

# Supplementary Material of Interactive and Explainable Region-guided Radiology Report Generation

Tim Tanida<sup>1,\*</sup> Philip Müller<sup>1,\*</sup> Georgios Kaassis<sup>1,2</sup> Daniel Rueckert<sup>1,3</sup>

<sup>1</sup>Technical University of Munich, <sup>2</sup>Helmholtz Zentrum Munich, <sup>3</sup>Imperial College London

{tim.tanida, philip.j.mueller, g.kaassis, daniel.rueckert}@tum.de

## A. Detailed analysis

### A.1. Ablation study

Tab. 1 shows the results of an ablation study on the region selection and abnormality classification modules. For full report generation, our method at minimum requires an object detector to extract region visual features and a language model to generate region-specific sentences, thus these two modules together form the base model. We investigate the effects of incorporating the abnormality classification and region selection module, respectively, into this base model by evaluating on BLEU-4, METEOR, and clinical efficacy (CE) metrics micro-averaged over five observations.

We observe that adding the abnormality classification module has a negligible effect on conventional natural language generation (NLG) metrics of BLEU-4 and METEOR, whilst substantially improving CE recall by +10.2% ( $\Delta+28.4\%$ ) at the slight expense of CE precision. This showcases that 1) conventional NLG metrics are ill-suited for evaluating the clinical accuracy of generated reports [1,8,13] and 2) the abnormality classification module effectively encodes abnormality information in the region visual features, as evidenced by the substantial increase in recall.

Incorporating the region selection module substantially boosts the performance of the base model across all metrics, likely due to the changed approach in training the language model once the region selection module is introduced to the base model. In the base model, the language model is trained with all reference sentences (*i.e.*, empty and non-empty) of all 29 regions per image, as the generated sentences of all 29 regions are concatenated to form the final report. Since there are 2.2 times more empty reference sentences than non-empty reference sentences (see weighted binary cross-entropy loss of the region selection module in Appendix C.2), the language model learns to often generate empty sentences for regions. Thus intuitively, the language model in the base model is not only tasked with generating region-specific sentences, but also with "deciding" which

regions require non-empty sentences. In addition, the generated reports of the base model are shorter than those of the base model + region selection module. This is because even though the language model in the base model generates sentences for all 29 regions (which are concatenated to form the final report), a lot of these generated sentences will be empty. We verify this by calculating the average number of tokens (using a Spacy tokenizer) in a generated report by the base model vs. base model + region selection module. While a generated report by the base model contains on average 39 tokens, incorporating the region selection module increases this to 52 tokens. Thus, the base model may not be generating sufficiently long reports containing region-specific sentences that accurately describe abnormalities, which may be reflected in the low CE recall score.

When the region selection module is incorporated into the base model, the language model is trained exclusively on region visual features with corresponding non-empty reference sentences. This removes the implicit task of "deciding" which regions need sentences, potentially allowing the language model more capacity to generate better region-specific sentences. This could explain the  $\Delta+9.6\%$  increase in the BLEU-4 score and  $\Delta+19.3\%$  increase in the METEOR score. Additionally, we can see that compared to the base model, the CE recall improves significantly by +19.2% ( $\Delta+53.5\%$ ), which is likely due to the increased capacity in generating better region-specific sentences, thus more abnormalities are correctly described in the final reports. However, we also observe a noticeable decrease in CE precision score compared to the baseline. This may be attributed to the low precision score of the region selection module w.r.t. normal regions (see Appendix A.3), leading to more normal regions being described in generated reports and thus increasing the likelihood of false positives (*i.e.*, normal regions being described as abnormal).

Finally, by combining the abnormality classification and region selection modules in the RGRG model (outlined in gray), we again see an increase in CE recall, verifying the effectiveness and relevance of both modules for the overall model performance.

\*Equal contribution

Dataset	Object detector	Abnormality classification	Region selection	Language model	BLEU-4	METEOR	$P_{\text{mic-5}}$	$R_{\text{mic-5}}$	$F_{1, \text{mic-5}}$
MIMIC-CXR	✓			✓	0.104	0.135	0.578	0.359	0.443
	✓	✓		✓	0.107	0.138	0.550	0.461	0.501
	✓		✓	✓	0.114	0.161	0.498	0.551	0.523
	✓	✓	✓	✓	<b>0.126</b>	<b>0.168</b>	0.491	<b>0.617</b>	<b>0.547</b>

Table 1. Ablation study on the abnormality classification and region selection modules. The performance is evaluated on two natural language generation metrics (BLEU-4 and METEOR) and clinical efficacy metrics micro averaged over five observations. Each module contributes to an increased performance (especially in recall) of the RGRG model.

