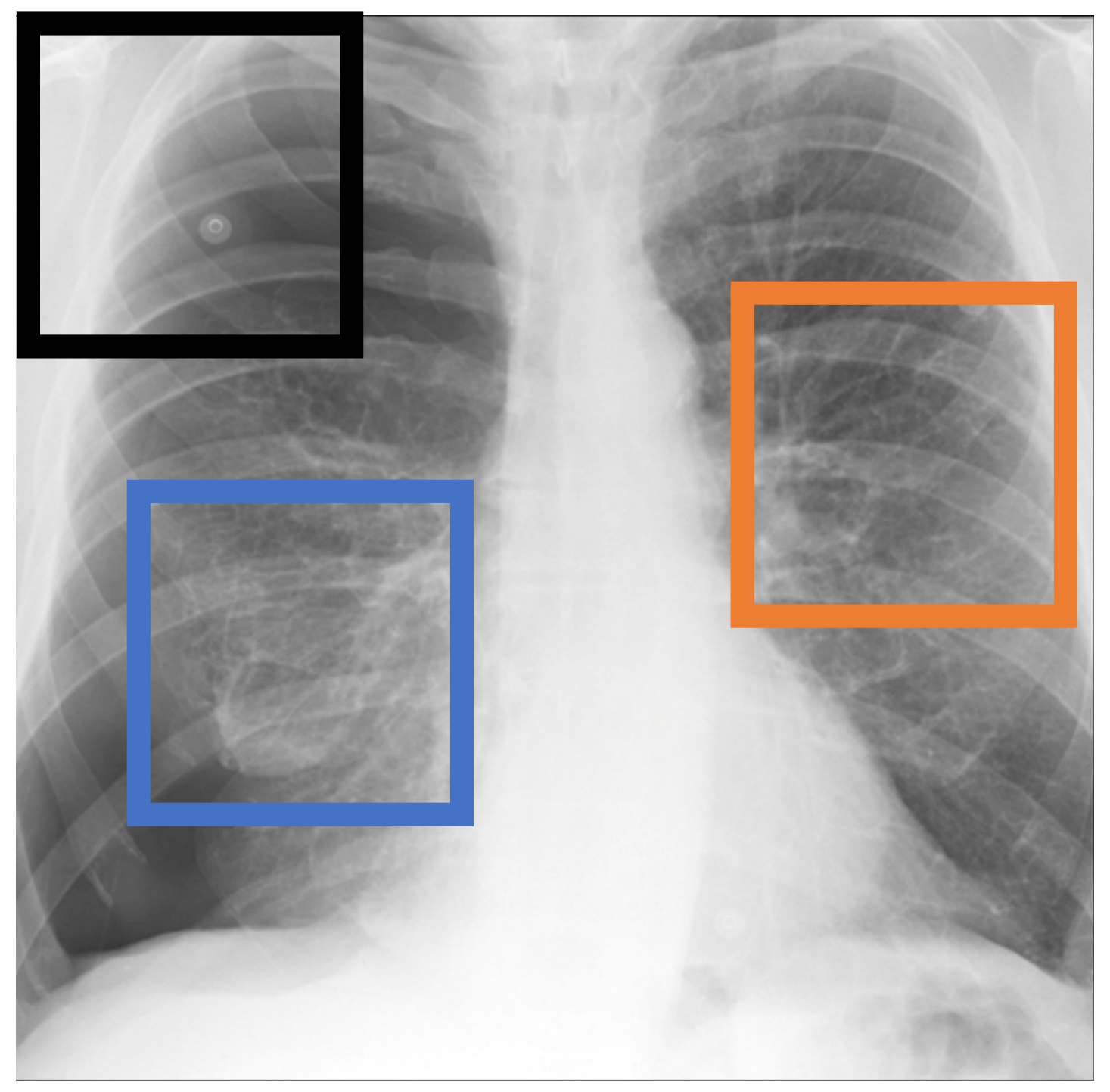
Using Multiple Instance Learning to Build Multimodal Representations

Peiqi Wang, William M. Wells, Seth Berkowitz, Steven Horng, and Polina Golland



"There is large right-sided pneumothorax." "The left lung is hyperinflated with evidence of emphysema."

