RadZero: Similarity-Based Cross-Attention for Explainable Vision-Language Alignment in Chest X-ray with Zero-Shot Multi-Task Capability

Jonggwon Park, Byungmu Yoon, Soobum Kim, Kyoyun Choi*
DEEPNOID Inc.
Seoul, South Korea
{jgpark, bmyoon, soobumk, kychoi}@deepnoid.com

Abstract

Recent advancements in multimodal models have significantly improved vision-language (VL) alignment in radiology. However, existing approaches struggle to effectively utilize complex radiology reports for learning and offer limited interpretability through attention probability visualizations. To address these challenges, we introduce **RadZero**, a novel framework for VL alignment in chest X-ray with zero-shot multi-task capability. A key component of our approach is **VL-CABS** (**V**ision-**L**anguage **C**ross-**A**ttention **B**ased on **S**imilarity), which aligns text embeddings with local image features for interpretable, fine-grained VL reasoning. RadZero leverages large language models to extract concise semantic sentences from radiology reports and employs multi-positive contrastive training to effectively capture relationships between images and multiple relevant textual descriptions. It uses a pre-trained vision encoder with additional trainable Transformer layers, allowing efficient high-resolution image processing. By computing similarity between text embeddings and local image patch features, VL-CABS enables zero-shot inference with similarity probability for classification, and pixel-level VL similarity maps for grounding and segmentation. Experimental results on public chest radiograph benchmarks show that RadZero outperforms state-of-the-art methods in zero-shot classification, grounding, and segmentation. Furthermore, VL similarity map analysis highlights the potential of VL-CABS for improving explainability in VL alignment. Additionally, qualitative evaluation demonstrates RadZero's capability for open-vocabulary semantic segmentation, further validating its effectiveness in medical imaging. Code is available at https://github.com/deepnoid-ai/RadZero.

1 Introduction

Recent advancements in deep learning have significantly impacted medical imaging, leading to numerous studies on computer-aided diagnosis [15, 8, 25, 36]. However, acquiring high-quality manual annotations remains a key challenge. In contrast, vision-language (VL) models (VLMs) in the natural image domain [29, 40, 39] have reduced reliance on manual labeling by learning from image-text pairs without explicit supervision, achieving strong zero-shot performance in tasks like classification and retrieval. Building on this progress, VLMs have been increasingly explored in medical imaging, including chest X-rays (CXRs). Several studies have demonstrated effective representation learning [37, 42, 43] and zero-shot capabilities [19, 24] without task-specific annotations.

Despite these advance, current medical VLMs underutilize the rich semantics of radiology image–report pairs. Prior methods—such as word-level alignment [12, 1], clinical entity extraction

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

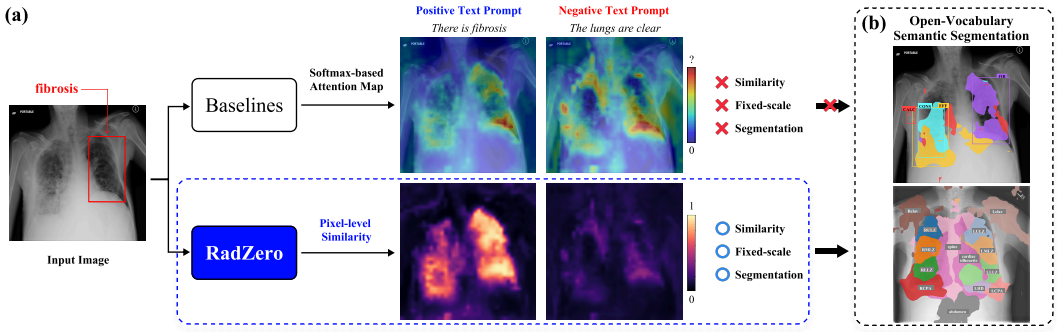

Figure 1: Comparison of attention maps and the proposed VL similarity map for visualizing VL alignment. (a) While traditional attention maps inevitably exhibit high values at certain points due to softmax activation, the proposed VL similarity maps yield low values for unrelated image-text pair. (b) Their fixed scale, originating from cosine similarity, enables open-vocabulary semantic segmentation through simple thresholding.

[42, 37], and using large language model (LLM) prompts [19]—face limitations, struggling with poor text embedding segmentation and inefficient training due to random sampling. An effective solution should 1) decompose reports into semantically minimal, clinically meaningful sentences, 2) embed each sentence independently, and 3) learn from multiple sentence–image pairs per study to fully leverage supervision.

Moreover, reliable explainability is critical for the clinical adoption of medical VLMs. Attention maps are used as explainable features by most of the recent research [37, 19, 42]. However, while an attention map (Figure 1 (a), top) can indicate where the model is focusing, it does not provide an explanation for why it is attending to those regions. This limitation can be addressed by computing pixel-level image–text similarity, which enables more fine-grained and transparent explanations.

To overcome these shortcomings, we propose **RadZero**, a novel VL alignment framework for chest X-ray with zero-shot multi-task capabilities. RadZero employs multi-positive contrastive learning [20] to incorporate multiple sentences per image-report pair. RadZero's core innovation is **V**ision-**L**anguage **C**ross-**A**ttention **B**ased on **S**imilarity (**VL-CABS**), which directly computes cosine similarity between text descriptions and local image patches. Unlike traditional attention maps, the resulting VL similarity maps offer clearer visual reasoning by maintaining low values for unrelated image-text pairs (Figure 1 (a), bottom). This enhances interpretability and enables open-vocabulary semantic segmentation via simple thresholding. RadZero also supports high-resolution inputs by freezing a pre-trained image encoder [39] and adding trainable Transformer layers [18], further boosting performance on fine-grained zero-shot tasks. Experiments on public chest radiograph benchmarks demonstrate that RadZero outperforms state-of-the-art (SOTA) models in various zero-shot tasks (Figure 2), while qualitative analyses reveal its enhanced explainability and potential for open-vocabulary semantic segmentation.

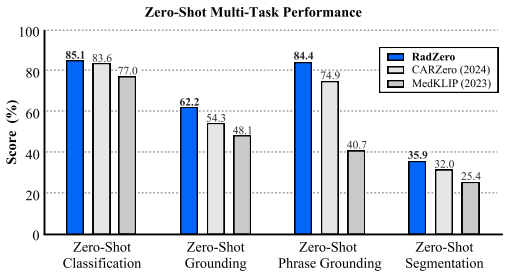

Figure 2: Zero-shot multi-task performance. Each score is averaged over multiple datasets per task.

2 Related Works

2.1 General vision-language alignment

Contrastive learning for vision–language alignment with large-scale image–text pairs has been actively studied. CLIP [29] demonstrated that directly aligning images and text enables strong zero-shot

classification. LiT [39] proposed freezing the pre-trained vision encoder during contrastive training, preserving fine-grained visual features and further enhancing zero-shot performance. `dino.txt` [18] extended this framework by adding Transformer [33] layers on top of a pre-trained DINOv2 [27], training only a lightweight module while keeping the vision encoder frozen. Additionally, it fused global and patch-averaged embeddings, enabling patch-level similarity computation with text and supporting open-vocabulary semantic segmentation. UniCLIP [20] introduced a multi-positive NCE (MP-NCE) loss, which independently evaluates multiple positive pairs per image. Building on these advances, our approach integrates a frozen, fine-grained vision encoder with trainable Transformer layers, following LiT and `dino.txt`. We also adopt MP-NCE loss to align images with multiple text representations effectively.

2.2 Vision-language alignment in chest X-ray

Since the introduction of contrastive learning in radiology [43], aligning CXR images with radiology reports has become an active research area. GLoRIA [12] focused on local alignment using cross-attention between word-level text embeddings and patch-level image features. MGCA [34] employed both report-level and token-level embeddings to extract multi-granular features, and BioViL-T [1] similarly relied on token-level embeddings. Nevertheless, segmenting reports into individual words or tokens often fails to capture their full semantic meaning.

Due to the complexity of medical image–report relationships, alignment interpretability is essential for clinical use and is commonly addressed using attention maps. MedKLIP [37] and KAD [42], for example, used RadGraph [14] to extract report features and employed attention maps for tasks such as grounding and segmentation. In addition to VL alignment, G2D [22] aggregated attention maps in addition to VL alignment to generate pseudo masks, which were used as pixel-wise pretext supervisory signals during pre-training. CARZero [19] also used attention maps when leveraging cross-attention alignment for zero-shot tasks, incorporating LLM-based prompt alignment to standardize reports. Despite their utility, attention maps have limitations: they often highlight irrelevant regions due to softmax activation, but removing softmax is not ideal as raw logits are unnormalized and uncentered. Additionally, variation in the norms of query and key embeddings leads to inconsistent similarity values across different image-text pairs. An example and a detailed discussion of the limited explainability of attention maps are provided in Appendix A. In contrast, our approach enhances explainability with VL-CABS, aligning visual patches and text embeddings. The resulting maps offer intuitive and consistent measures of fine-grained image-text similarity.

