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

Presenter: **Jonggwon Park**

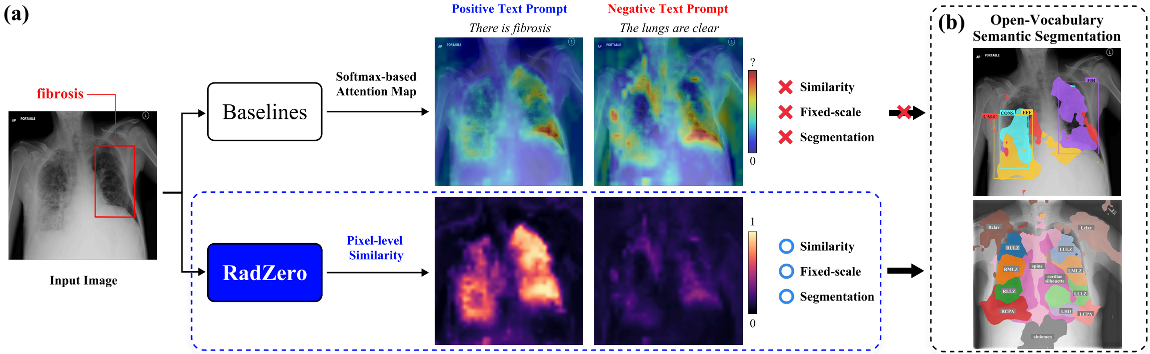
Jonggwon Park   Byungmu Yoon   Soobum Kim   Kyoyun Choi*

DEEPNOID Inc.
Seoul, South Korea

NEURAL INFORMATION PROCESSING SYSTEMS

DEEP NOID

Attention Maps vs. Vision-Language Similarity Maps



Problem Definition

Example: Chest X-ray radiology report

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.

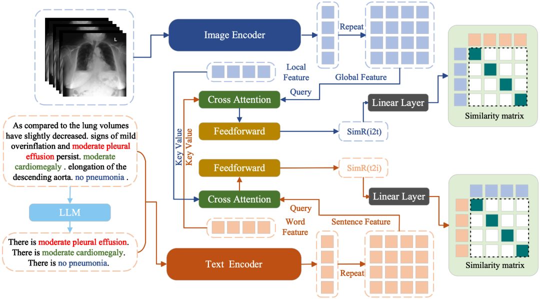
Characteristics of Radiology Reports

- Radiology reports contain diverse types of information - clinical history, observations, comparisons with prior exams, and diagnostic impressions.
- Each **finding description** in the report often corresponds to a **localized region** in the medical image.

Key Challenges

- Effectively utilizing these **complex and noisy free-text reports** as supervision signals
- Enabling the model to **learn the correspondence between textual descriptions and local image regions**

Related Works: CARZero (*CVPR* 2024)



Key Idea

- Reformulates reports into structured prompts, e.g., *"There is [disease]."*
- Proposes cross-attention alignment between vision and language features:
  - Global image embedding ↔ Local word embedding
  - Global sentence embedding ↔ Local image embedding

Limitations

- Uses **random selection** when forming image–sentence training pairs
- Utilizes **softmax attention probabilities** for visualization and grounding tasks

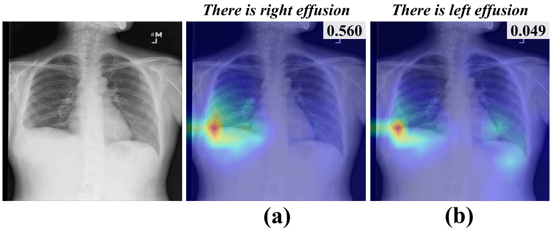
† H. Lai *et al.*, "CARZero: Cross-Attention Alignment for Radiology Zero-Shot Classification," *CVPR*, 2024.

Are Attention Maps of Medical VLMs Explainable?

Key Points

- Medical VLMs (e.g., CARZero, MedKLIP, KAD) often adopt **attention maps** as interpretable features.
- However, attention maps only reveal **where** the model focuses - not **why** it makes a certain decision.
- Without image-text similarity scores, such maps lack true interpretability and may depend on access to ground-truth labels.

