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

Anonymous ACL submission

Abstract

Pretraining multimodal models on Electronic Health Records (EHRs) provides a means to learn rich representations that might transfer to downstream tasks with minimal supervision. Recent multimodal models induce soft local alignments between modalities (image regions and sentences). This is of particular interest in the medical domain, where alignments could serve to highlight regions in an image relevant to specific phenomena described in free-text. Past work has presented example "heatmaps" as qualitative evidence that cross-modal soft alignments can be interpreted in this manner. However, there has been little quantitative evaluation of such alignments. Here we compare alignments from a state-of-the-art multimodal (image and text) model for EHR with human annotations that associate image regions with sentences. Our main finding is that the text has surprisingly little influence on the attention; alignments do not consistently reflect basic anatomical information. Moreover, synthetic modifications, such as substituting "left" for "right," do not substantially influence attention. We find that simple techniques such as masking out entity names during training show promise in terms of their ability to improve alignments without additional supervision.

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 have been shown to 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 such models for data from Electronic Health Records (EHRs). However, key limitations preclude practical adoption of multimodal models in this domain. One important consideration is interpretability. Naive architectures that map image-text pairs to shared representations are opaque; it is not clear a priori what attributes they encode, and how they model interactions between the modalities. Consequently, doctors have no means of interrogating models, which may be relying on artifacts rather than meaningful clinical signal (Zech et al., 2018).

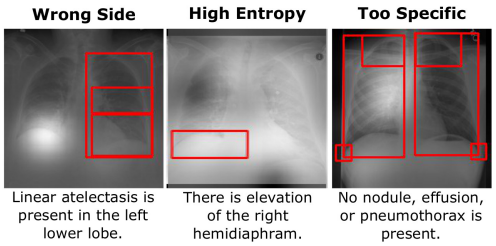

Figure 1: **Basic failure modes** of text-image alignment. We show smoothed attention saliency over images given texts. Red boxes are expert-provided target regions for the accompanying texts. (Note: imaging data is "mirrored", so right and left are effectively flipped.)

Recent work has proposed architectures that soft-align text snippets to image regions. This may afford one variety of interpretability by providing a means for practitioners to inspect what the model has "learned." Past work has provided *qualitative* evidence — illustrative multimodal "saliency" maps — suggesting that models can provide plausible looking outputs. But such alignments also run the risk of providing a false sense that the model "understands" more than it actually does.

Figure 1 illustrates a few obvious ways that a multimodal model may fail. It may simply focus on the wrong part of the image (e.g., the left rather than the right lung), or it might produce a high-entropy attention distribution, failing to meaningfully localize. Finally, the model may be myopic in its attention, missing the larger region of interest.

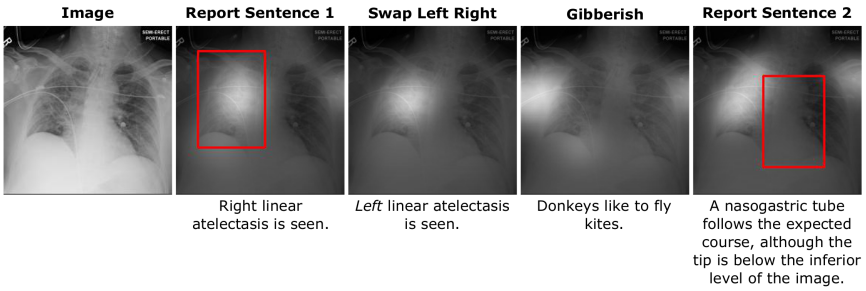

Figure 2: Alignment failures often occur when the model (overly) focuses on the image, largely ignoring the text.

Perhaps even more troubling, image attention may *appear* reasonable without actually reflecting both modalities. For example, see Figure 2. 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 over the image when provided text with radically different semantics (e.g., when swapping "right" with "left" or providing a nonsensical sentence), or when provided a different sentence altogether.

Recent work has sought to improve the ability 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. By contrast, our focus is on evaluating and improving localization itself, ideally without additional supervision.

The contributions of this work are as follows. (i) We critically appraise the interpretability of soft-alignments induced between images and texts by existing neural multimodal models for radiology. To our knowledge, this is the first such evaluation. We find that models that conditionally highlight regions relevant to a given text are often in fact more or less invariant to the text modality (Figure 2). (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 an existing multimodal model for EHR, namely 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. We next review details of this model, and then discuss the datasets we use to evaluate the alignments it induces.

