

ECT-CLIP: Resource-Efficient CT Vision-Language Pretraining via Negative Scaling

Hanwen Zhang¹, Yao Liu^{1*}, Jiaye Yang¹, Die Dai¹,
Qiao Liu¹, Yutong Xie², Peng Wang¹

¹University of Electronic Science and Technology of China, Chengdu, China

²Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
hanwenzhang1108@gmail.com, liuyao@uestc.edu.cn

Abstract

Vision-Language Pretraining (VLP) for 3D CT scans is severely constrained by the massive GPU memory required for high-resolution volumes and large contrastive batches. We propose **ECT-CLIP**, a resource-efficient CT VLP framework that achieves negative scaling—scaling up both the quantity and quality of negative samples—under limited computational resources. To scale the quantity, we exploit the memory asymmetry between CT and text encoding to decouple the image and text batch sizes, thereby enlarging the negative pool. To scale the quality, we design two complementary strategies: (1) deterministically separating false negatives from informative hard negatives via UMLS concept normalization, and (2) fusing multiple frozen medical text encoders to enrich semantic diversity without extra training cost. Trained on a single consumer-grade RTX 4090 GPU, ECT-CLIP achieves state-of-the-art zero-shot classification AUCs of 83.4% on CT-RATE and 75.0% on the external RAD-ChestCT benchmark, outperforming prior methods that require 4–8 high-performance GPUs. Extensive ablation studies validate the effectiveness of each negative-scaling component. Our code is available at <https://github.com/W11H08Z/ECT-CLIP>.

1 Introduction

Vision-Language Pretraining (VLP) has emerged as a powerful paradigm for learning transferable CT representations by leveraging paired radiology reports as supervision. The field’s trajectory reflects a shift toward increasing granularity: advancing from global-level alignment of entire volumes (Hamamci et al., 2026; Lai et al., 2025; Wald et al., 2026), to organ-level alignment via anatomical decomposition (Shui et al., 2025; Cao et al., 2025; Zhang et al., 2026), and recently to item-level alignment (Lyu et al., 2026) of discrete abnormal findings.

*Corresponding author.

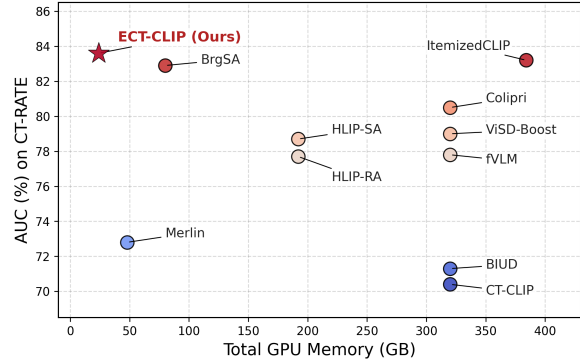


Figure 1: Zero-shot AUC (%) vs the total GPU memory capacity of the hardware used for training (GB) on CT-RATE. Colors indicate performance from blue (low) to red (high). ECT-CLIP (Ours, star) achieves the best AUC of 83.4% with only 24 GB, outperforming all baselines that require 48–384 GB.

However, current CT VLP methods are often hindered by prohibitive GPU memory consumption during training (Figure 1). First, the enormous voxel scale of 3D CT volumes generates a vast amount of intermediate activations (Li and Wu, 2026; Zeng et al., 2026), which significantly elevates GPU memory consumption. Second, the efficacy of contrastive learning is highly dependent on large batch sizes to ensure a sufficient number of negative samples (Radford et al., 2021; Zhai et al., 2023), which further intensifies the memory burden on the GPU. Consequently, these factors impose a severe hardware constraint, often limiting the training of high-performance CT VLP models.

To address these challenges, we propose **ECT-CLIP**, a resource-efficient CT VLP framework that achieves negative scaling—systematically expanding the supervision from negative samples along both quantity and quality dimensions under limited GPU memory. Specifically, we introduce three core modules that jointly realize this negative scaling:

To scale the quantity of negative supervision, we propose **Asymmetric Gradient Accumulation**.

Unlike conventional gradient accumulation (Goyal et al., 2017) or gradient caching (Gao et al., 2021) which assume symmetric memory costs across modalities, our method exploits the inherent memory asymmetry between CT and text encoding. By freezing the text encoder and pre-computing all text embeddings offline, we fix the image micro-batch size to 1 and draw a large number of cached text negatives per micro-batch. This drastically expands the negative pool with negligible visual-side memory increase.

To scale the quality of negative supervision, we introduce two complementary strategies. First, **UMLS-Guided Negative Mining** replaces heuristic embedding-similarity thresholds (Liu et al., 2023; Li et al., 2025) with a deterministic, explainable mechanism. By normalizing disease and anatomical mentions in each abnormal finding to Unified Medical Language System (Bodenreider, 2004) concepts, we precisely separate false negatives from truly informative hard negatives at the concept level. Second, **Multi-Expert Text Encoding** fuses representations from multiple frozen text encoders trained on distinct corpora. This enriches the semantic diversity of both positive and negative pairs, preserving complementary information without additional training cost.

We conduct extensive experiments on the CT-RATE and RAD-ChestCT benchmarks. Trained on a single consumer-grade RTX 4090 GPU, ECT-CLIP achieves state-of-the-art zero-shot classification performance, attaining AUC scores of 83.4% and 75.0%, respectively, and outperforming prior methods that require 4–8 high-performance GPUs. Ablation studies further validate the individual efficacy of each negative-scaling component. Our main contributions are as follows.

- We propose **ECT-CLIP**, a resource-efficient CT VLP framework that achieves negative scaling by jointly improving the quantity and quality of negative supervision, achieving state-of-the-art zero-shot classification performance (83.4% AUC and 75.0% AUC) using only a single RTX 4090 GPU.
- To scale the quantity of negative supervision, we exploit the memory asymmetry between CT and text encoding to enlarge the negative pool dramatically with negligible visual-side memory overhead.
- To scale the quality of negative supervision,

we design two complementary strategies: deterministically separating false negatives from informative hard negatives via UMLS concept normalization, and fusing multiple frozen medical text encoders to enrich semantic diversity without extra training cost.

2 Related Work

2.1 3D CT Vision-Language Pretraining

The evolution of 3D CT vision-language pretraining follows a clear path from coarse-grained to fine-grained alignment. Early works focus on global-level alignment, where entire 3D volumes and full reports are mapped into a shared feature space for contrastive learning (Hamamci et al., 2026; Lai et al., 2025; Wald et al., 2026; Blankemeier et al., 2026; Zhao et al., 2025). Recognizing that global representations lack fine-grained localization, subsequent work shifts to organ-level alignment. These methods pre-segment CT scans into anatomical regions and parse reports into organ-specific descriptions, explicitly aligning visual representations of individual organs with their corresponding text (Shui et al., 2025; Cao et al., 2025; Zhang et al., 2026; Lin et al., 2025). More recently, item-level alignment pushes granularity further by parsing reports into discrete abnormal findings. Employing cross-attention to produce text-conditioned visual features, these methods dynamically link individual findings to localized evidence in 3D volumes (Lyu et al., 2026). Our proposed framework builds upon this item-level paradigm.

