

# Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation

## Supplementary Material

### A. Experiments

#### A.1. Implementation Details

**1) MIMIC-CXR [5]:** In stage 1, the model is trained for 50 epochs with a learning rate of  $5e-5$  and a batch size of 32. In Stage 2, we train MLRG for another 50 epochs, using a batch size of 14. The learning rate is set to  $5e-6$  for parameters from Stage 1 and  $5e-5$  for the remaining parameters. **2) MIMIC-ABN [8] and Two-view CXR [7]:** Since most images are derived from the MIMIC-CXR dataset, we directly fine-tune the model from Stage 2 on MIMIC-CXR, using a learning rate of  $5e-6$  and a batch size of 12. **3) Common settings:** Early stopping with a patience of 15 is employed to prevent overfitting. The ReduceLROnPlateau scheduler and the AdamW optimizer are applied. The natural language generation (NLG) metrics are calculated with the pycocoevalcap<sup>1</sup>. For clinical efficacy (CE) metrics, Precision, Recall, and F1 score metrics are computed using the flchexbert<sup>2</sup> library, and the F1 RadGraph metric is calculated with the radgraph<sup>3</sup> library.

#### A.2. Clinical Accuracy of 14 Observations

Tables A1 and A2 show the clinical accuracy of 14 observations annotated by CheXpert [4] on the MIMIC-CXR, MIMIC-ABN, and Two-view CXR datasets. Results show that our MLRG outperforms SEI [6] on most observations. Even though MLRG is not specifically tailored for imbalanced observations, it still slightly surpasses the baseline on challenging observations like *Pneumothorax* and *Pleural Other*.

#### A.3. Performance of Generating “FINDINGS” and “IMPRESSION” Sections

Radiology reports typically consist of three key sections: “INDICATION”, which outlines the visit reasons or symptoms; “FINDINGS”, which details observations from current multi-view images and comparisons with the patient’s medical history; and “IMPRESSION”, which summarizes the key conclusions or diagnostic interpretations based on the “FINDINGS”. Table A3 presents the performance of generating “FINDINGS” and “IMPRESSION” sections, with specific examples in Figure A2. Since most existing methods focus primarily on generating the “FINDINGS” section, peer methods are not included in Table A3. The re-

sults indicate that our MLRG is capable of generating both sections with minor modifications. Specifically, we utilize special tokens, “[FINDINGS]” and “[IMPRESSION]”, before the respective section content to distinguish between them. These sections are then combined to form the final reference reports, with all other settings remaining identical to those used for the “FINDINGS” generation.

#### A.4. Qualitative Analysis

Figure A1 provides additional examples of the “FINDINGS” section generated by SEI [6] and our MLRG, while Figure A2 presents examples of both the “FINDINGS” and “IMPRESSION” sections from MLRG. These results suggest that 1) Our MLRG is highly competitive in generating both “FINDINGS” and “IMPRESSION” sections, as well as the “FINDINGS” section alone, for chest X-ray reports. 2) MLRG still has room for improvement in describing lesion attributes. For example, in Figure A1, MLRG incorrectly describes the “proximal parts of the stomach” as the “middle parts”. This occurs because MLRG has not fully learned region-level features. To improve this, we are exploring the use of saliency maps [11] to enhance regional feature learning and the accuracy of lesion descriptions.

#### A.5. Evaluation Using Large Language Models

Inspired by [12], the GREEN model [10] identifies six categories of clinical errors: (a) False report of a finding in the candidate; (b) Missing a finding present in the reference; (c) Misidentification of a finding’s anatomic location/position; (d) Misassessment of the severity of a finding; (e) Mentioning a comparison that isn’t in the reference; (f) Omitting a comparison detailing a change from a prior study. The GREEN score for the  $i^{th}$  sample is defined as:

$$GREEN_i = \frac{\#Matched Findings_i}{\#Matched Findings_i + \sum_{j=(a)}^{(f)} \#Error_{i,j}}, \quad (1)$$

where  $\sum_{j=(a)}^{(f)} \#Error_{i,j}$  represents the total clinically significant errors for the  $i^{th}$  sample across categories (a) to (f). “#Matched Findings” denotes the number of matched findings between generated and reference reports. Figure A3 illustrates the GREEN model’s output on the MIMIC-CXR test set. Furthermore, we compare our MLRG with R2Gen [2], CMN [3], CGPT2 [9], and SEI [6] in terms of “#Clinically Significant Errors” and “#Matched Findings”, as summarized in Table A4. The results reveal the follow-

