That's the Wrong Lung! Evaluating and Improving the Interpretability of Unsupervised Multimodal Encoders for Medical Data

Denis Jered McInerney
Northeastern University
mcinerney.de@northeastern.edu

Geoffrey Young
Brigham and Women's Hospital
gsyoung@bwh.harvard.edu

Jan-Willem van de Meent
University of Amsterdam
j.w.vandemeent@uva.nl

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

Pretraining multimodal models on Electronic Health Records (EHRs) provides a means of learning representations that can transfer to downstream tasks with minimal supervision. Recent multimodal models induce soft local alignments between image regions and sentences. This is of particular interest in the medical domain, where alignments might highlight regions in an image relevant to specific phenomena described in free-text. While past work has suggested that attention "heatmaps" can be interpreted in this manner, there has been little evaluation of such alignments. We compare alignments from a state-of-the-art multimodal (image and text) model for EHR with human annotations that link image regions to sentences. Our main finding is that the text has an often weak or unintuitive influence on attention; alignments do not consistently reflect basic anatomical information. Moreover, synthetic modifications — such as substituting "left" for "right" — do not substantially influence highlights. Simple techniques such as allowing the model to opt out of attending to the image and few-shot finetuning show promise in terms of their ability to improve alignments with very little or no supervision. We make our code and checkpoints open-source.[1]

1 Introduction

There has been a flurry of recent work on model architectures and self-supervised training objectives for multimodal representation learning, both generally (Li et al., 2019; Tan and Bansal, 2019; Huang et al., 2020; Su et al., 2020; Chen et al., 2020) and for medical data specifically (Wang et al., 2018; Chauhan et al., 2020; Li et al., 2020). These methods yield representations that permit efficient learning on various multimodal downstream tasks (e.g., classification, captioning).

Given the inherently multimodal nature of much medical data — e.g., in radiology images and text are naturally paired — there has been particular interest in designing multimodal models for Electronic Health Records (EHRs) data. However, one of the factors that currently stands in the way of broader adoption is interpretability. Neural models that map image-text pairs to shared representations are opaque. Consequently, doctors have no way of knowing whether such models rely on meaningful clinical signals or data artifacts (Zech et al., 2018).

Recent work has proposed models that soft-align text snippets to image regions. This may afford a type of interpretability by allowing practitioners to inspect what the model has "learned" or allow more efficient identification of relevant regions. Past work has presented illustrative multimodal "saliency" maps in which such models highlight plausible regions. But such highlights also risk providing a false sense that the model "understands" more than it actually does, and irrelevant highlights would be antithetical to the goal of a efficiency in clinical decision support.

Multimodal models may fail in a few obvious ways; they may focus on the wrong part of an image, fail to localize by producing a high-entropy attention distribution, or localize too much and miss a larger region of interest. However, even when image attention *appears* reasonable, it may not in actuality reflect both modalities. Figure 1 shows an example. Here the model ostensibly succeeds at identifying the image region relevant to the given text (left). One may be tempted to conclude the model has "understood" the text and indicated the corresponding region. But this may be misleading: We can see that the same model yields a similar attention pattern when provided text with radically different semantics (e.g., when swapping "right" with "left"), or when providing sentences referencing an abnormality in another region.

Our contributions are as follows. (i) We appraise the interpretability of soft-alignments induced between images and texts by existing neural multi-

[1] https://github.com/dmcinerney/gloria

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3626–3648
December 7-11, 2022

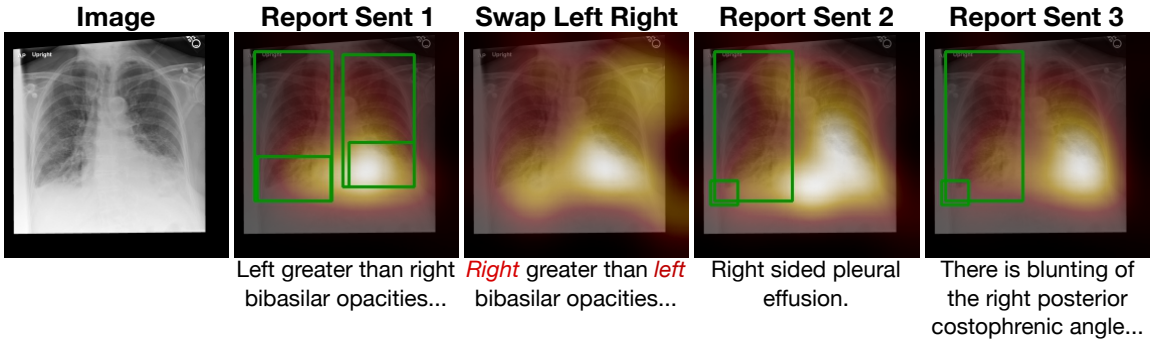

Figure 1: Alignment failures often occur when the model (overly) focuses on one aspect of the image, largely ignoring the text. (Note: images are "mirrored", so right and left are flipped.)

modal models for radiology, both retrospectively and via manual radiologist assessments. To the best of our knowledge, this is the first such evaluation. (ii) We propose methods that improve the ability of multimodal models for EHR to intuitively align image regions with texts.

2 Preliminaries

We aim to evaluate the localization abilities of multimodal models for EHR. For this we focus on the recently proposed GLoRIA model (Huang et al., 2021), which is representative of state-of-the-art, transformer-based multimodal architectures and accompanying pre-training methods. For completeness we also analyze (a modified version of) UNITER (Chen et al., 2020). We next review details of these models, and then discuss the datasets we use to evaluate the alignments they induce.

2.1 GLoRIA

GLoRIA uses Clinical BERT (Alsentzer et al., 2019) as a text encoder and ResNet (He et al., 2016) as an image encoder. Unlike prior work, GLoRIA does not assume an image can be partitioned into different objects, which is important because pre-trained object detectors are not readily available for X-ray images. GLoRIA passes a CNN over the image to yield local region representations. This is useful because a finding within an X-ray described in a report will usually appear in only a small region of the corresponding image (Huang et al., 2021). GLoRIA exploits this intuition via a local contrastive loss term in the objective.

We assume a dataset of instances comprising an image x_v and a sentence from the corresponding report x_t, and the model consumes this to produce a set of local embeddings and a global embedding per modality: $v_l \in \mathbb{R}^{M \times D}$, $v_g \in \mathbb{R}^D$, $t_l \in \mathbb{R}^{N \times D}$, and $t_g \in \mathbb{R}^D$. To construct the local contrastive loss, an attention mechanism (Bahdanau et al., 2014) is applied to local image embeddings, queried by the local text embeddings. This induces a soft alignment between the local vectors of each mode:

$$a_{ij} = \frac{\exp\left(t_{li}^T v_{lj}/\tau\right)}{\sum_{k=1}^{M} \exp\left(t_{li}^T v_{lk}/\tau\right)} \quad (1)$$

where t_i is the ith text embedding, v_j the jth image embedding, and τ is a temperature hyperparameter.

2.2 UNITER

Despite the challenges inherent to adopting "general-domain" multimodal models for this domain (discussed in Appendix A.1), we modify UNITER to serve as an additional model for analysis. We provide details regarding how we have implemented UNITER in Appendix A.2, but note here that *this requires ground-truth bounding boxes as inputs*, which means that (a) results with respect to most metrics (which measure overlap with target bounding boxes) for UNITER will be artificially high, and, (b) we could not use this method in practice, because it requires a set of reference bounding boxes as input (including at inference time). We include this for completeness.

2.3 Data and Metrics

Data Our retrospective evaluation of localization abilities is made possible by the MIMIC-CXR (Johnson et al., 2019a,b) and Chest ImaGenome (Wu et al., 2021) datasets. MIMIC-CXR comprises chest X-rays and corresponding radiology reports. ImaGenome includes 1000 manually annotated image/report pairs,[2] with bounding boxes for anatomical locations, links between referring sentences and image bounding boxes, and a set of conditions and

[2]Annotations were automatically derived then cleaned.

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

Table 1: Localization performance of GLoRIA.

positive/negative context annotations[3] associated with each sentence/bounding box pair.

Metrics We quantify the degree to which attention highlights the region to which a text snippet refers by comparing average attention over an input sentence $x_j = \frac{1}{N}\sum_{i=1}^{N} a_{ij}$ with reference annotated bounding boxes associated with the sentence.

We use several metrics to measure the alignment between soft attention weights and bounding boxes. We create scores $s \in \mathbb{R}^P$ for each of the P pixels based on the attention weight assigned to the image region the pixel belongs to. Specifically, for GLoRIA we use upsampling with bilinear interpolation to distribute attention over pixels. For UNITER, we score pixels by taking a max over attention scores for the bounding boxes that contain the pixel (scores for pixels not in any bounding boxes are 0). We use bounding boxes to create a segmentation label $\ell \in \mathbb{R}^P$ where $\ell_p = 1$ if pixel i is in any of the bounding boxes, and $\ell_p = 0$ otherwise. Given pixel-level scores s and pixel-level segmentation labels ℓ_p, we can compute the **AUROC**, **Average Precision**, and **Intersection Over Union** (IOU) at varying pixel percentile thresholds for the ranking ordered by s (See section A.4).

We also adopt a simple, interpretable metric to capture the accuracy of similarity scores assigned to pairs of images and texts. Specifically, we use a simpler version of the text retrieval task from (Huang et al., 2021): We report the percentage of time the similarity between an image and a sentence from the corresponding report is greater than the similarity between the image and a random sentence taken from a different report in the dataset. This allows us to interpret 50% as the mean value of a totally random similarity measure.

3 Are Alignment Weights Accurate?

We first use the metrics defined above to evaluate the pretrained, publicly available weights for GLoRIA (Huang et al., 2021). Table 1 reports the metrics used to evaluate localization on the gold split of the ImaGenome dataset.

AUROC scores are well over 50%, suggesting reasonable localization performance. IOU scores are small, which is expected as target bounding boxes tend to be much larger than the actual regions of interest and serve more to detect errors when highlighted regions are far from where they should be; this is further supported by the relatively high average precision scores.[4] However, while seemingly promising, our results below suggest that the attention patterns here may be less multimodal than one might expect.

[3]Here, context refers to whether the condition is negated in the text (negative) or not (positive).

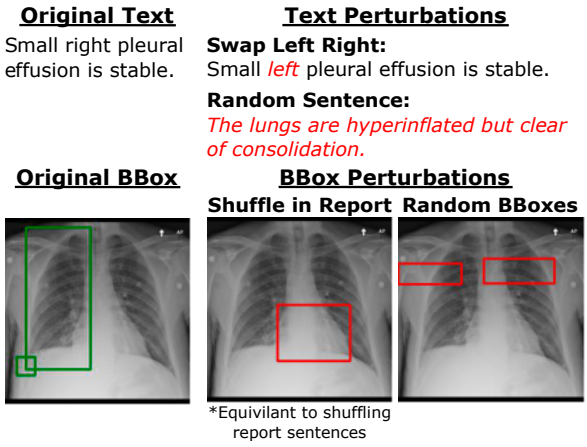

Figure 2: Examples of each **perturbation** for a given instance. (Synth w/ Swapped Conditions example in Appendix 10.)