2.2 Efficient Contrastive Learning

Contrastive pretraining (Radford et al., 2021; Zhai et al., 2023) relies on large batch sizes to provide sufficient in-batch negative samples. While gradient accumulation (Goyal et al., 2017) simulates large batches in standard supervised learning, it does not increase the number of effective negatives in contrastive learning. Gradient caching (Gao et al., 2021) breaks this limit by decoupling representation storage from backpropagation, yet existing methods typically assume symmetric computational costs across modalities. However, in CT vision-language pretraining, 3D CT encoding is substantially more memory-intensive than text encoding. By leveraging this asymmetric memory cost, we achieve large-scale negative supervision while avoiding the GPU memory overhead of caching intermediate representations.

2.3 Semantic-Aware Negative Mining

Existing negative mining methods in medical VLP can be broadly divided into two categories. (i) Similarity-based approaches reweight or filter negative pairs according to text–text or cross-modal embedding similarity (Li et al., 2025; Liu et al., 2023). (ii) Knowledge-injected noise suppression methods incorporate external clinical knowledge to estimate report-level semantic overlap and suppress noisy negatives through loss weighting or optimal transport (Chen et al., 2023; Cheng et al., 2025). A common limitation is that they treat semantic similarity as a continuous global score and rely on heuristic thresholds to decide whether a pair is “truly” negative. This lacks a deterministic, threshold-free classification of the relationship between two abnormal findings. Under our itemized supervision, where each report is decomposed into abnormal findings, we instead achieve deterministic and explainable negative classification based on normalized disease–anatomy concepts.

3 Methodology

3.1 Architecture Overview

As illustrated in Figure 2, we adopt an itemized text supervision strategy: each radiology report is parsed into a set of atomic abnormal findings, each describing a single disease at a single anatomical location (see Appendix A.1 for parsing details). Formally, for an image-report pair $(V^{(i)}, T^{(i)})$, the report is represented as $T^{(i)} = \{f_1^{(i)}, f_2^{(i)}, \dots, f_{n_i}^{(i)}\}$, where $f_j^{(i)}$ denotes an abnormal finding and n_i is the number of findings in $T^{(i)}$. As detailed in Section 3.3, each finding is further normalized to a UMLS concept tuple $z_j^{(i)}$ to enable semantic-aware negative mining.

Given an image $V^{(i)}$, the visual encoder E_V extracts a sequence of localized visual features $v^{(i)} = \{v_1^{(i)}, v_2^{(i)}, \dots, v_m^{(i)}\}$. For each abnormal finding $f_j^{(i)}$, the text encoder (Section 3.4) independently computes its representation as the [CLS] embedding $t_j^{(i)}$. To align visual features with specific abnormal findings, we compute a text-conditioned visual representation and its similarity with the corresponding abnormal finding. Following Lyu et al. (2026), we employ multi-head cross-attention to derive a text-conditioned visual representation. However, unlike their original formulation which computes similarity via cosine similarity, we compute

similarity via an unnormalized dot product:

$$\text{TCSim}(t_j^{(k)}, v^{(i)}) = t_j^{(k)} \cdot \text{CA}(t_j^{(k)}, v^{(i)}, v^{(i)}) \quad (1)$$

where CA (cross-attention) uses the abnormal finding representation $t_j^{(k)}$ as the Query and the visual feature sequence $v^{(i)}$ as the Key and Value to produce the text-conditioned visual representation.

To learn vision-language alignment, we employ a SigLIP objective (Zhai et al., 2023). The goal is to maximize $\text{TCSim}(t_j^{(k)}, v^{(i)})$ for positive pairs (where the j -th abnormal finding of the k -th report is present in the i -th image) and minimize it for negative pairs. Formally, the loss for a batch of N image-report pairs is:

$$\mathcal{L} = -\frac{1}{N} \sum_{i,j,k} \log \sigma(y_{i,j,k} \cdot (\tau \cdot s_{i,j,k} + b)) \quad (2)$$

where $s_{i,j,k} = \text{TCSim}(t_j^{(k)}, v^{(i)})$ is the similarity between the i -th image and the j -th finding of the k -th report, $y_{i,j,k} = 2 \cdot \mathbb{I}[i = k] - 1$ (where $\mathbb{I}[\cdot]$ is the indicator function that equals 1 if the condition holds and 0 otherwise), and τ, b are learnable temperature and bias parameters.

3.2 Asymmetric Gradient Accumulation

To enable large-scale negative sampling under limited GPU memory, we adopt an asymmetric gradient accumulation strategy. Since the memory consumption of the visual encoder is substantially higher than that of the text encoder, we fix the image-side micro-batch size to 1 and allocate the remaining memory budget to increase the number of text samples per optimization step.

Specifically, we first pre-compute and cache text embeddings for all abnormal findings in the dataset using the frozen text encoder E_T . During training, for each optimization step, we accumulate gradients over multiple micro-batches. In each micro-batch, we forward one CT image $V^{(i)}$ through E_V to obtain $v^{(i)}$, retrieve the cached embeddings of its corresponding abnormal findings as positives, and sample a subset of other abnormal findings as negatives (Section 3.3). The SigLIP loss is computed and back-propagated to accumulate gradients for the visual encoder. After accumulating gradients over multiple micro-batches, we perform a single parameter update and reset the gradients to zero.

By decoupling the image and text batch sizes and accumulating gradients across micro-batches, our method enlarges the effective negative pool with minimal additional image-side memory overhead.

