modal terms, with an average 8.1% improvement. MedTrim using all loss terms achieves the highest scores, with 10.8% improvement over the variant containing only within-modal terms and 2.9% improvement over the variant containing only cross-modal terms. Meanwhile, Supp. Table 3 lists performance of variants for disease classification tasks. We find that introducing each objective term leads to consistent performance improvements, indicating that both within- and cross-modal objectives contribute to representation learning. Averaged across datasets, the full MedTrim objective yields the highest performance, improving over the ablated configurations by 3.8% ACC, 3.7% F1, and 3.8% AUC on average. These results suggest that the different loss components play complementary roles, and that their joint use provides the most reliable performance gains.

Class-distribution of retrieval performance: A fundamental challenge in medical image-text retrieval is addressing class imbalance, where certain diseases are disproportionately represented compared to others. Notably, atelectasis and pleural effusion are among the most frequently occurring conditions, whereas others, such as enlarged cardiomeastinum and fracture, can be relatively rare. In learning-based retrieval, this imbalance can bias retrieved items towards either very common or rare disease classes, compromising precision for retrieval of remaining classes. Since consistent precision across disease classes is a crucial aspect of effective retrieval systems, we examined the distribution of retrieval performance across individual disease classes. Supp. Fig. 1 depicts MedTrim’s

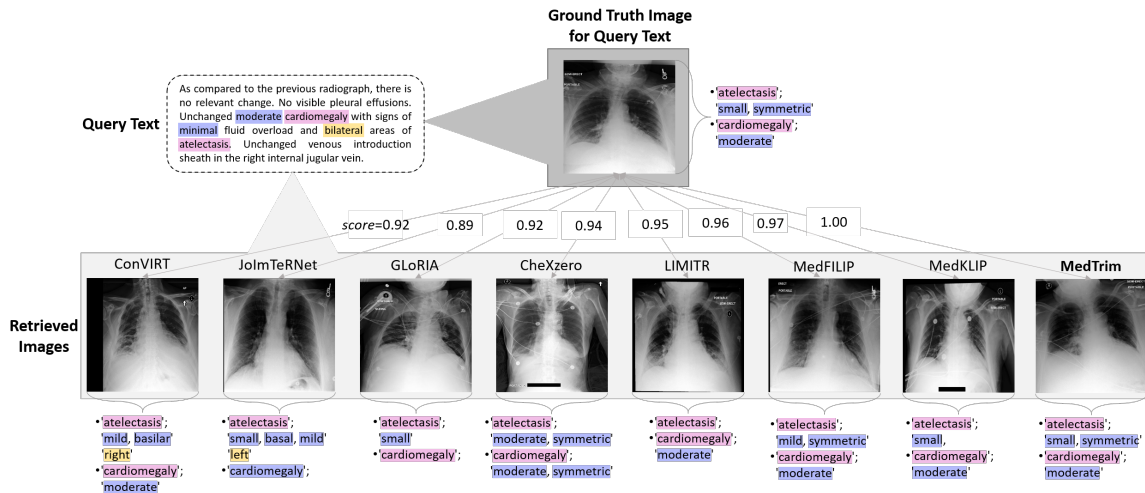


Fig. 3: Representative results for T2I retrieval. Given a query CXR report, the top-ranked images retrieved by competing methods are displayed, along with their similarity scores to the ground-truth CXR image. For each method, meta-entities extracted from the reports of the retrieved images are highlighted (magenta: disease class, purple: adjectival descriptor, yellow: directional descriptor).

