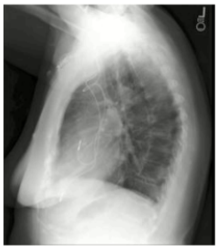
Image-aware Evaluation of Generated Medical Reports – Supplementary Material

1 Perturbed dataset

Following are examples from our perturbed dataset, for each perturbation described in the paper.

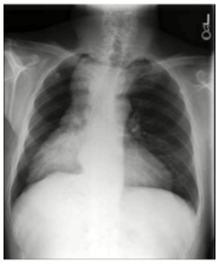
Removal of pathology sentence.

In this perturbation we modified the report by removing a sentence (marked in blue) describing a pathology from the ground truth report.



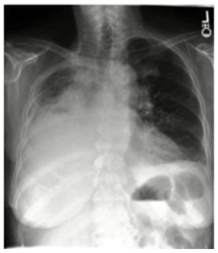
The lungs are relatively hyperinflated. There is no focal consolidation concerning for pneumonia. No pleural effusion or pneumothorax is detected. The pulmonary vasculature is not engorged and there is no overt pulmonary edema. The cardiac silhouette is top normal in size as before. A left pectoral pacemaker is in place with dual leads terminating in the right atrium and right ventricle. The mediastinal and hilar contours are within normal limits.

Example 1



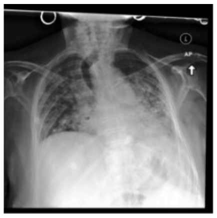
A large dilated <unk> possibly fluid filled esophagus is again appreciated abutting the right mediastinum in this patient with known achalasia. The finding appears more prominent as compared to the right study of but similar to. There is a questionable air-fluid level in the proximal thoracic esophagus. The possibility of progressed <unk> <unk> of the esophagus is raised. There is no evidence of aspiration. There is no pleural effusion or pneumothorax. The cardiac silhouette is difficult to assess.

Example 2



Comparison study of there is again extensive opacification involving much of the right hemithorax. This is consistent with a previous study showing substantial loculation of right pleural fluid collection with underlying extensive volume loss. Prominence of markings on the left most likely represents redistribution of blood flow to <unk> regions on the right.

Example 3



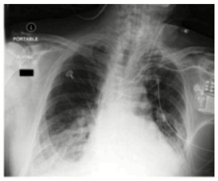
AP and lateral views of the chest. Low lung volumes are seen compatible with patients history of fibrosis. Diffusely increased interstitial markings are seen throughout the lungs but these appear overall slightly worse when compared to prior. Cardiomediastinal silhouette is grossly unchanged. No acute osseous abnormality is detected.

Example 4

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

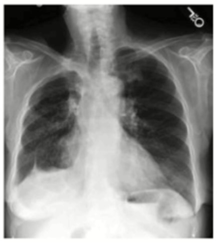
Removal of insignificant sentence.

In this perturbation, we modified the report by removing a sentence (marked in gray), which is irrelevant to the analysis of the given image. Such sentences may include common phrases that may or may not appear in reports, as well as phrases that describe information not present in the input image or relate to actions taken by the doctor.



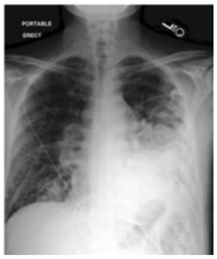
The endotracheal tube is too high at the thoracic inlet. This finding was called to the <unk> nurse at 500 pm. At the time of dictating this report by dr. Otherwise the appearance of the lungs is unchanged. Pacemaker and left ij line are unchanged.

Example 1



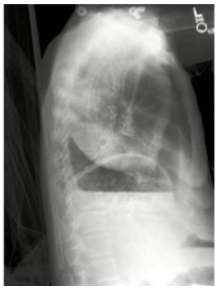
Frontal and lateral views of the chest were obtained. The patient is status post right upper chest wall resection right upper lobectomy with right apical scarring and upward traction of the right hilum from radiation fibrosis all unchanged. There is no pleural effusion or pneumothorax. The left lung is clear. Heart size is normal.