3 Methods

3.1 Finding-sentence extraction

Radiology reports contain diverse types of information, including clinical history, observations, comparative analysis with prior studies, and diagnostic impressions. Encoding the entire report into a single text embedding often fails to capture this complexity. CARZero [19] addressed this by using an LLM to extract relevant sentences and introducing a prompt alignment strategy based on the template "There is [disease]" for consistency between training and inference. Similarly, we use an LLM to extract such sentences, which we refer to as *finding-sentences*. These are generated using a prompt that follows a predefined structure, such as "There is [finding] of [location]," and are segmented into minimal semantic units containing the finding name, presence (or uncertainty), and location. The full prompt is provided in Appendix E.2. Each image is paired with multiple finding-sentences during training, as illustrated in Figure 3 (a). For zero-shot inference, we apply prompt alignment by prepending "There is" to text descriptions of findings and anatomical regions.

3.2 Vision-language alignment with similarity based cross-attention

3.2.1 Model architecture

To leverage the advantages of vision encoder pre-training, we adopt the approach of LiT [39] by freezing a pre-trained vision encoder in contrastive learning. In Vision Transformers [9] such as DINOv2 [27], interpolating the positional embeddings allows for increased input image resolution [31]. Building on this property, we train our model with high-resolution images. To embed the output

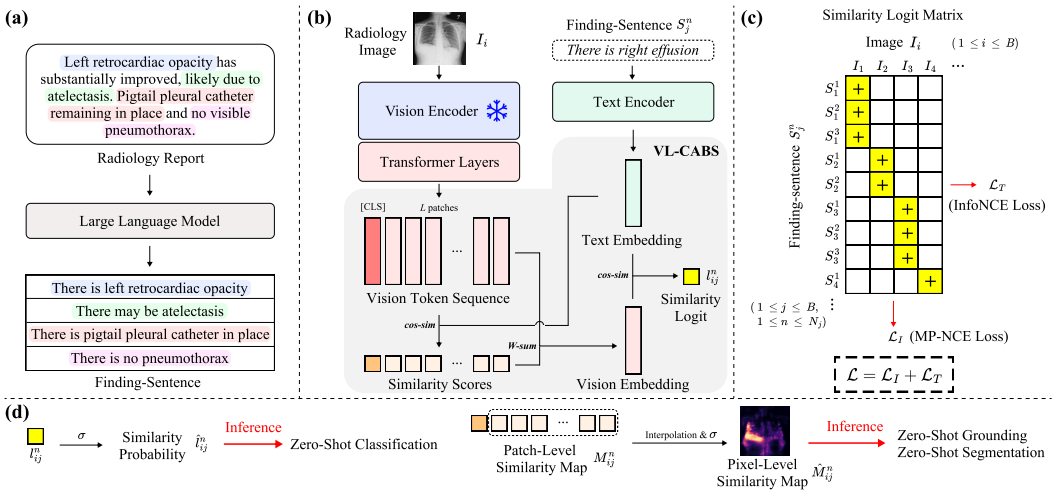

Figure 3: The overall framework of RadZero. (a) Finding-sentence extraction using an LLM. (b) Computation of the similarity logit, l_{ij}^n, between image I_i and finding-sentence S_j^n. ***W-sum*** and ***cos-sim*** denote weighted sum and cosine similarity, respectively. (c) Computation of MP-NCE loss ($\mathcal{L}_I$) and InfoNCE loss ($\mathcal{L}_T$) from the similarity logit matrix. (d) Zero-shot inference pipeline.

of the vision encoder, we add K trainable Transformer layers, as proposed by Jose et al. [18]. For the text encoder, we use a pre-trained Sentence-BERT [30], which is fine-tuned during training, to extract embeddings for each finding-sentence. The model architecture is illustrated in Figure 3 (b).

3.2.2 Vision-language cross-attention based on similarity

We propose VL-CABS (Vision-Language Cross-Attention Based on Similarity), a cosine similarity-based cross-attention mechanism for computing similarity logits. By directly employing cosine similarity between the text and visual patch embeddings, we obtain VL similarity scores that are well-defined in range and centered at zero. This consistent scaling allows for fair comparisons across different image–text pairs and significantly enhances explainability through the visualization of VL similarity maps. It also enables single thresholding, offering new possibilities for open-vocabulary semantic segmentation.

The proposed method operates on a mini-batch of size B, pairing each image I_i ($i = 1, \ldots, B$) with N_i associated finding-sentences $\{S_i^n\}_{n=1}^{N_i}$. Each image is processed by a vision encoder f_v followed by trainable layers f_a, yielding a sequence of embeddings $[\, v_{i0}, v_{i1}, \ldots, v_{iL} \,] = f_a(f_v(I_i))$, where $v_{i0} \in \mathbb{R}^D$ corresponds to the `[CLS]` token and $\{v_{ik}\}_{k=1}^{L} \subset \mathbb{R}^D$ are patch embeddings, with D denoting the embedding dimension and L the total number of patches. Each finding-sentence S_j^n is encoded into a sentence-level embedding $t_j^n = f_t(S_j^n) \in \mathbb{R}^D$ using a text encoder f_t.

To compute VL similarity, we $\ell 2$-normalize all embeddings as $\bar{v}_{ik} = v_{ik}/\|v_{ik}\|_2$ and $\bar{t}_j^n = t_j^n/\|t_j^n\|_2$, and calculate the scaled cosine similarity between each patch and sentence as $s_{ijk}^n =< \bar{v}_{ik}, \bar{t}_j^n > \cdot \exp(\tau) \quad (k = 0, \ldots, L)$, where τ is a learnable temperature. These scores are converted into attention weights using softmax over the patch index k: $a_{ijk}^n = \exp(s_{ijk}^n)/\sum_{m=0}^{L} \exp(s_{ijm}^n)$. A sentence-specific attended vision embedding is computed as the weighted sum $v_{ij}^n = \sum_{k=0}^{L} a_{ijk}^n v_{ik}$, which is then $\ell 2$-normalized as $\bar{v}_{ij}^n = v_{ij}^n/\|v_{ij}^n\|_2$. The global similarity logit between image I_i and sentence S_j^n is given by $l_{ij}^n =< \bar{v}_{ij}^n, \bar{t}_j^n > \cdot \exp(\tau)$. The corresponding patch-level similarity map is $M_{ij}^n = [\, s_{ij1}^n, \ldots, s_{ijL}^n \,]$.

3.3 Multi-positive contrastive learning

Although CARZero [19] also uses prompt templates for training, it suffers from instability due to randomly selecting one sentence for each image at every training step. To utilize all N finding-sentences matched to each image at every step, we adopt multi-positive NCE (MP-NCE) loss [20] which treats positive pairs independently in order to amplify the loss contributions from each positive

pair. A visualization of our contrastive loss is shown in Figure 3 (c). Let $N_T = \sum_{i=1}^{B} N_i$ be the total number of finding-sentences in a mini-batch. For the i-th image, the number of positive and negative finding-sentences are N_i and $N_T - N_i$, respectively. The MP-NCE loss can be computed as follows:

$$\mathcal{L}_I = -\frac{1}{N_T} \sum_{i=1}^{B} \sum_{n=1}^{N_i} \log \frac{\exp(l_{ii}^n)}{\exp(l_{ii}^n) + \sum_{j \neq i}^{B} \sum_{m=1}^{N_j} \exp(l_{ij}^m)} \tag{1}$$

For each finding-sentence S_i^n, there is one positive image I_i and $B-1$ negative images. The corresponding InfoNCE loss [26] is computed as follows:

$$\mathcal{L}_T = -\frac{1}{N_T} \sum_{i=1}^{B} \sum_{n=1}^{N_i} \log \frac{\exp(l_{ii}^n)}{\exp(l_{ii}^n) + \sum_{j \neq i}^{B} \exp(l_{j,i}^n)} \tag{2}$$

The final objective function is the sum of $\mathcal{L}_I$ and $\mathcal{L}_T$: $\mathcal{L} = \mathcal{L}_I + \mathcal{L}_T$.

3.4 Zero-shot inference

The similarity logit between an image I_i and a sentence S_j^n, denoted by l_{ij}^n, is converted into a similarity probability $\hat{l}_{ij}^n = \sigma(l_{ij}^n)$ via a sigmoid function, and used for zero-shot classification.

For grounding and segmentation, we reshape the patch-level similarity map $M_{ij}^n = [s_{ij1}^n, \ldots, s_{ijL}^n]$ into a $\sqrt{L} \times \sqrt{L}$ square map, and resize it to the original image resolution via bilinear interpolation. To account for preprocessing such as padding and resizing, this interpolation is applied accordingly. A final element-wise sigmoid activation is applied to obtain the pixel-level similarity map $\hat{M}_{ij}^n = \sigma(\text{bilinear}(M_{ij}^n))$, which we refer to as the *VL similarity map* and use for zero-shot grounding and segmentation. The VL similarity map is derived from the cosine similarity between vision patches and text embeddings, and since we do not modify the embedding space beyond applying $\ell 2$-normalization and adjusting the temperature, it can be directly interpreted as the similarity between each image pixel and the text. The zero-shot inference process is illustrated in Figure 3 (d).