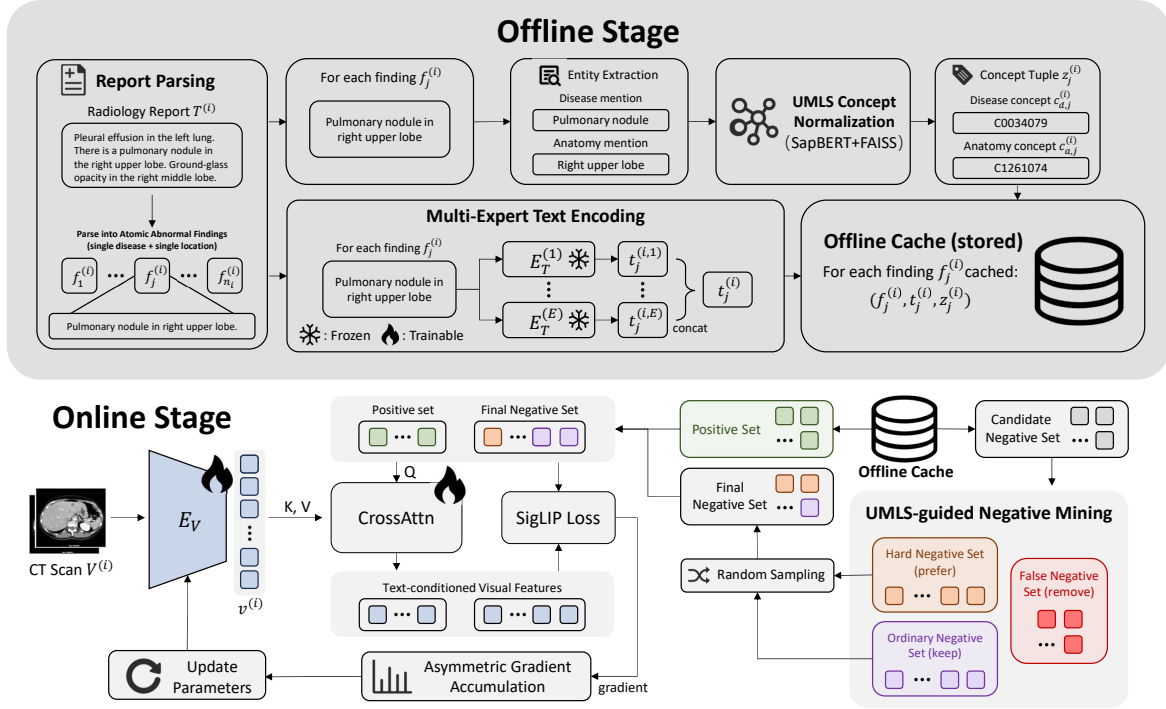


Figure 2: Overview of the proposed framework. The pipeline is divided into two main stages: (a) Offline Stage: Each radiology report is parsed into atomic abnormal findings, each representing a single disease at a single anatomical location. Disease and anatomical entity mentions are extracted and normalized into standard UMLS concept tuples $z_j^{(i)}$ via SapBERT and FAISS. Simultaneously, text embeddings $t_j^{(i)}$ for these findings are pre-computed using a frozen multi-expert text encoder and stored in an offline cache alongside their corresponding concept tuples. (b) Online Stage: During training, text-conditioned visual features are extracted from the input CT scan via cross-attention (CrossAttn), using the cached text embeddings as Queries and visual features as Keys/Values. A UMLS-guided negative mining strategy leverages the pre-computed concept tuples to remove false negatives and preferentially sample hard negatives. Finally, the model is optimized using a SigLIP loss paired with an asymmetric gradient accumulation strategy, enabling large-scale negative sampling under limited GPU memory.

3.3 UMLS-Guided Negative Mining

The above accumulation strategy yields a large effective negative pool, and here we refine its quality through UMLS-guided mining of false negatives and hard negatives.

Entity Extraction and Concept Normalization.

We employ DeepSeek-V3 to extract disease and anatomical entity mentions from each abnormal finding (see Appendix A.2). These mentions are normalized into standard UMLS concepts via a retrieval pipeline. Let $\mathcal{C}$ denote the set of candidate UMLS concepts for a target semantic type. Each concept $c \in \mathcal{C}$ is associated with a synonym set $\mathcal{S}_c$, and the full synonym vocabulary is $\mathcal{S} = \bigcup_{c \in \mathcal{C}} \mathcal{S}_c$. Disease concepts are restricted to semantic types Disease or Syndrome (DSYN) and Findings (FNDG), while anatomical concepts are restricted to Body Part, Organ, or Organ Component (BPOC).

Each synonym $s \in \mathcal{S}$ is encoded as $\mathbf{e}_s$ using SapBERT (Liu et al., 2021) with L2 normalization. We build a FAISS-based (Johnson et al., 2019) index over all $\mathbf{e}_s$. Given a mention m , we encode it as $\mathbf{e}_m$ and retrieve the top- L nearest synonyms:

$$\mathcal{N}_L(m) = \{s_1, s_2, \dots, s_L\} \subset \mathcal{S}, \quad (3)$$

ranked by descending inner product $\mathbf{e}_m \cdot \mathbf{e}_{s_l}$ (equivalent to cosine similarity after normalization). Let $\text{cui}(s_l) \in \mathcal{C}$ denote the concept associated with the retrieved synonym s_l . The vote count for each candidate concept is:

$$v(c) = \sum_{l=1}^L \mathbb{I}[\text{cui}(s_l) = c], \quad (4)$$

and the normalized concept is selected by voting:

$$c^* = \arg \max_{c \in \mathcal{C}} v(c). \quad (5)$$

In case of a tie, we select the concept with the smallest average retrieval rank:

$$c^* = \arg \min_{c \in \mathcal{C}^*} \frac{1}{v(c)} \sum_{l=1}^L l \cdot \mathbb{I}[\text{cui}(s_l) = c], \quad (6)$$

where $\mathcal{C}^* = \{c \in \mathcal{C} : v(c) = \max_{c'} v(c')\}$ is the set of top-voted concepts.

For each abnormal finding, we apply the above normalization procedure independently to its disease mention and anatomical mention. Finally, each normalized abnormal finding is represented as a concept tuple:

$$z_j^{(i)} = (c_{d,j}^{(i)}, c_{a,j}^{(i)}), \quad (7)$$

where $c_{d,j}^{(i)}$ and $c_{a,j}^{(i)}$ denote the normalized disease and anatomical concepts, respectively.

Negative Pair Definitions. Based on these concept tuples, we define two types of relationships between any two abnormal findings f and f' (with tuples $z = (c_d, c_a)$ and $z' = (c'_d, c'_a)$):

False Negatives. Findings with identical disease and anatomical concepts form false-negative pairs:

$$\text{FN}(z, z') = \mathbb{I}[c_d = c'_d \wedge c_a = c'_a] \quad (8)$$

Hard Negatives. Findings sharing exactly one concept—same disease at different locations or different diseases at the same location—form hard-negative pairs:

$$\begin{aligned} \text{HN}(z, z') = \mathbb{I} & \left[(c_d = c'_d \wedge c_a \neq c'_a) \right. \\ & \left. \vee (c_d \neq c'_d \wedge c_a = c'_a) \right]. \end{aligned} \quad (9)$$

Ordinary Negatives. Findings with distinct disease and anatomical concepts form ordinary-negative pairs:

$$\text{ON}(z, z') = \mathbb{I}[c_d \neq c'_d \wedge c_a \neq c'_a] \quad (10)$$

Online Negative Mining. For each image in a micro-batch, we treat all its associated abnormal findings as positive samples and construct the candidate negative set from the findings of all other reports. Using the pre-computed concept tuples, we evaluate the relationship between each candidate negative and every positive sample, partitioning the candidate pool into three subsets: the false-negative set, the hard-negative set, and the ordinary-negative set. The false-negative set is discarded to avoid semantic collision. We then randomly sample from

the hard-negative and ordinary-negative sets and mix them to form the final negative set, with the mixing ratio empirically determined through an ablation study (Figure 3).

3.4 Multi-Expert Text Encoding

Since different medical text encoders are trained on distinct corpora and objectives, they often capture complementary semantic information. Instead of relying on a single text encoder, we adopt a multi-encoder fusion strategy that aggregates representations from multiple frozen text encoders.

Given an abnormal finding $f_j^{(i)}$, we first obtain text embeddings from E frozen encoders:

$$t_j^{(i,e)} = E_T^{(e)}(f_j^{(i)}), \quad (11)$$

where $e \in \{1, \dots, E\}$ indexes the encoder. The fused text representation is constructed as:

$$t_j^{(i)} = [t_j^{(i,1)}; t_j^{(i,2)}; \dots; t_j^{(i,E)}] \quad (12)$$

where $[\cdot]$ denotes the concatenation.

3.5 Zero-Shot Inference

During zero-shot inference, given an image $V^{(i)}$ and a set of class descriptions $q_1, q_2, \dots, q_K$, we first encode them with the visual and text encoders (multi-expert text encoding; Section 3.4) to obtain the visual feature sequence $v^{(i)}$ and the text representations $t_{q_1}, t_{q_2}, \dots, t_{q_K}$. Our prediction logit for class k is then simply $s_k^{(i)} = \text{TCSim}(t_{q_k}, v^{(i)})$. For multi-label classification, a class-specific threshold is applied to all logits $s_k^{(i)}$ of the same class to determine the presence of each disease.