Example 2



Portable AP upright chest radiograph was obtained. Compared to the scout radiograph from a torso CT from. There is increased opacity in the left lower lung concerning for worsening effusion and consolidation. Extensive nodularity in the lungs is compatible with known metastatic disease. Heart size cannot be assessed. Bony structures appear unchanged.

Example 3

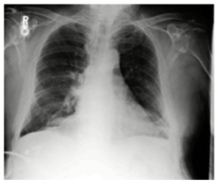


Right-sided chest tube has been removed. There is a hydropneumothorax in the inferior right chest. The amount of fluid has increased compared to the study from two days prior. The thick irregular pleural disease around the right lung is again visualized. The left lung is clear. Cardiac and mediastinal silhouettes are unchanged.

Example 4

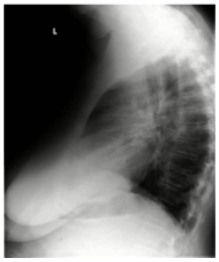
Modification of pathology severity.

In this perturbation, we modified a single word describing the severity of the pathology. This word is highlighted in the following examples, with the changed word indicated in parentheses.



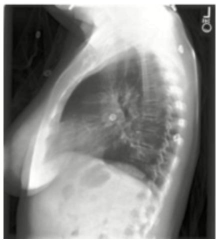
Since the prior examination there is little change. There is no evidence of pneumothorax. There is a moderate subpulmonic pleural effusion as better demonstrated on the prior lateral radiograph. There is a new small (large) left layering pleural effusion. There are no new focal opacities concerning for pneumonia. Cardiomediastinal and hilar contours are stable demonstrating mild tortuosity of the thoracic aorta. Heart size is within normal limits. Pulmonary vascularity is normal.

Example 1



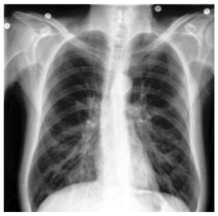
Both lungs are well expanded and clear. There are no lung opacities concerning for pneumonia or pulmonary edema. Heart size is mildly (severly) enlarged and stable since. Mediastinal and hilar contours are unchanged. There is no pleural effusion or pneumothorax.

Example 2



As compared to the previous radiograph the lung volumes have minimally decreased. In the retrocardiac lung areas there is a very subtle (extensive) parenchymal opacity that projects over the spine on the lateral radiograph. In the light of the clinical history this opacity is suspicious for pneumonia. There is no other lung parenchymal abnormality. No pulmonary edema. no pleural effusions. Normal hilar and mediastinal contours. At the time of dictation dr. was paged to notification at <unk> am.

Example 3



In comparison with the study of there is little overall change in the peribronchial thickening and impaction with extensive (very subtle) bibasilar bronchiectasis. This is again extremely well seen on the lateral radiograph. Hyperexpansion of the lungs is consistent with emphysema and the cardiac size is normal. No evidence of pulmonary edema. No evidence of acute focal pneumonia.

Example 4

Modification of pathology location.

In this perturbation, we modified a single word describing the location of the pathology. This word is highlighted in the following examples, with the changed word indicated in parentheses.

PA and lateral views of the chest were provided. When compared with multiple prior studies there is bilateral upper (lower) lung scarring with slight retraction of the bronchovasculature. There is no definite sign of new consolidation with relative opacity at the right heart border on the frontal view not convincing for pneumonia. Lung volumes are low. Heart and mediastinal contours appear stable. No effusion or pneumothorax.

Example 1

PA and lateral views of the chest were obtained. Patient is known to have extensive metastatic disease within the chest with loculated left (right) pleural effusion. Overall appearance of the chest appears essentially stable compared with multiple prior exams. Please note evaluation for subtle differences would be limited due to extensive underlying metastatic burden. Heart size cannot be readily assessed . Mediastinal contour appears grossly stable. No pneumothorax is seen. Imaged osseous structures appear grossly intact.

