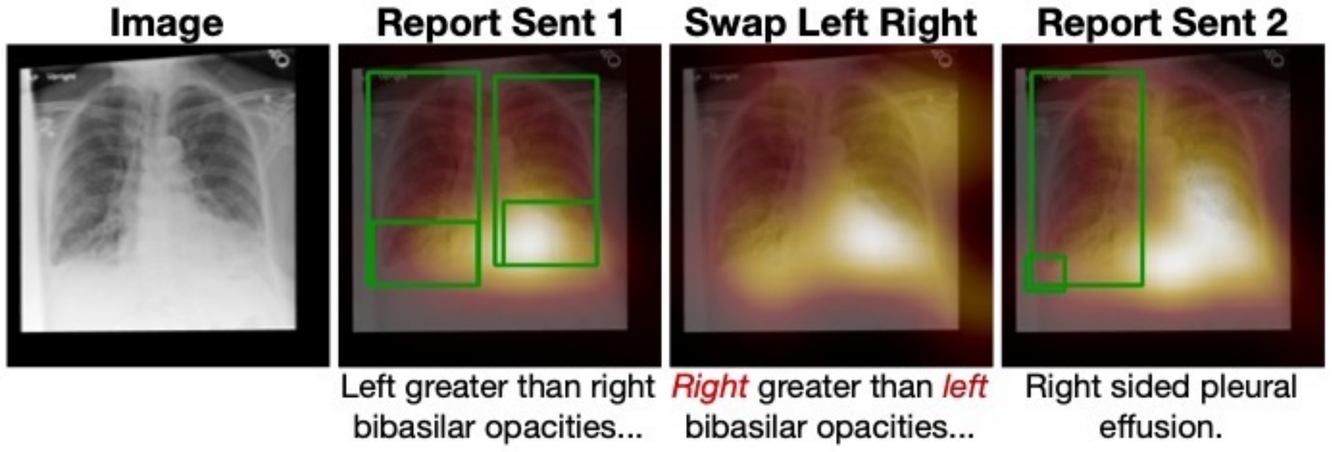
That's the Wrong Lung!: Evaluating and Improving the Interpretability of Unsupervised Multimodal Encoders

Denis Jered McInerney[1], Geoffrey Young[2], Jan-Willem van de Meent[3], Byron C. Wallace[1]
[1]Northeastern University, [2]Brigham and Women's Hospital, [3]University of Amsterdam

Introduction

Do contrastively trained image-text models yield **interpretable** alignments between image regions and text?

We find that text often has an **unintuitive or weak** effect on attention distributions over the image:



Automatic Evaluation

We measure the **localization ability of GLoRIA**—a SOTA multimodal encoder for Chest X-rays-–using metrics for classifying a pixel as within the ground truth ROI. GT bounding boxes come from the **ImaGenome** dataset.

AUROC	Avg. P	IOU@5/10/30%
69.07	51.68	3.79/6.69/20.10

This can be hard to interpret and misleading because it does not show how the models respond to changes in the text. So we:

1) Develop **text and label perturbations** that we expect to create a mismatch between the attention and labels:

Original Text

Small right pleural effusion is stable.

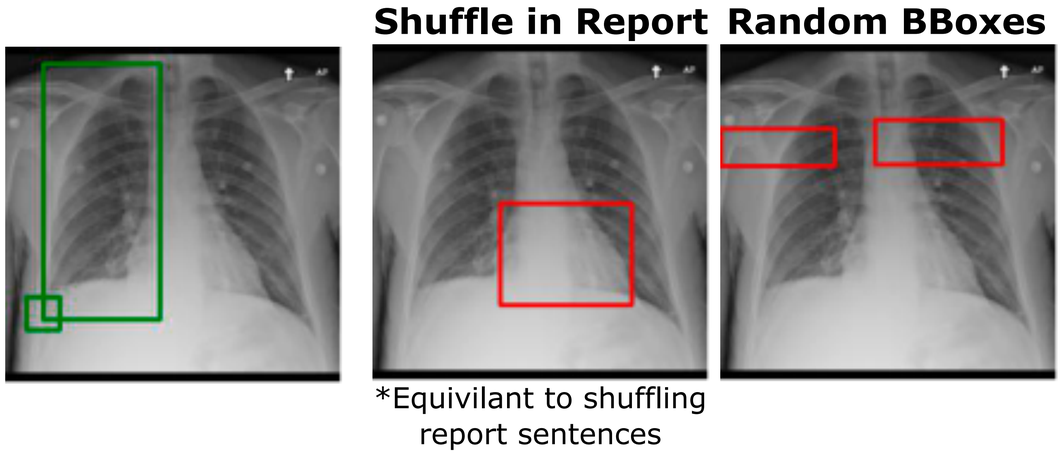
Text Perturbations

Swap Left Right:
Small *left* pleural effusion is stable.