We next focus on evaluating the degree to which these patterns actually reflect the associated text. To this end we perturb instances in ways that ought to shift the attention pattern (Section 3.1), e.g., by replacing "right" with "left" in the text. We then identify data subsets in Section 3.2 comprising "complex" instances, where we expect the image and text to be closely correlated at a local level.

3.1 Perturbations

Figure 2 shows examples of the perturbations that include: Swapping "left" with "right" (**Swap Left Right**); Shuffling the target bounding boxes for sentences within the same report at random (**Shuffle in Report**); Replacing *sentences* in a report with other sentences, randomly drawn from the rest of the dataset (**Random Sentences**); Replacing target *bounding boxes* with other bounding boxes randomly sampled from the dataset (**Random BBoxes**)[5], and; Swapping the correct conditions in a synthetically created prompt with random conditions **Synth w/ Swapped Conditions**. We in-

[4]In Section B.1, we address this with a modified evaluation that drops some large bounding boxes in the labels.

[5]**Shuffle in Report** bboxes will still correspond to valid and noteworthy anatomical regions, but **Random BBoxes** bboxes will not correspond to valid anatomical regions at all.

clude additional details about synthetic sentences and perturbations in Appendices A.3 and A.5.

Under these perturbations, we would expect a well-behaved model to shift its attention distribution over the image accordingly, resulting in a decrease in localization scores (overlap with the original reference bounding boxes). The **Random BBoxes** perturbation in particular targets the degree to which the attention relies specifically on the image modality, because here the "target" bounding boxes have been replaced with bounding boxes associated with *random other images*. By contrast, all other perturbations should measure the degree to which the model is sensitive to changes to the text (even **Shuffle in Report**, which is equivalent to shuffling the sentences in a report).

If attention maps reflect alignments with input texts, then under these perturbations one should expect large negative differences in performance (Δmetric) relative to observed performance using the unperturbed data. For all but **Random BBoxes**, if the performance does not much change (Δmetric ≈ 0), this suggests the attention maps are somewhat invariant to the text modality.

3.2 Subsets

We perform granular evaluations using specific data subsets, including: (1) **Abnormal** instances with an abnormality, (2) **One Lung** instances with only one side of the Chest X-ray (left or right) referenced, and (3) **Most Diverse Report BBoxes (MDRB)** instances with a lot of diversity in the labels for sentences in the same report. Details are in Appendix A.6.

Intuitively, some of the perturbations in Section 3.1 should mainly effect certain subsets: **Swap Left Right** should most impact the **One Lung** subset, **Shuffle in Report** should mainly effect **MDRB**, and **Random Sentences**, **Random BBoxes**, and **Synth w/ Swapped Conditions** should primarily effect **Abnormal** examples.

3.3 Annotations for Post-hoc Evaluation

We enlist a domain expert (radiologist) to conduct annotations to complement our retrospective quantitative evaluations. We elicit judgements on a five-point Likert scale regarding the recall, precision, and "intuitiveness" of image highlights induced for text snippets.[6] More details are in the Appendix, including annotation instructions (Section A.7) and a screenshot of the interface (Figure 11).

[6]For recall and precision, points on the Likert scale are intended to correspond to buckets of 0-20, 20-40, 40-60, 60-80, and 80-100 percent respectively.

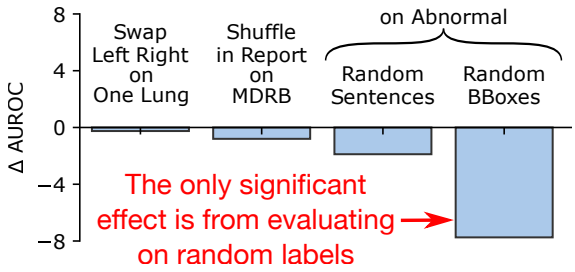

Figure 3: For each perturbation, we plot the change in localization performance (AUROC) of **GLoRIA**.

Subset	AUROC	Avg. P	IOU@5/10/30%
Abnormal	69.51	48.29	4.10/7.25/19.05
One Lung	65.48	38.68	4.43/8.05/20.54
MDRB	65.01	36.96	3.56/6.37/16.92

Table 2: GLoRIA Localization performance on subsets.

3.4 Results

We first evaluate performance on the subsets described in Section 3.2. This establishes a baseline with respect to which we can take differences observed under perturbations. We report results in Table 2. We observe that the model performs significantly worse on both the **One Lung** and **MDRB** subsets (which we view as "harder") in terms of AUROC and Average Precision, supporting this disaggregated evaluation.

Manual evaluation results of 3.1, 1.8, and 1.7 for recall, precision, and intuitiveness respectively indicate that GLoRIA produces unintuitive heatmaps that have poor precision and middling recall. Because GLoria was trained on the CheXpert dataset and we perform these evaluations on ImaGenome, the change in dataset may be one cause of poor performance; in Section 4 we report how retraining on the ImaGenome dataset affects these scores.

To measure the sensitivity of model attention to changes in the text, we report **differences in localization performance** in Figure 3. Specifically, this is the difference in model performance (ΔAUROC) achieved using (a) the original (unperturbed) sentences, and, (b) sentences perturbed as described in Section 3.1. We show results for each perturbation on the subsets they should most effect (Section 3.2), leaving the full results for the appendix (Figure 14).

The only real decrease in performance observed is under the **Random BBoxes** perturbation, which entails swapping out the target bounding box for an instance with one associated with some *other instance (image)*. Performance decreasing here (and

not for text perturbations) is consistent with the hypothesis that the attention map primarily reflects the image modality, but not the text. This is further supported by the observation that the model pays little mind to clear positional cue words such as "left" and "right" when constructing the attention map; witness the negligible drop in performance under the **Swap Left Right** perturbation. Finally, swapping in other sentences (even from different reports) yields almost no performance difference.

4 Can We Improve Alignments?

The above results indicate that image attention is unintuitive and less sensitive to the text modality than might be expected. Next we propose simple methods to try to improve image/text alignments.

4.1 Models

All models build on the GLoRIA architecture except the baseline **UNITER**, for which we perform no modifications except to re-train from scratch on the MIMIC-CXR/Chest ImaGenome dataset.[7] In the results, **GLoRIA** refers to weights fit using the CheXpert dataset, released by (Huang et al., 2021). We do not have access to the reports associated with this dataset so we do not use it for training or evaluation, but we do make comparisons to the original (released) **GLoRIA** model trained on it.

We also retrain our own **GLoRIA** model on the MIMIC-CXR/ImaGenome dataset; we call this **GLoRIA Retrained**. While the two datasets are similar in size and content, CheXpert has many more positive cases of conditions than MIMIC-CXR/ImaGenome (8.86% of CheXpert images are labeled as having "No Findings"; in the ImaGenome dataset, reports associated with 21.80% of train images do not contain a sentence labeled "abnormal"). Given this difference in the number of positive cases, we train a **Retrained+Abnormal** model variant on the subset of MIMIC-CXR/ImaGenome sentence/image pairs featuring an "abnormal" sentence.

We also train models in which we adopt masking strategies intended to improve localization, hypothesizing that this might prevent over-reliance on text artifacts that might allow the model to ignore text that localizes. Our **Retrained+Word Masking** model randomly replaces words in the input with `[MASK]` tokens during training with 30% probability.[8] For our **Retrained+Clinical Masking** model, we randomly swap clinical entity spans found using a `SciSpaCy` entity linker (Neumann et al., 2019) for `[MASK]` tokens with 50% probability.

Many sentences in a report will not refer to any particular region in the image. We therefore propose the **Retrained+"No Attn" Token** model, which concatenates a special "No Attn" token parameter vector to the set of local image embeddings just before attention is induced. This allows the model to attend to this special vector, rather than any of the local image embeddings, effectively indicating that there is no good match.

We also consider a setting in which we assume a small amount of supervision (annotations linking image regions to texts). We finetune a model to produce high attention on the annotated regions of interest, i.e., we supervise attention. We employ an alignment loss $\mathcal{L}_{\text{alignment}}(s, \ell) = \sum_p s_p \ell_p$ using the pixel-wise scores s derived from the attention[9] and the segmentation labels ℓ (Section 2.3). We train on a batch of 30 examples for up to 500 steps with early stopping on an additional 30-example validation set using a patience of 25 steps. This might be viewed as "few-shot alignment", where we use a small number of annotated examples to try to make the model more interpretable by improving image and text alignments.

Finally, as a point of reference we train **Retrained+Rand Sents** in the same style as the **Retrained** model except that all sentences are replaced with *random* sentences. This deprives the model of any meaningful training signal, which otherwise comes entirely through the pairing of images and texts. This variant provides a baseline to help contextualize results. For all models, we use early stopping with a patience of 10 epochs.[10]

4.2 Results and Discussion

4.2.1 Localization Metrics

Table 3 might seem to imply that **UNITER** performs best. However, we emphasize that this is not comparable to other models because, as discussed

[7] We re-train from scratch because: (1) Unlike in the original model, we are not feeding in features from Fast-RCNN, but instead using flattened pixels from a bounding box, and; (2) We would like a fair comparison to the GLoRIA variants which are also re-trained from scratch.

[8] We choose the high value of 30% here because without allowing hyperparameter tuning of this probability, we would like to see a significant impact when comparing to the baseline.

[9] In this case, we also renormalize again after upsampling so the pixel scores to sum to 1.

[10] For all models we report results on the last epoch before the early stopping condition is reached.

Model	AUROC	Avg. P
UNITER★	**84.92**	**68.57**
GLoRIA Retrained	55.84	41.22
+Word Masking	61.44	44.69
+Clinical Masking	54.67	40.61
+"No Attn" Token	57.00	41.80
+Abnormal	55.89	43.42
+30-shot Finetuned	**63.90**	**52.80**
+Rand Sents	38.88	30.55

Table 3: Localization performance for each retrained model. ★ UNITER here is not comparable because it uses ground truth bounding boxes as input. (Full results in Table 7.)

in 2.2, **UNITER**'s attention is defined over ground truth anatomical bounding boxes (rather than the entire image), of which the sentence bounding boxes are a subset; this dramatically inflates AUROC and average precision scores. (We have included **UNITER** despite this for completeness.)

Finetuning on a small set of ground truth bounding boxes (**+30-shot Finetuned**) substantially improves performance. Of the remaining (not explicitly supervised) approaches, **+Word Masking** fares best. This masking may serve a regularization function similar to dropout (Srivastava et al., 2014). Counter-intuitively, **+Clinical Masking** performs slightly worse than **Retrained**. Perhaps clinical masking blinds too much key information.

The **+"No Attn" Token** model also performs comparatively well, suggesting that allowing the model to not attend to any particular part of the image does increase performance.

To address a concern that the bounding boxes used as labels are bigger than the region of interest (Section 3), we try to improve our measure of precision by re-evaluating without some of the larger bounding boxes and find an overall drop in precision but with similar trends in relative model performance (see Section B.1).

4.2.2 Post-hoc Evaluation

The annotation results (Figure 4) of recall, precision, and intuitiveness are perhaps more revealing and do not necessarily align with our automatic metrics.[11] This is likely a product of the limitations of the ImaGenome bounding boxes. The **+"No Attn" Token** model scored highest in terms of intuitiveness and precision, which is promising given

[11]We do not include UNITER in this because the attention over the bounding boxes is very unintuitive and different from the other models' attention (See Appendix Figure 8).