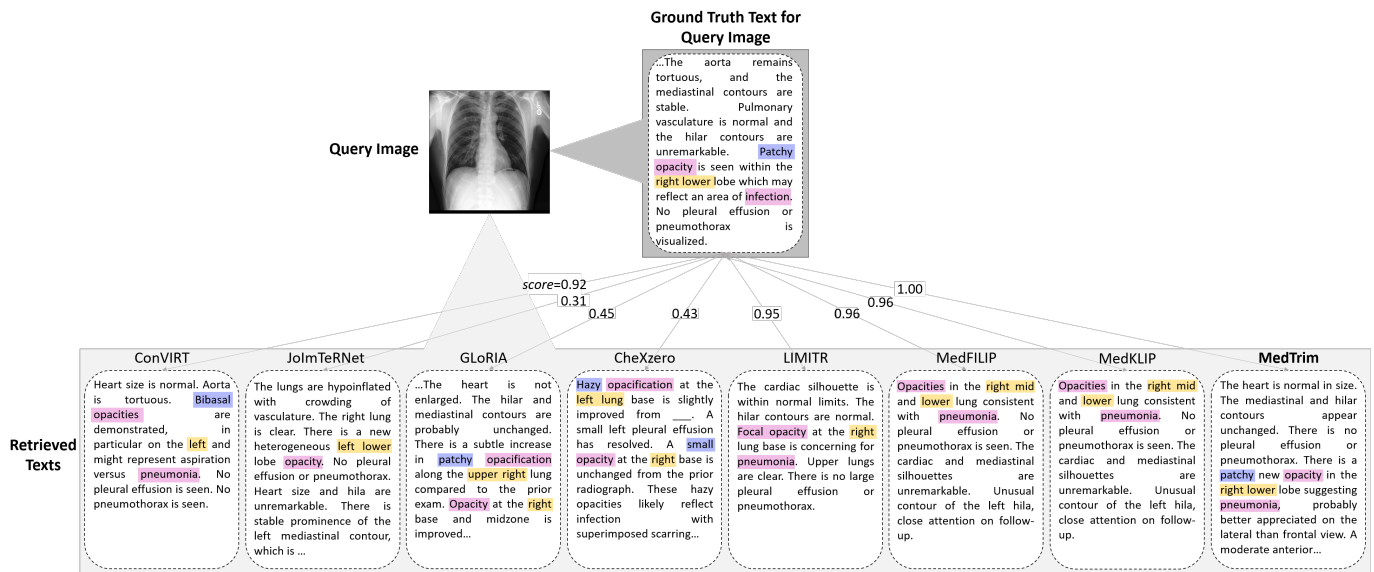


Fig. 4: Representative results for I2T retrieval. Given a query CXR image, the top-ranked reports retrieved by competing methods are displayed, along with their similarity scores to the ground-truth CXR report.

retrieval precision across disease classes for I2T and T2I tasks, against a contrastive variant that used the same objective albeit ablated f_{tri} , and a triplet variant that only used cross-modal loss terms. While all models show moderate variations, MedTrim attains consistently high precision across disease classes, and superior precision against variant models in moderately frequent disease classes such as edema, pneumonia, consolidation or lung lesion. These results highlight the importance of MedTrim’s multimodal triplet objective in achieving more equitable retrieval performance under class imbalances.

Semi-hard negative sampling: A critical factor that influences the ability of triplet mining to learn discriminative representations is the selection of negative samples. To improve learning of intra-class variations, MedTrim employs semi-hard negative samples by constraining the similarity scores between anchor and negative samples to a moderate range. This constraint serves to ensure that selected negatives share some overlap with the anchor in terms of disease classes (e.g., at least one shared disease) while differing in adjectival or directional attributes. We conducted two ablation studies

to examine the importance of semi-hard negative sampling in MedTrim. First, we built variant models trained on a mixture of semi-hard negative samples and easy negative samples. Supp. Fig. 2b displays precision in multimodal retrieval tasks for varying percentages of semi-hard samples in the training set. We find that increasing the proportion of semi-hard negatives significantly improves performance in both within-modal and cross-modal retrieval tasks. Second, we build variant models trained using negative samples with similarity scores constrained to different sets of $\tau_{min,max}$, including a variant trained on purely easy negatives ($\tau_{min,max} = (0.00, 0.08)$), and a variant trained on purely hard negatives ($\tau_{min,max} = (0.78, 1.00)$). Similarity levels for these variants were defined empirically, where easy negatives were selected from the least similar tail of τ distribution, and hard negatives were sampled from the top-10% similarity percentile to avoid the training instability observed when selecting the closest negatives. Supp. Table 4 lists average retrieval precision for variants. We observe that the proposed score range attains optimal performance among the examined ranges. We also

observe that retrieval precision remains relatively stable under moderate deviations from the proposed range, suggesting that the method shows a degree of robustness to changes in the hyperparameter values. Collectively, these results underscore the importance of using challenging but informative negatives for enhanced triplet learning.

Model capacity and optimization budget: Next, we examined how encoder capacity influences performance by varying the size of the transformer-based image/text backbones [65]. Retrieval precision for varying backbone sizes is listed in Supp. Table 5, along with computational efficiency metrics. Increasing the backbone from Tiny to Base yields consistent performance improvements, whereas scaling further to Large leads to reduced performance despite higher computational cost. This pattern suggests that, given the dataset sizes used here, the Base configuration provides a balanced trade-off between performance and efficiency. We also evaluated the effect of the optimization budget by varying the number of mined triplets used to train the alignment objective, which plays a crucial role in shaping the quality of the learned embedding space. MedTrim was trained using 0.1 to 2 million unique triplets derived from the training set. As shown in Supp. Fig. 2a, Following an initial inclination towards 0.25 million triplets, diminishing returns are observed in performance for further increases in triplet count, consistently across different retrieval tasks. Note that training time elevates from just under 20 minutes for 0.25 million triplets to over 2 hours for 2 million triplets. This substantial rise in computational cost, combined with only marginal performance gains, indicates that a moderate budget around 0.25 million triplets provides an effective operating point for MedTrim. Together with the backbone-size analysis, these results highlight that MedTrim achieves strong performance without requiring excessive model capacity or large-scale triplet enumeration.

Other hyperparameters: We conducted ablation studies to corroborate the cross-validated selection of the weighting parameter for cross-modal versus within-modal alignment objectives in the loss function (η in Eq. 18), and weighting parameters for disease class, adjectival qualifier, and directional descriptor entities in the score function ($\gamma_{1,2,3}$ in Eq. 1). Precision performance in retrieval tasks is listed for varying η values in Supp. Table 6, and for varying $\gamma_{1,2,3}$ values in Supp. Table 7. The results indicate that the selected hyperparameters in MedTrim yield near-optimal performance.