<sup>1</sup><https://github.com/tylin/coco-caption>

<sup>2</sup><https://pypi.org/project/flchexbert/>

<sup>3</sup><https://pypi.org/project/radgraph/>

Observation	MIMIC-CXR							Two-view CXR						
	%	SEI [6]			MLRG (Ours)			%	SEI [6]			MLRG (Ours)		
		P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑		P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑
ECM	10.0	<b>0.373</b>	0.208	0.267	0.370	<b>0.353</b>	<b>0.361</b>	10.4	0.345	0.259	0.296	<b>0.412</b>	<b>0.385</b>	<b>0.398</b>
Cardiomegaly	14.8	0.599	<b>0.633</b>	<b>0.616</b>	<b>0.629</b>	0.570	0.598	14.4	0.578	<b>0.602</b>	<b>0.589</b>	<b>0.627</b>	0.550	0.586
Lung Opacity	13.8	0.519	0.170	0.256	<b>0.594</b>	<b>0.317</b>	<b>0.413</b>	13.6	0.526	0.197	0.287	<b>0.549</b>	<b>0.295</b>	<b>0.384</b>
Lung Lesion	2.5	<b>0.462</b>	0.021	0.041	0.429	<b>0.046</b>	<b>0.082</b>	3.0	0.179	0.030	0.051	<b>0.297</b>	<b>0.045</b>	<b>0.078</b>
Edema	8.3	<b>0.526</b>	0.361	0.428	0.516	<b>0.448</b>	<b>0.480</b>	6.5	0.420	0.368	0.392	<b>0.457</b>	<b>0.455</b>	<b>0.456</b>
Consolidation	3.3	0.218	<b>0.194</b>	<b>0.205</b>	<b>0.259</b>	0.150	0.190	3.1	<b>0.296</b>	<b>0.192</b>	<b>0.233</b>	0.204	0.115	0.147
Pneumonia	4.4	0.174	0.065	0.095	<b>0.316</b>	<b>0.235</b>	<b>0.270</b>	4.0	0.255	0.174	0.207	<b>0.284</b>	<b>0.210</b>	<b>0.241</b>
Atelectasis	10.9	0.469	0.395	0.429	<b>0.499</b>	<b>0.475</b>	<b>0.487</b>	10.0	0.457	0.425	0.440	<b>0.496</b>	<b>0.444</b>	<b>0.469</b>
Pneumothorax	1.0	0.174	0.039	0.064	<b>0.426</b>	<b>0.230</b>	<b>0.299</b>	0.7	0.417	0.109	0.172	<b>0.457</b>	<b>0.291</b>	<b>0.356</b>
Pleural Effusion	12.4	0.683	<b>0.697</b>	<b>0.690</b>	<b>0.716</b>	0.641	0.676	10.4	0.723	<b>0.641</b>	<b>0.680</b>	<b>0.731</b>	0.612	0.666
Pleural Other	1.6	0.167	0.022	0.039	<b>0.231</b>	<b>0.054</b>	<b>0.087</b>	1.9	<b>0.250</b>	0.071	0.111	0.194	<b>0.083</b>	<b>0.116</b>
Fracture	1.8	0.000	0.000	0.000	<b>0.174</b>	<b>0.021</b>	<b>0.037</b>	2.4	0.000	0.000	0.000	<b>0.261</b>	<b>0.031</b>	<b>0.056</b>
Support Devices	12.8	0.763	0.708	0.734	<b>0.768</b>	<b>0.788</b>	<b>0.778</b>	9.3	<b>0.734</b>	0.572	0.643	0.703	<b>0.686</b>	<b>0.695</b>
No Finding	2.4	0.161	0.597	0.253	<b>0.233</b>	<b>0.629</b>	<b>0.340</b>	10.3	<b>0.509</b>	0.899	0.650	0.490	<b>0.933</b>	<b>0.643</b>
micro avg	-	0.523	0.410	0.460	<b>0.549</b>	<b>0.468</b>	<b>0.505</b>	-	0.522	0.447	0.481	<b>0.532</b>	<b>0.474</b>	<b>0.501</b>
macro avg	-	0.378	0.294	0.294	<b>0.440</b>	<b>0.354</b>	<b>0.364</b>	-	0.406	0.324	0.339	<b>0.440</b>	<b>0.367</b>	<b>0.378</b>

Table A1. Clinical accuracy on the MIMIC-CXR and Two-view CXR datasets. “ECM” refers to Enlarged Cardiomeastinum. “P”, “R”, and “F1” represent Precision, Recall, and F1 score, respectively.