4 Experiments

4.1 Training dataset

MIMIC-CXR [16] We train our model using the MIMIC-CXR dataset for VL alignment. MIMIC-CXR comprises 377,110 CXR images from 227,835 radiographic studies involving 65,379 patients. Each study includes a radiology report and one or more CXR images in either frontal or lateral views. Images are sourced from MIMIC-CXR-JPG [17], and only the 'findings' and 'impression' sections of reports are extracted using the official codebase[2]. All view positions are considered, and the official dataset split is followed. As described in Section 3.1, finding-sentence extraction is applied, with each study containing an average of 6.45 such sentences. Studies without extracted finding-sentences are discarded, resulting in 352,875 training images and 2,852 for validation.

4.2 Test datasets

Open-I (OI) [7] contains 3,851 radiology reports and 7,470 CXR images with multi-label annotations for 18 diseases. **PadChest (PC)** [4] comprises 160,868 CXR images from 67,000 patients, with 192 labels showing a long-tailed distribution. Following [19], we use 39,053 samples annotated by board-certified radiologists. Additionally, **PadChest20 (PC20)**, introduced in [19], serves as a test set for rare disease evaluation, consisting of 20 classes with fewer than 10 samples each. **ChestXray14 (CXR14)** [35] provides official test set with 22,433 images labeled for 14 diseases. **CheXpert (CXP)** [13] includes a test set of 500 patients' images annotated by five board-certified radiologists. Following [19], we evaluate classification on five observations: atelectasis, cardiomegaly, consolidation, edema,

[2] `https://github.com/MIT-LCP/mimic-cxr`

and pleural effusion. **ChestXDet10 (CXD10)** [23], a subset of CXR14, contains 542 images with bounding box annotations for 10 diseases in the official test set. **SIIM [38]** pneumothorax dataset provides segmentation masks for 11,582 CXRs; we adopt the test split from [34], which includes 1,704 images with 458 positives. **RSNA [32]** pneumonia dataset consists of 29,700 frontal CXRs with bounding box annotations; we use the test set from [37] containing 5,337 images, including 1,218 positives. **MS-CXR [2]** consists of 1,153 image-phrase-bounding box triplets, with images sourced from MIMIC-CXR. The bounding boxes annotated to specific phrases in the report enable more detailed grounding, referred to as *phrase grounding*. For fair evaluation on the test set of 167 images released by [6], where each phrase maps to a single bounding box, we exclude these images from the training set described in Sec. 4.1.

4.3 Evaluation metrics

AUC, or area under the ROC curve, is adopted to evaluate zero-shot classification on multi-label test datasets. **Pointing game [41]**, which determines whether the coordinates of the maximum value falls within the corresponding bounding box, is employed as the grounding metric. **Dice** score serves as a standard evaluation metric for segmentation. Following [37], we compute the Dice score using only positive samples and optimize the segmentation threshold on the test set to maximize the score. Threshold search intervals are 0.01 for sigmoid and 0.001 for softmax, depending on the feature map's activation function. **Pixel-wise AUC (Pix-AUC)** computes AUC at pixel-level to evaluate the quality of the segmentation probability map. To account for both sensitivity and specificity in mask prediction, we incorporate both positive and negative samples. For fine-grained tasks such as grounding and segmentation, predictions are interpolated back to the original image size before evaluation.

4.4 Implementation details

We adopt XrayDINOv2 [5] as the pre-trained vision encoder, which was trained in a unimodal setting using CXR images based on DINOv2 [27]. While the vision encoder was trained with an image resolution of 224, we increase it to 518 for our experiments. The patch size of 14×14 leads to 37×37 patches, yielding a vision patch length L of 1369. The text encoder is MPNet ("all-mpnet-base-v2") [30], initialized with pre-trained parameters and further fine-tuned during training. The trainable Transformer layers consist of two randomly initialized layers ($K = 2$), with a hidden dimension of 768, matching the embedding size of both the vision and text encoders. While the vision encoder remains frozen, all other parameters are trainable. Following [29], the learnable temperature parameter τ is initialized to $\log(1/0.07)$. The details of model training can be found in Appendix C.1. The LLM used for extracting finding-sentences is "Llama-3.3-70B-Instruct" [10], deployed in a private computing environment.

5 Results

5.1 Zero-shot evaluation

Classification. Table 1 compares RadZero with SOTA models on public test datasets. For the five datasets evaluated in CARZero [19], we report their published results. For SIIM and RSNA, we independently evaluated two open-source models. RadZero achieved new SOTA performance on OI and PC, irrespective of image resolution. In the long-tailed PC dataset with 192 classes, RadZero

Method	Open-I (OI)	PadChest (PC)	PadChest20 (PC20)	ChestXray14 (CXR14)	CheXpert (CXP)	ChestXDet10 (CXD10)	SIIM	RSNA
GLoRIA [12]	0.589	0.565	0.558	0.610	0.750	0.645	-	-
BioViL-T [1]	0.702	0.655	0.608	0.729	0.789	0.708	-	-
MedKLIP [37]	0.759	0.629	0.688	0.726	0.879	0.713	0.897	**0.869**
KAD [42]	0.807	0.750	0.735	0.789	0.905	0.735	-	-
CARZero [19]	0.838	0.810	0.837	**0.811**	**0.923**	**0.796**	**0.924**	0.747
RadZero (224px)	**0.851**	**0.841**	**0.879**	0.807	0.903	0.785	0.914	0.839
RadZero	0.847	**0.841**	0.871	0.804	0.900	0.787	**0.924**	0.834

Table 1: Zero-shot classification AUROC scores on public CXR datasets. For fair comparison, we also report the results of low-resolution (224×224) version of RadZero.

Method	Mean	ATE	CALC	CONS	EFF	EMPH	FIB	FX	MASS	NOD	PTX
GLoRIA [12]	0.367	0.479	0.053	0.737	0.528	0.667	0.366	0.013	0.533	0.156	0.143
KAD [42]	0.391	**0.646**	0.132	0.699	0.618	0.644	0.244	0.199	0.267	0.316	0.143
BioViL-T [1]	0.351	0.438	0.000	0.630	0.504	0.846	0.390	0.026	0.500	0.000	0.171
MedKLIP [37]	0.481	0.625	0.132	**0.837**	0.675	0.734	0.305	0.224	0.733	0.312	0.229
CARZero [19]	0.543	0.604	0.184	0.824	0.782	0.846	0.561	0.184	0.700	0.286	0.457
RadZero (224px)	0.537	0.604	0.211	0.806	0.813	0.795	0.451	0.197	**0.767**	0.325	0.400
RadZero	**0.622**	**0.646**	**0.368**	0.824	**0.857**	**0.872**	**0.585**	**0.250**	**0.767**	**0.506**	**0.543**

Table 2: Zero-shot grounding results (pointing game accuracy) on CXD10. Lesion abbreviations can be found in Appendix E.1.

outperformed CARZero by 3.1 percentage points, demonstrating strong generalization in zero-shot classification. Notable gains are also observed in PC20, which focuses on rare diseases, suggesting that VL-CABS is particularly effective for infrequent conditions. On datasets where RadZero failed to rank first, MedKLIP performed best on RSNA, while CARZero led on CXR14, CXP, and CXD10. However, MedKLIP underperformed on the latter datasets, and CARZero underperformed on RSNA. In contrast, RadZero showed results comparable to the top-performing models across all datasets. Interestingly, the lower-resolution RadZero (224px) even outperformed RadZero: potentially due to the pre-trained vision encoder, as discussed in Sec. 5.4.

The representative classification metric shown in Figure 2 is the average AUC across all datasets. RadZero established a new SOTA, outperforming CARZero by 1.5 percentage points, a gain attributable to our training strategy that incorporates multi-positive contrastive learning to enhance the diversity of both positive and negative samples per image.

Grounding. Table 2 presents zero-shot grounding results on CXD10. We adopted the pointing game scores reported by CARZero for all models except BioViL-T, which we evaluated using its released weights. RadZero achieved the highest average score across all diseases, outperforming CARZero by 0.079. Per-lesion analysis showed that RadZero achieved the best performance in all classes except consolidation, indicating that the proposed VL-CABS effectively captures local alignment between text and image patches regardless of disease type. Furthermore, RadZero efficiently supports higher input resolutions, enabling more precise localization.

Method	MS-CXR
BioViL-T [1]	0.719
MedKLIP [37]	0.407
CARZero [19]	0.749
RadZero (224px)	0.832
RadZero	**0.844**

Table 3: Zero-shot phrase grounding results (pointing game accuracy) on MS-CXR.

Table 3 reports zero-shot phrase grounding results, evaluating alignment at the phrase level in contrast to disease-level grounding in Table 2. Pointing game accuracy is used as the evaluation metric. We evaluated all baselines using publicly available models. RadZero achieved the highest score of 0.844, demonstrating accurate interpretation of text phrases. The strong performance of RadZero (224px) suggests that the gains are largely attributable to the effectiveness of VL-CABS rather than the higher input resolution.