Ontology extension: Lastly, we implemented an extended variant of MedTrim that incorporated disease course (acute vs. chronic) as an additional meta-entity. As shown in Supp. Table 8, this variant is able to recover disease course information with reasonable precision. However, we observe a modest reduction in precision for adjectival and directional attributes, reflecting the trade-off introduced when the score function integrates over a larger set of entities. Overall, these results suggest that ontology extensions are feasible within MedTrim, while their practical benefit may depend on how reliably and consistently the added attributes are expressed in the data.

B. Comparison Studies

1) Entity Extraction: Because MedTrim relies on structured pathology-related entities to guide its alignment objective, we compared the proposed OBER module with representative entity extractors from the literature, including the rule-based baselines NegBio [68] and CheXpert-Labeler [69], as well as the learned RadGraph-XL model [70]. Agreement between the entity tuples predicted by each method and radiologist-annotated reference tuples was quantified using an intersection-over-union (IoU) metric. The performance of competing extractors is summarized in Table III. OBER achieves higher extraction accuracy than baselines ($p < 0.05$), with particularly notable gains for adjectival and directional attributes that existing methods either do not capture or capture less precisely. We also evaluated entity extraction under more challenging conditions, including reports with terse descriptions (defined as the lowest 5% by character length) and reports in which components of entity tuples are distributed across multiple sentences (Table IV). In these settings, OBER remains competitive with baseline extractors ($p < 0.05$), indicating that its extraction performance does not degrade disproportionately under terse or structurally distributed descriptions. Overall, these results indicate that OBER provides reliable meta-entity extraction for MedTrim across both typical and less structured radiology reports.

2) Multimodal Retrieval: Next, we demonstrated MedTrim via comparisons against state-of-the-art alignment methods for med-VLMs. In particular, MedTrim was compared against global baselines (ConVIRT, and CheXzero), global-local baselines (JoImTeRNet, and GLoRIA), and knowledge-guided baselines (LIMITR, MedFILIP, and MedKLIP), which are all based on contrastive learning. Methods were evaluated on the MIMIC-CXR dataset to perform multimodal retrieval tasks. Precision, recall and mAP values are listed in Table V, separately evaluating metrics in terms of disease-class, adjectival-qualifier, and directional-descriptor entities. MedTrim achieves the highest retrieval performance metrics in all examined settings ($p < 0.05$), except for MedKLIP that yields higher adjectival P@1 in I2T and T2I, disease recall in I2I and adjectival mAP in T2I, and MedFILIP that yields higher directional P@1, disease P@10, disease recall and directional mAP in T2T. On average, MedTrim offers (Precision, Recall, mAP) improvements of (5.7%, 6.4%, 6.6%) over global baselines, (10.7%, 13.8%, 18.1%) over global/local baselines, and (2.6%, 2.9%, 3.1%) over knowledge-guided baselines. Representative retrieval results from competing methods are depicted in Fig. 3 for T2I and in Fig. 4 for I2T tasks. Compared to baselines, we observe that MedTrim retrieves items with higher similarity scores to the sought ground-truth information. In T2I, MedTrim retrieves an image whose report contains more closely matched meta-entities to the query report. In I2T, MedTrim retrieves a report whose meta-entities are more closely matched to the ground-truth report of the query image. These results demonstrate the superior ability of MedTrim to align and retrieve diagnostically-relevant cross-modal information in CXR data.

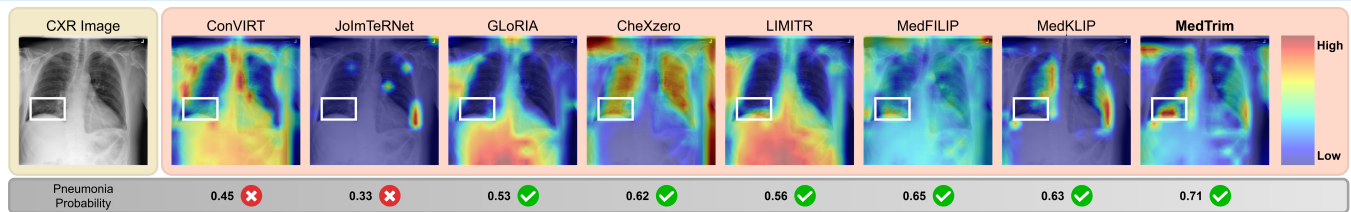


Fig. 5: Representative results for zero-shot disease classification. For an input CXR image, Grad-CAM heatmaps (see colorbar) highlighting regions with high contribution to the classification decision are displayed for competing methods, along with confidence scores (i.e., probability of pneumonia). White boxes denote the radiologist-annotated region-of-interest containing pathology in the CXR image.

TABLE VI: Zero-shot disease classification. Accuracy (ACC), area-under-the-curve (AUC) and F1 metrics are listed on CheXpert (5 classes, 20% chance level) and RSNA (2 classes, with 20% positive-class and 80% negative-class chance levels) datasets.