Region	RL	LL	SP	MED	CS	AB	Average	Avg. num. detected regions
IoU	0.925	0.920	0.950	0.870	0.837	0.913	0.887	28.792

Table 2. Object detector results micro averaged over all anatomical regions as well as 6 prominent regions: *right lung* (RL), *left lung* (LL), *spine* (SP), *mediastinum* (MED), *cardiac silhouette* (CS) and *abdomen* (AB). Almost all 29 anatomical regions are detected per image with adequate IoU scores.

Module	Regions	P	R	$F_1$
Region Selection	All	0.594	0.904	0.717
	Normal	0.459	0.903	0.608
	Abnormal	1.0	0.906	0.951
Abnorm. Classifier	All	0.354	0.911	0.510

Table 3. Results of the region selection and abnormality classification modules. Salient anatomical regions are selected for the final report with high recall for both normal and abnormal regions, at the expense of precision for normal regions. Anatomical regions are classified as abnormal with high recall but decreased precision.

## A.2. Object detector results

We evaluate the object detector via the Intersection over Union (IoU) metric, which we calculate as the sum of the intersection areas divided by the sum of union areas. We use the IoU metric instead of the (in object detection) more commonly used mean Average Precision (mAP) metric, since each anatomical region typically appears exactly once in an image, and never more than once. We report the micro average IoU score over all regions as well as for 6 prominent regions. Additionally, we report the average number of detected regions per image.

The IoU scores in Tab. 2 demonstrate that anatomical regions are detected adequately, with almost all 29 regions being detected per image with an average IoU score of 0.887. We noticed that the ground-truth bounding boxes in the Chest Imagenet dataset, which were automatically extracted by a bounding box pipeline, do not always precisely overlap with the real regions, which likely negatively impacted the IoU scores. However, since ultimately the goal is to generate consistent anatomy-related sentences (and not perfect object detection), we believe that imperfect object detection is acceptable.

## A.3. Region selection and abnormality classification results

We evaluate the binary classifiers of the two modules on precision, recall, and F1 score. For the region selection module, a region is deemed positive if it has a corresponding reference sentence, and for the abnormality classification module, a region is positive if it is abnormal as per ground-truth. For region selection, we additionally report the scores for the subsets of normal and abnormal regions.

Tab. 3 showcases the results. We observe that recall is high for both normal and abnormal regions for region selection, thus regions that are described in the reference report are also usually selected for the generated report. However, precision is low for normal regions, meaning usually more normal regions are selected for the generated report than are described in the reference report. As mentioned in the main paper, this can explain the low score for the ROUGE-L (F1) metric, since the generated report thus contains more information than the reference report, which in turn causes a lower ROUGE-L precision score. However, the decision to describe normal regions (*e.g.*, "There is no pleural effusion or pneumothorax.") lies with the radiologist and is arbitrary, since pathology-free regions are not required to be mentioned in a report. Thus, we believe that this rather subjective decision cannot be learned by a model and a low precision score for normal regions is expected.

Precision is 1.0 for abnormal regions since by default abnormal regions are always included in reference reports. Hence, there cannot be any false positives for the abnormal region subset. Consequently, the recall score for the normal and abnormal region subsets cannot be directly compared.

For the abnormality classifier, we observe high recall and low precision. Thus, abnormal regions are usually detected correctly while normal regions are sometimes misclassified.