4 Experiments

4.1 Datasets

We conducted experiments on two public chest CT datasets: CT-RATE (Hamamci et al., 2026) and RAD-ChestCT (Draelos et al., 2021). CT-RATE contains 50,188 chest CT scans collected from 21,304 patients. Following the official data split protocol (Hamamci et al., 2026), we used 20,000 patients for training and the remaining 1,304 patients for internal evaluation. To further evaluate the generalization capability of the proposed method, we additionally adopted RAD-ChestCT as an external test set. RAD-ChestCT consists of 3,630 chest CT volumes collected from a different clinical source. In all experiments, the model

Table 1: Zero-shot classification on CT-RATE and RAD-ChestCT. Baseline results are taken from Shui et al. (2025) and the respective original papers. The best result is highlighted in **bold**, and the second-best is underlined.

Model	# GPU	GPU Type	CT-RATE				RAD-ChestCT			
			AUC ↑	ACC ↑	F1 ↑	Prec ↑	AUC ↑	ACC ↑	F1 ↑	Prec ↑
CT-CLIP (Hamamci et al., 2026)	4	A100	70.4	65.1	69.1	30.6	63.2	59.9	64.8	33.9
CT-VocabFine (Hamamci et al., 2026)	4	A100	75.0	69.2	72.8	34.2	64.9	62.2	66.5	35.9
CT-LiPro (Hamamci et al., 2026)	4	A100	75.1	67.6	71.4	33.1	64.7	62.5	66.8	35.3
BIUD (Cao et al., 2024)	4	A100	71.3	68.1	71.6	33.8	62.9	60.6	65.2	33.7
Merlin (Blankemeier et al., 2026)	1	A6000	72.8	67.2	70.9	33.7	64.4	61.9	66.3	34.8
fVLM (Shui et al., 2025)	4	A100	77.8	71.8	75.1	37.9	68.0	64.7	68.8	37.4
HLIP-RA (Zhao et al., 2025)	4	A40	77.7	71.4	74.7	—	72.3	68.4	72.1	—
HLIP-SA (Zhao et al., 2025)	4	A40	78.7	72.4	75.5	—	71.7	67.7	71.4	—
ViSD-Boost (Cao et al., 2025)	4	A100	79.0	73.1	75.9	38.7	69.4	65.2	69.3	34.2
CA-GCL (Zhang et al., 2026)	4	H100	79.3	72.7	75.7	38.4	69.2	64.7	69.0	35.1
Colipri (Wald et al., 2026)	4	A100	80.5	—	—	—	73.2	—	—	—
BrgSA (Lai et al., 2025)	1	A800	82.9	77.0	79.3	<u>43.2</u>	74.2	68.6	72.0	42.2
ItemizedCLIP (Lyu et al., 2026)	8	L40S	<u>83.2</u>	76.5	78.9	—	<u>74.7</u>	<u>69.8</u>	<u>73.2</u>	—
ECT-CLIP (Ours)	1	RTX 4090	83.4	<u>76.7</u>	<u>79.2</u>	43.5	75.0	70.5	73.9	<u>38.9</u>

was trained solely on the CT-RATE training set and evaluated on both the CT-RATE test set and the external RAD-ChestCT benchmark.

4.2 Data Preprocessing

Image Preprocessing. We first applied intensity normalization to all CT volumes using a Hounsfield Unit (HU) window of $[-1150, 350]$, followed by min-max normalization into the range $[-1, 1]$. To suppress background noise and reduce computational overhead, we computed a 3D foreground bounding box based on the organ segmentation masks provided in Shui et al. (2025), and cropped the CT volume accordingly. After cropping, all volumes were zero-padded to a fixed spatial size of $112 \times 256 \times 352$. During training, random crops of size $96 \times 240 \times 336$ were sampled as model inputs, while center crops of the same size were used during evaluation. The same preprocessing pipeline was applied consistently to both training and evaluation data.

Text Preprocessing. For the text branch, we used DeepSeek-V3 (DeepSeek-AI et al., 2025) to extract all abnormal findings from the original radiology reports in CT-RATE. The prompt template is provided in Appendix A.1.

4.3 Implementation Details

We adopted a 3D ResNet-18 (Tran et al., 2018) as the visual encoder, initialized with publicly available pre-trained weights¹. For the text branch, we compared four frozen medical text encoders: RadBERT (Yan et al., 2022), BioLORD (Remy et al., 2024), CXR-BERT (Boecking et al., 2022), and MedCLIP-TextModel (Wang et al., 2022).

¹https://download.pytorch.org/models/r3d_18-b3b3357e.pth

All experiments were conducted on a single NVIDIA RTX 4090 GPU. Our model was trained for 5 epochs using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} . The learnable temperature parameter τ and bias term b were initialized to 3.0 and -10.0 , respectively. The asymmetric gradient accumulation was performed with a fixed accumulation step of 4. For the UMLS-based concept normalization, we retrieved the top- L synonyms with $L = 10$.

4.4 Evaluation Metrics

We evaluated the zero-shot classification performance using six commonly adopted metrics, including Area Under the ROC Curve (AUC), Accuracy (ACC), Specificity (Spec), Sensitivity (Sens), F1-score (F1), and Precision (Prec).

Following Shui et al. (2025); Cao et al. (2025); Lyu et al. (2026), we evaluated zero-shot classification on 16 disease categories for the CT-RATE dataset and 14 disease categories for the RAD-ChestCT dataset. The detailed category lists for both datasets are provided in Appendix B.1.

For the ablation study only, we employed 13 different prompt templates and report the mean and standard deviation across them to ensure robustness. For the main experiments, to maintain fair comparisons with prior methods, we did not use this averaging strategy but instead used a single standard prompt template. Detailed prompt templates are provided in the Appendix B.3.

4.5 Main Results

Table 1 reports the zero-shot classification results. Our method achieves the best or second-best performance on all metrics across both benchmarks, while using only a single RTX 4090 GPU—

Table 2: Ablation study on text encoder combinations. All results are reported on the CT-RATE test set under the zero-shot classification protocol. MedCLIP denotes MedCLIP-TextModel for brevity. The best results are in **bold**.

CXR-BERT	BioLORD	RadBERT	MedCLIP	AUC $\uparrow$	ACC $\uparrow$	Spec $\uparrow$	Sens $\uparrow$	F1 $\uparrow$	Prec $\uparrow$
$\checkmark$				80.8 ± 0.9	74.4 ± 0.8	74.2 ± 1.0	75.7 ± 1.3	77.2 ± 0.7	40.7 ± 1.0
	$\checkmark$			81.1 ± 0.9	74.6 ± 1.0	74.3 ± 1.0	76.8 ± 0.9	77.4 ± 0.9	41.3 ± 1.2
		$\checkmark$		80.3 ± 1.2	74.5 ± 0.8	74.3 ± 0.6	75.4 ± 1.6	77.1 ± 0.7	40.7 ± 1.1
			$\checkmark$	78.5 ± 0.8	72.1 ± 0.5	71.7 ± 0.6	74.2 ± 1.1	75.4 ± 0.5	38.3 ± 0.5
$\checkmark$	$\checkmark$			81.1 ± 0.8	74.8 ± 0.9	74.3 ± 1.1	76.6 ± 0.7	77.5 ± 0.8	41.4 ± 0.9
$\checkmark$		$\checkmark$		80.9 ± 0.7	74.4 ± 0.7	73.9 ± 0.9	76.3 ± 1.0	77.2 ± 0.6	40.5 ± 0.9
$\checkmark$			$\checkmark$	79.4 ± 1.2	73.4 ± 0.9	73.4 ± 0.9	74.6 ± 1.5	76.2 ± 0.8	39.6 ± 1.2
$\checkmark$	$\checkmark$	$\checkmark$		81.6 ± 0.6	75.1 ± 0.8	74.4 ± 0.9	76.9 ± 0.6	77.8 ± 0.6	41.1 ± 0.9

Table 3: Comparison of different hard negative mining strategies on the CT-RATE test set. All results are reported under the zero-shot classification protocol. The best results are in **bold**.