	CheXpert			RSNA		
	ACC	F1	AUC	ACC	F1	AUC
ConVIRT	36.4±2.5	36.7±2.4	64.1±1.4	71.9±1.6	68.6±1.2	55.8±3.1
JoImTeRNet	28.9±1.5	29.4±2.6	59.8±2.5	69.4±1.8	66.3±1.4	51.4±3.4
GLORIA	38.6±2.2	39.4±2.4	66.1±1.4	74.1±1.3	71.2±1.2	59.2±2.8
CheXzero	40.0±1.9	40.6±2.1	66.3±1.3	78.0±1.1	76.9±1.0	66.4±1.9
LIMITR	42.0±1.7	42.5±2.1	67.2±1.2	76.3±1.4	74.2±1.3	61.9±2.2
MedFILIP	41.8±2.1	42.4±2.2	66.7±1.1	77.8±1.5	76.6±1.3	62.3±1.8
MedKLIP	43.2±1.9	43.9±2.1	68.9±1.0	78.5±1.3	76.9±1.0	67.5±1.9
MedTrim	45.1±1.7	46.9±2.1	70.0±1.1	79.8±1.2	77.1±0.9	68.3±1.9

TABLE VII: Few-shot disease classification. Models were fine-tuned on 10% of labeled data and tested on CheXpert (5 classes) and RSNA (2 classes).

	CheXpert			RSNA		
	ACC	F1	AUC	ACC	F1	AUC
ConVIRT	58.4±2.2	59.7±2.4	85.6±0.9	75.6±1.5	73.8±1.5	80.3±1.3
JoImTeRNet	56.7±2.9	55.6±3.1	80.3±1.5	72.7±1.8	70.3±2.1	77.5±2.1
GLORIA	59.1±2.3	60.4±2.3	86.1±1.0	80.8±0.9	80.0±0.8	86.7±1.1
CheXzero	59.1±2.1	60.0±2.5	85.5±1.1	80.3±1.1	79.6±1.2	84.9±1.2
LIMITR	58.8±2.3	59.1±2.7	85.9±1.3	79.5±1.2	77.2±1.5	83.2±2.0
MedFILIP	61.0±2.0	62.5±1.8	86.4±1.0	81.5±0.8	80.3±1.0	86.5±0.9
MedKLIP	62.3±1.8	62.9±1.7	88.8±0.7	82.1±0.6	81.8±0.7	88.0±0.8
MedTrim	62.8±1.9	63.1±1.8	89.9±0.7	82.8±0.8	82.0±0.9	89.7±0.5

3) Disease Classification: We then demonstrated MedTrim on zero-shot and few-shot disease classification tasks. In zero-shot settings, disease prediction was performed by matching image embeddings to text embeddings of disease prompts via cosine similarity, without any classifier training. In few-shot settings, we adopted the linear probing protocol commonly used in prior alignment work: the image encoder was frozen and a linear classifier was trained on top of the fixed embeddings using the available data.

In zero-shot settings, models trained on the MIMIC-CXR dataset were directly evaluated on the CheXpert and RSNA Pneumonia datasets. Accuracy, AUC and F1 values for competing methods are listed in Table VI. MedTrim achieves the highest performance metrics among competing methods ($p < 0.05$). On average, MedTrim offers (ACC, AUC, F1) improvements of (5.9%, 6.3%, 6.0%) over global baselines, (9.7%, 10.4%, 10.0%) over global/local baselines, and (2.5%, 2.6%, 3.4%) over knowledge-guided baselines.

In few-shot settings, models initially trained on the MIMIC-CXR dataset were fine-tuned on the training splits of CheXpert and RSNA Pneumonia using only 10% of labeled data. This setup simulates realistic scenarios where only sparse annotations are available in target domains. As listed in Table VII, MedTrim again achieves the highest performance among competing methods ($p < 0.05$). On average, MedTrim offers (ACC, AUC, F1) improvements of (4.5%, 4.3%, 5.7%) over global

baselines, (5.5%, 6.0%, 7.2%) over global/local baselines, and (1.9%, 1.9%, 3.3%) over knowledge-guided baselines.

Representative classification results from competing methods, with Grad-CAM activation maps overlaid onto CXR images [71], are depicted in Fig. 5. We observe that the activation maps for MedTrim more closely focus on diagnostically-relevant regions of the CXR image containing pathology, whereas the activation maps are more diffusely distributed across the field-of-view for baselines. Overall, these results demonstrate the promising ability of MedTrim to detect disease-related features in CXR images, with little to no dataset-specific tuning.

4) Report Generation: Lastly, MedTrim was evaluated on a report generation task that leverages the shared image-text embedding space to generate reports for unseen CXR images [72]. For this purpose, a given CXR image was mapped into the shared embedding space via the image encoder, and three most closely aligned reports from the corpus (excluding the paired report for the image) were identified. A new report was then generated by synthesizing information from the identified reports using a GPT-based language model [73] (see prompt in Supp. Materials). The quality of the generated reports was evaluated with respect to the ground-truth reports using cosine similarity between GPT-encoded report-level text embeddings, and the GREEN metric that assesses clinical correctness of radiology reports [74]. As summarized in Table VIII, MedTrim achieves the highest performance among competing methods ($p < 0.05$), offering (Cosine sim., GREEN) improvements of (11.5%, 13.0%) over global baselines, (16.5%, 15.5%) over global/local baselines, and (7.7%, 6.7%) over knowledge-guided baselines. These findings indicate that the enhanced fidelity of MedTrim’s multimodal representations supports more accurate report generation from CXR images.