2.1 GLoRIA

GLoRIA uses Clinical BERT (Alsentzer et al., 2019) as its text encoder and ResNet (He et al., 2016) as its image encoder. Unlike prior work, GLoRIA does not assume an image can be partitioned into different objects. This is at least in part because pre-trained object detectors are not readily available for X-rays. GLoRIA instead passes a CNN over the image to yield intermediate representations of local regions. This is useful because — as noted elsewhere (Huang et al., 2021) — a finding within an X-ray described in a report will usually be evident in only a small region of the corresponding image. 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 for each 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.

This soft alignment suggests a natural means to facilitate interpretability: One can create multimodal saliency maps indicating the magnitude of the attention assigned jointly to image regions and text snippets. And indeed, in (Huang et al., 2021), the authors show an example where the multimodal attention pattern jointly highlights words describing a specific abnormality while plausibly illuminating the region of the image in which this

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}.

Table 1: 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").

abnormality appears. One may be tempted to conclude that the model has somehow "understood" the text semantically, and successfully linked this to the image modality. However, our experiments below caution against reading too much into these patterns ostensibly linking text to image regions.

2.2 Data and Metrics

Data Our novel 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,[1] which include: (1) Bounding boxes for anatomical locations; (2) Links between referring sentences and image bounding boxes; and (3) A set of conditions and positive/negative context annotations[2] associated with each sentence/bounding box pair.

To facilitate controlled experiments involving swapping out conditions — Section 3.1, **Synthetic w/ 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 create synthetic sentences using a set of rules translating the condition and positive/negative context annotations and the anatomical names for the corresponding bounding boxes into natural language, which we describe in Table 1.[3]

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

[1]These annotations were first derived automatically, and then cleaned manually.

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

[3]We present examples in the Appendix (Table 11).

Synth	AUROC	Avg. P	IOU@5/10/30%
✗	69.07	51.68	3.79/6.69/20.10
✔	69.35	52.30	4.92/9.12/23.82

Table 2: Localization performance of GLoRIA.

We consider several metrics to measure the alignment between soft attention weights and bounding boxes. First, we associate scores with each pixel based on the attention weight assigned to the image region to which it belongs. Specifically, we use upsampling with bilinear interpolation to distribute the attention at the pixel level, creating scores $s \in \mathbb{R}^P$ where P is the number of image pixels. We use the bounding boxes to create a segmentation label $\ell \in \mathbb{R}^P$ where $\ell_i = 1$ if pixel i is in any of the bounding boxes, and $\ell_i = 0$ otherwise. Given this pixel-level score s and the pixel-level segmentation label ℓ_i, we can compute the **AUROC**, **Average Precision**, and **Intersection Over Union** (IOU) at varying pixel percentile thresholds for the ranking ordered by s.

One thing we want to measure is the extent to which a model may be ignoring the text and relying almost entirely on the image to determine an attention pattern. To this end we introduce **Random Attention KL Divergence**, 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.

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 perform an initial evaluation of the pretrained, publicly available weights for GLoRIA (Huang et al., 2021). Table 2 reports the metrics used to evaluate localization on the gold split of the ImaGenome dataset.

The AUROC scores are well over 50%, indicating reasonable localization performance. The IOU scores are relatively small, but this can be ex-

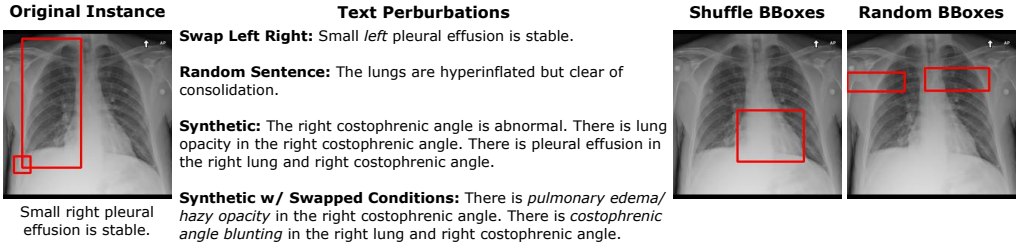

Figure 3: We present examples of each **perturbations** for a given instance.

pected because target bounding boxes tend to be much bigger than the actual regions of interest and serve more to detect errors when highlighted regions are far away from where they should be. This is validated by the relatively high average precision scores. (Arguably precision is more important than recall here because one would hope highlighted pixels are indeed relevant to the text, but would not necessarily expect this to be exhaustive.) However, while seemingly promising, our results below suggest that the attention patterns here may be less multimodal than one might expect.