Example 3

As compared to the previous radiograph the lung volumes have minimally decreased. In the retrocardiac (apical) lung areas there is a very subtle parenchymal opacity that projects over the spine on the lateral radiograph. In the light of the clinical history this opacity is suspicious for pneumonia. There is no other lung parenchymal abnormality. No pulmonary edema. No pleural effusions. normal hilar and mediastinal contours. At the time of dictation dr. was paged to notification at <unk> am.

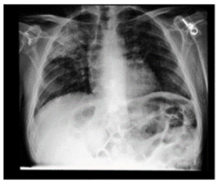
Example 2

In comparison with the study of there is little overall change in the peribronchial thickening and impaction with extensive bibasilar (biapical) bronchiectasis. This is again extremely well seen on the lateral radiograph. Hyperexpansion of the lungs is consistent with emphysema and the cardiac size is normal. No evidence of pulmonary edema . No evidence of acute focal pneumonia.

Example 4

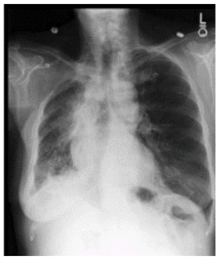
Modification of non-informative word.

In this perturbation, we modified a single non-informative word. This word is highlighted in the following examples, with the changed word indicated in parentheses.



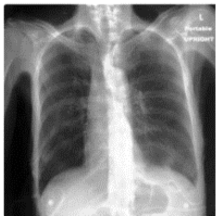
AP and lateral views of chest demonstrate a right upper lobe consolidation with some areas of air bronchogram. Background multifocal opacities with volume loss and chronic scarring are unchanged. There (<unk>) is no large pleural effusion. Cardiac size is normal.

Example 1



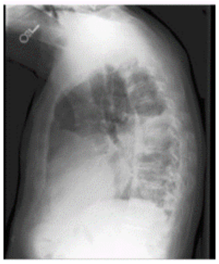
There is asymmetry and volume loss of the right hemithorax and mediastinal shift to the right and diffusely increased opacification of the (<unk>) right hemithorax which might represent early infection along with volume loss. There is no pneumothorax.

Example 2



Pigtail pleural catheters remain in place bilaterally. Small bilateral apical lateral pneumothoraces have slightly decreased in size since the (<unk>) prior study. Small left pleural effusion is again demonstrated.

Example 3



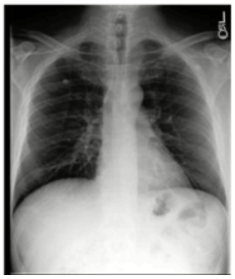
Again seen is a large pleural effusion with likely a loculated component on the right with compressive atelectasis of major portions of the right lower and middle lobes. There (<unk>) is no pneumothorax. The left lung is well expanded and clear. The cardiac size is within normal limits. The hilar and mediastinal contours are normal.

Example 4

Modification of reports without findings.

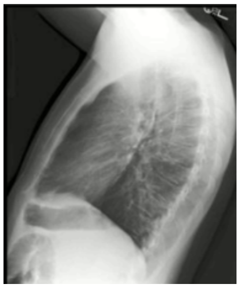
All studies were compared against the following report, which was constructed using common normal phrases across the dataset:

"Frontal and lateral views of the chest were obtained. There is no pleural effusion or pneumothorax. The heart size is normal. Bony structures are intact."



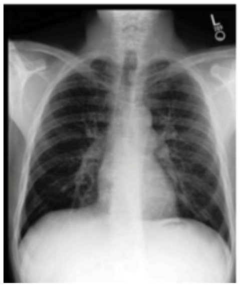
The lungs are clear. The cardiomediastinal silhouette is normal. No acute osseous abnormalities identified.