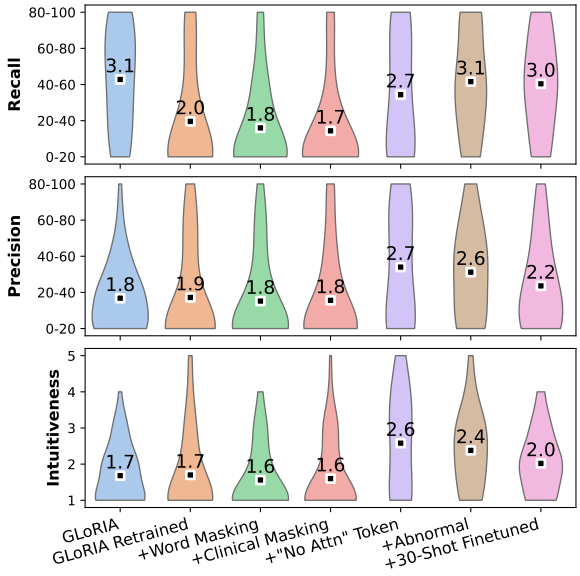

Figure 4: Annotations on Retrained Variants. We report means over 50 annotations. Recall and Precision scores 1, ..., 5 correspond to bins of 0-20, ..., 80-100 %. UNITER is not included in this because the attention over the bounding boxes is very unintuitive and different from the other models' attention.

that unlike the **+Abnormal** and **+30-shot Finetuned** models, this model does not require any additional training information (i.e., indications of training sample abnormalities, or ground truth bboxes). A simple modification to the architecture that allows it to pass on aligning a given text to the image yields a stark increase in performance with respect to the baseline **Retrained** model. The **Retrained** model performs about the same as **GLoRIA** in terms of precision and intuitiveness, although it incurs a significant drop in recall.

The **+30-shot Finetuned** model uses the bounding boxes as ground truth, but these are somewhat noisy. Better annotations of the regions of interest might improve intuitiveness further. When performing annotations, the radiologist also noticed that a large percentage of sentences in reports do not refer to anything focal, which indicates the necessity of looking at the subsets from Section 3.2—all of which should have more focal sentences—especially when it comes to the perturbations. This also may help explain the superior performance of the **+"No Attn" Token** model which explicitly handles these cases.

4.2.3 Perturbation Results

We next perform the perturbations introduced above (and assessed on GLoRIA) to the proposed variants to assess sensitivity to input texts (full re-

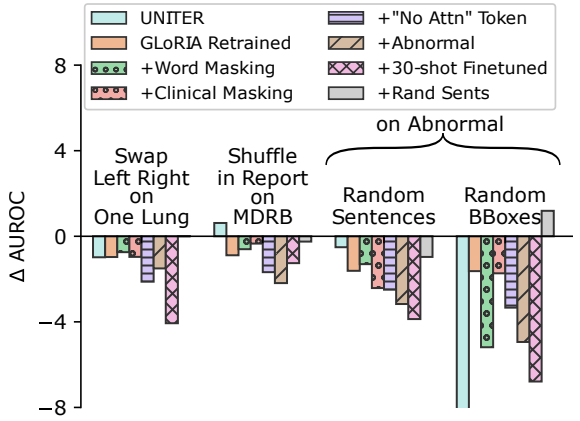

Figure 5: For each perturbation, we plot the change in localization performance (measured by AUROC), for each of the models we retrain from scratch on the respective subsets. Here, UNITER is effected most by the Random BBoxes perturbations because it uses the original ground truth as input.

sults in Figure 14 of the appendix). We observe that **+30-shot Finetuned**, **+"No Attn" Token**, and **+Abnormal**, in that order, are most affected when swapping left and right. These three models are also the most affected by shuffling bounding boxes within a report or swapping in a random sentence from the rest of the dataset, although for these perturbations, the **+Abnormal** model is more sensitive than the **+"No Attn" Token**.

The **Random BBoxes** perturbation serves mostly as a reference measure of how variable model scores can be when swapping in entirely wrong bounding boxes. But it also seems to suggest that for models affected more by this, the attention is more focused on precision. This indicates that besides UNITER, the **+30-shot Finetuned**, **+Word Masking**, **+Abnormal**, and **+"No Attn" Token**, in that order, are the most precise; this is in line with the average precision scores in Table 3 and the entropy scores in the appendix (Table 10).

Taken together these perturbation results suggest that **+"No Attn" Token**, **+Abnormal**, and **+30-shot Fine-tuned** are the models most intuitively sensitive to text. However, they remain less intuitive than they would ideally be.[12]

4.2.4 Contrastive Accuracy

Table 4 reports the accuracy of each model with respect to identifying the correct sentence from two candidates for a given image. These results indicate that performing comparatively well at identifying

[12]We discuss results for experiments in which we swap conditions in synthetic sentences in Appendix (B.4); these are inconclusive.

	All		Abnormal	
Model	local	global	local	global
UNITER	-	67.2	-	70.7
GLoRIA	55.2	70.3	43.3	77.0
GLoRIA Retrained	70.2	**82.9**	63.8	86.4
+Word Masking	**78.9**	81.6	**80.3**	**86.5**
+Clinical Masking	68.5	81.5	65.4	84.4
+"No Attn" Token	67.3	81.9	61.8	85.0
+Abnormal	72.1	76.6	73.1	84.1
+30-shot Finetuned	67.2	79.6	61.0	83.6
+Rand Sents	51.4	51.3	44.8	60.6

Table 4: Average accuracies with respect to discriminating between the sentence actually associated with an image and a sentence randomly sampled from the dataset. (See Appendix Table 9 for results on subsets.) Global and local refer to using only global or local embeddings for computing similarity.

the correct sentence does not necessarily correlate with intuitiveness or textual sensitivity, i.e., being able to discriminate between *sentences* given an image does not imply an ability to accurately *localize* within an image, given a sentence. In particular, **+Word Masking** performs best here, though we saw above that it is relatively unintuitive and its localization is somewhat invariant to perturbations. Further, the three best models in terms of textual sensitivity have relatively poor performance (with the possible exception of the **+Abnormal** variant).

4.2.5 Metric Correlations

To quantify the relationships between scores, we report correlations between them across instances for **+"No Attn" Token** (the best model in terms of manually judged intuitiveness) in Figure 6. Of the automatic metrics, IOU@10% has the strongest correlation with annotated intuitiveness. Avg. Precision and Precision at 10% have almost no correlation with intuitiveness and relatively weak correlation with annotated precision. We also show correlation with local and global similarity between two positive pairs.[13] Though the local similarity of positive pairs is somewhat correlated with each of the annotation ratings, the global similarity is only (weakly) correlated with annotated precision.

The "No Attn" score, which is what we use to refer to the attention score for the added "No Attn" token, has some interesting Pearson correlations. Unfortunately, its relationship with annotations is complicated by our user interface. Often the "No Attn" score (which we display in the corner of the image) will either be unnoticeable or it will saturate the heatmap, resulting in the radiologist assigning

[13]Because we only look at positive pairs, higher similarity scores are better.

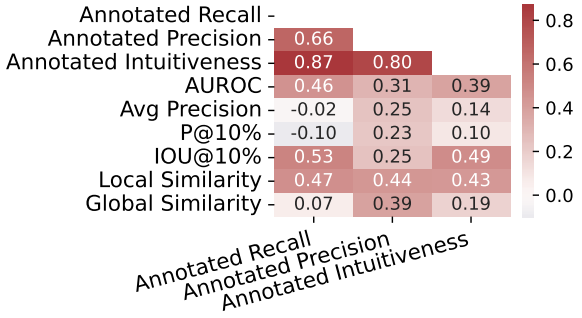

Figure 6: **Correlations** between metrics and annotations.

low scores (1s) for an instance. Therefore, we note that some negative correlations with annotations (Figure 22) may mostly reflect how the X-ray heatmap is displayed to the user. However, a -.30 correlation with IOU@10% and a -.47 correlation with whether an image contains an abnormality are significant. This demonstrates the potential for this score to identify situations where the model should abstain from displaying a heatmap altogether because either there is nothing abnormal to highlight or the model is not confident in its heatmap.

4.2.6 Qualitative Analysis

We note some interesting qualitative behavior discovered during the annotation that may also support the use of this "No Attn" architecture and score. When many of the models are incorrect, they tend to highlight image edges or corners. We hypothesize this occurs because the model attempts to find a static part of the image—one that is similar across most instances—on which to attend. This behavior is misleading and not quantifiable. The "No Attn" Token offers an alternative to this behavior, providing a means for the model to pass on inducing a heatmap altogether when appropriate.

We conclude with a qualitative impression of localization performance. Figure 7 shows model attention distributions for a (cherry-picked) instance and the accompanying **Swap Left Right** perturbation. This example was selected specifically to illustrate how models can fail to behave intuitively. In this example, the correct region of interest for the original prompt lies mostly centered on the small box, and the large box (corresponding to the left lung) is somewhat misleading as it covers more than the strict region of interest. This example demonstrates that though the anatomical locations discussed in the prompt are correctly highlighted by the bounding boxes, the *region of interest* is not always directly on those anatomical locations.

With the original prompt, **GLoRIA** yields a high-entropy map, **GLoRIA Retrained** and the **+Masking**, **+"No Attn" Token**, and **+Abnormal** are centered roughly correctly (some more intuitive than others), and finally, **+30-shot Finetuned** almost fully highlights the large box (even though this is not strictly the correct region of interest) and almost entirely ignores the small box (the real region of interest). The perturbation of swapping out "left" with "right" changes all of the models' heatmaps to varying degrees and with varying intuitiveness. In this example, the most intuitive heatmaps after the perturbation are given by the **+"No Attn" Token** and **+Abnormal** models, whereas other models still show significant emphasis on the original region and/or show emphasis on unintuitive and entirely irrelevant regions.

Summary of key findings Existing multimodal pretraining schemes beget models that accurately select the text that matches a given image (Table 4), and yield attention distributions that at least somewhat depend on the text. But these models are not found intuitive (Table 4) and perturbing texts does not cause not consistently yield changes in the attention patterns that one would expect (Figure 5). Simple changes to pre-training may improve this behavior. Specifically, adding the ability of the model to *not attend to any particular part of the image* may result in models that produce attention patterns which are more intuitive (Figure 4) *and* more reflective of input texts (Figure 5), although this may slightly harm performance on the pre-training task itself (Table 4).

5 Related Work

Work on multi-modal representation learning for medical data has proposed soft aligning modalities, but has focussed quantitative evaluation on the resultant performance that learned representations afford on downstream tasks (Ji et al., 2021; Liao et al., 2021; Huang et al., 2021). Model interpretability is often suggested using only qualitative examples; our work aims to close this gap.

A line of work in NLP evaluates the interpretability of neural attention mechanisms (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019). Elsewhere, work at the intersection of computer vision and radiology has critically evaluated use of saliency maps over images (Arun et al., 2021; Rajpurkar et al., 2018).