C. Class-Level Performance

To better understand model behavior across disease categories, class-level performance metrics were examined for multimodal retrieval (Supp. Fig. 3) and disease classification (Supp. Fig. 4). In retrieval, MedTrim achieves superior performance over baselines for low-prevalence classes ($\approx 2\%$ or less), including Enlarged Cardiomediastinum, Fracture, Lung Lesion, and Pleural Other ($p < 0.05$), with the exception that MedKLIP performs comparably for Lung Lesion and Pleural Other. Comparable gains are observed for most remaining categories; yet, knowledge-guided baselines achieve moderately higher precision for Consolidation and particularly Lung Opacity. In disease classification, MedTrim yields significant improvements over baselines for pneumonia detection on RSNA ($p < 0.05$). On CheXpert, it attains superior performance for Atelectasis, Cardiomegaly, Edema, and Pleural Effusion

TABLE VIII: Report generation. Cosine similarity and GREEN metrics are listed on MIMIC-CXR.

Method	Cosine Sim.	GREEN
ConVIRT	0.64±0.04	0.51±0.05
JoImTeRNet	0.58±0.09	0.46±0.06
GLoRIA	0.65±0.07	0.57±0.06
CheXzero	0.69±0.06	0.57±0.05
LIMITR	0.68±0.06	0.58±0.05
MedFILIP	0.72±0.04	0.61±0.03
MedKLIP	0.71±0.05	0.62±0.04
MedTrim	0.78±0.03	0.67±0.03

TABLE IX: Computational efficiency. Training time per epoch (min), inference time per sample (msec), and memory use (GB) are listed for models executed on an Nvidia RTX 4090 GPU.

	Train. time	Inf. time	Memory use
ConVIRT	20.1	13.1	4.2
JoImTeRNet	35.7	14.1	4.3
GLoRIA	31.6	12.7	6.7
CheXzero	30.3	12.3	6.9
LIMITR	22.7	12.7	4.1
MedFILIP	31.2	13.4	6.1
MedKLIP	33.5	13.6	6.2
MedTrim	35.1	14.2	6.5

($p < 0.05$), while MedKLIP and GloRIA perform comparably for Consolidation. These results suggest that MedTrim is particularly effective for low-prevalence and semantically-specific disease categories, whereas for classes with broad or overlapping definitions (e.g., Consolidation), explicit refinement via meta-entity guidance may provide more limited benefit.

D. Computational Efficiency

To assess computational efficiency, training time per epoch, inference time per sample, and GPU memory use were measured (Table IX). Inference time remains consistent across all approaches, with a small variance of ≈ 2 ms across models. While MedTrim’s training time and memory load (35.1 min/epoch; 6.5 GB) are higher than simpler baselines like ConVIRT, they remain within the typical range of recent methods such as JoImTeRNet (35.7 min) and CheXzero (6.9 GB). These results indicate that MedTrim delivers significantly stronger task performance while maintaining a computational footprint comparable to current state-of-the-art architectures.

VI. DISCUSSION

Our demonstrations on MedTrim highlight the potential of triplet learning for multimodal image-text alignment in med-VLMs, but several technical limitations remain, offering avenues for future work. A first avenue concerns the nature of multimodal data available for model training. The current implementation of MedTrim assumes access to a training dataset where CXR images and respective radiology reports are available for each subject. While such datasets exist publicly [57], [62], single-modality datasets containing only CXR images or only radiology reports are relatively more common [63]. To seek performance benefits by enabling training on broader datasets, triplet learning can be augmented with cycle-consistency losses [75]. For this purpose, a translation module

could be used that bidirectionally maps between images and text in the latent embedding space to permit capture of meaningful cross-modal representations.

A key component of MedTrim is extraction of pathology attributes as meta-entities from CXR reports, which guides triplet mining by selecting informative positive and semi-hard negative samples. In the current implementation, these attributes are extracted from radiology reports and used to align entire images and their associated text. This can be seen as a hybrid approach, enforcing global alignment for images while considering a common ground between global and local alignment for text [48]. Future work could explore enhanced localization by integrating region-identification or patch-based techniques on medical images [54], which have been shown to improve spatial precision in multimodal learning.

The meta-entities in the current implementation were based on adjectival qualifiers and directional descriptors, as these are among the most commonly used attributes in radiology reports for describing CXR findings. Additional structured information could also be incorporated into OBER when clinically meaningful and sufficiently supported by the data. Extending the ontology would involve several practical steps: collecting clinically relevant, domain-specific terms from sources such as radiology guidelines or medical dictionaries; validating their relevance and interpretability with expert radiologists; and consolidating synonymous expressions to account for variability in reporting. Once extended entity tuples are defined and annotated, their integration into MedTrim could follow the existing alignment framework, with similarities derived from additional meta-entities incorporated into the score function. The relative weighting of entity-level components could then be tuned on a validation set.