Segmentation. Table 4 summarizes zero-shot segmentation results on SIIM and RSNA. To benchmark against supervised models, we fine-tuned MGCA with varying proportions of training data; percentages in parentheses indicate the amount used.

Method	RSNA	SIIM	
	Dice	Dice	Pix-AUC
GLoRIA [12]	0.347*	-	-
BioViL [3]	0.439*	-	-
MedKLIP [37]	0.465*	0.044	0.648
G2D [22]	0.477†	0.051†	-
CARZero [19]	0.540	0.100	0.856
CARZero (logits)	0.529	0.081	0.928
RadZero (224px)	**0.562**	0.121	0.943
RadZero	0.546	**0.171**	**0.947**
MGCA [34] (1%)	0.513	0.144	0.752
MGCA (10%)	0.571	0.238	0.856
MGCA (100%)	0.578	0.305	0.976

Table 4: Zero-shot segmentation results. Values with * are from [37] and † from [22]

Among zero-shot models, RadZero achieved the highest Dice scores on both datasets. On SIIM, it outperformed CARZero by 71%, demonstrating superior segmentation capability. The smaller margin on RSNA is likely due to its coarser annotations—bounding boxes rather than pixel-level masks—which limit the advantages of RadZero's fine-grained VL similarity maps. As in phrase grounding, the results of RadZero (224px) suggest that the superior performance is not merely driven by higher resolution.

RadZero also remained competitive against fine-tuned models. It outperformed MGCA (1%) on both datasets, demonstrating the effectiveness of zero-shot

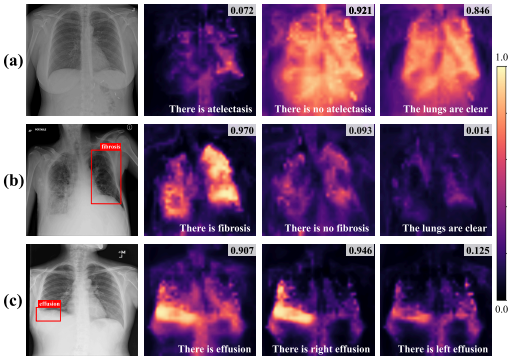

Figure 4: VL similarity maps of CXR images from CXD10, representing (a) normal, (b) fibrosis, and (c) effusion in the right lung. The value at the top-right corner represent the similarity probability $\hat{l}$ between each CXR image and the text prompt (bottom-right corner).

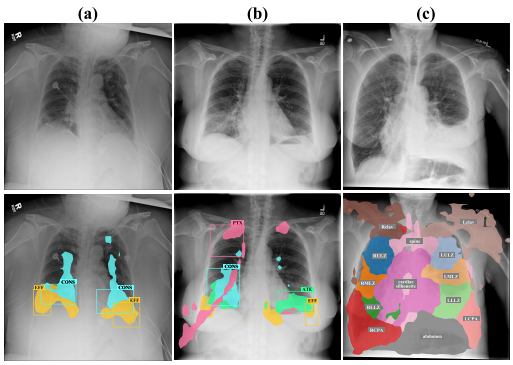

Figure 5: Open-vocabulary semantic segmentation results: (a), (b) for findings and (c) for anatomical regions. The CXR images and bounding box labels are from CXD10. The segmentation thresholds were set to 0.7 for (a) and (b), and 0.4 for (c).

segmentation. Although it did not surpass MGCA (10%) or (100%), RadZero requires no mask labels, enabling broader generalization beyond fixed vocabularies, as further discussed in Sec. 5.3.

SIIM's detailed annotations also allow for Pix-AUC evaluation. RadZero achieved the highest score, exceeding even MGCA (10%), indicating well-calibrated VL similarity maps that distinguish positive from negative regions. In contrast, MedKLIP and CARZero, both relying on attention maps, performed worse with scores of 0.648 and 0.856, respectively. For a fair comparison, we also evaluated CARZero's pre-softmax logits (CARZero (logits)), which improved performance but still fell short of RadZero. Notably, CARZero (logits) underperformed its own Dice score (0.081 vs. 0.100) despite threshold tuning, as expected from the inconsistent scaling of dot-product similarity. In contrast, VL-CABS directly encode pixel-level text–image similarity, allowing low values for negative samples. This contributed to its superior Pix-AUC, as further supported in Sec. 5.2.

5.2 VL similarity map analysis

Figure 4 illustrates that RadZero effectively aligned visual and textual representations with VL-CABS. The outputs, VL similarity map $\hat{M}$ and probability $\hat{l}$, offer both interpretable visualizations and quantitative metrics. For the normal image Figure 4 (a), the model assigned low similarity (0.072) to the prompt "There is atelectasis" with a dark VL similarity map, indicating weak alignment between vision tokens and the text embedding, In contrast, "There is no atelectasis" (0.921) and "The lungs are clear" (0.846) yielded bright activations across lung fields, reflecting strong alignment.

Figure 4 (b) shows fibrosis, and "There is fibrosis" resulted in high similarity (0.970) with strong activations in the affected lung. Prompts indicating normality received much lower scores (0.093 and 0.014) and darker VL similarity maps, clearly distinguishing abnormal from normal descriptions.

Figure 4 (c) highlights RadZero's ability to distinguish anatomical descriptions. For right-sided pleural effusion, the model assigned high similarity (0.907) to "There is effusion," with bright activations in the correct region. Notably, "There is right effusion" (0.946) scored even higher, indicating accurate localization, while "There is left effusion" scored much lower (0.125) and a dark VL similarity map, showing that the model correctly distinguishes between left and right lung regions.

Overall, these results underscore the explainability of VL-CABS. The similarity probability is verifiable at the pixel level, enabling spatially grounded explanations. By explicitly revealing how conclusions are derived, RadZero offers enhanced interpretability in the context of disease diagnosis.

5.3 Open-vocabulary semantic segmentation

Figure 5 presents open-vocabulary semantic segmentation results for both findings and anatomical regions. Segmentation masks were generated by thresholding the VL similarity map $\hat{M}$ for each text

Method	Similarity	Trainable layers	MP	Res.	Classification						Grounding		Segmentation
					OI	PC	PC20	CXR14	CXP	CXD10	CXD10	MS-CXR	SIIM
(a)	dot-product	Linear	✗	224	0.839	0.824	0.853	0.805	0.896	0.792	0.472	0.784	0.078
(b)	cos	Linear	✗	224	0.843	0.830	0.863	0.805	0.902	0.786	0.483	0.790	0.078
(c)	cos	2 Transformer	✗	224	0.845	0.832	0.860	**0.808**	0.895	**0.793**	0.539	0.838	0.099
(d) RadZero (224px)	cos	2 Transformer	✓	224	**0.851**	**0.841**	**0.879**	0.807	**0.903**	0.785	0.537	0.832	0.121
RadZero	cos	2 Transformer	✓	518	0.847	**0.841**	0.871	0.804	0.900	0.787	**0.622**	**0.844**	**0.171**
LiT [39]	-	Linear	✗	224	0.768	0.769	0.775	0.764	0.854	0.735	-	-	-
`dino.txt` [18]	-	2 Transformer	✗	224	0.834	0.816	0.837	0.797	0.901	0.770	0.121	0.174	0.021
CARZero [19]	-	Transformer Dec.	✗	224	0.827	0.815	0.877	0.795	0.889	0.770	0.437	0.743	0.072

Table 5: Ablation study of model architecture components. "MP" denotes multi-positive.

prompt; in cases of overlapping predictions, the prompt with the highest similarity was assigned. In Figure 5 (a), RadZero successfully localized lesions based on text queries, though some segmentation masks extended beyond ground truth boxes, indicating room for improvement. Notably, certain incorrect predictions captured clinically relevant features that were not explicitly annotated: in Figure 5 (b), a chest tube was reasonably associated with "pneumothorax." Figure 5 (c) further demonstrates RadZero's ability to segment anatomical structures without supervision, inferring approximate spatial regions from text despite imprecise boundaries. These results highlight the potential of VL-CABS for zero-shot open-vocabulary semantic segmentation and RadZero's capacity to align textual descriptions with medical imagery. Additional qualitative examples are provided in Appendix H.

5.4 Ablation Studies

Ablation study on RadZero components. **1) Similarity function:** Comparing (a) and (b) in Table 5 shows the effect of using cosine similarity instead of scaled dot-product [33] for VL alignment. Cosine similarity, which better aligns with inference-time VL similarity maps, improved classification and slightly enhances grounding. **2) Trainable parameters:** (b) and (c) compare a linear layer and a two-layer Transformer, with the image encoder frozen. Transformer layers yielded consistent gains across all tasks, particularly in grounding ($0.483 \rightarrow 0.539$) and segmentation ($0.078 \rightarrow 0.099$). **3) Multi-positive pairs:** The difference between (c) and (d) lay in the use of multi-positive contrastive pairs. (d) improved classification (e.g., PC20: $0.860 \rightarrow 0.879$) and segmentation ($0.099 \rightarrow 0.121$), highlighting the advantage of richer supervision. **4) Image resolution:** RadZero and (d) shared the same architecture, except that RadZero used higher resolution inputs (518 vs. 224). This change substantially improved grounding ($0.537 \rightarrow 0.622$) and segmentation ($0.121 \rightarrow 0.171$), showing the importance of high-resolution features for spatially localized tasks. In classification, (d) outperformed RadZero: likely due to the vision encoder (XrayDINOv2[5]) being pre-trained at 224 pixels, indicating that using an encoder pre-trained at higher resolutions may further enhance the performance.