Example 1



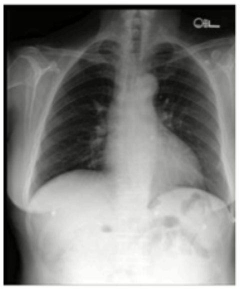
Frontal and lateral views of the chest are obtained. No focal consolidation, pleural effusion, or evidence of pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.

Example 2



Mediastinal and hilar contours are normal. Both lungs are clear with no focal consolidation, pleural effusion, or pneumothorax.

Example 3



The cardiac, mediastinal and hilar contours are normal. Both lungs are clear with no focal consolidation, pleural effusion or pneumothorax.

Example 4

2 SoTA model evaluation

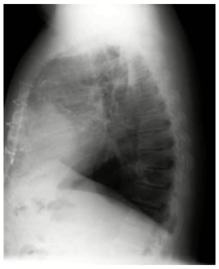
	NLG					CE			OURS
Model	B-1	B-4	M	R-L	BS	Cp	Cb	RG	VLScore
MSAT [32]	0.27	0.08	0.12	0.26	0.39	0.11	0.27	0.13	0.52
R2GEN-CMN [3]	0.36	0.11	0.15	0.28	0.41	0.23	0.35	0.18	0.53
RGRG [27]	0.19	0.03	0.13	0.14	0.32	0.24	0.33	0.16	0.49
COMG (w/o RL) [7]	0.36	0.12	0.14	0.29	0.40	0.01	0.26	0.17	0.21
XProNet [30]	0.32	0.10	0.13	0.27	0.40	0.23	0.35	0.17	0.57

Table 1: **Results of report generation models (MIMIC-CXR).** We compared several recent models that have published their code and allow their generated reports to be reproduced. For RGRG, we used the pretrained weights, while other works were trained using the official project repositories. All evaluations were done on the test set defined by [3]. We note that while [3] and [7] show similar performance in NLG metrics, there is a significant gap in CE metrics, especially in CheXpert. Our metric differentiates between them. For the other methods whose performance varies based on the metric, our score provides a unique ranking. Please refer to the additional qualitative comparison attached below.

Computational resources. All the experiments in the paper were performed on a single Nvidia A6000 GPU.

3 Qualitative comparison

Example 1



(a) Input image

There is no focal consolidation pleural effusion pneumothorax or pulmonary edema. Cardiomediastinal silhouette is unchanged and notable for tortuous aorta and mild cardiomegaly. Median sternotomy wires are present and intact. Clips are seen in the midline of the thorax. Bony structures are intact.

(b) Ground-truth report

There is no evidence of acute cardiopulmonary process. The mediastinal and hilar contours are unremarkable. There is no pneumothorax or pleural effusion. Moderate cardiomegaly is unchanged.

(c) RGRG report

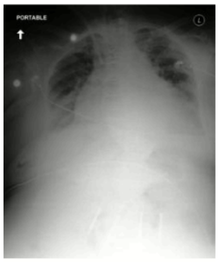
PA and lateral views of the chest provided. There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.

(c) MSAT report

	NLG					CE			OURS
	B-1	B-4	M	R-L	BS	Cp	Cb	RG	VLScore
RGRG	0.119	0	0.134	0.132	0.474	**1**	0.391	0.182	**0.652**
MSAT	**0.404**	**0.153**	**0.189**	**0.371**	**0.529**	0.461	**0.395**	**0.294**	0.521

The RGRG report succeeds in describing the pathology (cardiomegaly), while the MSAT report describes a radiograph without findings and omits this crucial information. In all metrics (except CheXpert), the MSAT report gets a higher score. Our metric captures this omission and assigns a higher score to the RGRG report.

Example 2



(a) Input image

There is no pneumothorax. Both lung volumes are very low .Bilateral right side more than left side moderate pulmonary edema has improved. Widened cardiomediastinal silhouette is more than it was on however this appearance could be exacerbation from low lung volumes. Patient is status post median sternotomy with intact sternal sutures.

(b) Ground-truth report

In comparison with the study of the monitoring and support devices remain in place .Continued low lung volumeswith enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases.