Recent work has sought to improve the ability

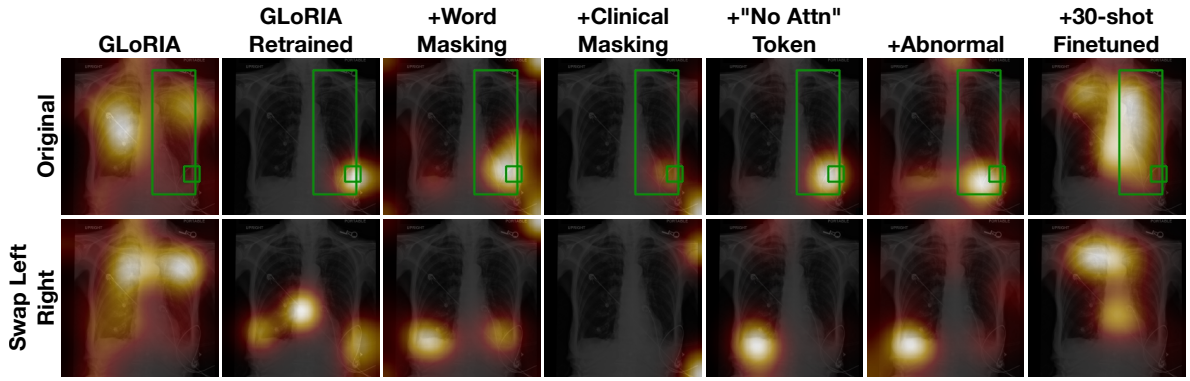

Figure 7: Attention for "Blunting of the left costophrenic angle suggests small effusion." (top), and perturbed version (bottom).

of these models to identify fine-grained alignments via supervised attention (Kervadec et al., 2020; Sood et al., 2021), but have focused on downstream task performance. This differs from our focus on evaluating and improving localization itself, especially within the medical domain. We also do not assume access to large amounts of supervision, which is commonly lacking in this domain.

6 Conclusions

We evaluated existing state-of-the-art unsupervised multimodal representation learning models for EHRs in terms of inducing fine-grained alignments between image regions and text. We found that the resultant heatmaps are often unintuitive and invariant to perturbations to the text that ought to change them substantially.

We evaluated a number of methods aimed at improving this behavior, finding that: (1) allowing the model to refrain from attending to the image, and; (2) finetuning the model on a small set of labels for interpretable heatmaps substantially improves performance. We hope that this effort motivates more work addressing the interpreteability of multimodal encoders for healthcare.

Limitations

This is a first attempt to investigate the interpretability of pre-trained multi-modals models for medical imaging data, and as such our work has important limitations. ImaGenome only annotates anatomical locations for each sentence and bounding boxes for each anatomical location; these may not correspond directly to regions of interest. In addition, these extracted anatomical locations highlight many levels of the hierarchy, so if a specific part of the lung is mentioned, the whole lung's bounding box may still be included. The annotations we collected for evaluation also have important limitations to consider, namely that we only used one radiologist annotator and only annotated 50 instances.

Finally, we did not try a more fine-grained UNITER model with more and smaller bounding boxes that form a grid (to avoid an object detector), primarily because this would incur a significantly higher computational cost due to the number of image vectors; future work might explore this option.

Ethics

There are significant risks associated with incorrectly interpreting models in the medical domain. Our aim in this work is to highlight gaps in the current technology and suggest avenues for future research and *not* to provide deployable models. These models are not ready for use in the field because they may mislead users about the underlying reasons for predictions and incorrectly inform resulting decisions. We hope this work facilitates advances in the interpretability of these models so that they may eventually provide meaningful guidance to radiologists. The MIMIC-CXR dataset used was licensed via the PhysioNet Credentialed Health Data License 1.5.0, and we properly comply with the PhysioNet Credentialed Health Data Use Agreement 1.5.0.

Acknowledgements

We acknowledge partial funding for this work by National Library of Medicine of the National Institutes of Health under award numbers R01LM013772 and R01LM013891. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work was also supported in part by the National Science Foundation (NSF) grant 1901117.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings. *ArXiv*, abs/1904.03323.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. 2021. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Geeticka Chauhan, Ruizhi Liao, William M. Wells, Jacob Andreas, Xin Wang, Seth J. Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. 2020. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 12262:529–539.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3942–3951.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv*, abs/2004.00849.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL*.

Zhanghexuan Ji, Mohammad Abuzar Shaikh, Dana Moukheiber, Sargur N. Srihari, Yifan Peng, and Mingchen Gao. 2021. Improving joint learning of chest x-ray and radiology report by word region alignment. In *MLMI@MICCAI*.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019a. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019b. Mimic-cxr: A large publicly available database of labeled chest radiographs. *ArXiv*, abs/1901.07042.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2020. Weak supervision helps emergence of word-object alignment and improves vision-language tasks. *ArXiv*, abs/1912.03063.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557.

Yikuan Li, Hanyin Wang, and Yuan Luo. 2020. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1999–2004.

Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth J. Berkowitz, Steven Horng, Polina Golland, and William M. Wells. 2021. Multimodal representation learning via maximization of local mutual information. In *MICCAI*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Ekta Sood, Fabian Kögel, Philippe Muller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2021. Multimodal integration of human-like attention in visual question answering. *ArXiv*, abs/2109.13139.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. *ArXiv*, abs/1908.08530.

Hao Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9049–9058.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *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 11–20, Hong Kong, China. Association for Computational Linguistics.

Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. 2021. Chest imagenome dataset for clinical reasoning. *ArXiv*, abs/2108.00316.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):1–17.

A More details

A.1 Challenges Applying General Domain Multimodal Models to Medical Data

Given recent progress made on open domain multimodal models, e.g., UNITER (Chen et al., 2020), it is reasonable to ask whether we can simply apply such models and pre-training schemes to multimodal medical data. However, a few key difficulties complicate straight-forward adaptation.

Necessity of Object Detectors. Many open domain models assume access to general *object detectors* during pre-processing. Such detectors are not readily available in the medical domain, and training object detection models requires large-scale, high-quality annotations for many different phenomena and/or anatomical regions. Further, one would need to collect such data for each domain in radiology (e.g., brain versus chest imaging).

In many multimodal models object detectors are used to produce bounding boxes, and are also tasked with inducing low-dimensional fixed-length vectors for significant regions, effectively taking care of region representation learning so that it need not be learned end-to-end. Open domain models often expect tens of bounding boxes in an image, but even a coarse segmentation of images (e.g., into a 19x19 grid as in GLoRIA) yields many more bounding boxes than this, exacerbating the mismatch between pre-trained general object detectors and the medical domain when the former are initialized from open-domain checkpoints.

Mismatch in Alignment Assumptions. UNITER uses **optimal transport** to align image and text vectors, but this assumes that each object (or salient part within an image) can be reasonably aligned to a segment of the corresponding text. This makes sense in the case of the general domain data like COCO (Lin et al., 2014) because usually we expect most detected objects to be mentioned in the caption. By contrast, in the medical domain we would expect that *most* parts of the image are unrelated to *any* portion of the corresponding text, and the task of the model is to identify salient regions of the text and match these with a particular image region. This is especially true when not using an object detector to identify the interesting regions as preprocessing step. The result of the optimal transport objective is that, averaging over tokens in the input text, each bounding box is equally important. In Section 2.2, we circumvent this problem by not using the optimal transport distribution itself (though that would be the natural choice), but instead using the attention mechanisms within UNITER.

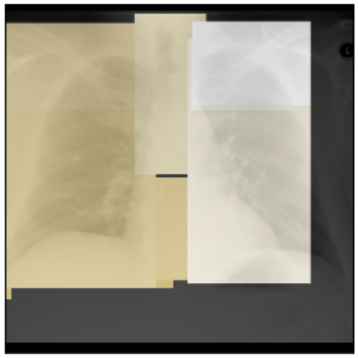

Figure 8: Here we show an example of what the UNITER attention over the ground truth bounding boxes looks like. It is easy to see why this attn usually has high localization with respect to the bounding boxes but also remains very unintuitive. It is still unclear if the UNITER model learns how to localize at all.

Despite these obstacles to re-purposing open domain multimodal models for this space, in Section A.2 we describe how we modify UNITER to serve as a baseline for our analysis for completeness.

A.2 UNITER Details

We use all reference anatomical bounding boxes available in the ImaGenome dataset. We reshape each bounding box to 45×45 pixels (enforcing fixed length), and then flatten and zero-pad the resultant vectors to be of length 2048 (the dimension UNITER expects). We train UNITER with a batch size of 4096 and 5 gradient accumulation steps on ImaGenome for 200k training steps.

For saliency we compute the mean attention over all 144 heads (12 layers $\times$ 12 heads) to produce pixel-wise scores (Gan et al., 2020). We take the mean of the attention when querying the text over the image, and when querying the image over the text; we normalize the resultant scores and treat these as analogous to a_{ij}, i.e., the saliency score relating the two modalities. We note that the absolute overlap scores between UNITER attention (just defined) and bounding boxes will be relatively high given that the UNITER attention is defined over the ground truth bounding boxes for all anatomical locations, and the bounding boxes used to evaluate the attention for a particular sentence are a subset of these same input bounding boxes. This also means we cannot use this approach in practice in the unsupervised setting in which we operate.

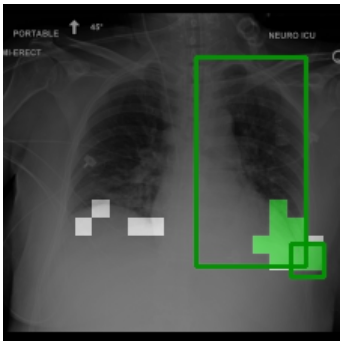

Figure 9: Visualization of Metrics. Attention map is thresholded, and true positives are shown in green.

A.3 Synthetic Sentences

In Section B, we include results involving synthetic sentences, which we describe here. To facilitate controlled experiments involving swapping out conditions — Section 3.1, **Synthetic+Swapped Conditions** — we also adopt a strategy for creating **synthetic sentences** using the labels from ImaGenome (Wu et al., 2021), and test our models on these sentences as well. Specifically, we construct these sentences using a set of rules to translate the condition and positive/negative context annotations and the anatomical names for the corresponding bounding boxes into natural language, as described in Table A.3.[14]

A.3.1 Rules for Creating Synthetic Sentences

Context	Condition (c)	Template
Pos	"Normal" or "Abnormal"	The {loclist} is/are {c}.
	Otherwise	There is {c} in the {loclist}.
Neg	-	There is no {c}.

Here we show the Rules for creating synthetic sentences. If there are multiple conditions in the sentence, we concatenate synthetic sentences for each of them. The "loclist" is created by turning the list of anatomical locations associated with the condition/context into a natural language list (e.g., "x," "x and y," or "x, y, and z"). We combine left- and right-side locations into one item ("left lung" and "right lung" is mapped to "lungs").

A.3.2 Synthetic Examples

In Table 5, we present examples of synthetic examples formed via the rules in Section A.3.

A.4 Metrics Figure

Figure 9 demonstrates what a thresholded (bilinearly upsampled) attention would look like and, for this specific threshold, which pixels are **true positives** (shown in green), false positives (shown in white), and false negatives (any other pixels inside either of the bounding boxes). For metrics such as **AUROC** and **Avg Precision**, statistics need to be computed while sliding through all possible thresholds.

[14]We present examples in the Appendix (Table 5).

A.5 Perturbations Details

Swap Left Right We replace every occurrence of the word "right" in the text with "left" and vice versa (ignoring capitalization). This is intended to probe the degree to which the attention mechanism relies on these two basic location cues. Of course, many sentences do not contain these words because conditions (or lack thereof) occur on both sides of the chest X-ray. Therefore, it is particularly important to look at the metrics on the "One Lung" subset (Section 3.2) for this perturbation.