Comparison among different VL alignment approaches. Table 5 compares RadZero with alternative VL alignment methods, keeping all settings identical except for VL feature fusion and loss computation. LiT, which uses a `[CLS]` embedding for alignment, showed limited classification performance and was incapable of grounding or segmentation. `dino.txt` improved classification through additional Transformer layers, but its mean pooling constrained grounding and segmentation performance. CARZero introduced a cross-attention decoder, enhancing performance on those tasks. However, when compared to the RadZero ablations, (b) outperformed CARZero across most metrics, showing that VL-CABS alone is sufficiently effective.

6 Conclusion

In this work, we introduced RadZero, a novel VL alignment model for chest X-ray that achieved strong zero-shot performance in classification, grounding, and segmentation. Central to RadZero is VL-CABS, which computes image-text similarity at the patch-level to improve interpretability. Combined with multi-positive contrastive training, VL-CABS enabled effective representation learning without pixel-level annotations, and the support for high-resolution inputs further boosted performance. Extensive evaluations on public chest radiograph benchmarks showed that RadZero outperformed SOTA methods. VL similarity map analysis highlighted the enhanced explainability of VL-CABS by providing transparent rationales for how conclusions are derived. Qualitative assessments further demonstrated RadZero's potential for open-vocabulary semantic segmentation.

Despite its impressive results, RadZero has limitations that indicate areas for future research. The observed performance degradation on specific datasets emphasizes the necessity of enhancing generalization capability. Its reliance on the pre-trained vision encoder may also restrict domain adaptability. In addition, the current study validates the proposed RadZero training framework only on chest X-ray datasets, which limits the scope of its generalization. Future work could explore extending RadZero to other medical imaging modalities such as CT and MRI, demonstrating its potential as a universal vision-language learning framework adaptable to diverse anatomical and visual characteristics. Moreover, applying VL-CABS to general imaging domains, for instance in open-vocabulary semantic segmentation, could be a meaningful direction toward building more interpretable VLMs.

Acknowledgments and Disclosure of Funding

This work was supported by the Technology Innovation Program (RS-2025-02221011, Development of Medical-Specialized Multimodal Hyperscale Generative AI Technology for Global Integration) funded by the Ministry of Trade Industry & Energy (MOTIE, South Korea).

References

[1] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15016–15027, June 2023.

[2] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. *Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing*, page 1–21. Springer Nature Switzerland, 2022. ISBN 9783031200595. doi: 10.1007/978-3-031-20059-5_1. URL http://dx.doi.org/10.1007/978-3-031-20059-5_1.

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

[4] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, December 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101797. URL http://dx.doi.org/10.1016/j.media.2020.101797.

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

[6] Zhihao Chen, Yang Zhou, Anh Tran, Junting Zhao, Liang Wan, Gideon Su Kai Ooi, Lionel Tim-Ee Cheng, Choon Hua Thng, Xinxing Xu, Yong Liu, et al. Medical phrase grounding with region-phrase context contrastive alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381. Springer, 2023.

[7] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Steven E Shooshan, Louis Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, Mar 2016. doi: 10.1093/jamia/ocv080.

[8] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4):198–211, 2007. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2007.02.002. URL https://www.sciencedirect.com/science/article/pii/S0895611107000262. Computer-aided Diagnosis (CAD) and Image-guided Decision Support.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[11] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

[12] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *ICCV*, pages 3942–3951, 2021.

[13] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, Jul. 2019. doi: 10.1609/aaai.v33i01.3301590. URL `https://ojs.aaai.org/index.php/AAAI/article/view/3834`.

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

[15] Mohammad Jamshidi, Ali Lalbakhsh, Jakub Talla, Zdeněk Peroutka, Farimah Hadjilooei, Pedram Lalbakhsh, Morteza Jamshidi, Luigi La Spada, Mirhamed Mirmozafari, Mojgan Dehghani, Asal Sabet, Saeed Roshani, Sobhan Roshani, Nima Bayat-Makou, Bahare Mohamadzade, Zahra Malek, Alireza Jamshidi, Sarah Kiani, Hamed Hashemi-Dezaki, and Wahab Mohyuddin. Artificial intelligence and covid-19: Deep learning approaches for diagnosis and treatment. *IEEE Access*, 8:109581–109595, 2020. doi: 10.1109/ACCESS.2020.3001973.

[16] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[17] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[18] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. *arXiv preprint arXiv:2412.16334*, 2024.

[19] Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. Carzero: Cross-attention alignment for radiology zero-shot classification. In *CVPR*, pages 11137–11146, 2024.

[20] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. *NeurIPS*, 35:1008–1019, 2022.

[21] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL `https://doi.org/10.1093/bioinformatics/btz682`.

[22] Che Liu, Cheng Ouyang, Sibo Cheng, Anand Shah, Wenjia Bai, and Rossella Arcucci. G2d: From global to dense radiography representation learning via vision-language pre-training. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 14751–14773. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/1ac14e44228aeadabb3c934822c1212a-Paper-Conference.pdf`.

[23] Jingyu Liu, Jie Lian, and Yizhou Yu. Chestx-det10: Chest x-ray dataset on detection of thoracic abnormalities. *arXiv preprint arXiv:2006.10550*, 2020.

[24] Dwarikanath Mahapatra, Behzad Bozorgtabar, and Zongyuan Ge. Medical image classification using generalized zero shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3344–3353, October 2021.

[25] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416/.

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

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

[28] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, January 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00965-w. URL http://dx.doi.org/10.1038/s42256-024-00965-w.

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

[30] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

[31] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. In *ICLR*, 2025. URL https://openreview.net/forum?id=Y2RW9EVwhT.

[32] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology. Artificial intelligence*, 1(1), 2019.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[34] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *NeurIPS*, 35:33536–33549, 2022.

[35] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

[36] Yancheng Wang, Rajeev Goel, Utkarsh Nath, Alvin C Silva, Teresa Wu, and Yingzhen Yang. Learning low-rank feature for thorax disease classification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=GkzrVxs9LS.

[37] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *ICCV*, pages 21372–21383, 2023.

[38] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. *https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation*, 2019.

[39] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022.

[40] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023.

[41] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[42] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.

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

[44] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, Jan 2022. ISSN 2522-5839. doi: 10.1038/s42256-021-00425-9. URL https://doi.org/10.1038/s42256-021-00425-9.

[45] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=w-x7U26GM7j.

A Are Attention Maps of Medical VLMs Fully Explainable?

Explainability is critical for clinical deployment of deep learning models, and medical VLMs have adopted attention maps as their de facto interpretable features [37, 19, 42]. However, attention maps alone reveal only *where* the model is focusing, not *why* it makes a particular prediction. Without the underlying image–text similarity scores, such heatmaps lack interpretability. Interpreting attention maps for complex text queries can be unintuitive and may depend on access to ground-truth labels.

To illustrate the challenges of interpreting attention maps, Figure 6 visualizes the attention outputs of CARZero [19] for different image–text pairs, with the overlaid values indicating the corresponding similarity probabilities. In (a), the prompt "There is right effusion" produces a heatmap focused on the right lower lung and a high similarity probability (0.560), intuitively linking the attended region to the predicted finding. In (b), when queried with "There is left effusion" on the same image, the attention map still highlights the right lower lung. While we can observe *where* the model attends, the attention map alone offers no clear explanation for *why* that region is relevant. The image–text similarity probability for the prompt "There is left effusion" is low (0.048), which—when considered alongside the attention map—can be interpreted as indicating that the attended right lung region does not support the presence of left-sided effusion. However, how should we interpret the fact that the model attends to the right lung when queried about *left* effusion? Making sense of this behavior requires access to the ground truth: effusion is present in the right lower lung but absent on the left. With this context, one might infer that the model correctly identifies effusion on the right, recognizes that it does not match the left-sided query, and therefore assigns a low similarity score.

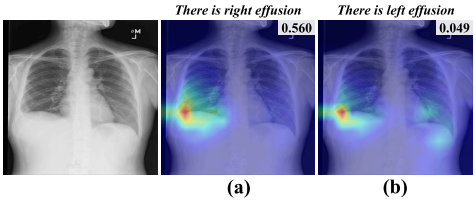

Figure 6: Attention maps from CARZero [19]. Image–text similarity probabilities are obtained by applying a sigmoid function to the classification logits.

This example underscores the need to interpret attention maps in conjunction with image–text similarity scores. For complex text queries, meaningful interpretation often depends on knowing the ground truth—a significant limitation in medical VLMs where such labels are frequently unavailable. This reliance undermines the standalone explainability of attention visualizations. To address these limitations, we propose vision–language cross-attention based on similarity (VL-CABS), a framework that enables transparent inspection of the vision–language decision process. Detailed examples and analysis of its application in RadZero are provided in Sec. 5.2.