## B. Detailed results

### B.1. Qualitative anatomy-based sentence generation results

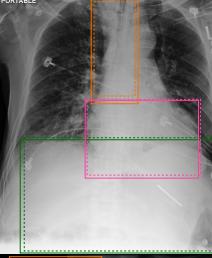
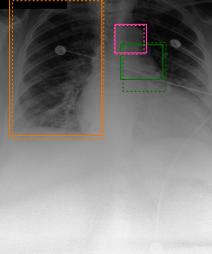
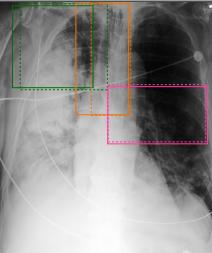
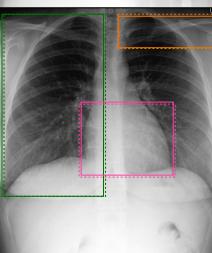
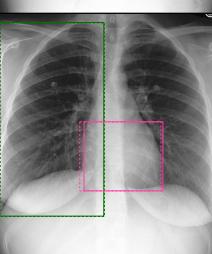
	<p><b>Abdomen:</b> Generated: NG tube tip is in the stomach. Reference: Two enteric tube tips terminate within the stomach.</p> <p><b>Trachea:</b> Generated: The endotracheal tube is in standard position. Reference: Endotracheal tube tip is slightly low lying, terminating approximately 2.3 cm from the carina.</p> <p><b>Cardiac silhouette:</b> Generated: The heart is borderline enlarged. Reference: Moderate cardiomegaly.</p>
	<p><b>Left hilar structures:</b> Generated: There is mild pulmonary edema. Reference: Moderate pulmonary edema is present.</p> <p><b>Right lung:</b> Generated: There is no pneumothorax or pleural effusion. Reference: There is no evidence of pneumothorax or pleural effusions.</p> <p><b>Aortic arch:</b> Generated: The aorta is tortuous. Reference: The thoracic aorta is slightly ectatic.</p>
	<p><b>Right upper lung zone:</b> Generated: The right upper lobe opacity is unchanged. Reference: Multifocal consolidations, worst in the right upper lobe.</p> <p><b>Trachea:</b> Generated: Endotracheal tube terminates approximately 4 cm above the carina. Reference: The patient is status post intubation with an endotracheal tube terminating approximately 5 cm from the carina.</p> <p><b>Left mid lung zone:</b> Generated: Left lower lobe atelectasis is unchanged. Reference: There are also patchy left lower lobe opacities.</p>
	<p><b>Right lung:</b> Generated: No acute intrathoracic process. There is no focal consolidation, effusion, or pneumothorax. Reference: No acute intrathoracic process. There is no focal consolidation, effusion, or pneumothorax.</p> <p><b>Left clavicle:</b> Generated: There are no acute osseous abnormalities. Reference: Imaged osseous structures are intact.</p> <p><b>Cardiac silhouette:</b> Generated: The cardiomedastinal silhouette is normal. Reference: The cardiomedastinal silhouette is normal.</p>
	<p><b>Right lung:</b> Generated: No acute intrathoracic process. The lungs are clear without focal consolidation, effusion, or pneumothorax. Reference: The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen.</p> <p><b>Cardiac silhouette:</b> Generated: The cardiomedastinal silhouette is normal. Reference: The cardiac and mediastinal silhouettes are unremarkable, as are the hilar contours.</p>

Figure 1. Anatomy-based sentence generation for 5 test set images. The upper three images depict abnormal cases, while the lower two depict normal cases. We show predicted (dashed boxes) and ground-truth (solid boxes) anatomical regions and color sentences accordingly.

### B.2. Qualitative full report generation results

Fig. 2 showcases generated full reports for three test set images. The left image shows a healthy chest X-ray image devoid of any pathologies. Based on the matching colors between generated and reference reports, we can see that all information contained in the reference report is also included in the generated report. In particular, the generated report correctly describes the placement of the endotracheal tube (colored in yellow), although with a slightly wrong numerical value. Also, the nasogastric tube is correctly mentioned in the generated report. The generated report describes, clinically correctly, four additional negative observations (i.e. non-present pathologies), which are however not mentioned in the reference report. As discussed in Appendix A.3, the region selection module has a low precision score for normal regions, since the decision to mention normal regions in a report is arbitrary and cannot be effectively learned. Thus, typically more regions are described in generated reports than in the corresponding reference reports, which in turn lowers the ROUGE-L score. However, the additionally described observations in the generated report of the left image are all clinically accurate, thus we believe that the region selection module selecting more regions than are described in the reference report is not detrimental to the quality of the generated reports.

In the middle image, we can see that the generated report correctly describes the pleural effusion in the right lung (colored in blue). However, while the generated sentence erroneously specifies a decrease in the effusion in comparison to a previous radiograph, the reference describes an increase. As mentioned in the limitations section of the main paper, our method considers each chest X-ray in isolation, and, as illustrated by this case, cannot correctly generate sentences that depend on previous radiographs. Thus, incorporating the information of localized comparison relations for anatomical regions between sequential exams into our method may be required to improve the generation of such sentences. In addition, there are some duplicate mentions of observations in the generated report, but since they are consistent with each other and clinically accurate, this is acceptable from a clinical point of view. However, the generated report misses a potential small pleural effusion on the left side (*“presence of a small pleural effusion cannot be excluded”*), illustrating the need for interactivity and transparency during report generation, which is simplified by our method.

The right image shows another chest X-ray with pathological findings, which were mainly captured by the generated report. However, we can see that all sentences in the reference report refer to previous radiographs, highlighting again the importance of incorporating sequential information in the method.