(c) R2GenCMN report

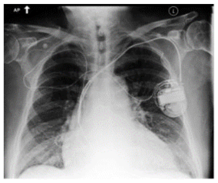
There is no pneumothorax or pleural effusion. Moderate cardiomegaly and bibasilar atelectasis are unchanged. There is no pulmonary edema. The endotracheal tube is in standard position. The cardiomediastinal silhouette is unremarkable. Moderate cardiomegaly with mild pulmonary edema. NG tube tip is in the stomach.

(c) RGRG report

			NLG				CE		OURS
	B-1	B-4	M	R-L	BS	Cp	Cb	RG	VLScore
R2GenCMN	0.124	0	0.071	0.124	0.212	0.52	0.062	**0.2**	**0.646**
RGRG	**0.158**	0	**0.118**	**0.154**	**0.348**	**0.71**	**0.186**	0.165	0.293

R2GenCMN's report correctly describes the pathologies (low lung volumes, enlarged cardiac silhouette and edema) described in the ground-truth study, although it additionally describes incorrect pathologies (effusion, atelectasis). The RGRG report fails to describe either of the pathologies in the ground-truth report, describes incorrect pathologies (cardiomegaly, atelectasis), and contains contradictions (regarding edema). All other metrics, except RadGraph F1, fail to capture the incorrectness of the RGRG report in that study and assign to it higher or equal scores. However, our metric captures this difference in quality and assigns a higher score to the report which shares more information with the ground-truth.

Example 3



(a) Input image

Frontal and lateral views of the chest were obtained. Mild cardiomegaly is similar to prior. There is mild pulmonary congestion without overt pulmonary edema. No focal pulmonary consolidation pleural effusion or pneumothorax is seen. The osseous structures are unremarkable. The leads of an icd are in similar position to prior.

(b) Ground-truth report

Single portable view of the chest is compared to previous exam from. There has been interval placement of a right-sided central venous catheter with tip projecting over the right atrium. Left chest wall dual lead pacing device is again seen. There are low lung volumes. There is mild pulmonary vascular congestion. The heart is mildly enlarged. Osseous and soft tissue structures are unremarkable.

(c) XProNet report

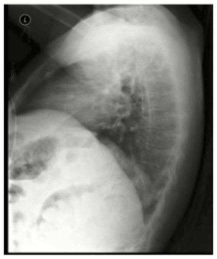
PA and lateral views of the chest provided. Midline sternotomy wires and mediastinal clips are again noted. The heart is mildly enlarged. The lungs are clear without focal consolidation large effusion or pneumothorax. No signs of congestion or edema. Mediastinal contour is stable. Bony structures are intact. No free air below the.

(c) MSAT report

			NLG				CE		OURS
	B-1	B-4	M	R-L	BS	Cp	Cb	RG	VLScore
XProNet	0.369	0.116	0.184	0.299	0.416	**1**	0.468	0.117	**0.555**
MSAT	**0.491**	**0.148**	**0.186**	**0.363**	**0.504**	0.71	**0.759**	**0.172**	0.412

The XProNet report describes all the pathologies outlined in the input study, while the MSAT report only identifies one of them (cardiomegaly). In all metrics except ours & CheXpert, the MSAT report achieves higher scores, while our metric recognize the absence of other two pathologies (pacemaker and edema).

Example 4



(a) Input image

Frontal and lateral views of the chest are compared to previous exam from. Again seen is biapical fibrotic changes. Previously seen perihilar and right basilar opacities have resolved. There is no effusion or new consolidation. The cardiomediastinal silhouette is stable. Orthopedic hardware projects over the right glenoid fossa.

(b) Ground-truth report

The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.