Goal: Find better local (region, sentence) correspondence

Contributions

1. Connects contrastive learning with multiple instance learning.
2. A conceptual framework to think about existing contrastive learning approaches.
3. A concrete instantiation, LSE+NL, that achieves strong performance.

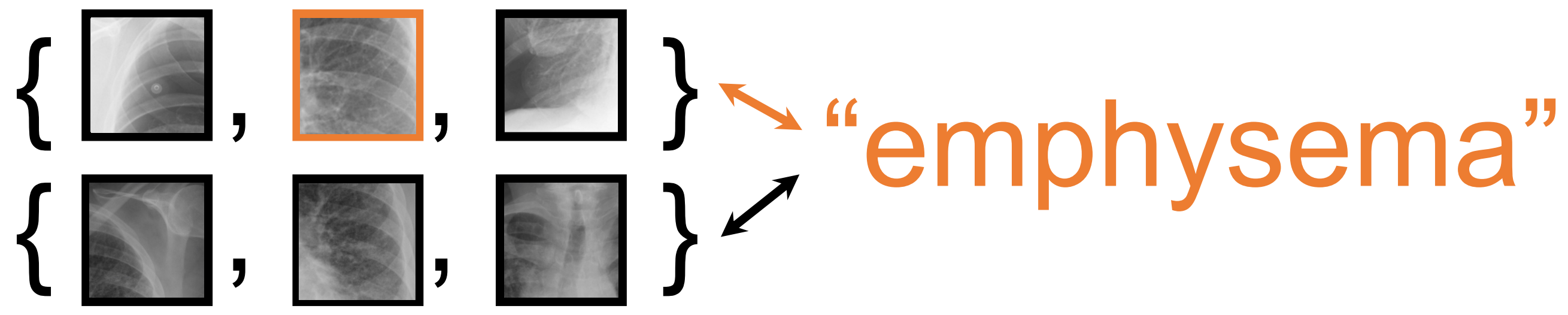
Idea: Connect contrastive learning with multiple instance learning!

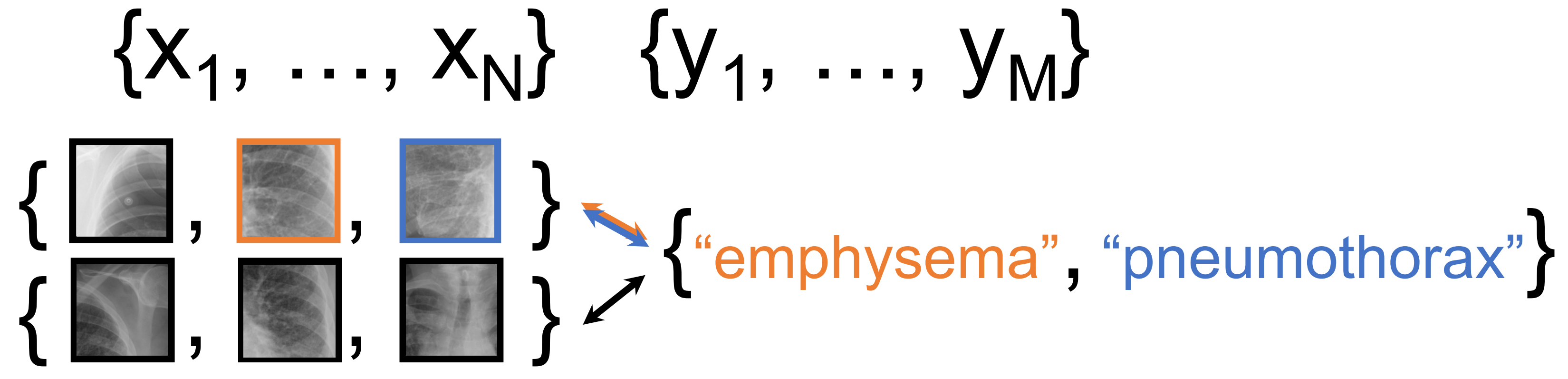
Multiple instance learning

$\{x_1, x_2, x_3\}$, +1

$\{x_1, x_2, x_3\}$, -1

Contrastive learning





Method: Compute image-document score $S(\{x_n\}, \{y_m\})$ and maximize the likelihood of correctly matched pairs.