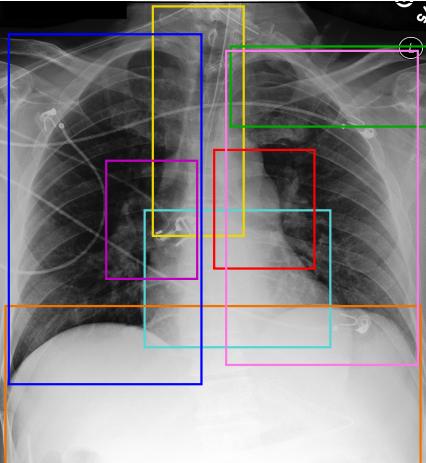
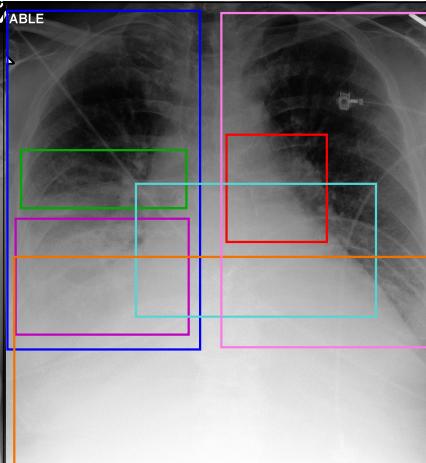
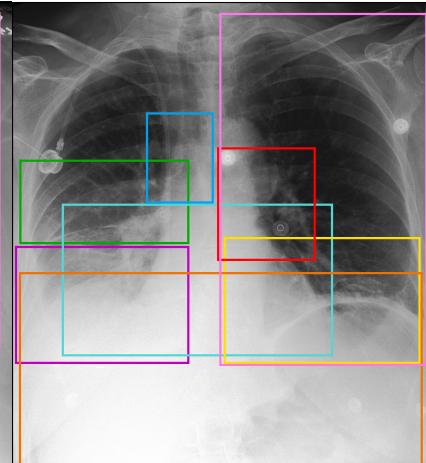
		
<p>Generated report: The lungs are clear. There is no pleural effusion or pneumothorax. There is no pulmonary edema. Mediastinal and hilar contours are unremarkable. The endotracheal tube terminates approximately 4 cm above the carina. There are no acute osseous abnormalities. The cardiomedastinal silhouette is within normal limits. NG tube tip is in the stomach.</p> <p>Reference report: Tip of endotracheal tube terminates 5.5 cm above the carina, and a nasogastric tube courses below the diaphragm. Cardiomedastinal contours are within normal limits, and lungs are clear.</p>	<p>Generated report: As compared to the previous radiograph, there has been a decrease in the extent of the right pleural effusion. There is no pneumothorax. Moderate right pleural effusion and moderate right lower lobe atelectasis are unchanged. There is mild pulmonary edema. Moderate cardiomegaly and mild pulmonary edema are unchanged. Moderate cardiomegaly persists. NG tube tip is out of view below the diaphragm.</p> <p>Reference report: As compared to the previous radiograph, the pre-existing right pleural effusion has minimally increased. The resulting atelectasis at the left and right lung bases as well as the cardiomegaly with mild pulmonary edema persist. Blunting of the left costophrenic sinus, so that the presence of a small pleural effusion cannot be excluded. No new parenchymal opacity. No pneumothorax.</p>	<p>Generated report: There is no pneumothorax. Right lower lobe atelectasis is unchanged. Moderate right pleural effusion is unchanged. Bibasilar atelectasis is unchanged. There is no evidence of pulmonary edema. Right subclavian line ends in the mid SVC. Heart size is normal. No free air below the right hemidiaphragm.</p> <p>Reference report: Portable chest radiograph demonstrates unchanged mediastinal, hilar, and cardiac contours. There has been interval development of bibasilar opacities likely reflecting atelectasis, though cannot exclude developing infectious process. Additionally, there has been interval increase in small right-sided pleural effusion.</p>

Figure 2. Full report generation for 3 test set images. Detected anatomical regions (solid boxes), corresponding generated sentences, and semantically matching reference sentences are colored the same. The generated reports mostly capture the information contained in the reference reports, as reflected by the matching colors. The left image shows a healthy chest X-ray image devoid of any pathologies, while the other two images depict abnormalities.

### B.3. Qualitative selection-based sentence generation results

<p><b>Ground-truth:</b> Generated: Right lower lobe pneumonia. Reference: Interval worsening of right lower lobe pneumonia.</p> <p><b>Position 1:</b> Generated: The right upper lobe opacity has improved since ____.</p> <p><b>Position 2:</b> Generated: Upper lungs are clear.</p>	<p><b>Ground-truth:</b> Generated: Right lower lobe pneumonia. Reference: Interval worsening of right lower lobe pneumonia.</p> <p><b>Aspect ratio 1:</b> Generated: Right lower lobe pneumonia.</p> <p><b>Aspect ratio 2:</b> Generated: Right upper lobe pneumonia.</p>	<p><b>Ground-truth:</b> Generated: Right lower lobe pneumonia. Reference: Interval worsening of right lower lobe pneumonia.</p> <p><b>Scale 1:</b> Generated: Right lower lobe pneumonia.</p> <p><b>Scale 2:</b> Generated: Right upper lobe pneumonia.</p>

Figure 3. Visualizing selection-based sentence generation for a test set image with pneumonia pathology. The solid red bounding box indicates the ground-truth anatomical region containing the pathology. Various dashed and colored bounding boxes represent radiologist-drawn bounding boxes, deviating from the ground-truth in terms of position, aspect ratio, or scale. The generated sentences demonstrate heightened sensitivity to bounding box position, while maintaining robustness towards variations in aspect ratio and scale.