B Ablation Studies

We conducted ablation studies to assess the impact of key design choices by selectively modifying parts of our approach. The zero-shot tasks for evaluation included classification (*class.*), grounding (*ground.*), phrase grounding (*phrase.*), and segmentation (*seg.*). Classification was tested on the PadChest dataset, known for its highly imbalanced (long-tailed) label distribution, with AUC as the evaluation metric. Grounding and phrase grounding were evaluated using the pointing game on the ChestXDet10 and MS-CXR test sets, respectively. Segmentation performance was measured by the Dice score on the SIIM dataset. For the ablation study, the default batch size and maximum number of epochs were set to 128 and 10, respectively.

View position. Table 6 shows the performance variation based on the view position of CXR images. We compared two models: one trained exclusively on frontal view images from MIMIC-CXR and another trained on both frontal and lateral views. The model trained on all view positions consistently outperformed the frontal-only model, suggesting that it effectively learned to interpret lateral images, enhancing overall robustness.

View Position	*class.*	*ground.*	*phrase.*	*seg.*
Frontal	0.831	0.604	0.838	0.161
All View	**0.841**	**0.622**	**0.844**	**0.171**

Table 6: Impact of view position.

Vision Encoder	Image Resolution	*class.*	*ground.*	*phrase.*	*seg.*
DINOv2 [27]	518	0.825	0.606	0.814	0.100
RadDINO [28]	518	**0.850**	0.610	**0.844**	0.144
XrayDINOv2 [5]	224	0.841	0.548	0.832	0.118
XrayDINOv2	518	0.841	**0.622**	**0.844**	**0.171**

Table 7: Impact of vision encoder and resolution.

Text Encoder	*class.*	*ground.*	*phrase.*	*seg.*
BioBERT	**0.842**	0.582	0.832	0.127
MPNet [30]	0.841	**0.622**	**0.844**	**0.171**

Table 9: Impact of text encoder.

Model	*class.*	*ground.*	*phrase.*	*seg.*
Linear	0.826	0.549	0.826	0.100
1 Transformer layer	0.835	0.585	0.832	0.158
2 Transformer layers	**0.841**	**0.622**	**0.844**	**0.171**

Table 8: Impact of trainable vision layer.

Batch size	*class.*	*ground.*	*phrase.*	*seg.*
64	0.835	0.583	0.826	0.165
128	0.840	0.594	**0.850**	**0.177**
256	**0.841**	**0.622**	0.844	0.171

Table 10: Impact of batch size.

Vision encoder and resolution. Table 7 presents the impact of the vision encoder and image resolution on model performance. We compared DINOv2 [27], RadDINO [28] and XrayDINOv2 [5] using image resolutions of 224 and 518. DINOv2, which was trained on natural images rather than X-rays, exhibited relatively lower performance, as expected due to the domain mismatch. Comparing XrayDINOv2 at resolutions of 224 and 518, we observe that higher image resolution improves fine-grained tasks such as grounding and segmentation. RadDINO and XrayDINOv2 showed similar performance, suggesting that our approach is effectively applied to models trained with the DINOv2 strategy on chest X-ray images.

Trainable vision layer architecture. Table 8 presents the impact of different trainable layers in the image encoder. The commonly used linear layer showed relatively lower performance across tasks. In contrast, two Transformer layers achieved the best results across all tasks. Based on this observation, RadZero was designed with two Transformer layers added to the vision encoder. This improvement is likely due to the Transformer's ability to attend to all patch embeddings, capturing richer semantic information.

Text encoder. Table 9 presents the performance of different text encoders used during training. We compared MPNet [30] and BioBERT [21], where BioBERT was fine-tuned on clinical reports by CARZero [19]. While MPNet showed slightly lower performance in classification, it achieved notable improvements in phrase grounding and segmentation, demonstrating its effectiveness in tasks requiring fine-grained text-image alignment.

Batch size. Table 10 presents the impact of batch size on model performance during training. To ensure a fair comparison, we maintained a consistent total number of training steps by adjusting the number of epochs: 5 for a batch size of 64, 10 for 128, and 20 for 256. We observed that a batch size of 64 resulted in lower performance across all tasks. While the model trained with a batch size of 128 performed reasonably well, its zero-shot grounding performance was notably lower than that of the 256 batch size model. As a result, we selected 256 as the final batch size. This trend aligns with the well-known impact of batch size in contrastive learning, where larger batch sizes generally improve representation learning by providing more diverse negative samples, leading to better alignment and discrimination.

C Model Training and Computational Details

C.1 Training Configuration

RadZero is trained for 20 epochs with an early stopping patience of 5 epochs, selecting the best model based on validation loss. We employ the AdamW optimizer with a learning rate of 0.0001, following a cosine decay scheduler, with 50 warm-up steps, a weight decay of 0.05, and gradient clipping set to 1.0. Training is conducted with a global batch size of 256 using distributed data parallel (DDP) on four Nvidia H100 GPUs for 13 hours.

C.2 Resolution Trade-off Analysis

RadZero employs high-resolution images of 518 px instead of 224 px, which naturally increases the computational cost. Table 11 summarizes the trade-off between performance and resource usage across different image resolutions. The columns with downward arrows (↓) indicate that lower values are better.

Method	**GPU Memory**↓ (GB)		**Training**	**Latency**↓	**Throughput**	**Cls**	**Grnd**	**Seg**
	Training	Inference	**GPU hour**↓	(ms/img)	(img/s)	**AUC**	**ACC**	**DICE**
RadZero (224px)	25.13 × 4	2.20	40	60.18	733.56	0.852	0.537	0.342
RadZero	73.36 × 4	2.38	52	70.05	94.93	0.851	0.622	0.359

Table 11: Performance–cost trade-off across different image resolutions.

Both models were trained using four GPUs with a batch size of 64 per device (total batch size of 256). For inference, memory usage and latency were measured with a batch size of 1. Throughput (images per second) was measured using the largest power-of-two batch size that fits into GPU memory for each model: 4096 for the 224 px model and 256 for the 518 px model. Note that throughput (img/s) and latency (ms/img) are not exact reciprocals because throughput is measured under a large-batch setting where computation is parallelized across samples, while latency reflects the time required to process a single image without such parallelism. All computational cost measurements were conducted on Nvidia H100 GPUs using the ChestXray14 dataset. The reported performance represents the average over all datasets for each task. The results indicate that the additional computational cost is a worthwhile trade-off given the substantial improvement in fine-grained performance.

D Detailed Classification Analysis

D.1 Per-Finding Classification Performance

Tables 12-14 compare RadZero and CARZero [19] in terms of per-finding classification AUCs on the OpenI, PadChest, and ChestXray14 datasets, respectively. For PadChest, which has a large number of classes, we evaluated on five representative categories commonly used. The full names for each abbreviation are provided in Table 16.

Method	Mean	ATE	CARD	EFF	INFL	MASS	NOD	PNA	PTX	EDE	EMPH	FIB	PLTH	HERN	FX	OPAC	LES	CG	LG
CarZero[19]	0.838	**0.859**	0.933	**0.938**	**0.776**	0.887	0.612	0.877	0.921	0.900	0.899	**0.917**	0.822	0.953	**0.726**	**0.784**	**0.976**	0.658	0.621
RadZero(224px)	**0.851**	0.850	**0.939**	0.937	0.774	**0.891**	**0.653**	**0.881**	**0.952**	**0.910**	0.925	0.900	**0.837**	**0.989**	0.720	**0.784**	0.970	0.700	0.700
RadZero	0.847	0.857	0.933	0.933	0.775	0.886	0.630	0.870	0.949	0.902	**0.926**	0.904	0.820	0.982	0.701	0.781	0.929	**0.731**	**0.734**

Table 12: Class-wise disease classification results on the OpenI dataset.

Method	Mean	ATE	CARD	CONS	EDE	PNA
CarZero[19]	0.810	0.835	0.906	**0.903**	0.971	0.841
RadZero(224px)	**0.841**	**0.839**	0.917	0.902	**0.973**	**0.846**
RadZero	**0.841**	**0.839**	**0.920**	0.899	0.972	0.831

Table 13: Class-wise disease classification results on the PadChest dataset.

Method	Mean	ATE	CARD	EFF	INFL	MASS	NOD	PNA	PTX	CONS	EDE	EMPH	FIB	PLTH	HERN
CarZero[19]	**0.811**	**0.819**	0.852	**0.873**	0.670	0.854	0.718	0.737	0.871	**0.786**	**0.884**	**0.808**	0.788	0.770	0.928
RadZero(224px)	0.807	0.796	**0.864**	0.857	**0.672**	**0.859**	**0.742**	**0.772**	0.873	0.784	0.883	0.631	**0.807**	**0.789**	**0.963**
RadZero	0.804	0.792	0.863	0.854	0.669	0.837	0.727	0.766	**0.875**	0.784	0.881	0.655	0.806	0.781	**0.963**

Table 14: Class-wise disease classification results on the NIH ChestXray14 dataset.