Example: CARZero Attention Behavior



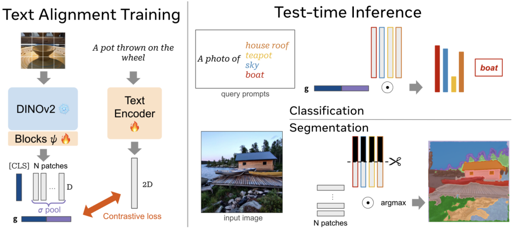
Attention maps and image-text probabilities from CARZero

- Query: *"There is right effusion"* (**0.560**) → highlights right lung
- Query: *"There is left effusion"* (**0.049**) → still attends to right lung
- **Ground truth:** right effusion
- Attention maps show **where** the model focuses, but not **why** - interpretation requires ground truth.

Related Works: dino.txt (*CVPR* 2025)

Key Idea

- **CLIP**-style models align text and images but learn *global* alignment, thus weak on dense understanding.
- **SSL** models (e.g., DINOv2) capture fine-grained details but lack a language interface.
- Uses shallow transformer blocks and *average pooling* over patch embeddings to connect SSL features to text for alignment.
- Enables **open-vocabulary** classification and segmentation at test time.



Limitations

- Pooling patch embeddings → loss of **locality** (spatial information).
- In medical imaging, simple average pooling is insufficient to capture local context (empirically observed).

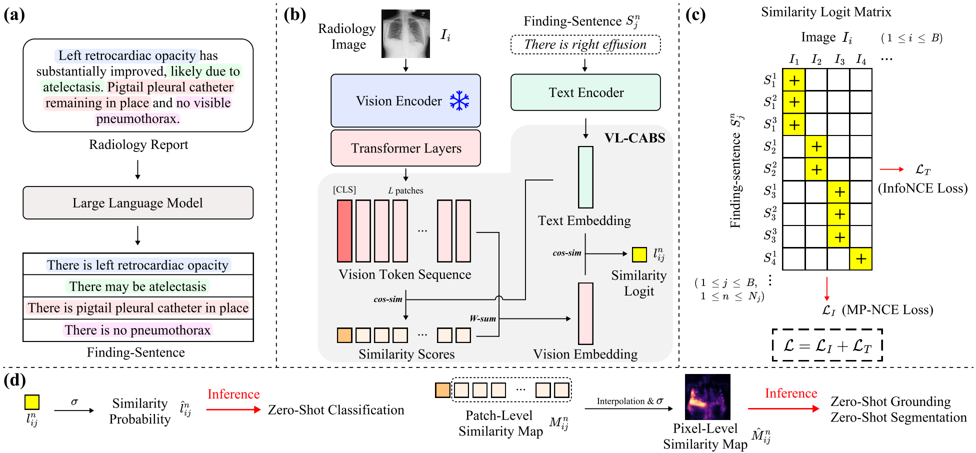
†C. Jose *et al.*, "DINOv2 Meets Text: A Unified Framework for Image- and Pixel-Level Vision-Language Alignment," *CVPR*, 2025.

Contributions

- We propose **RadZero**, a novel vision-language (VL) alignment framework for chest X-rays, designed with **zero-shot multi-task capability**.
- We introduce **VL-CABS** (Vision-Language Cross-Attention Based on Similarity), which computes **cosine similarity** directly between text descriptions and local image patches, producing **interpretable VL similarity maps**.
- We employ **multi-positive contrastive learning**, incorporating multiple sentences per image–report pair to provide **richer supervision**.
- RadZero achieves **state-of-the-art zero-shot performance** across public chest X-ray benchmarks, while enhancing **explainability** and enabling **open-vocabulary semantic segmentation**.