For meta-entity extraction, OBER operates at the sentence level, partitioning complex sentences at discourse markers (e.g., “however”, “but”, “which”) to identify clauses that qualify or negate a finding. This prevents qualifiers from being incorrectly propagated across unrelated clauses while preserving the local linkage between a finding and its related entities. A potential limitation of sentence-based processing is that elements of a meta-entity tuple (disease, direction, adjective) could be spread across multiple sentences. In the reports analyzed here, descriptions extended across multiple sentences in 11.2% of cases; however, in the majority of these reports the disease label was explicitly repeated, allowing OBER to correctly associate related entities. In a smaller subset of reports (3.0%), the disease was mentioned once and subsequent sentences provided additional qualifiers without repetition, in which case OBER correctly identified the disease but left the remaining entities unspecified. While this behavior is not ideal, it did not lead to incorrect entity assignments. Moreover, comparisons with a learning-based extractor that performs report-level processing did not reveal performance loss due to sentence-level processing. Extending OBER with sliding-window aggregation across neighboring sentences or co-reference checks represents a natural direction for future work to better handle more sparsely distributed descriptions. Note that because OBER relies on deterministic, ontology-guided extraction, it may be less susceptible to hallucinated

entities than learning-based extractors. A systematic, entity-level analysis of this behavior might also be explored in future work using a dedicated experimental design.

Additional practical challenges include semantic ambiguity in medical terminology and variability in report quality. For example, the term “opacity” may refer to diffuse air-space or interstitial changes, or to focal nodular findings suggestive of a lesion. In this work, OBER follows the MIMIC-CXR disease taxonomy, which distinguishes these cases by mapping diffuse or non-nodular opacities to the lung opacity class and nodular opacities to the lung lesion class [62]. By grounding extraction in this standardized taxonomy, OBER reduces reliance on free-text semantics and provides a consistent basis for interpreting such terms. OBER does not enforce a single disease label per report; when multiple cues are present, multiple disease-level entities may be extracted, avoiding the need for explicit disambiguation at extraction time. While report quality can vary, in the datasets analyzed here we did not observe non-standard abbreviations beyond the common variants already covered by the CheXpert lexicon used by OBER [57], nor incomplete sentence structures that interfered with entity extraction. Our evaluations further indicate competitive extraction performance in terse or structurally distributed reports, suggesting that grounding extraction in a structured taxonomy while allowing co-occurring findings to be represented provides a degree of robustness in practice. At the same time, more systematic analyses under controlled variations in semantic ambiguity and report quality remain an important direction for future work.

Implementation of MedTrim in the current study, similar to most VLM alignment methods, relied on transformer encoders to derive image and text embeddings. This choice reflects a broadly adopted balance in the literature between model complexity and representational power. From a practical standpoint, MedTrim is computationally comparable to existing transformer-based med-VLMs and remains within the operational limits of common research and clinical settings. Notably, all experiments were conducted on a single consumer-grade GPU. Furthermore, the additional overhead of triplet mining is restricted to the training phase; during deployment, inference requires only a single forward pass through the encoders, maintaining similar viability to simpler architectures. Nevertheless, the modular nature of our method permits adaptation to more efficient encoders when necessary, such as state-space models [76], which could help enhance model scalability.

In this study, we focused on zero-shot deployment of aligned med-VLMs for downstream retrieval and classification tasks, where MedTrim demonstrated stronger generalization across tasks and datasets than competing methods. Prior work suggests that generalization to new datasets, particularly those containing previously unseen disease classes or attributes, may further benefit from few-shot adaptation using a small number of target-domain samples [35], [43], [65]. In this context, explicitly computing semantic similarities between disease-descriptive entities and incorporating them into the triplet learning objective could provide additional guidance [10], [41], potentially improving robustness for novel disease cases.

Beyond retrieval, classification, and generation, other down-

stream tasks may benefit from the improved multimodal alignment achieved by MedTrim. In particular, task-specific models trained on the aligned embeddings could support applications such as segmentation, severity grading, and report summarization [5], [9], [52]. Given the central role of multimodal interpretation in imaging-based clinical workflows, the framework could also be extended to other modalities, including MRI or CT, to enable broader applications in medical imaging.

VII. CONCLUSION

Representational alignment between medical images and radiology reports is key to building performant medical vision-language models (med-VLMs). Here, we introduced a novel entity-aware triplet learning technique, MedTrim, to enable explicit guidance from pathology attributes in refining sample selection, so as to derive image-text embeddings that capture fine-grained intra-class variability, while reliably segregating different classes. Our experiments indicate that this approach elicits superior retrieval and classification performance on CXR data over state-of-the-art alignment methods. Therefore, MedTrim holds great promise for training reliable med-VLMs with precise image-text alignment and thereby improved sensitivity in downstream tasks.