Abstract Framework

similarity function

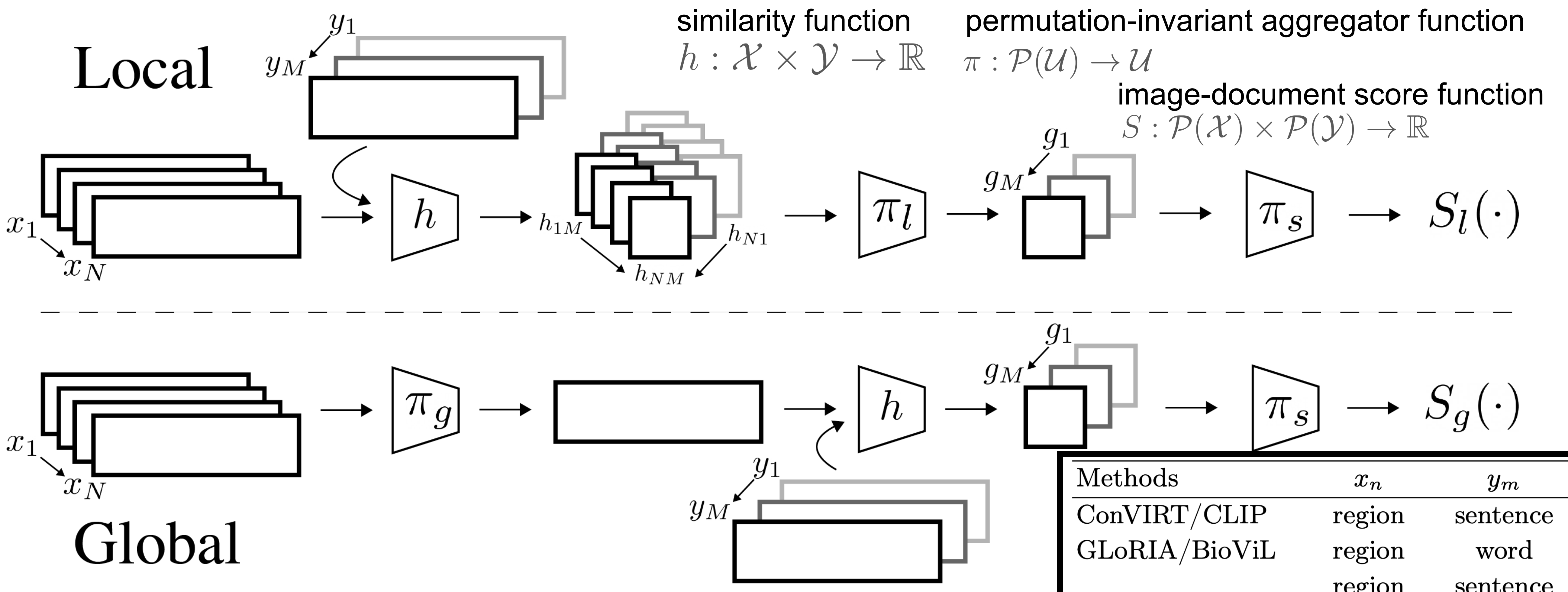
$h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

permutation-invariant aggregator function

$\pi : \mathcal{P}(\mathcal{U}) \to \mathcal{U}$

image-document score function

$S : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$



Concrete Instantiation: LSE+NL

Fix h be cosine similarity and π_s be average.

Log-Sum-Exp (LSE):

$$\pi_l(\{h_n\}) = \log \sum_{n=1}^{N} \exp(h_n)$$

Non-Local (NL):

$$\pi_g(\{x_n\}) = \sum_{n=1}^{N} \frac{\exp(\langle x_n, x_k \rangle)}{\sum_{n'=1}^{N} \exp(\langle x'_n, x_k \rangle)} x_n$$

k is index of the critical region, i.e., $k = \arg\max_n h(x_n, y_m)$

Methods	x_n	y_m	π_l	π_g	π_s
ConVIRT/CLIP	region	sentence	-	NeuralNet ∘ Avg	Id
GLoRIA/BioViL	region	word	-	Cross Attention	LSE
	region	sentence	-	Avg	Id
LSE+NL (Ours)	region	sentence	LSE	-	Avg
	region	sentence	-	NL	Avg

Table: Taxonomy of related methods for image-language representation learning in out multiple instance learning inspired framework.