(c) MSAT report

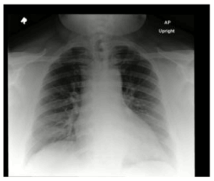
Frontal and lateral views of the chest were obtained. Low lung volumes result in bronchovascular crowding. There is no focal consolidation pleural effusion or pneumothorax. Heart size is normal. Mediastinal silhouette and hilar contours are normal allowing for lung volumes. No acute osseous abnormality is identified. There is no free air under the diaphragm.

(c) R2GenCMN report

	NLG					CE			OURS
	B-1	B-4	M	R-L	BS	Cp	Cb	RG	VLScore
MSAT	0.155	0	0.084	0.240	0.381	0.461	**0.282**	0.069	**0.693**
R2GenCMN	**0.410**	**0.163**	**0.155**	**0.352**	**0.405**	**0.461**	0.264	**0.071**	0.444

In every metric except ours and CheXbert, the R2GenCMN report attains superior scores despite depicting inaccurate pathologies (such as low lung volumes and edema) that are absent in the input study. While the MSAT report overlooks certain pathologies in the radiograph, it refrains from introducing false diagnoses, thus achieving a higher score in our metric.

Example 5



(a) Input image

Lungs are clear. There is no consolidation effusion or edema. The cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities.

(b) Ground-truth report

The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.

(c) R2GenCMN report

AP upright and lateral views of the chest provided. Lung volumes are low. There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.

(c) MSAT report

	NLG					CE			OURS
	B-1	B-4	M	R-L	BS	Cp	Cb	RG	VLScore
R2GenCMN	**0.479**	0	0.175	0.406	0.496	0.481	**0.786**	0.243	**0.913**
MSAT	0.391	**0.169**	**0.253**	**0.505**	**0.589**	**0.481**	0.780	**0.311**	0.748

The ground truth report describes no findings in the input image. R2GenCMN succeeds in capturing this information, while the MSAT report describes a non-existing pathology (in red). Our metric penalizes for this mistake, while most of the other metrics assign a higher or equal score to the inaccurate report.

Example 6

(a) Input image

As compared to the previous radiograph there is unchanged evidence of bilateral parenchymal opacities constant on the right and minimally improving on the left. Unchanged cardiomegaly and small bilateral pleural effusions. Subsequent areas of basal atelectasis. Unchanged position of the endotracheal tube and right-sided central venous access line.

(b) Ground-truth report

Moderate right pleural effusion and moderate right lower lobe atelectasis are unchanged. Moderate cardiomegaly and mild pulmonary edema are unchanged. There is mild bibasilar atelectasis. The endotracheal tube terminates approximately 4 cm above the carina. The right internal jugular vein catheter tip is in the right atrium. The cardiomediastinal silhouette is unremarkable. Right internal jugular line tip is at the level of mid SVC. Right pectoral pacemaker leads terminate in the right atrium and right ventricle.

(c) RGRG report

As compared to the previous radiograph the patient has been intubated. The tip of the endotracheal tube projects 4 cm above the carina. The course of the nasogastric tube is unremarkable the tip of the tube projects over the middle parts of the stomach. There is no evidence of complications notably no pneumothorax. The appearance of the lung parenchyma and the cardiac silhouette is constant.

(c) R2GenCMN report

			NLG				CE		OURS
	B-1	B-4	M	R-L	BS	Cp	Cb	RG	VLScore
RGRG	0.176	0	0.155	0.117	0.299	**0.844**	0.294	**0.166**	**0.648**
R2GenCMN	**0.328**	**0.126**	**0.156**	**0.286**	**0.363**	0.714	**0.326**	0.111	0.515

The RGRG report describes most of the pathologies outlined in the input study, while the R2GenCMN report only identifies one of them. In all metrics except ours, CheXpert and RadGraph, the R2GenCMN report achieves higher scores, while RadGraph, CheXpert and our metric recognize the absence of all other pathologies (pleural effusion, atelectasis, cardiomegaly, and support devices).