Observation	%	SEI [6]			MLRG (Ours)		
		P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑
Enlarged Cardiomeastinum	5.7	0.146	0.074	0.099	<b>0.242</b>	<b>0.264</b>	<b>0.252</b>
Cardiomegaly	12.7	0.515	0.627	0.566	<b>0.547</b>	<b>0.785</b>	<b>0.644</b>
Lung Opacity	20.2	0.640	0.342	0.446	<b>0.649</b>	<b>0.512</b>	<b>0.572</b>
Lung Lesion	5.0	0.333	0.035	0.063	<b>0.357</b>	<b>0.052</b>	<b>0.090</b>
Edema	7.1	<b>0.464</b>	0.524	<b>0.492</b>	0.441	<b>0.547</b>	0.489
Consolidation	3.3	<b>0.359</b>	<b>0.383</b>	<b>0.371</b>	0.354	0.270	0.306
Pneumonia	5.9	0.300	0.222	0.255	<b>0.318</b>	<b>0.307</b>	<b>0.312</b>
Atelectasis	10.5	0.381	0.441	0.409	<b>0.445</b>	<b>0.578</b>	<b>0.503</b>
Pneumothorax	0.0	-	-	-	-	-	-
Pleural Effusion	8.9	0.590	0.685	0.634	<b>0.698</b>	<b>0.723</b>	<b>0.710</b>
Pleural Other	3.2	<b>0.158</b>	0.056	0.082	0.135	<b>0.081</b>	<b>0.101</b>
Fracture	3.9	0.000	0.000	0.000	0.000	0.000	0.000
Support Devices	10.6	0.705	0.591	0.643	<b>0.715</b>	<b>0.840</b>	<b>0.772</b>
No Finding	3.0	0.175	<b>0.540</b>	0.265	<b>0.262</b>	0.466	<b>0.335</b>
micro avg	-	0.466	0.408	0.435	<b>0.513</b>	<b>0.517</b>	<b>0.515</b>
macro avg	-	0.341	0.323	0.309	<b>0.369</b>	<b>0.387</b>	<b>0.363</b>

Table A2. Clinical accuracy on the MIMIC-ABN dataset. “P”, “R”, and “F1” represent Precision, Recall, and F1 score, respectively.

ing: 1) Our MLRG achieves the highest “#Matched Findings” and GREEN score, with the fewest total clinically significant errors. This further confirms the effectiveness of our MLRG in generating clinically accurate radiology reports. 2) MLRG performs best in category “(f) Omitting a

comparison detailing a change from a prior study”, suggesting its ability to effectively extract temporal features from multi-view longitudinal data. 3) Although MLRG shows higher error counts than the baselines in categories (c) and (d), its total clinically significant errors across categories

Generated Section(s)	NLG Metrics $\uparrow$						CE Metrics $\uparrow$			
	B-1	B-2	B-3	B-4	MTR	R-L	RG	P	R	F1
FINDINGS	0.411	0.277	0.204	0.158	0.176	0.320	0.291	0.549	0.468	0.505
FINDINGS+IMPRESSION	0.402	0.270	0.197	0.152	0.172	0.327	0.289	0.558	0.468	0.509

Table A3. Performance of generating “FINDINGS” and “IMPRESSION” sections on the MIMIC-CXR dataset.

Method	#Clinically Significant Errors $\downarrow$						$\sum_{j=(a)}^{(f)} \#Error_j \downarrow$	#Matched Findings $\uparrow$	GREEN $\uparrow$
	(a)	(b)	(c)	(d)	(e)	(f)			
R2Gen [2]	1.310	3.089	<b>0.103</b>	<b>0.201</b>	<u>0.082</u>	0.142	4.926	1.803	0.283
CMN [3]	1.383	2.963	<u>0.127</u>	<u>0.228</u>	<b>0.081</b>	0.161	4.942	1.935	0.297
CGPT2 [9]	<b>1.150</b>	2.881	0.146	0.234	0.103	0.153	<u>4.666</u>	1.967	0.313
SEI [6]	1.391	<u>2.610</u>	0.154	0.273	0.108	<u>0.132</u>	4.668	<u>2.101</u>	<u>0.326</u>
MLRG (Ours)	<u>1.277</u>	<b>2.469</b>	0.199	0.284	0.091	<b>0.130</b>	<b>4.451</b>	<b>2.261</b>	<b>0.353</b>

Table A4. Performance comparison of our MLRG and four baselines on the MIMIC-CXR test set in terms of “#Clinically Significant Errors” and “#Matched Findings”. The best and second-best values are marked in **bold** and underlined, respectively.

Welch’s t-test	B-2	B-4	MTR	R-L	RG
p-value	0.0000	0.0078	0.0002	0.0045	0.0000

Table A5. P-values for all metrics between “w/ MV” and “w/o MV”.