Fig. 3 showcases the sensitivity of selection-based sentence generation to the position, aspect ratio, and scale of manually drawn bounding boxes within a test set image featuring pneumonia pathology.

The left image in the figure demonstrates variations in the position of the manually drawn bounding boxes. It is evident that the position is crucial, as the generated sentence for position 1 (slightly above the pathology) already misses the pathology and only describes an upper lobe opacity (which we believe is accurate). However, the generated sentence for position 2, which is even higher, completely misses the pathology and states that the upper lungs are clear (which we again believe to be accurate). Consequently, radiologists must be cautious to accurately position the bounding box to ensure correct pathology detection.

The middle image in the figure displays variations in the aspect ratio of the bounding boxes. For aspect ratios 1 and 2, both generated sentences correctly identify pneumonia. However, the sentence for aspect ratio 2 erroneously indicates that the pneumonia is located in the right upper lobe,

rather than the lower lobe. This mistake is understandable, as the bounding box lacks sufficient surrounding information to accurately determine the relative position within the lung (*i.e.*, upper or lower lobe).

The right image in the figure showcases variations in scale. The generated sentences exhibit robustness, as both scale variations correctly identify pneumonia. However, similar to the aspect ratio case, the sentence for scale 2 inaccurately describes an upper lobe pneumonia. Again, this error can be attributed to the insufficient surrounding information in the small-scaled bounding box.

In conclusion, selection-based sentence generation introduces additional flexibility into the clinical workflow by allowing radiologists to draw bounding boxes around areas of interest anywhere in the image. The primary caveat is the importance of correct positioning for the bounding box, which, if possible, should contain enough surrounding information to enable the model to generate accurate sentences.

#### B.4. Detailed clinical efficacy metrics results

Dataset	Observation	RGRG			
		P	R	F <sub>1</sub>	acc.
MIMIC-CXR	Micro Average	0.524	0.474	0.498	0.849
	Atelectasis	0.402	0.853	0.546	0.602
	Cardiomegaly	0.577	0.679	0.624	0.770
	Consolidation	0.132	0.055	0.078	0.919
	Edema	0.504	0.524	0.514	0.859
	Pleural Effusion	0.700	0.467	0.560	0.826
	Enlarged Cardiomediastinum	0.360	0.001	0.003	0.811
	Fracture	0.0	0.0	0.0	0.0
	Lung Lesion	0.217	0.004	0.007	0.957
	Lung Opacity	0.517	0.181	0.268	0.730
	No Finding	0.554	0.735	0.632	0.805
	Pleural Other	0.200	0.001	0.002	0.975
	Pneumonia	0.240	0.122	0.162	0.880
	Pneumothorax	0.189	0.138	0.159	0.950
	Support Devices	0.732	0.687	0.709	0.838

Table 4. Detailed results for the clinical efficacy (CE) metrics (see Appendix D.3 for details) for each observation as well as micro averaged over all 14 observations. The first five observations listed from the top are those used in calculating the  $P_{mic-5}$ ,  $R_{mic-5}$ , and  $F_{1,mic-5}$  scores in Tab. 2 of the main paper. The observation of fracture has a score of 0.0 (outlined in gray), since there are no sentences describing fractures in the Chest ImaGenome dataset. Thus, as mentioned in the limitations section of the main paper, a hybrid model that uses image-level features and sentences describing observations such as fractures from the MIMIC-CXR dataset may be required to further improve clinical accuracy.

#### B.5. Detailed anatomy-level results

Dataset	Anatomical Region	METEOR	IoU	Anatomical Region	METEOR	IoU
Chest ImaGenome	Abdomen	0.119	0.913	Right Apical Zone	0.157	0.863
	Aortic Arch	0.127	0.759	Right Atrium	0.237	0.755
	Cardiac Silhouette	0.110	0.837	Right Clavicle	0.290	0.849
	Carina	0.229	0.542	Right Costophrenic Angle	0.264	0.819
	Cavoatrial Junction	0.171	0.616	Right Hemidiaphragm	0.147	0.826
	Left Apical Zone	0.157	0.873	Right Hilar Structures	0.104	0.882
	Left Clavicle	0.294	0.841	Right Lower Lung Zone	0.051	0.897
	Left Costophrenic Angle	0.270	0.858	Right Lung	0.104	0.925
	Left Hemidiaphragm	0.074	0.796	Right Mid Lung Zone	0.083	0.893
	Left Hilar Structures	0.108	0.875	Right Upper Lung Zone	0.066	0.920
	Left Lower Lung Zone	0.054	0.881	Spine	0.165	0.950
	Left Lung	0.105	0.920	SVC	0.162	0.790
	Left Mid Lung Zone	0.089	0.894	Trachea	0.144	0.857
	Left Upper Lung Zone	0.049	0.922	Upper Mediastinum	0.162	0.881
	Mediastinum	0.119	0.870			