Consistent with the average performance, RadZero generally outperforms or matches CARZero on OpenI and PadChest. On ChestXray14, our model performs comparably or better on most pathologies, but a notable drop on emphysema (EMPH) accounts for the overall underperformance on the dataset. However, RadZero outperforms CARZero on EMPH in OpenI, indicating that the drop is dataset-driven rather than due to lesion-specific modeling issues. These results reaffirm that while RadZero demonstrates strong zero-shot classification, further improvements in generalization remain an important direction.

D.2 Comparison with Supervised Baselines

Table 15 compares our zero-shot classification performance against supervised baselines reported in prior work on the CheXpert and ChestXray14 datasets. For the baselines, the percentages in parentheses indicate the proportion of training data used for supervised training. Since RadZero is trained solely on MIMIC-CXR, its external test results on CheXpert and ChestXray14 are reported as single zero-shot scores rather than separate results for 1%, 10%, and 100% of the training data. Note that the results for CheXpert differ from those presented in Table 1, as a different test set was used. Specifically, we followed the test split adopted in [45] to ensure a fair comparison with other models. Although both RadZero and RadZero(224px) underperform some fully supervised models trained with 10% or 100% of labeled data, they consistently outperform models trained with 1% supervision. These results demonstrate that RadZero achieves competitive zero-shot classification performance, consistent with the segmentation results in Table 4.

Method	CheXpert			ChestXray14		
	(1%)	(10%)	(100%)	(1%)	(10%)	(100%)
ConVIRT [43]	0.870	0.881	0.881	-	-	-
REFERS [44]*	0.872	0.881	0.882	-	-	-
M3AE [11]*	0.862	0.873	0.879	-	-	-
MGCA [34]	0.888	0.891	0.897	-	-	-
MRM [45]	0.885	0.885	0.887	0.794	0.840	0.859
RadZero(224px)	0.888			0.807		
RadZero	0.889			0.804		

Table 15: Comparison of zero-shot classification performance of RadZero against supervised baselines on CheXpert and ChestXray14 datasets. Results for models marked with an asterisk (*) are taken from [45], whereas the results for all other models are reported in their respective papers. The percentages in parentheses indicate the proportion of training data used for supervised training.

E Additional Details

E.1 Abbreviations

Table 16 lists the abbreviations used in this paper for lesions and anatomical regions. The left column shows the abbreviated terms, and the right column gives their description.

Abbreviation	Description
ATE	Atelectasis
CALC	Calcification
CARD	Cardiomegaly
CG	Calcified Granuloma
CONS	Consolidation
EDE	Pulmonary Edema
EFF	Effusion
EMPH	Emphysema
FIB	Fibrosis
FX	Fracture
HERN	Hernia
INFL	Infiltration
LES	Lesion
LG	Lung Granuloma
MASS	Mass
NOD	Nodule
OPAC	Opacity
PLTH	Pleural Thickening
PNA	Pneumonia
PTX	Pneumothorax

(a) Lesion abbreviations

Abbreviation	Description
UL	Upper Lobe
ML	Mid Lobe
LL	Lower Lobe
Rclav	Right Clavicle
Lclav	Left Clavicle
RULZ	Right Upper Lung Zone
RMLZ	Right Mid Lung Zone
RLLZ	Right Lower Lung Zone
LULZ	Left Upper Lung Zone
LMLZ	Left Mid Lung Zone
LLLZ	Left Lower Lung Zone
RCPA	Right Costophrenic Angle
LCPA	Left Costophrenic Angle
HD	Hemidiaphragm
RHD	Right Hemidiaphragm
LHD	Left Hemidiaphragm

(b) Anatomical region abbreviations

Table 16: Abbreviations for lesions and anatomical regions.

E.2 Prompt for finding-sentence extraction.

As shown in Figure 7, the prompt instructs the LLM to extract clinically relevant minimal semantic units in the form of sentences from radiology reports. Finding-sentences are standardized through prompt alignment to follow a "There is" format, with a one-shot example enhancing extraction accuracy and guiding the model to identify both findings and their corresponding anatomical locations in a structured manner.

```
You are an expert medical assistant AI specializing in understanding and analyzing chest x-ray radiology reports.

Your task is to extract the medically significant and meaningful findings from the given chest x-ray report, focusing on identifying phrases or expressions that describe notable conditions or abnormalities. Note that the report may reference previous studies, but we only need an interpretation based on the current chest x-ray. Therefore, remove and rewrite terms like "new", "improved", "unchanged", "worsened" or "consistent" to reflect the current status in a way that indicates the condition exists as observed in this image, without implying any comparison to prior images or studies.

The template format includes:
"There is [finding] of [location]."
"There may be [finding] of [location]."
"There is no [finding] of [location]."

[finding] represents the extracted key findings from the radiology report, and [location] represents the anatomical location mentioned in the report. If no location is provided, do not include it in the output. Adhere strictly to the following JSON format for the final output, using examples as a guideline for the desired analysis structure. Do not provide any explanations; output only in JSON format. If the report does not contain any findings, output an empty list (example: {"finding_sentence": []}).

[Example]
INPUT:
Cardiomegaly is accompanied by improving pulmonary vascular congestion and decreasing pulmonary edema. Left retrocardiac opacity has substantially improved, likely a combination of atelectasis and effusion. A more confluent opacity at the right lung base persists, and could be due to asymmetrically resolving edema, but pneumonia should be considered in the appropriate clinical setting. Small right pleural effusion is likely unchanged, with pigtail pleural catheter remaining in place and no visible pneumothorax.

OUTPUT:
{
  "finding_sentence" : [
    "There is cardiomegaly with pulmonary vascular congestion", "There is pulmonary edema", "There is left retrocardiac opacity",
    "There may be atelectasis", "There may be effusion", "There is right lung base opacity", "There is right lung base opacity suggestive of possible pneumonia",
    "There may be small right pleural effusion", "There is pigtail pleural catheter in place", "There is no pneumothorax"
  ]
}
```

Figure 7: Prompt design for extracting finding-sentences with LLM.

F Linguistic Robustness and Generalization Analysis

To evaluate the linguistic robustness of RadZero to variations in report templates and phrasing patterns, we conducted a series of qualitative analyses using similarity search within a vector database of training sentences encoded by RadZero's text encoder. In Table 17 (a), given a query sentence "the lungs are clear", the model retrieved multiple semantically similar sentences with varied phrasing and syntactic structures. These results indicate that the encoder captures semantic equivalence beyond fixed textual templates, such as recognizing "the chest is clear" or "the airways are clear" as close variants.

Query	Retrieved Sentences (Top-k)	Similarity
(a) **the lungs are clear**	The lungs are clear	1.000
	Lungs are clear	0.994
	The airways are clear	0.993
	There is clear lungs	0.985
	There are clear lungs	0.983
	The chest is clear	0.982
(b) **pleural effusion in the right lower lung**	There is pleural effusion in the right lower lung	0.993
	There is a pleural effusion in the right lower lung	0.978
	There is pleural effusion of the right lower lung	0.976
	There is effusion in the right lower lung	0.968
	There is a pleural effusion of the right lower lung	0.965
	There is right lower lung pleural effusion	0.953
(c) **there is fibrosis**	There is fibrosis	1.000
	There is probable fibrosis	0.937
	There is lung fibrosis	0.937
	There is chronic fibrosis	0.934
	There is a component of fibrosis	0.924
	There is fibrotic disease	0.917
(d) **pleural effusion in RLL (unseen)**	There is pleural effusion with drainage	0.753
	There is slight decrease in pleural fluid	0.725
	There is collecting pleural fluid	0.708

Table 17: Examples of similarity search results using RadZero's text encoder. Cell colors indicate cosine similarity: high (blue) ≥ 0.95, medium (yellow) ≥ 0.85, and low (red) < 0.85.

We further tested syntactic diversity using queries with complex structures (e.g., "pleural effusion in the right lower lung"). As shown in Table 17 (b), the retrieved results included expressions such as "right lower lung pleural effusion" and "pleural effusion of the right lower lung," demonstrating the model's robustness to grammatical variations. Similarly, queries evaluating lexical flexibility (e.g., "there is fibrosis" in Table 17 (c)) showed that RadZero identifies semantically related expressions such as "chronic fibrosis" and "fibrotic disease," suggesting strong generalization across vocabulary variations. These

findings collectively suggest that RadZero is not confined to specific textual templates or vocabularies and generalizes well across diverse linguistic patterns, thereby mitigating the risk of overfitting to particular report styles.

However, we observed that embeddings may be less accurate for report expressions that were rare or absent in the training data. One notable case involves abbreviations-for example, "RLL" (right lower lung)-which occurred infrequently in the original reports. As shown in Table 17 (d), such unseen abbreviations tend to yield weaker semantic alignment. To address this limitation, non-standardized or abbreviated expressions can be rewritten into complete finding-sentences using LLM-based normalization, which helps maintain robustness across diverse report formats.

G Statistical Significance of Main Results

To assess run-to-run variability and verify the statistical reliability of the reported performance, we conducted additional experiments across five random seeds for all tasks. For each model and task, we report the mean and standard deviation of the evaluation metrics.