Strategy	AUC $\uparrow$	ACC $\uparrow$	F1 $\uparrow$	Prec $\uparrow$
Random	80.2 ± 1.1	73.7 ± 0.9	76.5 ± 0.8	40.0 ± 1.2
TextSim Weighting	80.1 ± 1.1	73.8 ± 1.0	76.6 ± 0.8	39.9 ± 1.4
TextSim Mixing	79.9 ± 1.1	73.8 ± 0.8	76.5 ± 0.8	39.9 ± 1.3
TextSim Filtering	78.9 ± 1.5	73.6 ± 1.2	76.4 ± 1.0	39.6 ± 1.2
UMLS-NM (Ours)	80.8 ± 0.9	74.4 ± 0.8	77.2 ± 0.7	40.7 ± 1.0

substantially fewer resources than competing methods that require 4–8 high-end GPUs (A100, A40, L40S, A800). Notably, ECT-CLIP attains state-of-the-art AUC scores of 83.4% on CT-RATE and 75.0% on RAD-ChestCT, outperforming Itemized-CLIP (8 L40S, 83.2%) and BrgSA (1 A800, 82.9%) on the internal test set, and surpassing all baselines by a larger margin on the external RAD-ChestCT benchmark (+0.3% AUC over Itemized-CLIP and +0.8% over BrgSA). On RAD-ChestCT, our method also secures the best ACC (70.5%) and F1 (73.9%), underscoring its strong generalization to out-of-domain clinical data. While BrgSA achieves marginally higher ACC and F1 on CT-RATE (+0.2% and +0.1%, respectively), it requires an A800 datacenter GPU and lags behind on RAD-ChestCT by substantial margins, indicating weaker cross-institutional transferability. These results demonstrate that the proposed approach delivers superior zero-shot classification capability and cross-domain robustness with minimal computational overhead. A detailed per-disease performance breakdown is provided in Appendix B.2.

4.6 Ablation Study

Impact of Negative Mining Strategy. Table 3 compares our UMLS-NM against random sampling and three text-similarity-based variants (see Appendix B.4 for implementation details). Random sampling achieves 80.2 AUC and 76.5 F1. TextSim

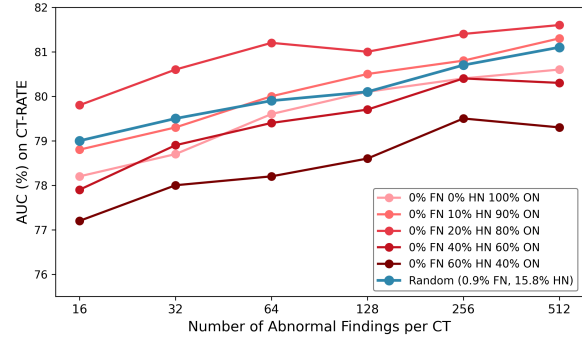


Figure 3: Average zero-shot AUC (%) on CT-RATE over 13 prompt templates, versus the number of abnormal findings per CT. FN: False Negatives; HN: Hard Negatives; ON: Ordinary Negatives. The “Random” curve denotes standard random sampling (0.9% FN, 15.8% HN). The remaining curves correspond to model performance with varying HN proportions.

Weighting and Mixing remain on par with the random baseline (80.1 and 79.9 AUC, respectively), while TextSim Filtering results in a noticeable drop to 78.9 AUC. In contrast, UMLS-NM achieves the best performance across all metrics, with 80.8 AUC and 77.2 F1, outperforming all competing strategies. These results demonstrate the effectiveness of our proposed negative mining strategy.

Impact of Hard Negative Ratio. Figure 3 reports the zero-shot AUC on CT-RATE across varying numbers of abnormal findings per CT. The 20% HN configuration consistently outperforms all other strategies across the entire range. Interestingly, the random baseline falls slightly below the 10% HN curve despite naturally containing 15.8% hard negatives and 0.9% false negatives, suggesting that even a small proportion of false negatives hampers learning. Moreover, the performance margin between the 20% HN strategy and the random baseline is largest under smaller numbers of abnormal findings per CT, indicating that our mined hard negatives are highly informative and provide dis-

Table 4: Progressive ablation of key components. **AGA**: Asymmetric Gradient Accumulation; **UMLS-NM**: UMLS-Guided Negative Mining; **METE**: Multi-Expert Text Encoding. The best results are in **bold**.

AGA	UMLS-NM	METE	AUC $\uparrow$	ACC $\uparrow$	Spec $\uparrow$	Sens $\uparrow$	F1 $\uparrow$	Prec $\uparrow$
✓			80.2 $\pm$ 1.1	73.7 $\pm$ 0.9	73.3 $\pm$ 0.9	75.4 $\pm$ 1.6	76.5 $\pm$ 0.8	40.0 $\pm$ 1.2
✓	✓		80.8 $\pm$ 0.9	74.4 $\pm$ 0.8	74.2 $\pm$ 1.0	75.7 $\pm$ 1.3	77.2 $\pm$ 0.7	40.7 $\pm$ 1.0
✓	✓	✓	81.6 $\pm$ 0.6	75.1 $\pm$ 0.8	74.4 $\pm$ 0.9	76.9 $\pm$ 0.6	77.8 $\pm$ 0.6	41.1 $\pm$ 0.9

criminative signals. These results demonstrate the effectiveness of the hard negatives and false negatives identified by our negative mining strategy.

Impact of Text Encoders. Table 2 reports the ablation study on text encoder combinations. Among single encoders, BioLORD achieves the best AUC (81.1), while MedCLIP performs the worst (78.5). Combining multiple encoders consistently improves performance, with the three-encoder ensemble (CXR-BERT, BioLORD, RadBERT) reaching the highest AUC of 81.6. We exclude MedCLIP from the final ensemble because it degrades performance in all pairwise combinations (e.g., CXR-BERT + MedCLIP drops to 79.4 AUC), suggesting its feature space is less compatible with the other medical encoders for item-level alignment.

Component-wise Ablation. Table 4 validates the contribution of each proposed component. The baseline with only Asymmetric Gradient Accumulation (AGA) achieves an AUC of 80.2. Adding UMLS-Guided Negative Mining (UMLS-NM) improves AUC by 0.6 points (80.8), demonstrating that UMLS-based hard negatives provide more informative supervision than random sampling. Incorporating Multi-Expert Text Encoding (METE) further boosts AUC by 0.8 points to 81.6, confirming that ensemble medical text encoders yield richer semantic representations for fine-grained alignment. Each component contributes complementary gains across all metrics.