Text-to-image Contrastive Loss:

Minimize $\mathcal{L}(s_l) + \mathcal{L}(s_g)$ with $s = (s^+, s_1^-, \cdots, s_K^-)$ image-document score vector

$$\mathcal{L}(s) = -log \frac{\exp(s^+)}{\exp(s^+) + \sum_{k=1}^{K} \exp(s_k^-)}$$

Results: Our proposed method LSE+NL achieves state-of-the-art performance on zero-shot image classification, visual grounding, and cross-modal retrieval.

Image Classification on RSNA Pneumonia

- zero-shot or fine-tuned with 1% or 100% of labels.
- ACC is accuracy; AUC is area under curve.
- LSE+NL compares favorably to BioViL[1]

Method	Zero-Shot		1%		100%	
	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑
BioViL[1]	0.73	0.83	0.81	**0.88**	0.82	**0.89**
LSE+NL	**0.80**	**0.84**	**0.84**	0.87	**0.85**	**0.89**

Visual Grounding on MS-CXR

- CNR or contrast-noise ratio is a measure of discrepancy between region-sentence scores (heatmap) inside vs. outside the bbox; mIoU measures how well the thresholded heatmap overlap with the bbox.
- LSE+NL outperforms BioViL[1] on both measures.

Method	CNR↑	mIoU↑
BioViL[1]	1.14	0.17
LSE+NL	**1.44**	**0.19**

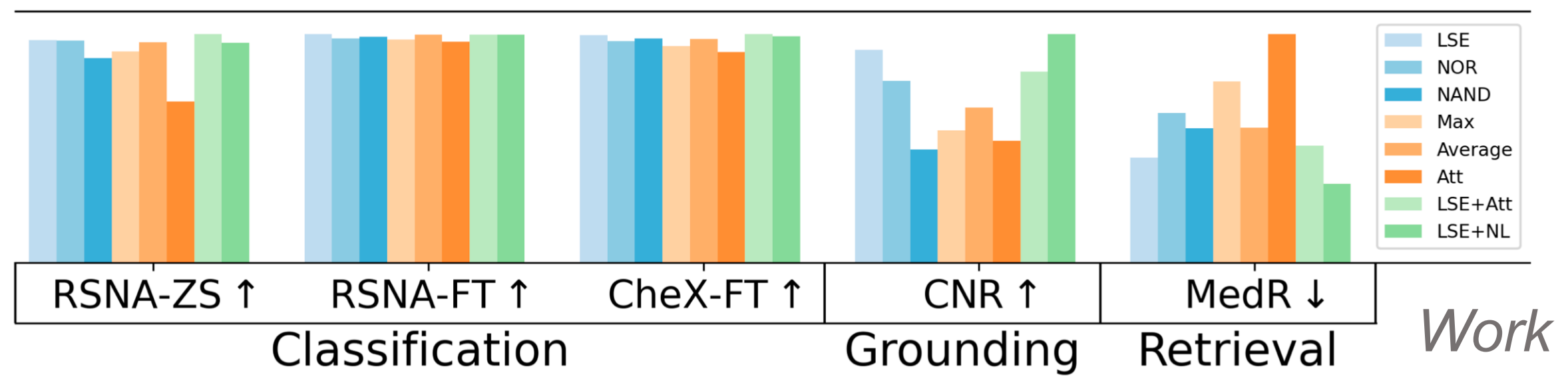
Cross-Modal Retrieval on MS-CXR

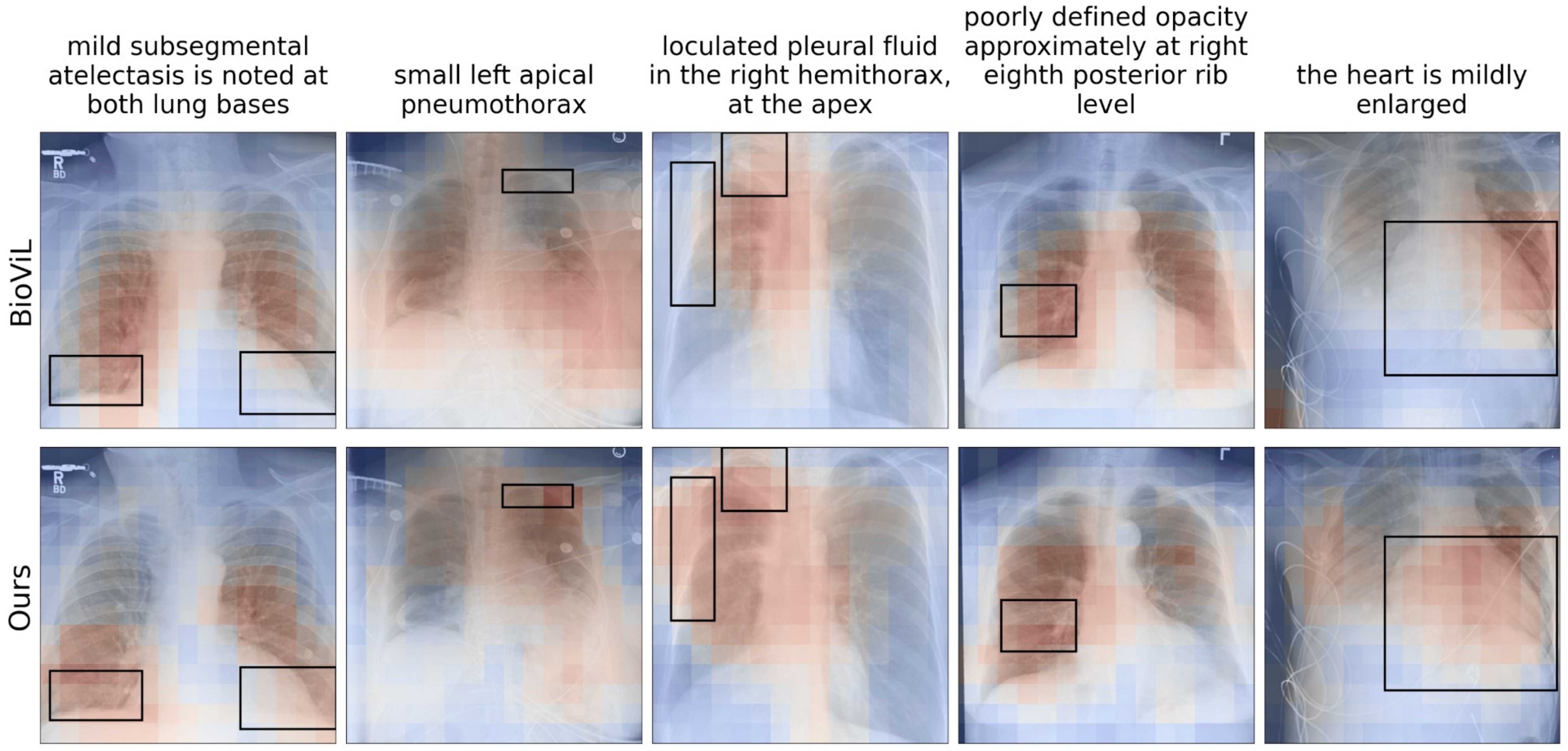
- R@K is the fraction of times the correct item was found in the top K results; MedR is the median rank of the correct item in the ranked list.
- LSE+NL outperforms BioViL[1] and GLoRIA[2].

Method	Region → Sentence				Sentence → Region			
	R@10↑	R@50↑	R@100↑	MedR↓	R@10↑	R@50↑	R@100↑	MedR↓
GLoRIA[2]	0.06	0.21	0.37	162	0.06	0.21	0.34	183
BioViL[1]	0.07	0.26	0.40	151	0.08	0.26	0.40	146
LSE+NL	**0.11**	**0.29**	**0.45**	**119**	**0.11**	**0.36**	**0.51**	**97**

Effects of aggregator choice on performance

- Effects image classification much less than other tasks.
- High performance variations within each group.
- Combination approaches do well on all tasks.




mild subsegmental atelectasis is noted at both lung bases
small left apical pneumothorax
loculated pleural fluid in the right hemithorax, at the apex
poorly defined opacity approximately at right eighth posterior rib level
the heart is mildly enlarged
BioViL
Ours

Black box: ground truth bounding box. **Heatmap**: up-sampled region-sentence score.

[1] Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing.
[2] GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition

Work supported by MIT Jclinic, Philips, and Wistron.