REFERENCES

- [1] A. Chaudhari, “Vision-language models for radiology AI,” in *Emerging Top Artif Intell*, vol. 13118. SPIE, 2024, p. 1311806.
- [2] G. Dawidowicz *et al.*, “Limit: Leveraging local information for medical image-text representation,” in *ICCV*, 2023, pp. 21 108–21 116.
- [3] C. Shu *et al.*, “Miter: Medical image-text joint adaptive pretraining with multi-level contrastive learning,” *Exp Syst App*, vol. 238, p. 121526, 2024.
- [4] Z. Wang *et al.*, “Automated radiographic report generation purely on transformer: A multicriteria supervised approach,” *IEEE Trans Med Imaging*, vol. 41, no. 10, pp. 2803–2813, 2022.
- [5] X. Cai *et al.*, “Chestxraybert: A pretrained language model for chest radiology report summarization,” *IEEE Trans Multimed*, vol. 25, pp. 845–855, 2023.
- [6] B. P. Veasey *et al.*, “Low-rank adaptation of pre-trained large vision models for improved lung nodule malignancy classification,” *IEEE Open J Eng Med Biol*, vol. 6, pp. 296–304, 2025.
- [7] S. Bannur *et al.*, “Learning to exploit temporal structure for biomedical vision-language processing,” in *CVPR*, 2023, pp. 15 016–15 027.
- [8] K. Vilouras *et al.*, “Zero-shot medical phrase grounding with off-the-shelf diffusion models,” *IEEE J Biomed Health Inf*, pp. 1–10, 2024.
- [9] Z. Yi *et al.*, “A survey on multimodal large language models in radiology for report generation and visual question answering,” *Information*, vol. 16, no. 2, 2025.
- [10] C. Tung *et al.*, “Exploring vision language pretraining with knowledge enhancement via large language model,” in *TAIH*, 2024, pp. 81–91.
- [11] D. Edirisinghe *et al.*, “Chest x-ray report generation using abnormality guided vision language model,” *IEEE Access*, 2025.
- [12] D. Gu *et al.*, “Anatomy-vlm: A fine-grained vision-language model for medical interpretation,” *arXiv:2511.08402*, 2025.
- [13] Y. Ibrahim *et al.*, “Enhancing medical vision-language models with rich textual descriptions and multiple alignments for chest x-ray diagnosis,” in *IEEE ICIP*, 2025, pp. 2079–2084.
- [14] S.-C. Huang *et al.*, “Multimodal foundation models for medical imaging - a systematic review and implementation guidelines,” *medRxiv* 2024.10.23.24316003, 2024.
- [15] P. Müller *et al.*, “Joint learning of localized representations from medical images and reports,” in *ECCV*, 2022, p. 685–701.
- [16] Z. Chen *et al.*, “Mapping medical image-text to a joint space via masked modeling,” *Med Image Anal*, vol. 91, p. 103018, 2024.
- [17] W. Huang *et al.*, “Enhancing Representation in Medical Vision-Language Foundation Models via Multi-Scale Information Extraction Techniques,” *arXiv:2401.01583*, 2024.