Scaling Law and Data Efficiency. Figure 4 compares scaling behavior against CA-GCL (Zhang et al., 2026) on {1%, 10%, 25%, 50%, 100%} of CT-RATE training data. AUC scores are averaged over 13 prompt templates with error bars denoting standard deviation. Both methods follow a power-law AUC = αD^β , yet ours achieves a higher intercept ($\alpha = 0.734$ vs. 0.692) and steeper exponent ($\beta = 0.0248$ vs. 0.0228). At 1% data, we outperform CA-GCL by 3.5 points (72.7% vs. 69.2%); at 100% data, the gap increases to 4.5 points (81.6% vs. 77.0%), demonstrating markedly higher sample

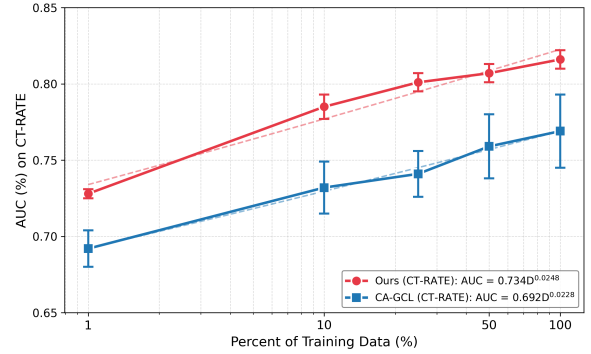


Figure 4: Zero-shot AUC (%) vs. training data percentage on CT-RATE. Each marker denotes the mean AUC across 13 prompt templates, with error bars indicating standard deviation. Dashed lines show fitted power-law curves ($\text{AUC} = \alpha D^\beta$).

efficiency under only 24GB GPU memory.

5 Conclusion

In this work, we introduced ECT-CLIP, an efficient CT VLP framework that achieves negative scaling—systematically expanding negative supervision along both quantity and quality dimensions—under severely limited GPU memory. To scale the quantity, our asymmetric gradient accumulation strategy leverages modality memory asymmetry to drastically enlarge the negative pool with negligible visual-side overhead. To scale the quality, UMLS-guided negative mining provides deterministic, concept-level separation of false negatives from hard negatives, while multi-expert text encoding enriches semantic representations by fusing complementary frozen encoders at no extra training cost. Trained on a single consumer-grade RTX 4090 GPU, ECT-CLIP achieves state-of-the-art zero-shot classification on both CT-RATE (83.4% AUC) and the external RAD-ChestCT benchmark (75.0% AUC), outperforming methods that rely on 4–8 high-performance GPUs. Extensive ablation studies confirm the individual contribution of each negative-scaling component. These results demonstrate that high-quality medical vision-language pretraining is achievable under strict resource con-

straints, paving the way for more accessible and scalable 3D medical image analysis.

Limitations

Although ECT-CLIP achieves strong performance under limited computational resources, several limitations remain. First, our framework relies on itemized abnormal findings extracted by a large language model, and the quality of these parsed findings may affect the reliability of both contrastive supervision and UMLS-guided negative mining. Second, the current negative mining strategy depends on deterministic concept normalization and therefore may not fully capture subtle semantic relationships between clinically related findings that do not share identical UMLS concepts. Third, due to hardware constraints, our experiments are conducted using relatively lightweight visual backbones and limited training schedules. Scaling the proposed framework to larger vision encoders and larger-scale datasets may further improve representation quality. Finally, our study focuses primarily on zero-shot classification for chest CT analysis, and future work could investigate broader downstream applications such as report generation, retrieval, and dense lesion localization.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and constructive feedback during the review process. This work was supported by the National Natural Science Foundation of China (U22B2061), the National Key R&D Program of China (2022YFB4300603), and Natural Science Foundation of Sichuan, China (2026NSFSC0429).

References

- Akhila Abdunazar, Roland Roller, Stefan Schulz, and Markus Kreuzthaler. 2024. [Unsupervised sapbert-based bi-encoders for medical concept annotation of clinical narratives with snomed ct](#). *DIGITAL HEALTH*, 10:20552076241288681.
- Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, and 1 others. 2026. Merlin: a computed tomography vision-language foundation model and dataset. *Nature*, pages 1–11.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and 1 others. 2022. Making the most of text semantics to improve biomedical vision-language processing. In *European conference on computer vision*, pages 1–21. Springer.
- Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, and 1 others. 2025. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23041–23050.
- Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. 2024. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11238–11247.
- Xiaofei Chen, Yuting He, Cheng Xue, Rongjun Ge, Shuo Li, and Guanyu Yang. 2023. Knowledge boosting: Rethinking medical contrastive vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer.
- Pujin Cheng, Yijin Huang, Li Lin, Junyan Lyu, Kenneth Kin-Yip Wong, and Xiaoying Tang. 2025. Mirage: Medical image-text pre-training for robustness against noisy environments. *Medical Image Analysis*, page 103912.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. 2021. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857.
- Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321.
- Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.
- Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihn Simsek, Sevval Nil

- Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, and 1 others. 2026. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547.
- Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809.
- David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. [A comprehensive evaluation of biomedical entity linking models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14462–14478, Singapore. Association for Computational Linguistics.
- Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Weifu Lv, Wei Wei, and Shaohua Kevin Zhou. 2025. Bridged semantic alignment for zero-shot 3d medical image diagnosis. *IEEE Journal of Biomedical and Health Informatics*.
- Jintao Li and Xinming Wu. 2026. Memory-efficient full-volume inference for large-scale 3d dense prediction without performance degradation. *Communications Engineering*.
- Yongxin Li, Ying Cheng, Yaning Pan, Wen He, Qing Wang, Rui Feng, and Xiaobo Zhang. 2025. Semantic-aware hard negative mining for medical vision-language contrastive pretraining. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3133–3142.
- Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Kai Cao, Le Lu, Jiebo Luo, and Ling Zhang. 2025. [Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios](#). *Preprint*, arXiv:2404.15272.
- Bo Liu, Donghuan Lu, Dong Wei, Xian Wu, Yan Wang, Yu Zhang, and Yefeng Zheng. 2023. Improving medical vision-language contrastive pretraining with semantics-aware triage. *IEEE Transactions on Medical Imaging*, 42(12):3579–3589.
- Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. [Learning domain-specialised representations for cross-lingual biomedical entity linking](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574, Online. Association for Computational Linguistics.
- Yiwei Lyu, Chenhui Zhao, Soumyanil Banerjee, Shixuan Liu, Akshay Rao, Akhil Kondepudi, Honglak Lee, and Todd C. Hollon. 2026. [Learning complete and explainable visual representations from itemized text supervision](#). *Preprint*, arXiv:2512.11141.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- François Remy, Kris Demuynck, and Thomas De-meester. 2024. [BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights](#). *Journal of the American Medical Informatics Association*, page ocae029.
- Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and 1 others. 2025. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*.
- Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.
- Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel C. F. Codella, Maria Teodora Wetscherek, Klaus H. Maier-Hein, Panagiotis Korfiatis, Valentina Salvatelli, Javier Alvarez-Valle, and Fernando Pérez-García. 2026. [Comprehensive language-image pre-training for 3d medical image understanding](#). *Preprint*, arXiv:2510.15042.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jiemeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887.
- An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.
- Sen Zeng, Hong Zhou, Zheng Zhu, and Yang Liu. 2026. Tokenseg: Efficient 3d medical image segmentation via hierarchical visual token compression. *arXiv preprint arXiv:2601.04519*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Hanwen Zhang, Yao Liu, Die Dai, Jiaye Yang, Qiao Liu, Yutong Xie, and Peng Wang. 2026. [Ca-gcl: Cross-anatomy global-local contrastive learning for robust 3d medical image understanding](#). *Preprint*, arXiv:2605.13544.