Example 7

(a) Input image

A right upper lobe consolidation with air bronchograms is similar to. Focal tubular lucency within the opacity is new and may reflect cavitation dilated airways or spared lung parenchyma. Opacity in the right lower lobe has progressed since the prior study. There is no effusion or pneumothorax. Cardiac and mediastinal contours are normal. There is mild thickening of the left major fissure.

(b) Ground-truth report

PA and lateral views of the chest provided. Airspace consolidation is noted within the right upper lobe concerning for pneumonia. No large effusion or pneumothorax. Cardiomediastinal silhouette is stable. Bony structures are intact.

(c) R2GenCMN report

PA and lateral views of the chest. The lungs are clear. There is no focal consolidation pleural effusion or pneumothorax. The heart size is normal. The mediastinal and hilar contours are normal.

(c) COMG report

The R2GenCMN report captures the pathology described in the ground-truth study: consolidation. However, COMG fails to capture the pathologies present in the radiograph and describes a healthy study instead. All metrics, except CheXpert and RadGraph F1, fail to discern this difference in quality

	NLG					CE			OURS
	B-1	B-4	M	R-L	BS	Cp	Cb	RG	VLScore
R2GenCMN	0.232	0.056	0.135	0.314	0.449	**0.813**	0.493	**0.199**	**0.737**
COMG	**0.415**	**0.130**	**0.162**	**0.330**	**0.472**	0.440	**0.587**	0.181	0.133

and assign a higher score to the COMG report. Our metric captures the absence of the pathologies' description and assigns a higher score to the R2GenCMN report.

4 Subgroup analysis

The results of Table 2 from the paper present the average score acorss all pathology sentences (in our metric the average score is 0.69). For a finer-grained evaluations for the removal of pathology experiment, we report the scores for each pathology:

- Atelectasis: 0.77
- Cardiomegaly: 0.74
- Consolidation: 0.71
- Edema: 0.69
- Enlarged Cardiomediastinum: 0.75
- Fracture: 0.72
- Lung Lesion: 0.57
- Lung Opacity: 0.69
- Pleural Effusion: 0.69
- Pleural Other: 0.71
- Pneumonia: 0.83
- Pneumothorax: 0.51
- Support Device: 0.65

We observe that these scores are similar for most pathologies.

5 Synonym analysis

The influence of modifying words with synonyms experimentally demonstrates a robustness of our method to these perturbations. For example, in Figure 1b in the main paper, changing the word "suggests" to "indicates" results in a VLScore of 0.96; changing "a site" to "an area" yields a VLScore of 0.95; and changing "seems" to "appears" results in a VLScore of 0.98. These scores, which are very close to 1, indicate the desired robustness to exact wording.

6 Comparison to RadCliq (linear combination of reported metrics)

The following table shows that VLScore is more sensitive than RadCliq for all perturbations (sentence removal and word change). For example, our metric shows a difference of 0.15 for pathology versus insignificant sentence removal, while RadCliq shows only 0.06. This is expected, as RadCliq is a linear combination of other metrics we report in the paper (BertScore, CheXbert, and RadGraph F1). Additionally, for reports without findings, our metric captures their similarity with an average score of 0.85, whereas RadCliq assigns a low score of 0.434 to these pairs, showcasing the robustness of our metric in this aspect.

Table 2: **Result of VLScore vs. RadCliq on the perturbed dataset**. For each perturbation (sentence removal or word change), we report the difference between the scores obtained for significant changes versus non-significant changes. For sentence removal, we compare the removal of an insignificant sentence to the removal of a pathology sentence. For word changes, we compare two cases: a location word versus a non-informative word, and a severity word versus a non-informative word. The higher Δ's obtained by VLScore indicates greater sensitivity to significant information compared to RadCliq.

	RadCliq	VLScore
Sentence Removed		
Insignificant	0.92	0.84
Significant (pathology)	0.86	0.69
Δ	0.06	**0.15**
Changed word		
Non-informative	0.97	0.91
Location	0.96	0.72
Δ	0.01	**0.19**
Severity	0.96	0.79
Δ (to non-informative)	0.01	**0.12**