Next we further probe multimodal attention distributions, with a specific focus on evaluating the degree to which these patterns actually reflect the associated text. To this end we first propose perturbing instances in ways that ought to shift the attention pattern, e.g., by replacing "right" with "left" in the text (Section 3.1). 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 3 shows examples of the perturbations we perform, which we describe in detail below.

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 have no mention of these words because conditions (or lack there of) occur on both sides of the chest X-ray. Therefore, it is particularly important to look at the metrics on the "One Lung" subset (see Section 3.2) in this setting.

Shuffle BBoxes Here the sets of bounding boxes for different sentences in the same report are shuffled 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 with (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 w/ Swapped Conditions This is performed on the synthetic sentences instead of the original sentences. This is because swapping out conditions can only be done in the case where we generate the sentence programmatically. To swap conditions in the sentence, we simply follow the same rules for generating the synthetic sentence with a different condition randomly sampled from a set of other possible conditions. We define these other possible conditions as any condition (excluding the current) that occurs in the same anatomical locations anywhere else in the gold dataset.[4] This perturbation should measure the impact of conditions on the model's attention mechanism.

Under these perturbations, we would expect a well-behaved model to shift its attention distribu-

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

tion over the image accordingly, and as a result we would expect the localization scores (overlap with the original reference bounding boxes) to decrease. 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 BBoxes**, which is equivalent to shuffling the sentences in a report).

If the assumption is that attention saliency maps reflect alignments with the input texts — as one might expect — then under these perturbations one should expect large negative differences in performance (Δs) relative to observed performance using the unperturbed data. For all but **Random BBoxes**, if the performance does not much change ($\Delta\text{s} \approx 0$), this suggests the attention maps are more or less invariant to the text modality.

3.2 Subsets

We also consider specific data subsets to perform more granular evaluations, enumerated below.

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 2.2) 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, 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.

Subset	Synth	AUROC	Avg. P	IOU@5/10/30%
Abnormal	✗	69.51	48.29	4.10/7.25/19.06
	✔	70.22	49.93	7.17/13.28/29.12
One Lung	✗	65.48	38.67	4.42/8.05/20.55
	✔	66.48	41.36	7.56/12.95/27.68
MDRB	✗	65.03	36.96	3.56/6.36/16.92
	✔	66.26	38.08	4.90/8.55/20.29

Table 3: Localization performance on different subsets.

3.3 Results

We first evaluate performance on the subsets described in Section 3.2 without perturbations so that we can later calculate differences in performances observed on these subsets following perturbations. We report results in Table 3. 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 the use of this disaggregated evaluation. Synthetic sentences still yield similar (even slightly improved) performance compared to the original sentences, suggesting the validity of our process for constructing these.

To measure the sensitivity of model attention to changes in the text, we report **differences in localization performance** in Figure 4. 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 over the full dataset and the subsets.[5]

The only real decrease in performance observed — even on the **One Lung** set — is under the **Random BBoxes** perturbation, which entails swapping out the target bounding box for an instance with

[5]We report results on the **Abnormal** and **MDRB** subsets in the Appendix; we include **One Lung** results here because these are important for interpreting the **Swap Left Right** perturbation results.

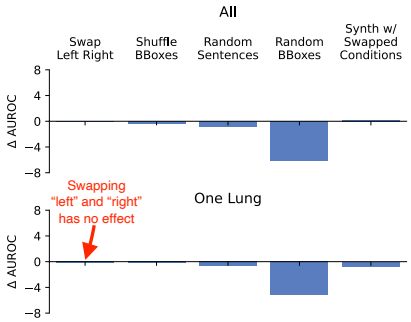

Figure 4: For each perturbation, we plot the change in localization performance (AUROC) for **GLoRIA** on the full reference set (top) and the **One Lung** subset (bottom).

one associated with some *other instance (image)*. Performance decreasing here 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 it seems 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. Swapping in synthetic sentences with the *wrong conditions* also results in only a marginal performance drop, which suggests conditions do not much influence the attention. Finally, swapping in other sentences (even from a different report) yields almost no performance difference.

4 Can We Improve Alignments?

The above results indicate that alignments between modalities are sub-optimal, e.g., image attention is less sensitive to the text modality than would be expected. Here we propose simple methods to try and improve image/text alignment performance, and specifically increase the model's sensitivity to changes in the text. We introduce (Section 4.1) and then evaluate (Section 4.2) our model variants.