Chenhui Zhao, Yiwei Lyu, Asadur Chowdury, Edward Harake, Akhil Kondepudi, Akshay Rao, Xinhai Hou, Honglak Lee, and Todd Hollon. 2025. Towards scalable language-image pre-training for 3d medical imaging. *arXiv preprint arXiv:2505.21862*.

A Text Processing Details

A.1 Prompt for Abnormal Finding Extraction

We employed DeepSeek-V3 (DeepSeek-AI et al., 2025) to decompose each radiology report into a set of atomic abnormal findings. The complete system prompt, illustrated in Figure 5, instructs the model to retain only thoracic abnormalities, split complex descriptions, and preserve key clinical attributes while discarding normal statements.

A.2 Prompt for Entity Extraction

After obtaining each abnormal finding, we further extracted disease mentions and anatomical location mentions using DeepSeek-V3 with the prompt shown in Figure 6. The extracted textual spans were subsequently normalized to UMLS concepts for deterministic negative mining.

B Experimental Details

B.1 Disease Categories

Table 5 lists the disease categories used for zero-shot classification on the CT-RATE and RAD-ChestCT datasets. CT-RATE comprises 16 disease categories, while RAD-ChestCT covers 14 categories. Among these, 14 categories are shared across both datasets, enabling consistent cross-dataset evaluation. Two categories are exclusive to CT-RATE (Coronary artery wall calcification and Arterial wall calcification), and one category is unique to RAD-ChestCT (Calcification).

B.2 Per-Disease Performance Analysis.

Table 6 breaks down the zero-shot classification performance of ECT-CLIP across all 16 disease categories in CT-RATE. The model achieves near-perfect discrimination on pathologies with distinct and localized visual signatures, including Pleural effusion (96.9% AUC), Cardiomegaly (93.7%), Coronary artery wall calcification (93.6%), and Arterial wall calcification (93.5%). These conditions

Table 5: Disease categories for CT-RATE (16 classes) and RAD-ChestCT (14 classes).

CT-RATE	RAD-ChestCT
Emphysema	Emphysema
Atelectasis	Atelectasis
Lung nodule	Lung nodule
Lung opacity	Lung opacity
Pulmonary fibrotic sequela	Pulmonary fibrotic sequela
Pleural effusion	Pleural effusion
Peribronchial thickening	Peribronchial thickening
Consolidation	Consolidation
Bronchiectasis	Bronchiectasis
Interlobular septal thickening	Interlobular septal thickening
Cardiomegaly	Cardiomegaly
Pericardial effusion	Pericardial effusion
Hiatal hernia	Hiatal hernia
Mosaic attenuation pattern	Mosaic attenuation pattern
Coronary artery wall calcification	–
Arterial wall calcification	–
–	Calcification

Table 6: Per-disease zero-shot classification performance (%) of ECT-CLIP on the CT-RATE test set.

Disease Category	AUC ↑	ACC ↑	F1 ↑	Prec ↑
Emphysema	78.7	72.1	74.6	38.5
Atelectasis	74.6	65.9	68.4	37.8
Lung nodule	68.5	64.3	64.4	59.7
Lung opacity	81.5	76.1	76.4	65.4
Pulmonary fibrotic sequela	71.2	66.5	68.0	42.2
Pleural effusion	96.9	94.1	94.5	69.3
Peribronchial thickening	81.8	69.3	74.8	25.0
Consolidation	91.2	82.9	84.2	53.5
Bronchiectasis	75.1	72.9	77.6	23.2
Interlobular septal thickening	86.9	78.3	82.8	24.4
Cardiomegaly	93.7	85.8	87.9	41.8
Pericardial effusion	87.8	82.1	85.9	26.1
Hiatal hernia	73.4	64.3	69.8	23.2
Mosaic attenuation pattern	86.4	78.8	83.2	25.1
Coronary artery wall calcification	93.6	86.7	87.3	67.3
Arterial wall calcification	93.5	87.8	88.1	73.6
Mean	83.4	76.7	79.2	43.5

typically present as large, high-contrast, or anatomically well-defined abnormalities, which naturally align with the itemized text supervision strategy. Conversely, performance drops on finer-grained or more subtle lesions such as Lung nodule (68.5% AUC), Pulmonary fibrotic sequela (71.2%), and Hiatal hernia (73.4%), where small lesion sizes, lower visual contrast, or greater similarity to normal structures increase classification difficulty. Despite these challenging categories, the model maintains a mean AUC of 83.4%, indicating balanced generalization across the full disease spectrum.

B.3 Zero-Shot Prompt Templates

For the ablation study (Section 4.6), we averaged results over 13 prompt templates to reduce prompt sensitivity. Each template consists of a negative prompt (indicating absence of the disease) and a positive prompt (indicating presence). The templates are listed in Table 7.

Convert the following medical imaging report into concise descriptions suitable for image-text contrastive learning, following these rules:

1. **Only retain thoracic-related content**, including but not limited to: lungs, trachea and bronchi, heart, aorta, pulmonary arteries, pulmonary veins, mediastinum, lymph nodes, thymus, superior vena cava, etc.
 - * Explicitly exclude non-thoracic organs, including but not limited to: liver, gallbladder, pancreas, spleen, kidneys, adrenal glands, stomach, intestines, etc. If a sentence contains both thoracic and abdominal organs, keep only the thoracic part.
2. **Extract only abnormal or pathological findings**; ignore normal statements.
 - * If there are no abnormal descriptions, simply return an empty list.
3. **Split long or complex sentences into shorter ones.**
4. **Each sentence should describe:**
 - * One organ (or a specific anatomical region)
 - * One primary abnormality
5. **Preserve key medical details whenever possible**, including:
 - * Anatomical location (e.g., "right upper lung", "mediastinum")
 - * Lesion type (e.g., "nodule", "ground-glass opacity")
 - * Attributes (e.g., "5 mm", "round", "well-defined")
6. **Maintain natural and fluent medical language.**
7. **Reduce redundancy, but do not over-simplify into rigid or unnatural phrases.**
8. **Each sentence should be short and simple.** Avoid using commas or complex sentence structures.
 - * If a sentence contains multiple pieces of information, split it into multiple independent sentences.

Figure 5: System prompt used with DeepSeek-V3 for extracting concise, item-level abnormal findings from radiology reports.

Extract anatomical locations and pathological findings from clinical text. Return JSON.

Entity Types:

- **Organ locations:** Anatomical sites (organs, lobes, segments, positional qualifiers [left/right/upper/lower], fissures)
- **Abnormal findings:** Pathological changes (nodules, masses, infiltrates, consolidations, calcifications, fibrosis, effusions, lesions, opacities)

Figure 6: System prompt for extracting disease mentions and anatomical location mentions from each abnormal finding, returning JSON-formatted entity lists.

Table 7: 13 zero-shot prompt templates used in ablation experiments.