Shuffle in Report We shuffle the sets of bounding boxes for different sentences in the same report at random. One would expect that performance would decrease significantly, because the resultant bounding boxes associated with given a sentence are (probably) wrong. However, sentences within the same report *might* be talking about similar regions. Therefore, for this perturbation it is important to look at the instances where the overlap between (a) the region of interest for the sentence and (b) the regions associated with *other* sentences in the report is low. We look at results for such cases explicitly using the **Most Diverse Report BBoxes (MDRB)** subset (Section 3.2).

Random Sentences We replace sentences in an instance with other sentences, randomly drawn from the rest of the dataset. Here too we expect performance to decrease significantly because the sampled text will refer to an entirely different image.

Random BBoxes We replace the set of bounding boxes for a sentence with a different set of bounding boxes randomly selected from the rest of the dataset. This differs from the **Random Sentences** perturbation in that the bounding boxes here are not only unrelated to the sentences, *but also unrelated to the image*. Therefore, we expect that this will have the poorest performance of all the settings, especially under the hypothesis that the attention is mostly a function of the image.

Synthetic+Swapped Conditions This is performed on the synthetic, rather than original, sentences because swapping out conditions can only be done reliably when we generate sentences. To swap

Original Sentence	Condition	Context	Location	Synthetic Sentence
Bulging mediastinum projecting over the left main bronchus and aortopulmonic window could be due to fat deposition exaggerated by low lung volumes.	low lung volumes	✔	left lung, right lung	There is low lung volumes in the lungs.
In the upper lobes, there is the suggestion of emphysema.	abnormal	✔	left mid lung zone, left upper lung zone, left lung, right mid lung zone, right upper lung zone	The left lung, upper lung zones, and mid lung zones are abnormal. There is copd/emphysema in the lungs, upper lung zones, and mid lung zones.
	copd/emphysema	✔	left mid lung zone, left upper lung zone, left lung, right mid lung zone, right upper lung zone	
Small left pleural effusion with atelectasis.	atelectasis	✔	left costophrenic angle	There is atelectasis in the left costophrenic angle.
No focal consolidation concerning for pneumonia.	pneumonia	✗	left lung, right lung	There is no pneumonia. There is no consolidation.
	consolidation	✗	right lung	
Mild bibasilar atelectasis.	abnormal	✔	left lower lung zone, left lung, right lung, right lower lung zone	The lungs and lower lung zones are abnormal. There is atelectasis in the lungs and lower lung zones. There is lung opacity in the lungs and lower lung zones.
	atelectasis	✔	left lower lung zone, left lung, right lung, right lower lung zone	
	lung opacity	✔	left lower lung zone, left lung, right lung, right lower lung zone	

Table 5: Examples of **Synthetic Sentences**.

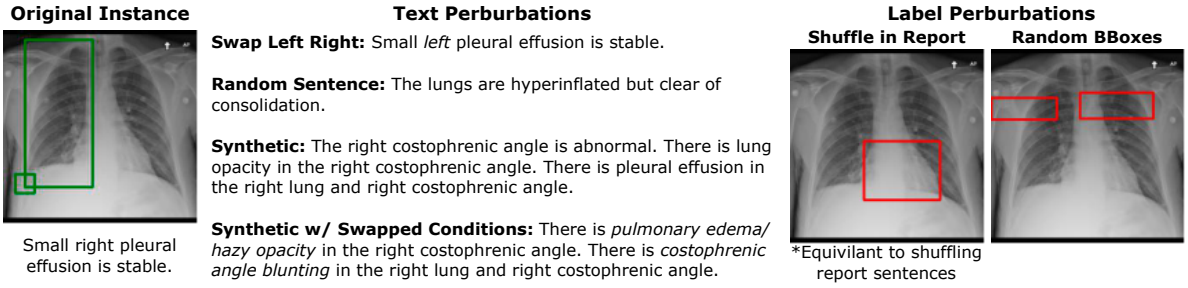

Figure 10: Examples of each **perturbation** (including **Snyth w/ Swapped Conditions**) for a given instance.

conditions, we follow the same rules for generating the synthetic sentence with a different condition randomly sampled from a set of (other) possible conditions. Possible conditions are defined as any condition (excluding the current) that occurs in the same anatomical locations anywhere else in the gold dataset.[15] This perturbation should measure the impact of conditions on model attention.

A.6 Subset Details

Abnormal Image/sentence pairs where there is an "abnormal" label associated with the sentence. This occurs if any conditions are mentioned in a *positive* context, i.e., where the radiologist believes the patient has said condition. This targets "interesting" examples where the attention should ideally highlight the region relevant to the condition described.

One Lung Image/sentence pairs where the bounding boxes corresponding to the sentence contain a bounding box of either the left or right lung, but not both. This subset allows us to evaluate how the model performs when the attention should only be on one side of the image.

Most Diverse Report BBoxes Instances where the overlap in the sets of bounding boxes for sentences within the same report is minimal. Specifically, we calculate the mean intersection over union (IOU; Section 2.3) of the segmentation labels ℓ_1, ℓ_2 for pairs of sentences in the same report. We then take the 10% of instances within reports with the smallest mean IOU. This subset is intended to include examples within reports where multiple distinct regions of interest discussed in different sentences.

These first two subsets are important because in many examples there is nothing abnormal, and the reports contain sentences such as "No effusion is present." For these types of sentences, the bounding boxes are commonly over both lungs because the evidence for the sentence is that nothing abnormal is in either lung. In these situations, it seems as though it might be easier for the model to realize higher scores for two reasons: 1) lungs take up most of the image, so attention is likely to fall in the bounding boxes, and 2) the lungs are a pretty good guess for the "important" regions of any image,

[15] If there are no other conditions, we leave the condition as is and the synthetic sentence is not perturbed.

independent of the text. The last subset is important because it comprises examples which contain a set of target bounding boxes and associated texts which cover mostly distinct image regions.

A.7 Annotations

In Figure 11, we present our user interface for collecting annotations created using streamlit. In Section A.7.1 (below), we show the annotation instructions.

A.7.1 Instructions

Our aim here is to collect judgements (*annotations*) concerning the interpretability and possible usefulness of alignments between text snippets and image regions induced by neural network models. More specifically, we will ask you to evaluate "heatmaps" output by different unsupervised (or minimally supervised) models which attempt to align natural language (sentences) and image regions (within accompanying chest X-rays). We ask three specific questions to asses these heatmaps; each question is 5-way multiple choice, and each of the answers are described below. In each round of annotation collection, we aim to collect annotations for multiple models with respect to a shared set of text snippets. That is, for each image, we ask for multiple assessments (across models) for the quality of alignments performed for a particular sentence. You will not be told which model generated which "heatmaps", and model aliases are randomly selected for every instance.

Prompts

You can choose the natural language sentences fed to the model—which we refer to as "prompts"—by either selecting a sentence from the list of sentences in the associated radiology report, or by writing your own "custom" prompt. We ask you to complete one round of annotations for report sentences, followed by one round in which you evaluate the alignments generated by the model for custom prompts (i.e., text you enter). For the report sentences round, we ask you to select one sentence that you think is interesting from the list of report sentences (prior to looking at any heatmaps). More specifically, you should, when possible, select a sentence with a focal abnormality that has strong clinical relevance. If one is not present, you can select a sentence that has a more diffuse abnormality or a negative statement that is still relatively focal. You will then annotate or judge the alignments induced by all models for this particular sentence. For instances that you do not think have any appropriate sentences or for instances where you can think of a better prompt, we ask you to write a prompt to annotate using the "custom prompt" option in addition to annotating the best sentence from the report.

Annotations

Bellow we list the questions and what each of the possible answers would mean.

1. **The heatmap includes what percentage of the region of interest from the prompt?**
    - 0-20 – The heatmap is focused on entirely the wrong part of the image, does not highlight any part of the image strongly, or has very minimal intensity on the region of interest.
    - 20-40
    - 40-60 – The heatmap comes close to covering the region of interest, or does cover the region of interest but with not too much intensity.
    - 60-80
    - 80-100 – The prompt refers to a region that is within a high-intensity part of the heatmap.
2. **What percentage of the heatmap represents an area of interest?**
    - 0-20 – This heatmap is all over the place or highlights a large portion of the image.
    - 20-40
    - 40-60 – The focus includes the relevant region(s) but also other irrelevant regions (either adjacent or elsewhere in the image).
    - 60-80
    - 80-100 – The heatmap is very targeted to only the parts of the image most relevant to the prompt.
3. **Rate how intuitive the heatmap is on a scale from 1-5 (1 being the worst, 5 being the best).**
    - 1 – The heatmap is completely unhelpful, counterintuitive, or misleading.
    - 2 – The heatmap might have something in common with an intuitive one, but very little.

- 3 – The heatmap does show a region of tiniest, but has some stray parts or does not catch all relevant regions.
- 4 – The heatmap is reasonably intuitive and contains mostly (though not exclusively) the regions I would expect.
- 5 – The heatmap is almost exactly what you might draw to represent the region of interest.

Ground truth bounding boxes

You have the ability to see *ground truth* bounding boxes from the dataset associated with the particular sentence you have selected from the report; these were manually drawn to match the corresponding sentence. We suggest that you use these bounding boxes when annotating the heatmaps associated with the report sentences. No such bounding boxes are available for the custom prompts that you will author.

A.8 No Attn Model Saturating Attn Map

Figure 12 depicts what happens when the model attends very highly to the "No Attn" token.

B Full/Additional Results

Here we include full/expanded results for the tables in the main paper and some additional results from which we may not yet have a takeaway.

B.1 Dropping Large Bounding Boxes for Evaluation

As discussed in sections 3 and 4.2.1, we noticed that in many cases in the ImaGenome dataset, the bounding boxes cover more than the true region of interest. We argue the original labels still serve us well in understanding when highlighted regions are far from where they should be, but to get a better sense of the precision of the models, we also trim out some of the larger boxes and repeat the evaluation from Tables 1 and 3 with the modified labels in Table 6.

Specifically for a sentence's label, we delete the "right lung"/"left lung" bounding box when there exist another bounding box within the label that contains the word "right"/"left". If no other box exists on the same side, then we still keep the full lung bounding box. As an example, in Figure 7 the larger of the two bounding boxes would be deleted from the label.

Model	AUROC	Avg. P
GLoRIA	**64.67**	**32.79**
GLoRIA Retrained	54.56	26.00
+Word Masking	59.88	28.62
+Clinical Masking	53.75	25.42
+"No Attn" Token	57.10	28.43
+Abnormal	57.83	29.42
+30-shot Finetuned	**61.98**	**35.27**

Table 6: Localization performance with large bounding boxes trimmed.

B.2 Custom Prompts

We also let annotators chose to write (and annotate) a fitting prompt if one was not present in the report. Figure 13 shows the annotations for these "custom" prompts for **GLoRIA** and the **+"No Attn" Token** models and indicates that even in this small, potentially out of domain setting, the scores are consistent with the in-domain annotations.

B.3 Localization performance for all models on all subsets.

Table 7 reports additional results to those in Table 3, describing localization performance on each subset individually.