4.1 Models

All models build on the GLoRIA architecture. 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 **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 w/ Abnormal** model variant on the subset of MIMIC-CXR/ImaGenome sentence/image pairs featuring an "abnormal" sentence.

Finally, we train a model in which we adopt a focused masking strategy intended to improve localization. Specifically, we use a clinical entity linker from `SciSpaCy` (Neumann et al., 2019) to find and remove entities in the text, replacing them with `[MASK]` tokens. We refer to this as the **Retrained w/ Masking** model.

Motivating this approach is the hypothesis discussed above, i.e., that attention patterns learned under existing regimes may largely function as an *unconditioned* attention map highlighting generally salient image regions, rather than accurately aligning image regions to specific accompanying texts. We mask biomedical entities specifically to try and prevent the model from using "shortcuts" where knowing the condition allows the model to leave incorporation of text information until computing a final (aggregated) similarity score. For example, if pneumonia is present anywhere in an image the model can discriminate between paired and unpaired texts based on whether "pneumonia" appears at all. If instead the text reads "`[MASK]` in the right lung", the model would be forced to specifically attend to the right lung to identify any abnormality to predict whether the text is correct.

Finally, we train **Retrained w/ Rand Sents** in the same style as the **Retrained** model except that all sentences are replaced with *random* sentences. This effectively deprives the model of any meaningful training signal, which otherwise comes entirely through the pairing of images and texts. Computing the metrics on this variant therefore provides a baseline from which to view the other models. For all models, we use early stopping with a patience of 10 epochs.[6] We will make code to reproduce all results available upon publication.

4.2 Results and Discussion

We observe in Figure 4 that **Retrained w/ Abnormal** performs better than **Retrained** in terms of overall localization scores, indicating that there is

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

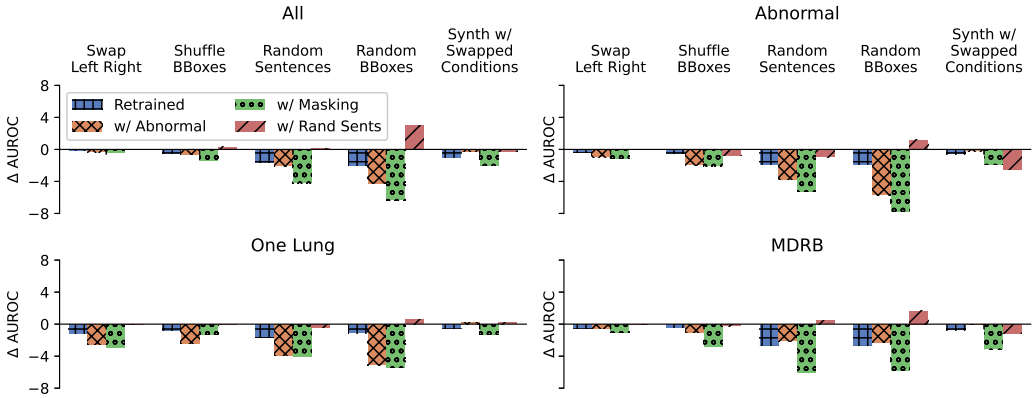

Figure 5: For each perturbation, we plot the change in localization performance (as measured by AUROC), for each of the models we retrain from scratch on the respective subsets.

Model	Synth	AUROC	Avg. P
Retrained	✗	58.44	43.31
	✔	58.14	43.18
w/ Abnormal	✗	61.18	45.78
	✔	59.05	44.84
w/ Masking	✗	**64.71**	**51.40**
	✔	**63.69**	**51.63**
w/ Randsent	✗	38.88	30.55
	✔	36.09	29.15

Table 4: Localization performance for each retrained model.

GLoRIA	7.99
Retrained	4.98
Retrained w/ Abnormal	6.74
Retrained w/ Masking	**10.63**
Retrained w/ Rand Sents	0.01

Table 5: Average **Random Attention KL Divergences** (Section 2.2). (See Appendix Table 8 for results on subsets.)

a benefit to having a higher concentration of abnormal examples, even if it means using less data. This may be due to the contrastive objective: Consider, e.g., the case where all instances are *not* abnormal. The reports may not feature any signal available with which to pair images to specific texts, providing little to learn from. Though 78.20% of training reports have some sentence indicating an abnormality, only 42.33% of all training *sentences* indicate an abnormality. Therefore, most of the image-text pairs in a batch have sentences with no abnormality, which may impact learning and localization.