Method	Classification								Grounding		Segmentation	
	OI	PC	PC20	CXR14	CXP	CXD10	SIIM	RSNA	CXD10	MS-CXR	SIIM	RSNA
RadZero	0.847	0.841	0.871	0.804	0.900	0.787	0.924	0.834	0.622	0.844	0.171	0.546
Mean (±)	0.848 (±)	0.840 (±)	0.869 (±)	0.802 (±)	0.899 (±)	0.787 (±)	0.918 (±)	0.843 (±)	0.601 (±)	0.845 (±)	0.164 (±)	0.549 (±)
std	0.0016	0.0011	0.0040	0.0016	0.0019	0.0043	0.0063	0.0072	0.0130	0.0143	0.0080	0.0023

Table 18: Mean and standard deviation of main results over five runs.

As shown in Table 18, the variance across runs is relatively small, indicating that the observed improvements are consistent and not due to random fluctuations. These results confirm the stability and robustness of RadZero across different random initializations.

H Additional Visualization Results

Figure 8 depicts segmentation of anatomical regions, which, while not perfect, generally align with appropriate locations. Figure 9 presents examples demonstrating RadZero's potential for open-vocabulary semantic segmentation, including additional lesion types such as mass, fibrosis, and calcification. Full names of lesion and anatomical region abbreviations are provided in Table 16.

Figure 10 presents VL similarity maps for 10 different findings of the ChestXDet10 dataset, following the pipeline in Sec. 5.2. The brightest regions in the map align well with the bounding boxes, even for multiple or small lesions. The similarity probability was above 0.5 for all findings except calcification. While the model correctly localized calcifications, the activated regions appeared as small bright spots, leading to a lower similarity probability of 0.45 due to the weighted sum calculation. This highlights a limitation of RadZero, suggesting the need for further refinement in future work.

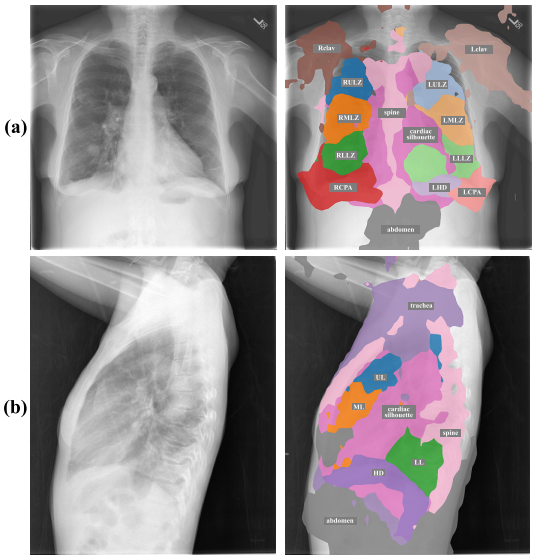

Figure 8: Open-vocabulary semantic segmentation for anatomical regions. The CXR images are sourced from Open-I. The segmentation threshold was set to 0.4.

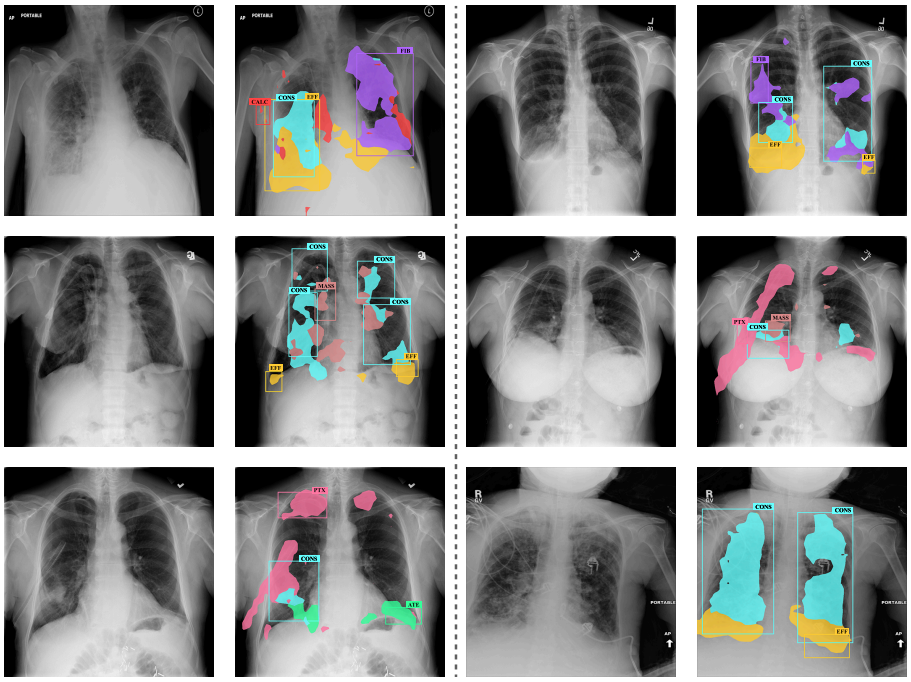

Figure 9: Open-vocabulary semantic segmentation for findings. The CXR images and bounding box labels are sourced from ChestXDet10. The segmentation threshold was set to 0.7.

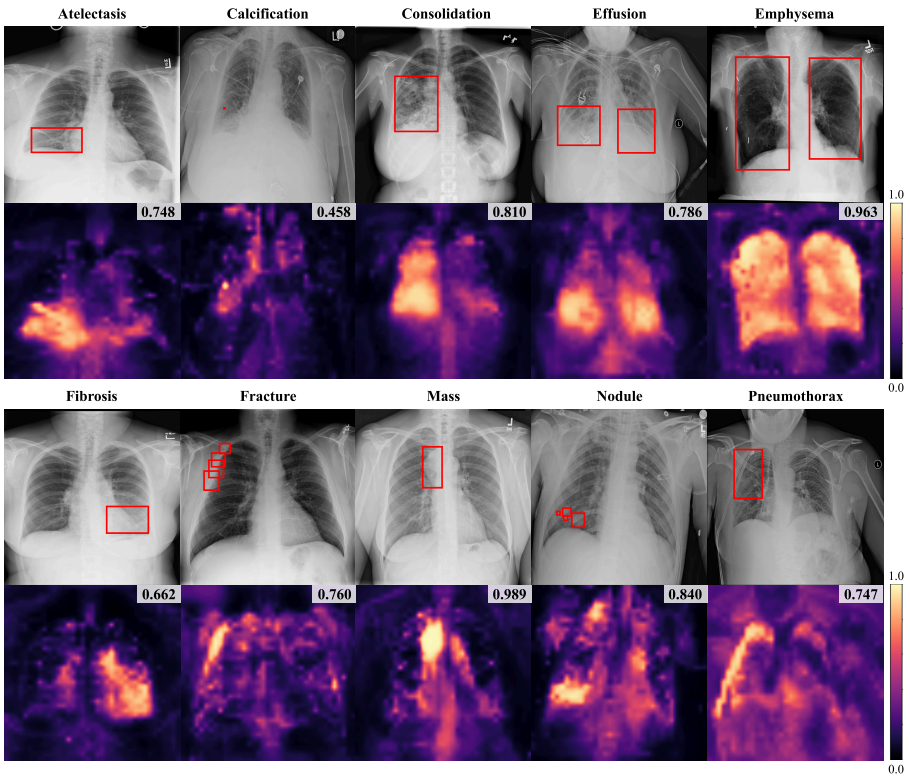

Figure 10: Visualization of VL similarity maps generated by RadZero for 10 findings on the ChestXDet10 dataset. Red boxes indicate ground truth bounding boxes. The similarity probability $\hat{l}$ is shown in the top-right corner of each map.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's scope by highlighting RadZero and its core component, VL-CABS, which computes text–image similarity for interpretable, fine-grained vision-language alignment. They also clearly describe RadZero's zero-shot capability across multiple tasks.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses RadZero's limitations in each section including Sec. 6 Conclusion: performance gaps on certain datasets (Sec. 5.1), prediction of regions extending beyond the ground truth bounding box in open-vocabulary segmentation (Sec. 5.3), and reliance on a pre-trained vision encoder (Sec. 5.4), suggesting directions for future improvement.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This work mainly includes empirical contributions.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental configurations in Sections 4.4 and Appendices C.1, E.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our experiments are all conducted on publicly accessible datasets, and the details about the datasets used are described in Sections 4.1 (training data) and 4.2 (test data). For experiment implementation, we follow the official code of exisiting works, all code can be found in their official GitHub repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed for experimental configurations in Section 4.4 and Appendices C.1, E.2, and for the data split in Sections 4.1 (training data) and 4.2 (test data).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean and standard deviation across five random seeds in Appendix G, ensuring statistical significance and reliability of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information on the computation resources is described in Appendix C.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research was conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential societal impacts are mentioned in Sections 1 and 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work does not release data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please refer to Sections 4.1 and 4.2.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There is no new assets released in this work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work has no human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work has no human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This work have used a publicly available LLM for data processing. To prevent any potential misuse, the LLM was executed on a secure, private server.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.