DEEP NOID

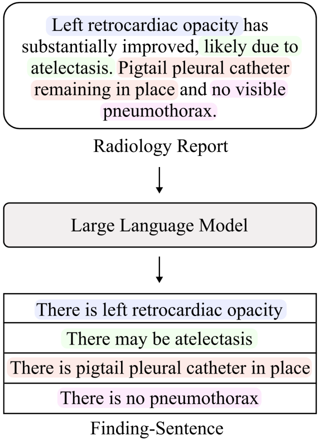
Method: RadZero Overview



DEEP NOID

Method: Finding-Sentence Extraction with LLM

(a)



Finding-Sentence Extraction with LLM

- **Input:** Radiology report → extract **finding-sentences**

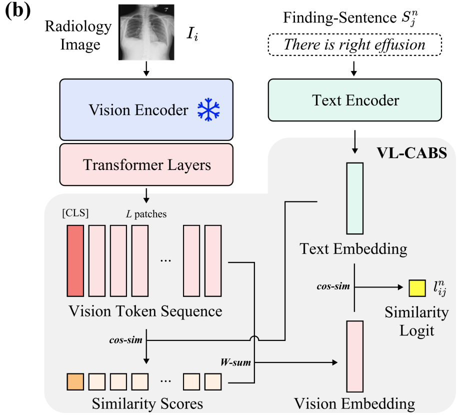
Sentence Templates:

- *There is* {`finding, location`}
- *There may be* {`finding, location`}
- *There is no* {`finding, location`}

Prompting

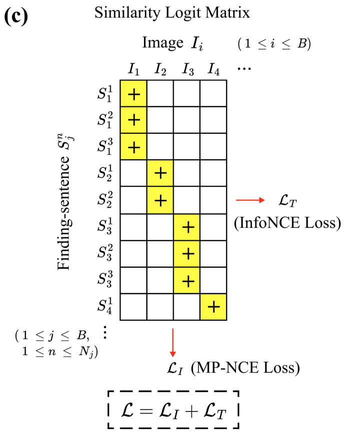
- The LLM is guided to **decompose reports into minimal semantic units** corresponding to distinct clinical findings.
- Temporal expressions (*e.g., new, improved, unchanged, worsened, consistent*) are explicitly excluded to ensure static descriptions.

Method: Vision–Language Cross-Attention Based on Similarity (VL-CABS)



- In a mini-batch of size B, each image I_i $(i = 1, \ldots, B)$ is paired with its finding sentences $\{S_i^n\}_{n=1}^{N_i}$.
- Encode features using vision and text encoders:
  - Vision token sequence: $[v_{i0}, v_{i1}, \ldots, v_{iL}] = f_a(f_v(I_i))$ where f_v is the **vision encoder** and f_a denotes the **transformer layers**.
  - Text embedding: $t_j^n = f_t(S_j^n)$ where f_t is the **text encoder**.
- Compute the **scaled cosine similarity** for each image patch: $s_{ijk}^n = \cos(v_{ik}, t_j^n)/\tau$, where τ is the temperature parameter.
- Apply **softmax** over patches to obtain attention probabilities: $a_{ijk}^n = \exp(s_{ijk}^n)/\sum_{m=0}^{L} \exp(s_{ijm}^n)$.
- Form a **text-aware vision embedding** by the weighted sum of patch tokens: $v_{ij}^n = \sum_{k=0}^{L} a_{ijk}^n v_{ik}$.
- Compute the final **similarity logit**: $l_{ij}^n = \cos(v_{ij}^n, t_j^n)/\tau$.

Method: RadZero Loss Function



- For each image I_i, there are multiple positive finding-sentences S_i^n (N_i positives in total).
  - Total number of finding-sentences in the batch: $N_T = \sum_{i=1}^{B} N_i$
  - Negatives for each image: $N_T - N_i$
- Apply **multi-positive NCE (MP-NCE)**† loss, treating each positive sentence independently:

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

- For each finding-sentence S_i^n, one positive image I_i is used, following the standard **InfoNCE** loss:

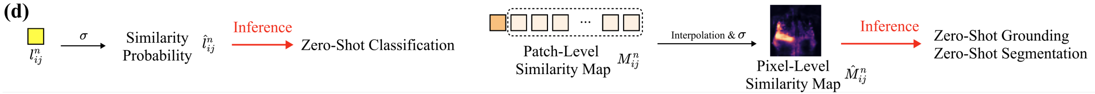
$$\mathcal{L}_T = -\frac{1}{N_T} \sum_{i=1}^{B} \sum_{n=1}^{N_i} \log \frac{\exp(l_{ii}^n)}{\exp(l_{ii}^n) + \sum_{j \neq i}^{B} \exp(l_{ji}^n)}$$

- The overall training objective combines both terms:

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_T$$

† J. Lee *et al.*, "UniCLIP: Unified Framework for Contrastive Language-Image Pre-training," *NeurIPS*, 2022.