Table 5. Detailed results of the anatomy-based sentence generation (evaluated using METEOR) and object detection (evaluated using the IoU score) for each of the 29 anatomical regions.

## C. Method

### C.1. Module details

**Object detector.** Readers familiar with the Faster R-CNN [15] architecture may wonder why our method does not use RoI feature vectors (which are extracted from the RoI feature maps through fully connected layers in Faster R-CNN) directly as our region visual features, since instead we apply 2D average pooling and a linear transformation to the RoI feature maps to extract the region visual features. We found that taking the RoI feature vectors directly as the region visual features hurt the object detector’s performance, which we suspect is due to the coupling of features between the object detector’s subsequent classifier and regressor and the report generation model’s subsequent modules.

**Language model.** For the language model, we use the GPT-2 implementation from the huggingface library (transformers 4.19.2) [19] with the following checkpoint [12]: <https://huggingface.co/healx/gpt-2-pubmed-medium>.

### C.2. Training

For the overall training loss (Eq. 3 of the main paper), we specified that  $\mathcal{L}_{\text{select}}$  and  $\mathcal{L}_{\text{abnormal}}$  are weighted binary cross-entropy losses for the region selection and abnormality classification modules. Based on statistics computed on the training dataset, these weights for the positive examples are set to 2.2 for  $\mathcal{L}_{\text{select}}$  and 6.0 for  $\mathcal{L}_{\text{abnormal}}$  to account for class imbalances between regions with/without sentences and that are abnormal/normal, respectively.

As mentioned in the main paper, the model is trained in three stages:

1. Object detector
2. Object detector + region selection module  
+ abnormality classification module
3. Full model end-to-end

During all three stages, we train on a single NVIDIA A40 with PyTorch 1.12.1 in native mixed precision. The total training took about 45 hours and up to 48 GB of GPU memory was required. We refer to the code for more specifications of dependencies and versions. We use the AdamW [9] optimizer with a weight decay of 1e-2, reduce the learning rate by a factor of 0.5 if the total validation loss has plateaued or decreased (compared to the best epoch), and apply early stopping. In the first training stage, we use a batch size of 16, an initial learning rate of 1e-3, and train for 6 epochs. In the second training stage, we use a batch size of 16, an initial learning rate of 5e-4, and train for 9 epochs. In the third training stage, we use a batch size of 2, an initial learning rate of 5e-5, and train for 2 epochs. All batch sizes are (gradient) accumulated to 64.

### C.3. Inference

**Sentence generation.** We employ beam search with a width of 4 for sentence generation and use a BERTScore [22] threshold of 0.9 (based on best validation set performance) to remove similar generated sentences in radiology report generation. The high BERTScore value ensures robust duplicate removal, as only highly similar sentences are deduplicated, minimizing the risk of eliminating relevant information. For BERTScore, we use the uncased base version of DistilBERT [16] (*distilbert-base-uncased*).

## D. Experimental Setup

### D.1. Dataset and pre-processing

We use the recently released Chest ImaGenome v1.0.0 [3, 20, 21] dataset for training and evaluation of our proposed model. The MIMIC-CXR [4, 5] dataset, from which the Chest ImaGenome dataset is automatically constructed, consists of 377,110 chest X-ray images corresponding to 227,835 free-text radiology reports. The Chest ImaGenome contains automatically constructed scene graphs for 242,072 of those MIMIC-CXR images. For the images themselves, we use the MIMIC-CXR-JPG v2.0.0 [6, 7] dataset, which is fully derived from MIMIC-CXR and conveniently offers the images in JPG format.

The following image data augmentations are applied with 50% probability (each) during training:

- Color jitter of 20% brightness and contrast (saturation and hue jittering are not used as chest X-rays are single-channel greyscale images)
- Gaussian noise of zero mean and variance in the range [10, 50]
- Affine transformation with translation up to  $\pm 2\%$  of the image height/width and rotation up to  $\pm 2^\circ$

For the sentences of the Chest ImaGenome dataset, we always remove redundant whitespaces (as mentioned in the main paper). In some cases, we noticed that sentences assigned to regions contained superfluous, introductory phrases (such as “*UPRIGHT PORTABLE AP CHEST RADIOGRAPH:*”), which do not contain any relevant information. We assumed that these phrases were erroneously extracted by the Chest ImaGenome dataset from the MIMIC-CXR radiology reports and assigned to regions, thus they are also removed.