- [18] T. van Sonsbeek *et al.*, "X-TRA: Improving Chest X-ray Tasks with Cross-Modal Retrieval Augmentation," *arXiv:2302.11352*, 2023.
- [19] M. Li *et al.*, "Enhancing medical vision-language contrastive learning via inter-matching relation modelling," *arXiv:2401.10501*, 2024.
- [20] J. Wang *et al.*, "Cross-modal prototype driven network for radiology report generation," in *ECCV*, 2022, pp. 563–579.
- [21] G. Ding *et al.*, "Semantic extension for cross-modal retrieval of medical image-diagnosis report," in *NLPCC*, 2023, pp. 442–455.
- [22] B. Boecking *et al.*, "Making the most of text semantics to improve biomedical vision-language processing," in *ECCV*, 2022, pp. 1–21.
- [23] S. Wang *et al.*, "Fine-grained medical vision-language representation learning for radiology report generation," in *EMNLP*, 2023, pp. 15 949–15 956.
- [24] Y. Zhang *et al.*, "Contrastive learning of medical visual representations from paired images and text," *Proc Mach Learn Res*, vol. 182, pp. 1–24, 2022.
- [25] J. Lu *et al.*, "Deep metric learning for visual understanding: An overview of recent advances," *IEEE Signal Process Mag*, vol. 34, no. 6, pp. 76–84, 2017.
- [26] H. Xuan *et al.*, "Improved Embeddings with Easy Positive Triplet Mining," in *IEEE WACV*, 2020, pp. 2463–2471.
- [27] G. Kertész *et al.*, "A survey on novel applications of deep metric learning," in *IEEE INES*, 2025, pp. 371–376.
- [28] A. Hermans *et al.*, "In defense of the triplet loss for person re-identification," *arXiv:1703.07737*, 2017.
- [29] B. C. Kalpélibé *et al.*, "Vision language models in medicine," *arXiv:2503.01863*, 2025.
- [30] H. Lin *et al.*, "Taming vision-language models for medical image analysis: A comprehensive review," *arXiv:2506.18378*, 2025.
- [31] W. Fan *et al.*, "Medslip: Medical dual-stream language-image pre-training with pathology-anatomy semantic alignment," *arXiv:2403.10635*, 2025.
- [32] M. Li *et al.*, "Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation," in *CVPR*, 2023, pp. 3334–3343.
- [33] X. Wang *et al.*, "Self-supervised Image-text Pre-training With Mixed Data In Chest X-rays," *arXiv:2103.16022*, 2021.
- [34] F. Wang *et al.*, "Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning," in *NeurIPS*, 2022, pp. 1–14.
- [35] Z. Wan *et al.*, "Med-UniC: Unifying Cross-Lingual Medical Vision-Language Pre-Training by Diminishing Bias," *arXiv:2305.19894*, 2023.
- [36] Z. Chen *et al.*, "CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation," *arXiv:2401.12208*, 2024.
- [37] V. A. Nezhad *et al.*, "Generative autoregressive transformers for model-agnostic federated mri reconstruction," *arXiv:2502.04521*, 2025.
- [38] S. Lee *et al.*, "Cxr-llava: a multimodal large language model for interpreting chest x-ray images," *Eur Radiol*, pp. 1–13, 2025.
- [39] R. Tanno *et al.*, "Collaboration between clinicians and vision-language models in radiology report generation," *Nat Med*, pp. 1–10, 2024.
- [40] M. Monajatipoor *et al.*, "Berthop: An effective vision-and-language model for chest x-ray disease diagnosis," in *MICCAI*, 2022, pp. 725–734.
- [41] C. Wu *et al.*, "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis," in *ICCV*, 2023, pp. 21 315–21 326.
- [42] X. Zhang *et al.*, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nat Commun*, vol. 14, no. 1, p. 4542, 2023.
- [43] B. Liu *et al.*, "Improving medical vision-language contrastive pretraining with semantics-aware triage," *IEEE Trans Med Imaging*, vol. 42, no. 12, pp. 3579–3589, 2023.
- [44] S.-C. Huang *et al.*, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *ICCV*, 2021, pp. 3922–3931.
- [45] C. Liu *et al.*, "Utilizing Synthetic Data for Medical Vision-Language Pre-training: Bypassing the Need for Real Images," *arXiv:2310.07027*, 2023.
- [46] J. Liu *et al.*, "MLIP: Medical Language-Image Pre-training with Masked Local Representation Learning," *arXiv:2401.01591*, 2024.
- [47] S. Zhang *et al.*, "Visual prior-based cross-modal alignment network for radiology report generation," *Comput Biol Med*, vol. 166, p. 107522, 2023.
- [48] Y. Li *et al.*, "Unify, align and refine: Multi-level semantic alignment for radiology report generation," in *ICCV*, 2023, pp. 2851–2862.
- [49] J. Zhang *et al.*, "Vision-language models for vision tasks: A survey," *IEEE Trans Pattern Anal Mach Intell*, pp. 1–20, 2024.
- [50] Z. Wang *et al.*, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," *arXiv:2210.10163*, 2022.
- [51] Q. Li *et al.*, "Anatomical Structure-Guided Medical Vision-Language Pre-training," *arXiv:2403.09294*, 2024.
- [52] H. Yin *et al.*, "KIA: Knowledge-guided implicit vision-language alignment for chest X-ray report generation," in *ICCL*, 2025, pp. 4096–4108.
- [53] S. Ozturk *et al.*, "Content-based medical image retrieval with opponent class adaptive margin loss," *Inf Sci*, vol. 637, p. 118938, 2023.
- [54] Z. Ji *et al.*, "Improving joint learning of chest X-Ray and radiology report by word region alignment," *Mach Learn Med Imaging*, vol. 12966, pp. 110–119, 2021.
- [55] J. Chen *et al.*, "Detecting and Evaluating Medical Hallucinations in Large Vision Language Models," *arXiv:2406.10185*, 2024.
- [56] S. Mate *et al.*, "Ontology-based data integration between clinical and research systems," *PLOS ONE*, vol. 10, no. 1, pp. 1–20, 2015.
- [57] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI CAI*, vol. 33, no. 01, 2019, pp. 590–597.
- [58] O. Toporkov *et al.*, "On the role of morphological information for contextual lemmatization," *Comput Ling*, pp. 1–35, 2024.
- [59] J. Bertels *et al.*, "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice," in *MICCAI*, 2019, pp. 92–100.
- [60] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.
- [61] R. Wang *et al.*, "Spd manifold deep metric learning for image set classification," *IEEE Trans Neural Netw Learn Syst*, pp. 1–15, 2024.
- [62] A. E. Johnson *et al.*, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Sci Data*, vol. 6, no. 1, p. 317, 2019.
- [63] G. Shih *et al.*, "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia," *Rad Artif Intell*, vol. 1, no. 1, p. e180041, 2019.
- [64] E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," *arXiv:1904.03323*, 2019.
- [65] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, vol. 139, 2021, pp. 8748–8763.
- [66] E. Tiu *et al.*, "Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning," *Nat Bio Eng*, vol. 6, no. 12, pp. 1399–1406, 2022.
- [67] X. Liang *et al.*, "Medfilip: Medical fine-grained language-image pre-training," *IEEE J Biomed Health Inf*, 2025.
- [68] Y. Peng *et al.*, "Negbio: a high-performance tool for negation and uncertainty detection in radiology reports," *arXiv:1712.05898*, 2017.
- [69] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *arXiv:1901.07031*, 2019.
- [70] J.-B. Delbrouck *et al.*, "RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports," in *ACL*, 2024, pp. 12 902–12 915.
- [71] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.
- [72] Z. Chen *et al.*, "Generating radiology reports via memory-driven transformer," in *EMNLP*, 2020, pp. 1439–1449.
- [73] M. Ranjit *et al.*, "Retrieval Augmented Chest X-Ray Report Generation using OpenAI GPT models," in *PMLR*, vol. 219, 2023, pp. 650–666.
- [74] S. Ostmeier *et al.*, "GREEN: Generative radiology report evaluation and error notation," in *EMNLP*, 2024, pp. 374–390.
- [75] M. Ozbey *et al.*, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Trans Med Imaging*, vol. 42, no. 12, pp. 3524–3539, 2023.
- [76] B. Kabas *et al.*, "Physics-driven autoregressive state space models for medical image reconstruction," *arXiv:2412.09331*, 2024.