It is also clear that **Retrained w/ Masking** performs the best out of the retrained models, suggesting the potential of masking to aid with localization. The pre-trained **GloRIA** mostly performs better than our retrained variant; we attribute this to differences in the training datasets given that we use the same code and architecture as (Huang et al., 2021) for training.[7] Therefore, differences between the retrained models offer evidence for the utility of both limiting to abnormal examples and using masking, although further exploration is warranted.

We next perform the perturbations introduced above to the proposed variants to assess sensitivity to input texts. We report results in Figure 5. **Retrained w/ Abnormal** is on average more sensitive to perturbations than **Retrained**, and **Retrained w/ Masking** outperforms all models (including the original **GLoRIA**[8]) in terms of being affected by the perturbations. Of particular note are results under the **Swap Left Right** perturbation on the **One Lung** subset, which is the subset that should show the most change under this perturbation.

In addition to being most sensitive to the perturbations considered, **Retrained w/ Masking** has the highest Random Attention KL Divergence (Table 5); defined in Section 2.2. This table also indicates that while the perturbations do not impact the localization scores much, cross-modal attention *does* still change for the vanilla models as a function of the text, much more than the model trained on random sentences. In other words, while the standard **GLoRIA** model is less sensitive to specific perturbations to the text, the attention distributions are not *entirely* invariant to the text like the model trained on random image and sentence pairs is.[9]

Table 6 reports the accuracy of each model in terms of identifying the correct sentence from two candidates for a given image. Perhaps surprisingly, these results indicate that performing comparatively well at identifying the correct sentence does

[7] In fact, some of this difference in performance may be explained by the higher percentage of abnormalities in the CheXpert dataset as discussed in Section 4.1.

[8] We note that the ΔAUROCs for **GLoRIA** are comparable to our variant **Retrained** except that **GLoRIA** experiences a larger drop under the **Random BBoxes** perturbation.

[9] The latter model is extreme because the only signal to learn from in pre-training is the pairing of images and texts.

	All		Abnormal	
Model	local	global	local	global
GLoRIA	54.9	71.0	41.3	77.3
Retrained	72.5	82.8	69.7	86.5
Retrained w/ Abnormal	68.7	76.4	79.3	85.4
Retrained w/ Masking	66.6	83.3	64.0	86.5
Retrained w/ Rand Sents	51.4	51.3	44.8	60.6

Table 6: 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.

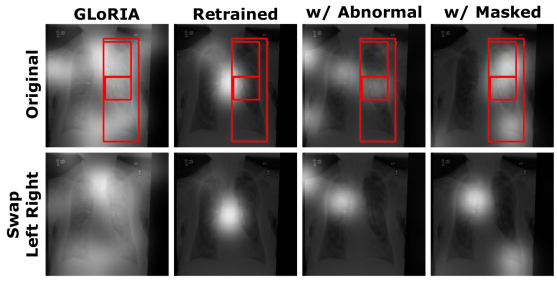

Figure 6: Model attention for the sentence: "Previously noted left upper lobe nodular opacity is not distinctly visualized on the current exam." (top), and perturbed version (bottom).

not necessarily correlate with localization ability. For example, the **Retrained** model performs best here, though we saw above that its localization is somewhat invariant to the perturbations considered. Note that **Retrained w/ Abnormal** was trained on a small subset for this task, comprising *only* abnormal examples; this may explain its middling accuracy on the full set (it does well on the **Abnormal** set). **Retrained w/ Masking** was similarly trained on a distribution that differs from the test set here (it is used to seeing `MASK` tokens). The original **GLoRIA** weights perform somewhat poorly here, perhaps owing to a domain shift.

We conclude with a qualitative impression of localization performance. Figure 6 shows model attention distributions for a cherry-picked instance and the accompanying **Swap Left Right** perturbation. In this example (selected as an illustrative case where the masking variant improves model behavior), **GLoRIA** yields a high-entropy map, while **Retrained** delineates roughly the correct region. The perturbation does not affect either of these models' attention distributions much at all. The **Retrained w/ Abnormal** model is also almost focused in the right area, but it shifts attention more to the right lung following the perturbation. Finally, the **Masked w/ Abnormal** variant identifies the correct area in the original instance, and switches as expected to the right lung after the perturbation.