Not shown in the main paper, we can see here that synthetic sentences perform comparably to real sentences, validating our method for constructing synthetic sentences. In fact, on **+30-shot Finetuned**, there is a significant jump in performance when using synthetic sentences.

B.4 Deltas of all models on all subsets

In Figure 14 we report results analogous to those in Figures 3 and 5, but on all subsets, all models, and all perturbations at once.

The results from swapping conditions in synthetic sentences, which were not shown in the main paper, vary across data subsets (Figure 14). The most telling subset for this perturbation is probably the **Abnormal** set. The results here are difficult to interpret because the **+Rand Sents** model seems to be considerably effected, which is counter-intuitive as we would expect this model to be invariant to the text by construction (note that the other perturbation results are consistent with this). Given this, we do not draw any particular conclusions from the swapped conditions experiment at present.

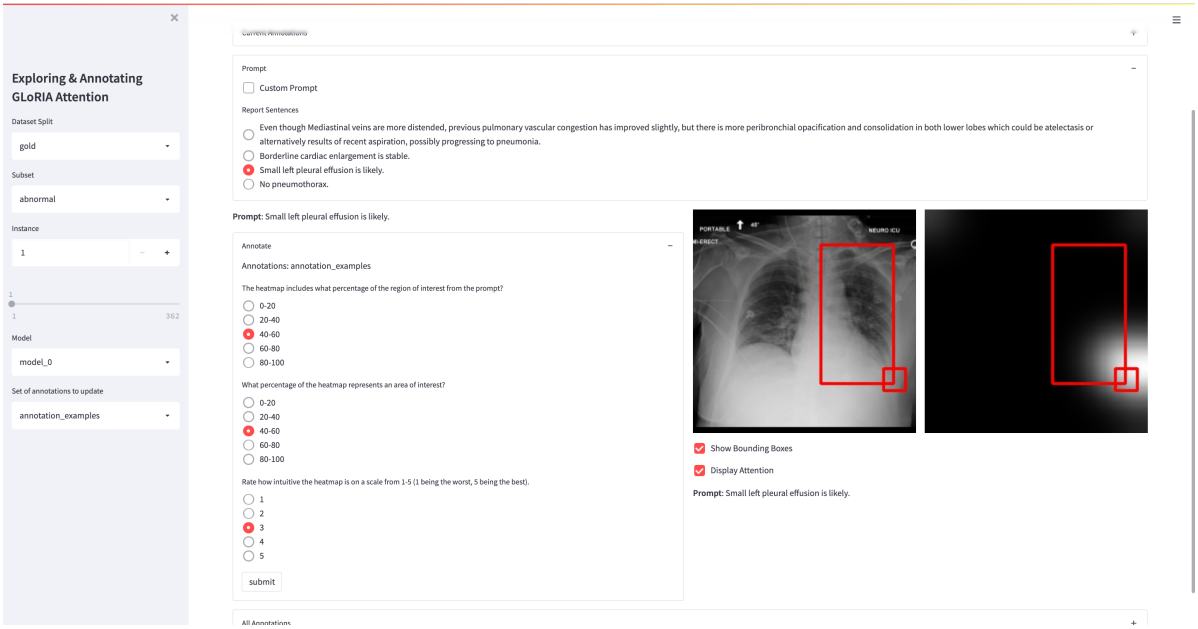

Figure 11: Interface

Model	Synth	All		Abnormal		One Lung		MDRB	
		AUROC	Avg. P	AUROC	Avg. P	AUROC	Avg. P	AUROC	Avg. P
UNITER*	✗	84.92	68.57	83.47	66.33	76.86	57.71	80.49	56.12
	✔	84.87	68.80	83.68	67.10	76.61	57.64	79.96	56.11
GLoRIA	✗	69.07	51.68	69.51	48.29	65.48	38.68	65.01	36.96
	✔	69.28	52.17	70.30	49.93	66.62	41.29	66.24	37.95
GLoRIA Retrained	✗	55.84	41.22	55.11	37.01	53.45	28.87	55.14	30.36
	✔	54.98	41.05	53.39	36.59	52.59	28.67	54.95	30.22
+Word Masking	✗	61.44	44.69	61.80	41.42	58.14	31.95	60.23	32.54
	✔	59.28	43.36	58.47	39.32	56.00	30.62	57.61	30.87
+Clinical Masking	✗	54.67	40.61	54.94	37.30	52.78	28.73	54.27	29.20
	✔	54.57	40.60	53.62	36.52	51.70	28.22	53.91	28.99
+"No Attn" Token	✗	57.00	41.80	57.32	39.20	56.47	32.65	56.76	31.08
	✔	56.29	41.66	56.62	38.92	56.09	32.69	56.18	30.84
+Abnormal	✗	55.89	43.42	57.59	42.20	54.68	33.01	55.33	32.32
	✔	52.78	41.86	54.05	39.96	51.22	30.48	53.15	31.01
+30-shot Finetuned	✗	63.90	52.80	65.28	50.44	61.61	40.79	62.16	39.91
	✔	68.38	56.05	73.14	55.57	67.92	45.00	66.26	42.28
+Rand Sents	✗	38.88	30.55	41.10	28.16	41.15	22.45	41.47	21.60
	✔	36.09	29.15	39.84	27.73	36.81	20.76	39.73	20.77

Table 7: **Localization performance** for each retrained model on the subsets. This also includes results on synthetic sentences (Section A.3).

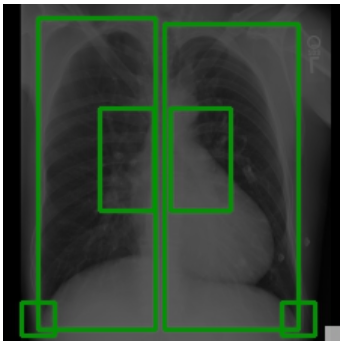

Figure 12: Example of when the attention map produced by the **+ "No Attn" Token** model is fully saturated by having strong attention on the "No Attn" token. The bottom right corner depicts the strength of the attention on the "No Attn" token, and the rest of the attention map is invisible because it has little attention in comparison.

B.5 Δ Average Precision

In Figure 15, we plot the analogous plot to Figure 14 for the changes in Average Precision as opposed to AUROC. Average Precision seems to tell a similar story to AUROC in terms of which models have greater changes for each perturbation. The only major difference is that for Average Precision, all models show a positive change for the Random BBoxes perturbation in the MDRB subset. This is likely because picking a random bounding box from the whole dataset when in this subset means that the random bounding box will likely be bigger than the original because the bounding boxes in this subset tend to be small. Having a larger bounding box as a label would therefore likely improve precision in general. This makes it harder to interpret this particular perturbation in this subset.

B.6 Random Attention KL Divergences

To measure the extent to which a model eschews the text and relies mostly on the image to induce an attention pattern, we introduce **Random Attention KL Divergence**. This is the symmetric Kullback–Leibler (KL) divergence for an instance between (a) the attention distribution induced given the original text, and (b) the attention over the same image but paired with random text. In Table 8, we show the mean **Random Attention KL Divergence** for each subset.

B.7 Candidate Selection Accuracy for other subsets

In 9, we extend Table 4 to the remaining subsets.

B.8 Entropy

In Table 10 we present results for the entropy attention mechanisms for each model for the entire dataset as well as the subsets.

B.9 Performance across Specific Abnormalities

In Figure 16, we present Intuitiveness for all models on examples with specific abnormalities.

B.10 Correlations

In Figures 18, 19, 20, 21, 22, 23, and 24, we present the pairwise pearson correlation over instances for a few different values for each model's outputs on the full gold split.

Most of the localization metrics here seem to be somewhat correlated, although not as much as one might expect. IOU seems to be generally more correlated with AUROC than with Average Precision.

Of particular note is the correlation between Attention Entropy and the global and local similarities: Attention Entropy is usually slightly positively correlated with Global Similarity and slightly negatively correlated with Local Similarity. Though it is still unclear why this is, it may have to do with a model's ability to localize seeing as this is more pronounced in models that perform better localization.

Finally, it is interesting that **+Abnormal** model has a somewhat negative correlation between Attention Entropy and all of the localization metrics, potentially indicating a connection between examples of abnormalities and Attention Entropy, but more work should be done to probe this further.

B.11 Precision and IOU at different Thresholds

Finally, we present Precision (Table 11) and IOU (Table 12) at different thresholds to get a better sense for the differences in the attention between each model. (Some IOU scores for GLoRIA are repeated here to allow for an easier comparison.) It is also clear that the Masking Model performs the best when only taking the top 5 or 10 percent, but GLoRIA starts producing similar or better scores at less strict thresholds. The precision scores above 70% here for **+Masking**, which far exceed any other model's scores at any threshold, give the sense that this model is quite effective at localization, but the dropoff when looking at the subsets do indicate the need for future work in this area.

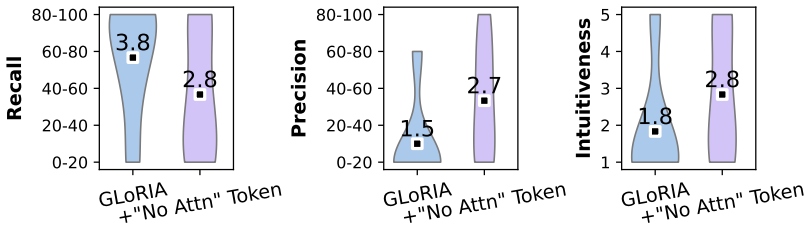

Figure 13: Custom annotation results (means over 6 instances).

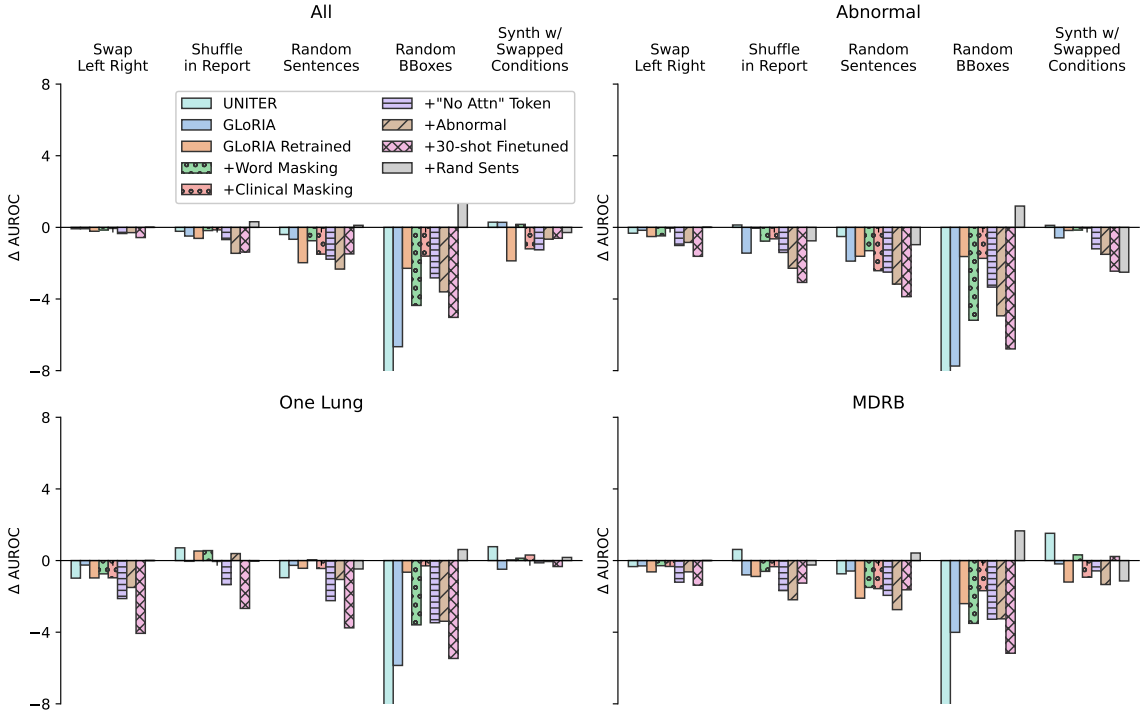

Figure 14: **ΔAUROC** for all models and subsets.