### D.2. Reference reports and processing

As described in the main paper, we use the *findings* section of MIMIC-CXR radiology reports as our reference reports. To extract these sections, we use a text

extraction tool provided by the MIMIC-CXR dataset authors: <https://github.com/MIT-LCP/mimic-cxr/tree/master/txt>.

We emphasize that we do not apply any further processing to these extracted reports. In contrast, some papers, such as the two papers [11, 18] from 2022 in Tab. 1 of the main paper, most likely applied additional processing to these extracted reports, including lowercasing all words. While [11] details the applied processing, [18] does not provide this information, and no code is available for verification. However, their qualitative analysis showcases lowercased reference reports, leading us to believe that they did employ lowercasing.

Lowercasing can significantly impact natural language generation (NLG) scores, particularly BLEU scores [14]. We discovered that when lowercasing reference reports, our method produces these BLEU scores: *BLEU-1*: 0.400, *BLEU-2*: 0.266, *BLEU-3*: 0.187, *BLEU-4*: 0.135 ( $\Delta+8.9\%$  against best baseline). METEOR and CE scores remain unchanged, as lowercasing does not affect them.

We believe that this highlights another reason why NLG metrics are ill-suited for evaluating radiology reports, as scores heavily depend on the specific processing applied to reference reports (since NLG metrics count matching n-grams). In contrast, CE-metrics are processing-invariant, as they compare disease presence status between reference and generated reports, independent of sentence structure or casing. Thus, CE metrics allow for a fairer comparison between methods while also capturing the diagnostic accuracy of generated reports. Consequently, we encourage future radiology report generation research to place greater emphasis on CE metrics when evaluating generated reports.

### D.3. Clinical efficacy metrics

Clinical efficacy (CE) metrics capture how semantically coherent the generated and corresponding reference reports are w.r.t. an array of prominent clinical observations. To ensure comparability of results, we specifically follow [10] in calculating the CE scores micro averaged over five observations, and [11] in calculating the CE scores example-based averaged over all 14 observations as follows: CheXbert [17] - a BERT [2]-based information extraction system - is first used to classify the status of 14 observations as either *positive*, *negative*, *uncertain*, or *no mention* for each generated report and corresponding reference report. The observations consist of 12 types of diseases as well as "*Support Devices*" and "*No Finding*". Next, these multi-class classifications are converted to binary-class. [11] performs this conversion by considering *positive* as the positive class, and *negative*, *uncertain* and *no mention* as the negative class. In contrast, [10] considers *positive* and *uncertain* the positive class, and *negative* and *no mention* the negative class. Finally, [11] calculates the example-based precision, recall,

and F1 scores over all 14 observations by comparing the classifications for each generated report and corresponding reference report. In contrast, [10] calculates the micro average precision, recall, and F1 scores over a subset of 5 observations: *atelectasis*, *cardiomegaly*, *consolidation*, *edema*, and *pleural effusion*. We follow each approach respectively when comparing our results with the two works.

### D.4. Variation sampling for evaluation of selection-based sentence generation

The variation sampling experiments for the evaluation of selection-based sentence generation, as showcased in Fig. 4 and with results shown in Fig. 5 from the main paper, were conducted as follows. First, we select the first 1000 samples from the test set to reduce the required computational resources. We then use our (trained) RGRG model for selection-based sentence generation inference (see the third paragraph of Sec. 3.4 in the main paper) on this subset. Instead of letting radiologists manually draw bounding boxes, we randomly modify the ground-truth bounding boxes from those samples and use them during inference (*i.e.*, pass them through RoI pooling). We investigate three types of variations independently: position, aspect ratio, and scale of the bounding boxes. For each of these cases, we run several experiments with different degrees of random variations. For a specific type of variation (*i.e.*, position, aspect ratio, or scale) and a degree of variation as defined by the  $1-\sigma$  interval (*i.e.*, one standard deviation, as used in the  $x$ -axis of Fig. 5 in the main paper), an experiment corresponds to a single inference pass through all of the 1000 samples. We compute the micro-averaged per-anatomy METEOR score for each experiment and compare it to the default case without any variations, *i.e.* inference on the 1000 samples using the ground-truth bounding boxes.