Summary of Key Findings Training with randomly paired images and sentences yields models that produce attention maps that are *completely* independent of the text (Tables 5 and 6). This results in high-entropy distributions because there is nothing for the model to learn.[10] Existing multimodal pretraining schemes beget models that accurately select the text that matches a given image (Table 6), and yield attention distributions that at least somewhat depend on the text (Table 5). But perturbing texts does not cause the changes in attention patterns one would intuitively expect, and do not much affect localization performance (Figure 5).

Some simple changes to the pre-training process may mitigate this behavior. In particular, masking tokens during training seems to result in models that produce attention patterns which more intuitively depend on input texts (Figure 5 and Table 5), although this may slightly harm performance on the pre-training task itself (Table 6).

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).

6 Conclusions

We evaluated an existing state-of-the-art unsupervised multimodal representation learning model for EHRs in terms of inducing fine-grained alignments between image regions and text. We found that the resultant heatmaps are often invariant to perturbations to the text that ought to change them substantially, which seems problematic. We proposed two methods that somewhat improved this model behavior: training with abnormal examples and training with masking. We hope that this effort motivates more work addressing the interpreteability of multimodal encoders for healthcare.

[10] See Appendix Table 10

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*.

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*.

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.

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 Appendix

A.1 GLoRIA Deltas on other Subsets

In Figure 7 we report results analogous to those in 4, but on the other subsets introduced.

A.2 Localization performance for each retrained model on subsets.

Table 7 reports additional results to those in Table 4, describing localization performance on each subset individually.

A.3 Random Attention KL Divergences for subsets

In Table 8, we extend Table 5 to show the mean **Random Attention KL Divergence** for each subset.

A.4 Candidate Selection Accuracy for other subsets

In 9, we extend Table 6 to the remaining subsets.

A.5 Entropy

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

A.6 Δ Average Precision

In Figures 8 and 9, we plot the analogous plots to Figures 4, 7, and 5 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.

A.7 Synthetic Examples

In Table 11, we present examples of synthetic examples formed via the rules in Table 1.

A.8 Correlations

In Table 12 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 Retrained w/ 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.

A.9 Precision and IOU at different Thresholds

Finally, we present Precision (Table 13) and IOU (Table 14) 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 **Retrained w/ 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.

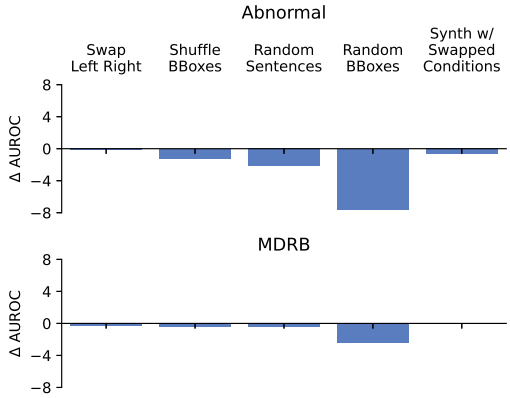

Figure 7: Same as Figure 4 for the last two subsets.

Model	Synth	Abnormal		One Lung		MDRB	
		AUROC	Avg. P	AUROC	Avg. P	AUROC	Avg. P
Retrained	✗	57.68	39.40	55.84	30.88	58.60	32.91
	✔	57.18	39.08	56.52	31.17	57.87	32.29
w/ Abnormal	✗	62.24	43.63	60.52	36.71	59.08	33.56
	✔	58.20	41.80	55.98	34.39	57.44	32.74
w/ Masking	✗	**65.70**	**49.78**	**62.07**	**40.41**	**64.73**	**39.95**
	✔	**64.37**	**49.62**	**61.51**	**41.44**	**63.56**	**39.76**
w/ Randsent	✗	41.10	28.16	41.15	22.45	41.47	21.60
	✔	39.84	27.73	36.81	20.76	39.73	20.77

Table 7: **Localization performance** for each retrained model on the subsets.

Model	**Abnormal**	**One Lung**	**MDRB**
GLoRIA	7.71	8.09	8.68
Retrained	5.31	4.63	5.28
Retrained w/ Abnormal	6.89	6.01	6.46
Retrained w/ Masking	10.54	9.94	11.27
Retrained w/ Rand Sents	0.01	0.01	0.01

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

Model	**One Lung**		**MDRB**	
	local	global	local	global
GLoRIA	40.4	74.4	53.6	76.6
Retrained	74.4	88.1	74.6	86.5
Retrained w/ Abnormal	83.2	84.9	70.2	76.2
Retrained w/ Masking	61.4	86.7	67.5	84.5
Retrained w/ Rand Sents	44.6	59.6	50.8	48.4