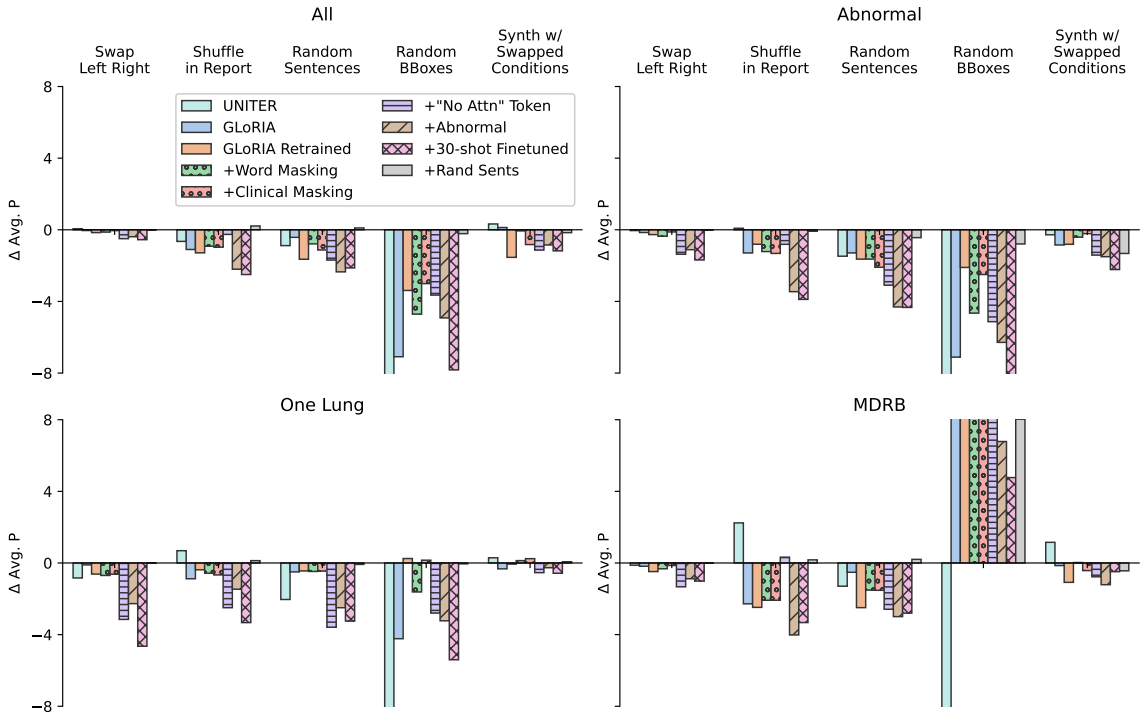

Figure 15: **ΔAverage Precision** for all models and subsets.

Model	All	Abnormal	One Lung	MDRB
UNITER	0.04	0.04	0.04	0.04
GLoRIA	0.08	0.08	0.08	0.09
Retrained	0.04	0.04	0.03	0.04
+Word Masking	0.05	0.04	0.04	0.05
+Clinical Masking	0.03	0.03	0.02	0.03
+"No Attn" Token	0.04	0.05	0.04	0.04
+Abnormal	0.11	0.11	0.10	0.11
+30-shot Finetuned	**0.17**	**0.16**	**0.15**	**0.17**
+Rand Sents	0.00	0.00	0.00	0.00

Table 8: Average **Random Attention KL Divergences** on the subsets

Model	One Lung		MDRB	
	local	global	local	global
UNITER	-	70.1	-	65.5
GLoRIA	38.9	72.3	53.6	73.8
GLoRIA Retrained	62.8	86.7	75.4	**84.1**
+Word Masking	**82.5**	**88.4**	**78.6**	81.7
+Clinical Masking	60.0	83.2	67.9	82.5
+"No Attn" Token	70.9	83.9	69.0	80.2
+Abnormal	72.3	85.6	73.0	75.4
+30-shot Finetuned	59.6	84.9	65.9	79.0
+Rand Sents	44.6	59.6	50.8	48.4

Table 9: **Candidate Selection Accuracy** for other subsets.

Model	**All**	**Abnormal**	**One Lung**	**MDRB**
UNITER*	**1.777**	**1.668**	**1.644**	**1.721**
GLoRIA	5.828	5.841	5.833	5.822
GLoRIA Retrained	5.857	5.863	5.872	5.862
+Word Masking	5.841	5.848	5.858	5.846
+Clinical Masking	5.864	5.866	5.876	5.868
+"No Attn" Token	5.849	5.855	5.861	5.856
+Abnormal	5.803	5.816	5.825	5.806
+30-shot Finetuned	**5.677**	**5.729**	**5.748**	**5.692**
+Rand Sents	5.889	5.889	5.889	5.889

Table 10: **Attention Entropy**

Model	Synth	**All**	**Abnormal**	**One Lung**	**MDRB**
UNITER*	✗	63.08/66.66/63.82	60.16/63.33/58.51	47.73/47.62/45.97	50.83/52.21/46.34
	✔	63.18/66.59/63.86	61.69/63.96/58.69	47.74/47.98/45.96	49.68/50.12/46.04
GLoRIA	✗	58.56/59.20/54.98	53.63/54.60/51.57	42.70/43.57/39.89	41.00/41.48/37.90
	✔	58.70/58.82/55.23	57.46/56.77/51.09	50.53/47.57/39.18	42.80/42.37/38.51
GLoRIA Retrained	✗	34.08/37.81/40.04	32.82/33.56/35.18	25.63/26.73/27.86	26.05/26.35/27.81
	✔	34.12/37.08/39.61	29.32/31.86/34.81	22.00/25.58/27.95	25.76/26.21/27.64
+Word Masking	✗	20.69/36.06/45.14	26.36/34.84/40.59	19.87/27.67/31.53	16.73/26.55/31.99
	✔	18.38/34.34/43.72	22.91/33.24/38.38	15.16/25.96/30.36	13.45/24.64/30.38
+Clinical Masking	✗	27.79/35.72/40.07	30.70/33.16/35.67	21.71/26.05/28.07	21.80/26.08/27.88
	✔	24.41/35.07/40.37	24.83/31.75/35.55	17.27/24.87/27.96	18.47/24.89/27.96
+"No Attn" Token	✗	37.93/38.97/40.19	40.48/35.93/35.75	35.38/28.68/28.34	28.24/27.59/28.12
	✔	36.37/37.67/40.17	39.44/34.43/35.94	36.70/28.19/28.79	28.09/26.50/28.29
+Abnormal	✗	42.95/33.30/39.20	48.32/37.29/36.36	40.43/30.22/28.08	34.69/25.58/27.47
	✔	35.11/26.45/37.95	33.62/23.35/36.04	25.90/16.99/27.77	26.31/19.12/27.06
+30-shot Finetuned	✗	73.15/69.35/39.67	67.45/64.35/37.87	52.55/49.84/30.49	53.46/49.25/28.10
	✔	70.95/70.05/46.92	62.81/62.99/51.54	52.04/50.24/39.39	50.95/49.66/34.66
+Rand Sents	✗	14.54/14.98/23.22	15.66/15.37/22.26	11.66/11.61/17.05	9.31/10.18/16.61
	✔	8.68/8.94/20.00	13.62/12.92/21.15	4.78/4.32/12.75	4.67/5.68/14.81

Table 11: **Precision** at 5/10/30%

Model	Synth	All	Abnormal	One Lung	MDRB
UNITER*	✗	2.57/7.17/33.61	2.56/6.95/34.88	2.83/7.81/30.78	3.14/9.21/30.33
	✔	2.71/7.54/34.13	2.76/8.36/35.53	2.88/8.03/31.85	3.73/9.77/30.50
GLoRIA	✗	3.79/6.69/20.10	4.10/7.25/19.05	4.43/8.05/20.54	3.56/6.37/16.92
	✔	4.89/8.96/23.62	7.20/13.25/29.30	7.55/12.82/27.69	4.83/8.24/19.84
GLoRIA Retrained	✗	2.51/3.80/4.21	3.10/3.86/4.08	2.29/2.87/3.14	3.27/4.68/4.82
	✔	2.75/4.21/4.74	3.43/3.89/4.21	2.33/2.77/3.21	3.25/4.82/5.01
+Word Masking	✗	1.79/2.60/3.48	2.77/3.69/4.34	2.43/3.19/3.90	2.14/2.83/3.40
	✔	1.55/2.26/3.06	2.50/3.11/3.48	1.81/2.31/2.88	1.54/2.24/2.65
+Clinical Masking	✗	1.63/2.18/2.54	2.66/3.12/3.26	1.36/1.58/1.73	1.89/2.54/2.67
	✔	1.65/2.17/2.72	2.19/2.45/2.70	1.14/1.45/1.78	2.18/2.59/2.93
+“No Attn” Token	✗	3.13/4.17/4.32	5.22/6.54/6.59	6.01/7.43/7.64	3.82/5.19/5.29
	✔	3.09/4.06/4.29	5.04/5.97/6.22	6.07/6.96/7.26	3.24/4.80/4.93
+Abnormal	✗	5.45/8.08/9.08	6.60/10.48/10.49	6.28/9.45/9.76	6.12/7.94/8.05
	✔	4.71/6.20/7.24	5.51/6.72/6.91	4.46/5.09/6.00	4.90/5.97/6.09
+30-shot Finetuned	✗	9.53/18.24/30.05	9.91/18.83/31.24	9.30/17.50/27.86	9.01/16.09/25.24
	✔	9.44/18.01/34.38	9.23/17.54/38.35	9.37/16.98/32.41	8.59/15.57/27.89
+Rand Sents	✗	0.35/0.76/5.51	0.36/0.62/4.85	0.43/0.76/4.68	0.16/0.59/4.46
	✔	0.45/0.94/7.35	0.47/0.75/5.90	0.66/1.11/7.18	0.22/0.70/6.22

Table 12: **IOU** at 5/10/30%

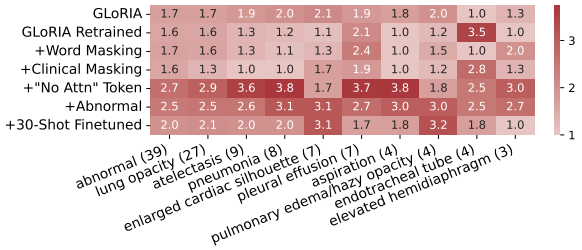

Figure 16: **Intuitiveness** on subsets of the annotations corresponding to the top 10 most frequent abnormalities.