Random Sentence:
The lungs are hyperinflated but clear of consolidation.

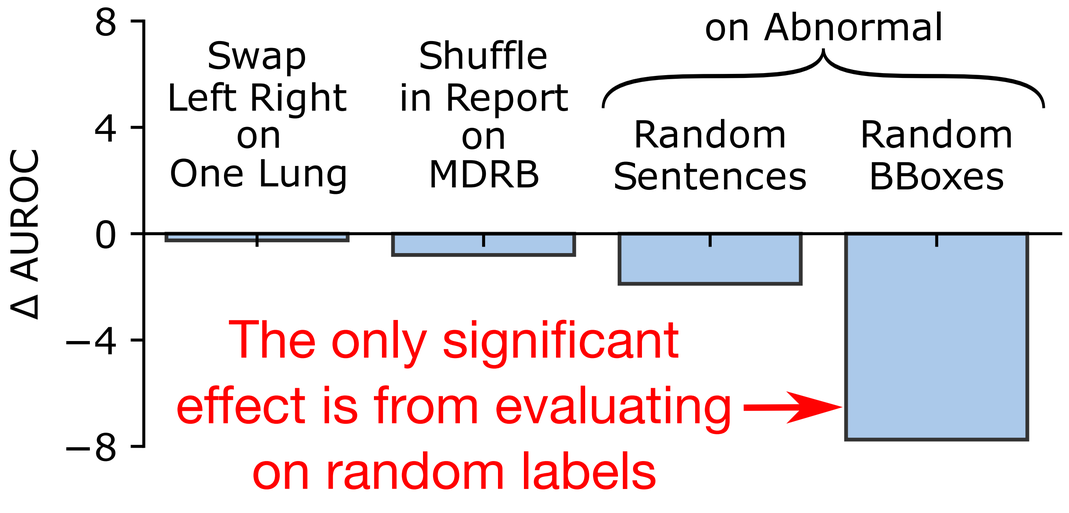
Original BBox

BBox Perturbations



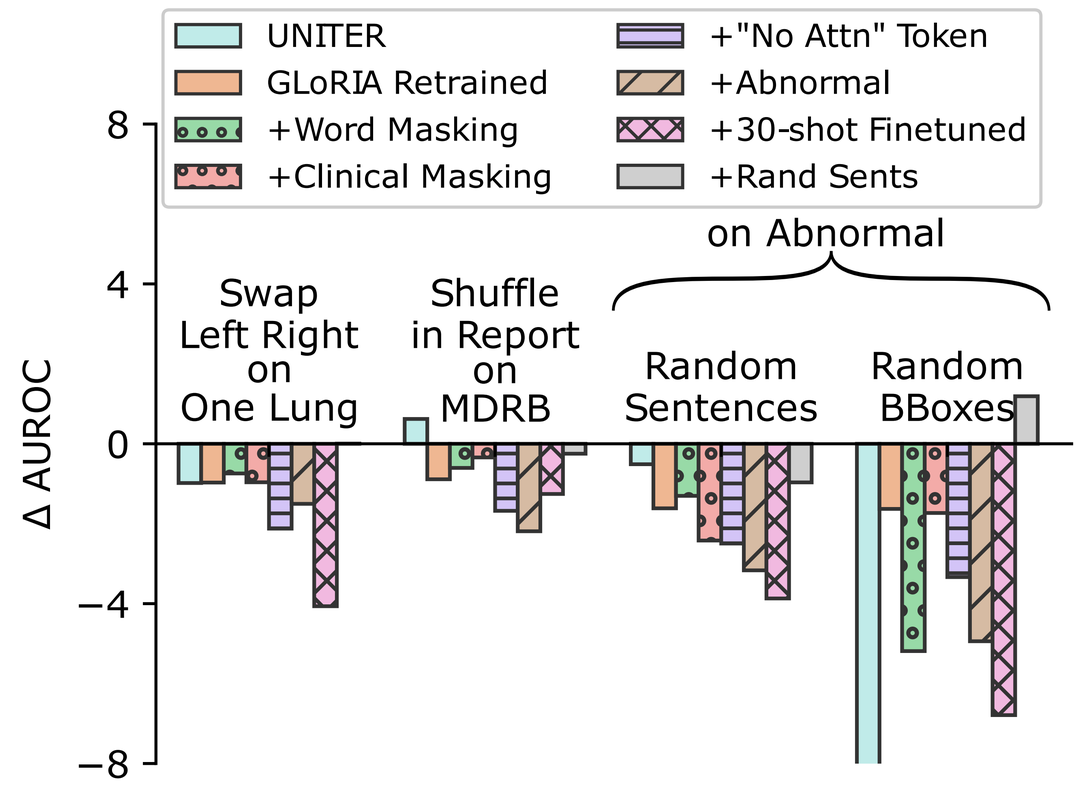
*Equivilant to shuffling report sentences