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

Model	**All**	**Abnormal**	**One Lung**	**MDRB**
GLoRIA	5.828	5.841	5.833	5.822
Retrained	5.825	5.833	5.841	5.827
Retrained w/ Abnormal	5.839	5.840	5.849	5.845
Retrained w/ Masking	5.791	5.809	5.815	5.793
Retrained w/ Rand Sents	5.889	5.889	5.889	5.889

Table 10: **Attention Entropy**

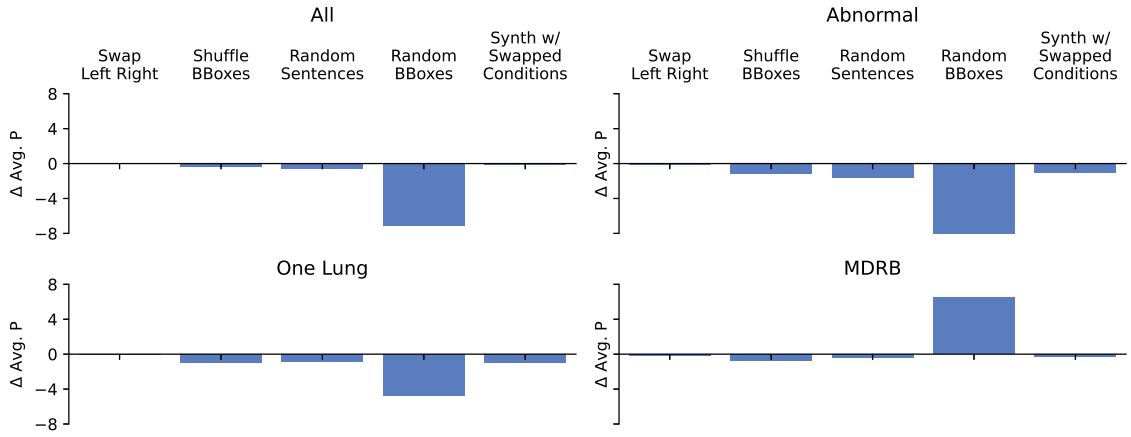

Figure 8: Δ **Average Precision** for GLoRIA

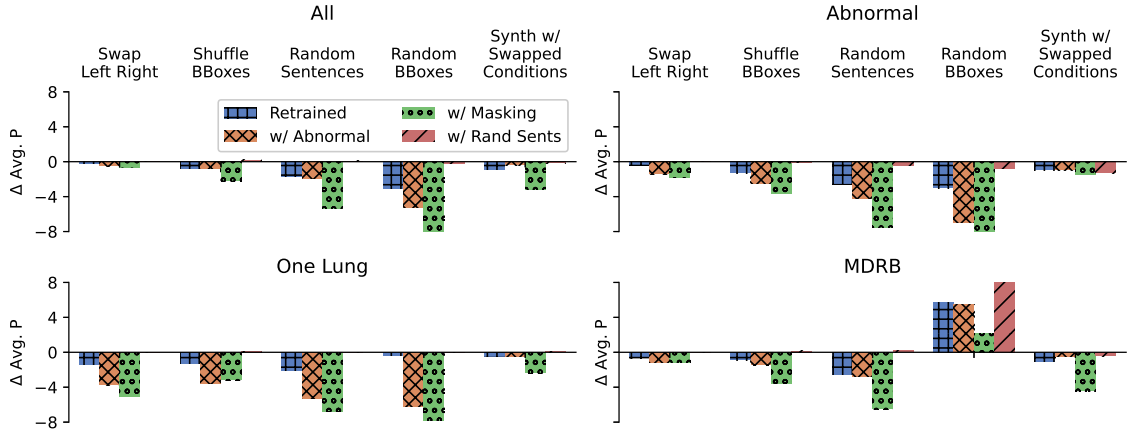

Figure 9: Δ **Average Precision** for retrained models.

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 11: Examples of **Synthetic Sentences**.