Negative Template	Positive Template
Not {disease}	{disease}
Ruled out {disease}	Confirmed {disease}
No findings of {disease}	Findings suggestive of {disease}
Negative for {disease}	Positive for {disease}
No evidence of {disease}	Findings consistent with {disease}
No sign of {disease}	Sign of {disease}
No indication of {disease}	Indication of {disease}
{disease} is not present	{disease} is present
Absence of {disease}	Presence of {disease}
No diagnosis of {disease}	Diagnosis of {disease}
No imaging findings of {disease}	Imaging findings of {disease}
No features of {disease}	Features of {disease}
No manifestation of {disease}	Manifestation of {disease}

B.4 Details of Text-Similarity-Based Negative Mining Variants

In the ablation study on negative mining strategies (Section 4.6), we compared our UMLS-guided negative mining (UMLS-NM) against three heuristic baselines based on text embedding similarity. All variants used the CXR-BERT (Boecking et al., 2022) text encoder. For each image, we construct a contrastive pool of $N = 512$ text samples, consisting of all positive abnormal findings of that image and a set of abnormal findings randomly drawn from other reports. Let $\bar{t}^+$ denote the mean embedding of all positive text instances associated with the image. The three strategies modify how the

negatives within this fixed-size pool are utilized.

TextSim Weighting. For each negative embedding t_k^- , we compute its cosine similarity to $\bar{t}^+$ as $s_k = \cos(\bar{t}^+, t_k^-)$. The logit of the negative pair is then multiplied by a weight $w_k = 1 + s_k$, which increases the contribution of harder (more similar) negatives while leaving dissimilar negatives largely unchanged. No additional normalization is applied, as the learnable temperature τ in the SigLIP loss can accommodate the scaling.

TextSim Mixing. Inspired by Kalantidis et al. (2020), we first identify all negatives with a cosine similarity greater than 0.9 to $\bar{t}^+$. From this high-similarity subset, we generate 64 synthetic embeddings via random convex combinations of pairs: $\tilde{t} = \lambda t_i^- + (1 - \lambda) t_j^-$, where $\lambda \sim \text{Uniform}(0, 1)$. These 64 synthetic negatives are added to the pool, replacing 64 randomly selected original negatives. Consequently, the initial negative count is $512 - n_{\text{pos}}$ (with n_{pos} positives in the pool), and after replacement the total pool size remains 512.

TextSim Filtering. Inspired by Liu et al. (2023), based on the cosine similarity s_k , each candidate negative is relabeled as follows:

- If $s_k > 0.97$, the pair is considered a false-negative and treated as a positive sample ($y = 1$).

Table 8: Cross-template zero-shot comparison on CT-RATE. We report the mean and standard deviation over the 13 prompt templates in Table 7. All baselines are evaluated using their official checkpoints and preprocessing pipelines. The best results are in **bold**.

Method	AUC $\uparrow$	ACC $\uparrow$	F1 $\uparrow$	Prec $\uparrow$
Colipri	78.7 $\pm$ 1.4	71.4 $\pm$ 2.2	74.6 $\pm$ 2.3	39.0 $\pm$ 0.9
BrgSA	78.9 $\pm$ 2.5	73.3 $\pm$ 2.0	76.1 $\pm$ 1.8	39.1 $\pm$ 2.2
ItemizedCLIP	80.1 $\pm$ 2.9	73.4 $\pm$ 2.4	76.1 $\pm$ 2.3	39.2 $\pm$ 2.7
ECT-CLIP	81.6 $\pm$ 0.6	75.1 $\pm$ 0.8	77.8 $\pm$ 0.6	41.1 $\pm$ 0.9

- If $s_k < 0.8$, the pair is kept as a negative sample ($y = -1$).
- If $0.8 \leq s_k \leq 0.97$, the pair is deemed ambiguous and excluded from the loss computation (neutral).

This filtering aims to remove likely false negatives and avoid training on uncertain pairs, relying solely on embedding-space proximity.

C Additional Analyses

C.1 Cross-Template Zero-Shot Comparison

To verify that the improvements of ECT-CLIP over the strongest baselines are consistent rather than prompt-specific, we evaluated Colipri (Wald et al., 2026), BrgSA (Lai et al., 2025), and Itemized-CLIP (Lyu et al., 2026) on CT-RATE across all 13 prompt templates (Table 7) using their official checkpoints and preprocessing pipelines. As shown in Table 8, ECT-CLIP achieves the highest mean AUC with the lowest standard deviation (± 0.6 vs. ± 1.4 – 2.9), confirming that the improvement is both consistent and reliable.

C.2 Visual Backbone Ablation

To examine whether our conclusions generalize beyond the lightweight 3D ResNet-18 backbone, we ablate three 3D ResNet variants of increasing depth. For a fair comparison, all variants are trained from scratch under identical settings, since pretrained weights are publicly available only for 3D ResNet-18. As shown in Table 9, AUC improves modestly with deeper backbones, suggesting that the gains of our method are not tied to a specific backbone depth. 3D ResNet-50, while slightly stronger, incurs higher memory and training cost—a trade-off that depends on the practitioner’s resources. Our negative scaling strategy is orthogonal to backbone choice and applicable to any vision encoder.

Table 9: Zero-shot classification performance with different 3D ResNet backbones on CT-RATE. All variants are trained from scratch under identical settings; results are averaged over 13 prompt templates.

Backbone	AUC $\uparrow$	ACC $\uparrow$	F1 $\uparrow$	Prec $\uparrow$
3D ResNet-18	79.3 $\pm$ 1.1	73.0 $\pm$ 0.9	75.8 $\pm$ 0.9	38.6 $\pm$ 1.0
3D ResNet-34	79.6 $\pm$ 0.8	73.3 $\pm$ 0.7	76.2 $\pm$ 0.7	39.1 $\pm$ 0.8
3D ResNet-50	80.0 $\pm$ 0.8	73.8 $\pm$ 0.5	76.6 $\pm$ 0.5	39.5 $\pm$ 0.6

Table 10: Evaluation of LLM-based abnormal finding extraction on CT-RATE, measured against dataset-provided ground-truth labels.

Model	Accuracy	Micro F1	Macro F1
DeepSeek-V3	96.87	91.85	91.77
DeepSeek-V4-Flash	97.10	92.44	92.27
DeepSeek-V4-Pro	97.69	94.59	94.04

C.3 Extraction and Normalization Reliability

Extraction Quality of DeepSeek-V3. To quantify the quality of LLM-based finding extraction, we designed an evaluation protocol: we feed CT reports to the model, ask it to return structured JSON outputs containing disease labels, parse them, and compare them against the dataset-provided ground-truth labels. As shown in Table 10, all three evaluated DeepSeek versions achieve over 96% accuracy and 91% F1. We note that the ground-truth labels themselves contain human annotation noise, so the true extraction quality may be even higher than these numbers suggest.

Robustness of UMLS Normalization. The reliability of our SapBERT+FAISS normalization pipeline has been independently validated in prior work. Abdunazar et al. (2024) report that SapBERT with FAISS retrieval achieves a weighted F1 of 0.765 for medical concept normalization on the n2c2 dataset in a fully unsupervised setting; notably, the *Disorder* semantic type—the most relevant to radiological findings—achieves the highest F1 of 0.871. Their error analysis further shows that normalization failures stem primarily from contextual span errors (32.6%) and granularity mismatches (14.3%) rather than surface-form variation, confirming that SapBERT reliably handles diverse phrasing. In addition, Kartchner et al. (2023) conducted the largest biomedical entity linking comparison to date (9 models across 8 datasets) and concluded that SapBERT was arguably the best-performing alias matching method.