Method: Zero-shot Inference



Similarity Probability

- Compute similarity logit between image I_i and sentence S_j^n: $l_{ij}^n = \cos(v_{ij}^n, t_j^n)/\tau$
- Sentences are constructed from simple templates, e.g., *"There is* $\{$*finding*$\}$*"*.
- Convert to probability for **zero-shot classification**: $\hat{l}_{ij}^n = \sigma(l_{ij}^n)$
- Represents the confidence of image–text alignment.

Pixel-level Similarity Map

- Patch-level similarities: $M_{ij}^n = [\, s_{ij1}^n, \ldots, s_{ijL}^n \,]$
- Reshape into a $\sqrt{L} \times \sqrt{L}$ grid, upsample via **bilinear interpolation**, and apply element-wise sigmoid to obtain: $\hat{M}_{ij}^n = \sigma(\text{bilinear}(M_{ij}^n))$
- This produces a **pixel-level vision–language similarity map**.
- Enables **zero-shot grounding** and **zero-shot segmentation**.

Experiments

Training

- **Dataset: MIMIC-CXR**
  - 377,110 chest X-ray images
  - 227,835 studies, 65,379 patients
  - On average, **6.45 finding-sentences per image**

Implementation Details

- Vision encoder: **XrayDINOv2** (CXR-pretrained DINOv2, frozen), pretrained at 224^2, used at 518^2 resolution
- Text encoder: **MPNet** (pretrained SentenceBERT)
- Transformer layers: **2 layers**

Test Dataset

- Public benchmark datasets:
  - Open-I, PadChest, ChestXray14, CheXpert, ChestXDet10, MS-CXR, SIIM, RSNA
- Includes classification, bounding box, and segmentation labels

Evaluation Metrics

- AUC (classification)
- Pointing Game (localization accuracy)
- Dice Score (segmentation overlap)
- Pixel-wise AUC (dense correspondence; alignment between similarity map and mask)

Results: Zero-shot Classification

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 [36]	0.759	0.629	0.688	0.726	0.879	0.713	0.897	**0.869**
KAD [41]	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

- **RadZero** achieves the **best performance** on **Open-I**, **PadChest**, and **SIIM** benchmarks.
- Especially, it shows a **remarkable improvement on PadChest20**, which evaluates long-tail 20-class disease distributions.

DEEP NOID

Results: Zero-shot Grounding

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 [41]	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 [36]	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**

Zero-shot Grounding on ChestXDet10

Method	MS-CXR
BioViL-T [1]	0.719
MedKLIP [36]	0.407
CARZero [19]	0.749
RadZero (224px)	0.832
RadZero	**0.844**

Zero-shot Phrase Grounding on MS-CXR

- On **ChestXDet10**, RadZero records the highest mean Pointing Game accuracy (**0.622**), indicating more precise localization for disease classes.
- On **MS-CXR**, it achieves **0.844**, demonstrating strong zero-shot phrase grounding ability.

DEEP NOID

Results: Zero-shot Segmentation

Method	RSNA	SIIM	
	Dice	Dice	Pix-AUC
GLoRIA [12]	0.347*	-	-
BioViL [3]	0.439*	-	-
MedKLIP [36]	0.465*	0.044	0.648
G2D [21]	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 [33] (1%)	0.513	0.144	0.752
MGCA (10%)	0.571	0.238	0.856
MGCA (100%)	0.578	0.305	0.976

- **RadZero** shows the highest zero-shot segmentation performance on both **RSNA** and **SIIM**.
- On **SIIM**, it achieves a Pixel-level AUC of **0.947**, surpassing the fully-supervised **MGCA (10%)** model.
- Thanks to the **VL-CABS** structure, RadZero attains strong pixel-level performance **without any pixel-level supervision.**