	Local Similarity	Global Similarity	Attn Entropy	AUROC	Avg Precision	P@10%	IOU@10%
GLoRIA							
Global Similarity	0.059						
Attn Entropy	**-0.275**	**0.220**					
AUROC	0.053	0.040	0.089				
Avg Precision	**0.116**	0.004	-0.006	**0.592**			
P@10%	**0.154**	-0.028	-0.041	**0.593**	**0.970**		
IOU@10%	**-0.168**	-0.012	**-0.238**	**0.152**	**0.177**	**0.168**	
Is Abnormal	**-0.307**	-0.007	**0.158**	0.033	-0.092	**-0.100**	0.059
Retrained							
Global Similarity	**0.410**						
Attn Entropy	-0.023	**0.189**					
AUROC	**0.384**	**0.106**	**-0.117**				
Avg Precision	-0.050	-0.028	-0.076	**0.203**			
P@10%	-0.027	-0.013	-0.067	**0.163**	**0.927**		
IOU@10%	**0.431**	**0.232**	0.000	**0.588**	-0.012	-0.090	
Is Abnormal	**-0.119**	**0.163**	0.099	-0.080	**-0.131**	**-0.116**	0.083
Retrained w/ Abnormal							
Global Similarity	**0.488**						
Attn Entropy	**-0.212**	0.086					
AUROC	**0.309**	**0.204**	**-0.298**				
Avg Precision	**0.135**	**0.169**	**-0.203**	**0.499**			
P@10%	**0.206**	**0.219**	**-0.206**	**0.562**	**0.821**		
IOU@10%	**0.378**	**0.296**	**-0.259**	**0.557**	**0.282**	**0.489**	
Is Abnormal	**0.286**	**0.188**	0.020	**0.107**	-0.064	0.025	**0.303**
Retrained w/ Masking							
Global Similarity	**0.138**						
Attn Entropy	**-0.324**	**0.250**					
AUROC	**-0.162**	**0.229**	-0.046				
Avg Precision	**0.251**	0.038	**-0.175**	**0.211**			
P@10%	**0.223**	**0.151**	**-0.156**	**0.236**	**0.927**		
IOU@10%	**-0.129**	**0.116**	-0.057	**0.435**	**0.113**	**0.260**	
Is Abnormal	-0.062	**0.109**	**0.144**	0.080	-0.052	0.031	**0.221**

Table 12: **Correlations** for positive pairs. Any number over 0.1 is bolded.

Model	Synth	All	Abnormal	One Lung	MDRB
GLoRIA	✗	58.58/59.20/**54.93**	53.63/54.60/**51.54**	42.65/43.55/**39.82**	40.98/41.46/**37.90**
	✔	58.88/59.19/**55.23**	57.28/57.06/**50.95**	50.87/48.12/**38.80**	42.77/43.30/**38.39**
Retrained	✗	46.54/41.07/39.78	46.53/37.20/35.20	35.86/29.05/28.03	38.89/30.61/27.73
	✔	45.01/40.51/39.92	43.04/35.72/35.74	33.20/29.09/29.10	34.84/29.73/27.91
w/ Abnormal	✗	40.13/39.58/44.11	49.27/40.56/39.76	45.22/35.10/31.44	32.08/28.84/30.55
	✔	35.78/35.19/43.24	42.34/33.56/38.29	39.90/29.20/29.73	28.52/25.73/29.96
w/ Masking	✗	**71.38/63.68**/44.90	**74.19/65.01**/39.72	**64.26/53.01**/31.19	**55.22/48.93**/32.01
	✔	**73.97/64.99**/44.76	**74.59/63.65**/39.52	**66.36/53.28**/31.74	**56.73/48.83**/31.76
w/ 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 13: **Precision** at 5/10/30%

Model	Synth	All	Abnormal	One Lung	MDRB
GLoRIA	✗	3.79/6.69/20.10	4.10/7.25/19.06	4.42/8.05/20.55	3.56/6.36/16.92
	✔	4.92/9.12/**23.82**	7.17/13.28/**29.12**	7.56/12.95/**27.68**	4.90/8.55/**20.29**
Retrained	✗	5.59/7.61/7.99	6.48/8.54/8.64	5.95/7.15/7.37	5.98/8.57/8.84
	✔	5.50/7.11/7.65	6.24/7.39/7.53	5.47/6.31/6.55	5.60/7.64/8.03
w/ Abnormal	✗	4.27/6.48/7.18	5.87/9.30/9.51	6.92/10.31/10.24	4.53/6.57/7.08
	✔	4.21/5.66/6.41	5.60/7.49/7.61	6.66/8.15/8.35	4.41/5.77/6.11
w/ Masking	✗	**7.76/15.43/20.60**	**9.01/18.12/22.76**	**8.98/17.62/20.76**	**8.40/14.93/18.83**
	✔	**9.41/16.74**/21.61	**11.35/19.12**/22.00	**10.73/18.29**/20.26	**9.45/15.70**/19.25
w/ 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 14: **IOU** at 5/10/30%