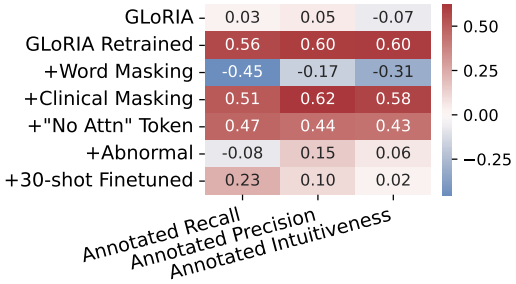

Figure 17: **Correlations with local similarity** (from heatmaps below).

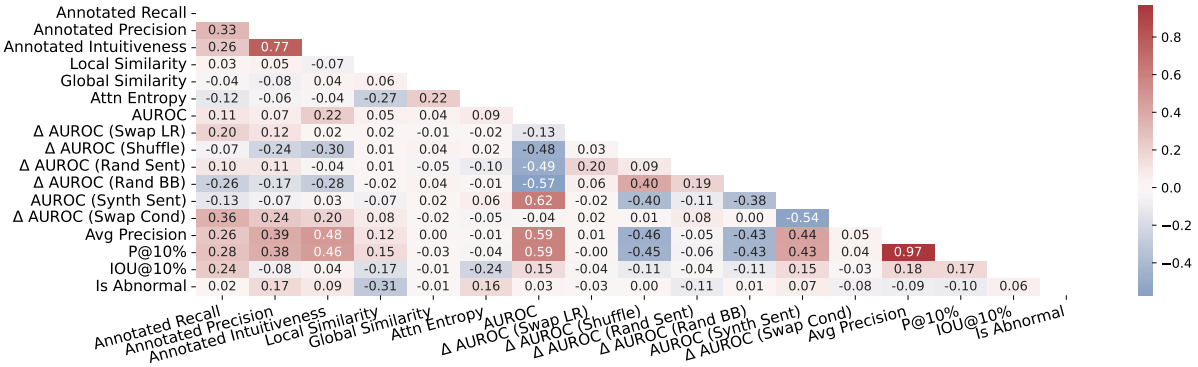

Figure 18: GLoRIA

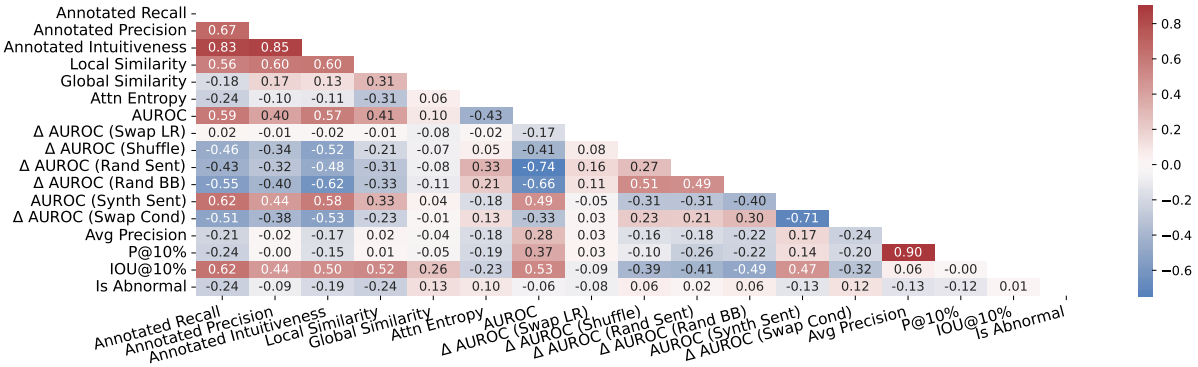

Figure 19: GLoRIA Retrained

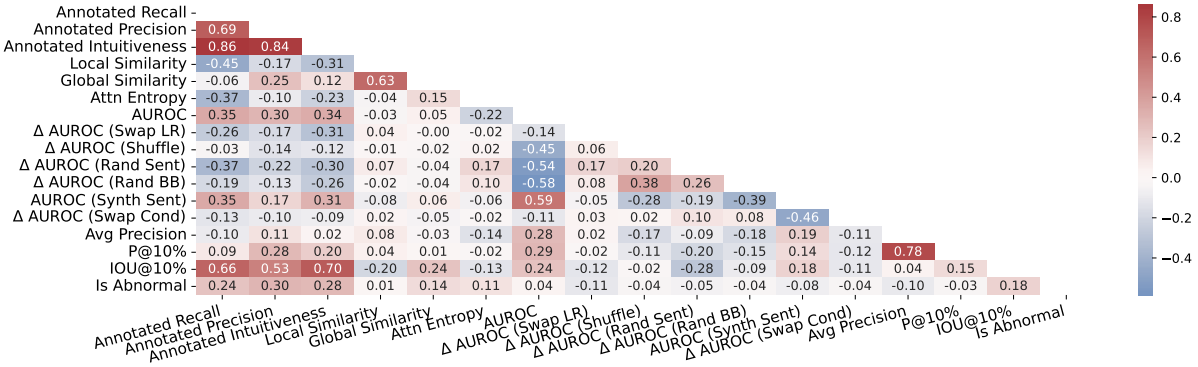

Figure 20: +Word Masking

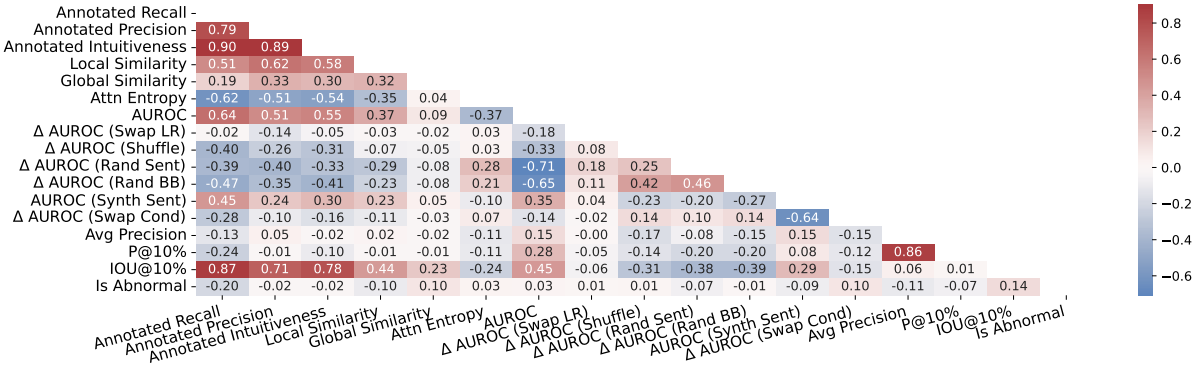

Figure 21: +Clinical Masking

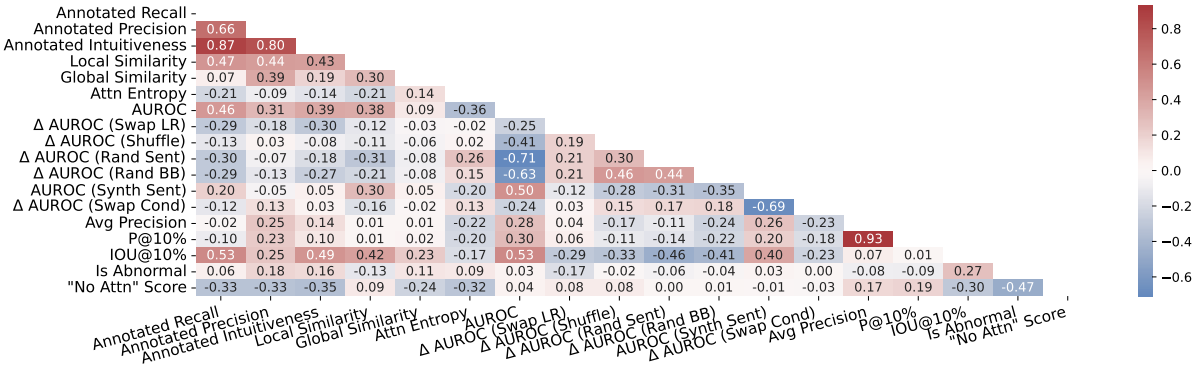

Figure 22: +"No Attn" Token

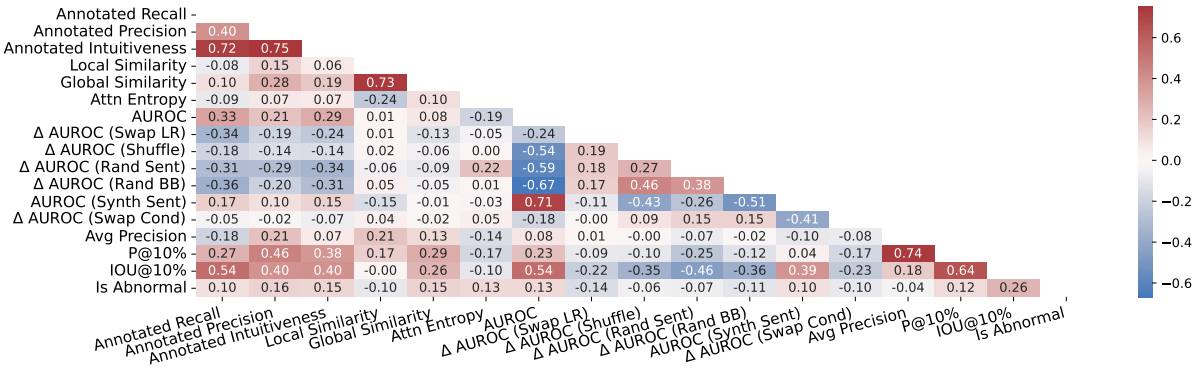

Figure 23: +Abnormal

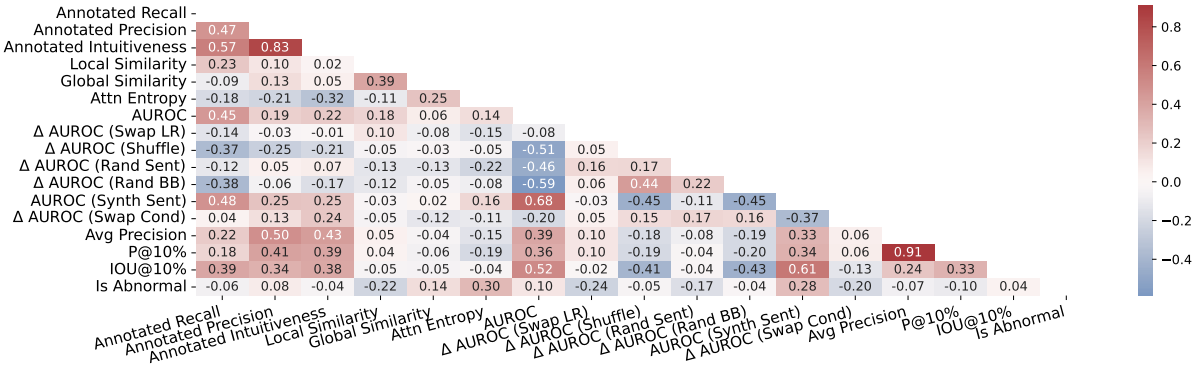

Figure 24: +30-shot Finetuned