Ablation Studies

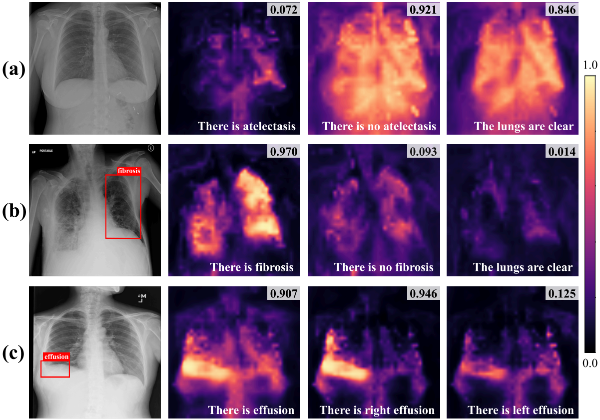
Method	Similarity	Trainable layers	MP	Res.	Classification						Grounding		Segmentation
					OI	PC	PC20	CXR14	CXP	CXD10	CXD10	MS-CXR	SIIM
(a)	dot-product	Linear	✗	224	0.839	0.824	0.853	0.805	0.896	0.792	0.472	0.784	0.078
(b)	cos	Linear	✗	224	0.843	0.830	0.863	0.805	0.902	0.786	0.483	0.790	0.078
(c)	cos	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

- **Similarity Function:** Cosine similarity $\Rightarrow$ better alignment with vision-language similarity maps.
- **Trainable Layers:** 2-layer Transformer (frozen encoder) $\Rightarrow$ improved grounding and segmentation.
- **Multi-Positive Pairs:** richer supervision $\Rightarrow$ enhanced classification and segmentation performance.
- **Image Resolution:** 518px inputs $\Rightarrow$ substantial gains in grounding and segmentation.

Comparison with prior VL alignment:

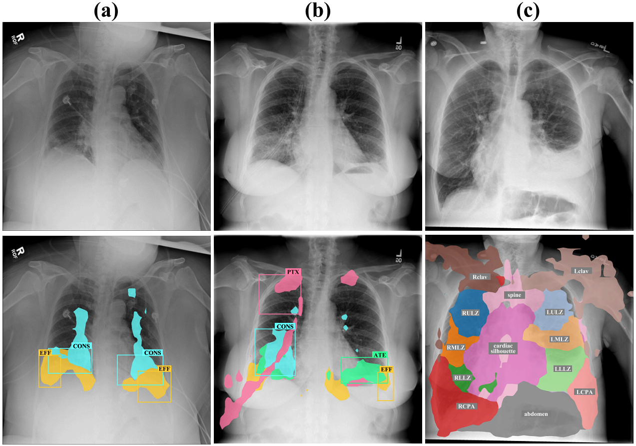
- **LiT:** low overall performance.
- **dino.txt:** improved classification, but limited spatial grounding.
- **CARZero:** overall performance improved, but still notably below RadZero.

Pixel-level Vision–Language Similarity Map Analysis



- **VL-CABS** produces **pixel-level similarity maps** aligning visual features with textual descriptions.
- Enables spatially grounded and interpretable reasoning without pixel-level supervision.
- Improves transparency by explicitly showing how text descriptions align with image regions.

Open-vocabulary Semantic Segmentation



(a), (b): Disease segmentation,
(c): Anatomical region segmentation

- **Pixel-level similarity maps** enable segmentation through simple thresholding.
- Characterized by the ability to infer disease and anatomical locations **without pixel-level supervision**.
- While the predictions are **approximate**, **RadZero** show strong potential for **open-vocabulary segmentation**.

DEEP NOID

Conclusion and Future Work

Summary

- **RadZero:** a novel vision–language alignment model for chest X-rays.
- **VL-CABS:** computes patch-level image–text similarity for interpretable alignment.
- Achieves strong zero-shot performance on classification, grounding, and segmentation.
- Provides explainable visual–text alignment without pixel-level supervision.

Limitations

- Depends on a **pretrained vision encoder**.
- **Limited to chest X-ray** datasets.

Future Work

- Extend RadZero beyond vision–language **alignment** toward **pretraining**.
- Verify **RadZero framework's generalizability** to other modalities such as **CT** and **MRI**.

Thank you!

Paper

https://openreview.net/pdf?id=WQq5JPGQ0C

Code

https://github.com/deepnoid-ai/RadZero