In a single experiment, we sample the variation for each anatomical region in each sample independently. Assume the ground-truth box for an anatomical region is defined by its upper left  $(x_1, y_1)$  and lower right  $(x_2, y_2)$  corners and has width  $w = x_2 - x_1$  and height  $h = y_2 - y_1$ . We sample the (additive) **position variations**  $\Delta x \in (-\infty, +\infty)$  and  $\Delta y \in (-\infty, +\infty)$  for the given standard deviation  $\sigma$  from a zero-mean normal distribution  $\mathcal{N}$  as

$$\Delta x \sim \mathcal{N}(0, \sigma^2), \quad \Delta y \sim \mathcal{N}(0, \sigma^2), \quad (1)$$

and then compute the modified box  $(\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$  by varying the original box additively in relation to its size as

$$\begin{aligned} \hat{x}_1 &= x_1 + \Delta x \cdot w, & \hat{x}_2 &= x_2 + \Delta x \cdot w, \\ \hat{y}_1 &= y_1 + \Delta y \cdot h, & \hat{y}_2 &= y_2 + \Delta y \cdot h. \end{aligned} \quad (2)$$

Aspect ratio and scale are varied multiplicatively and we, therefore, sample from the normal distribution in log-space (of aspect ratio or scale variations). In other words, the

variations are Lognormal distributed. The (multiplicative) **aspect ratio** variation  $\Delta a \in (0, +\infty)$  is sampled as

$$\Delta a \sim \text{Lognormal}(0, \sigma^2), \quad (3)$$

i.e.

$$\ln(\Delta a) \sim \mathcal{N}(0, \sigma^2), \quad (4)$$

and similarly, the (multiplicative) **scale variation**  $\Delta s \in (0, +\infty)$  is sampled as

$$\Delta s \sim \text{Lognormal}(0, \sigma^2). \quad (5)$$

The  $1-\sigma$  interval for both cases is therefore defined as  $[e^{-\sigma}, e^{\sigma}]$ . Given a sampled aspect ratio variation  $\Delta a$  and a ground-truth box  $(x_1, y_1, x_2, y_2)$  with aspect ratio  $a = \frac{w}{h}$  and area  $A = w \cdot h$ , we first compute the modified aspect ratio  $\hat{a}$  as

$$\hat{a} = \Delta a \cdot a, \quad (6)$$

then compute the modified width  $\hat{w}$  and height  $\hat{h}$  using the unmodified area  $A$  as

$$\hat{w} = \sqrt{A \cdot \hat{a}}, \quad \hat{h} = \sqrt{\frac{A}{\hat{a}}}, \quad (7)$$

to finally compute the modified box as

$$\begin{aligned} \hat{x}_1 &= x_1 + \frac{w - \hat{w}}{2} & \hat{x}_2 &= x_2 - \frac{w - \hat{w}}{2}, \\ \hat{y}_1 &= y_1 + \frac{h - \hat{h}}{2} & \hat{y}_2 &= y_2 - \frac{h - \hat{h}}{2}. \end{aligned} \quad (8)$$

Similarly, given a sampled scale variation  $\Delta s$ , we first compute the updated width and height as

$$\hat{w} = \Delta s \cdot w, \quad \hat{h} = \Delta s \cdot h, \quad (9)$$

and then again use (8) to compute the modified box.

## References

- [1] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation. In *ML4H Workshop*, pages 126–140, 2020. 1
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 8
- [3] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. 7
- [4] 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. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019. 7
- [5] 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. Mimic-cxr database (version 2.0.0). *PhysioNet*, 2019. 7
- [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. Mimic-cxr-jpg - chest radiographs with structured labels (version 2.0.0). *PhysioNet*, 2019. 7
- [7] 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. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 7
- [8] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghazsemi. Clinically accurate chest x-ray report generation. In *MLHC*, pages 249–269, 2019. 1
- [9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 7
- [10] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *NAACL*, pages 5288–5304, 2021. 8
- [11] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm-starting. *arXiv preprint arXiv:2201.09405*, 2022. 8
- [12] Yannis Papanikolaou and Andrea Pierleoni. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*, 2020. 7
- [13] Pablo Pino, Denis Parra, Pablo Messina, Cecilia Besa, and Sergio Uribe. Inspecting state of the art performance and nlp metrics in image-based medical report generation. *arXiv preprint arXiv:2011.09257*, 2020. 1
- [14] Matt Post. A call for clarity in reporting bleu scores. In *WMT*, pages 186–191, 2018. 8
- [15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 28, 2015. 7
- [16] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 7
- [17] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *EMNLP*, pages 1500–1519, 2020. 8
- [18] Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. An inclusive task-aware framework for radiology report generation. In *MICCAI*, pages 568–577, 2022. 8

- [19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *EMNLP: system demonstrations*, pages 38–45, 2020. 7
- [20] Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, et al. Chest imangenome dataset for clinical reasoning. In *NIPS*, 2021. 7
- [21] Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imangenome dataset (version 1.0.0). *PhysioNet*, 2021. 7
- [22] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2019. 7