2) Measure the resulting **change in performance**:

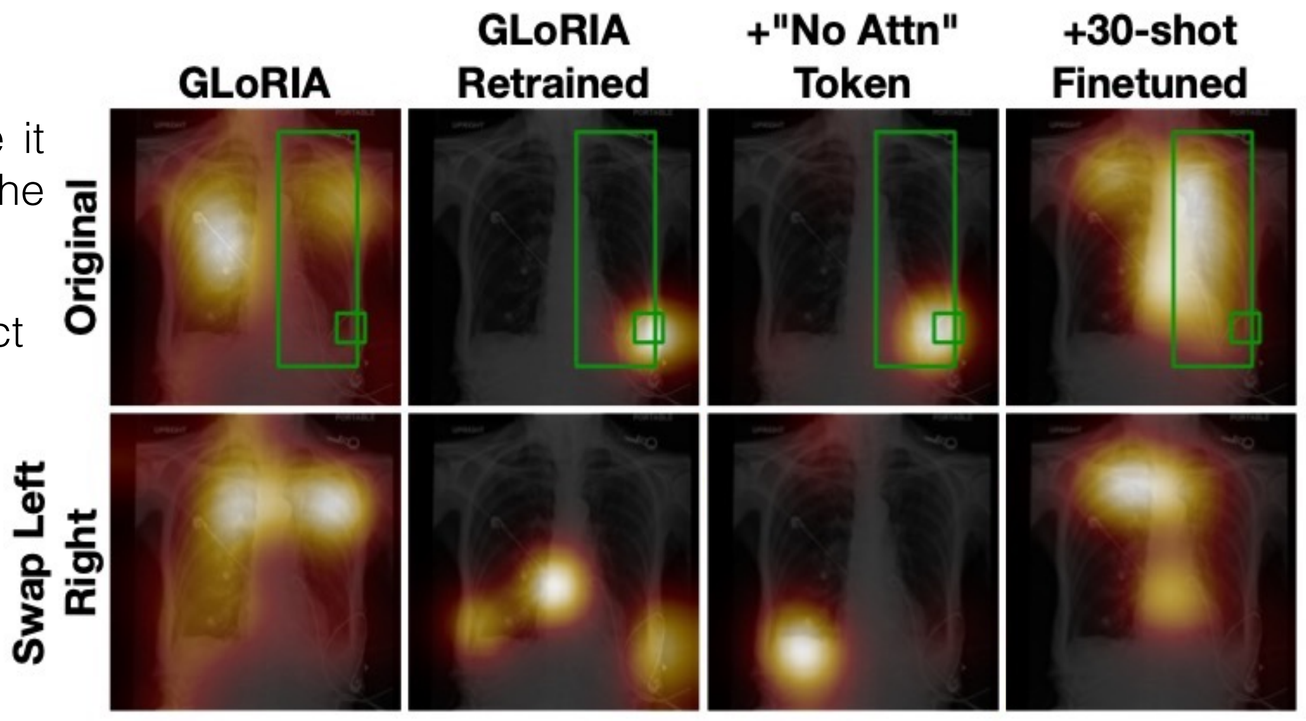


Improving Alignment

We try to improve on previous work **with little to no supervision** with new model variants, some of which do better respond to the perturbations.



Particularly, without any additional supervision, we achieve much better results by allowing the model to attend to an extra **"no attention" token** instead of the image. We provide a cherry-picked example below.



The fact that **only 30 shots of supervision** can change attention drastically highlights the potential of this method.

Manual Evaluations

We collect annotations of the model outputs from an experienced board-certified radiologist on the precision and recall (with respect to the perceived region of interest) and the intuitiveness of the attention.