Model	B-2 $\uparrow$	B-4 $\uparrow$	MTR $\uparrow$	R-L $\uparrow$	RG $\uparrow$
no semantic	0.249	0.136	0.162	0.298	0.288
with semantic	0.248	0.134	0.162	0.298	0.286
special tokens	0.277	0.158	0.176	0.320	0.291

Table A6. Ablation study on the effect of special tokens.

(a) to (f) remain lower than those of all baselines. To improve the accuracy of severity assessments and anatomical location descriptions, we are exploring the integration of saliency maps [11] and MIMIC-CXR-VQA [1] data to learn region-based features, aiming to generate more precise descriptions of findings.

## A.6. Gains from Multi-view Images

Table 4 presents quantitative evidence that multi-view input (“w/ MV”) outperforms single-view input (“w/o MV”). Additionally, we use Welch’s t-test to analyze the significance of the difference between “w/ MV” and “w/o MV”. Results in Appendix Table A5 show that integrating multi-view input significantly improves model performance.

## A.7. Effect of Tokenized Absence Encoding Technique

Our tokenized absence encoding technique employs special tokens to represent missing patient-specific prior knowledge, ensuring consistent and systematic handling of such cases. When prior knowledge is unavailable, these special tokens guide the model to rely exclusively on medical images; otherwise, both images and patient-specific prior knowledge are considered. To investigate potential bias introduced by special tokens, we evaluate two alternative methods for handling missing data on the MIMIC-CXR test set: 1) using a semantically neutral empty string (denoted as “no semantic”). 2) using a semantically meaningful string (specifically, “previous report/indication unavailable”, denoted as “with semantic”). As shown in Table A6, our tokenized absence encoding technique achieves superior performance without introducing detectable bias, outperforming both alternative methods.

## References

- [1] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. In *Advances in Neural Information Processing Systems*, pages 3867–3880. Curran Associates, Inc., 2023. 3
- [2] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020. 1, 3
- [3] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan.

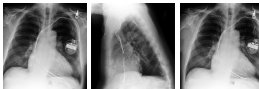


 <table><tr><td>B-4↑</td><td>RG↑</td><td>Ind</td><td>PI</td><td>PR</td></tr><tr><td>0.356/0.158</td><td>0.750/0.500</td><td>×</td><td>✓</td><td>×</td></tr></table>	B-4↑	RG↑	Ind	PI	PR	0.356/0.158	0.750/0.500	×	✓	×	<p><b>Reference report:</b> in comparison with study of <u>    </u>, there is again enlargement of the cardiac silhouette with a pacer device in place. no definite vascular congestion, raising the possibility of underlying cardiomyopathy or pericardial effusion. no acute focal pneumonia. <u>the right pice line has been removed.</u></p>	<p><b>Generated report by SEI:</b> in comparison with the study of <u>    </u>, there is little change and no evidence of acute cardiopulmonary disease. no pneumonia, vascular congestion, or pleural effusion. dual - channel pacer device remains in place.</p>	<p><b>Generated report by MLRG (Ours):</b> in comparison with the study of <u>    </u>, there is little overall change. <u>again there is enlargement of the cardiac silhouette</u> with triple-channel pacer device in place. no evidence of vascular congestion, pleural effusion, or acute focal pneumonia.</p>
B-4↑	RG↑	Ind	PI	PR									
0.356/0.158	0.750/0.500	×	✓	×									
 <table><tr><td>B-4↑</td><td>RG↑</td><td>Ind</td><td>PI</td><td>PR</td></tr><tr><td>0.383/0.284</td><td>0.667/0.400</td><td>✓</td><td>×</td><td>×</td></tr></table>	B-4↑	RG↑	Ind	PI	PR	0.383/0.284	0.667/0.400	✓	×	×	<p><b>Reference report:</b> as compared to the previous radiograph, the patient has received a nasogastric tube. the course of the tube <u>through</u> the esophagus is unremarkable. the tip of the tube appears to project over the very proximal parts of the stomach, the tube should be advanced by approximately 5 cm. otherwise, there is no relevant change. no complications such as pneumothorax.</p>	<p><b>Generated report by SEI:</b> as compared to the previous radiograph, the patient has received a dobbhoff catheter. the course of the catheter is unremarkable, the tip of the catheter is not included on the image. no evidence of complications, notably no pneumothorax. otherwise, the radiograph is unchanged.</p>	<p><b>Generated report by MLRG (Ours):</b> as compared to the previous radiograph, <u>the patient has received a nasogastric tube.</u> the course of the tube is unremarkable, the tip of the tube projects over the middle parts of the stomach. no evidence of complications, notably no pneumothorax. otherwise, unchanged radiograph.</p>
B-4↑	RG↑	Ind	PI	PR									
0.383/0.284	0.667/0.400	✓	×	×									
 <table><tr><td>B-4↑</td><td>RG↑</td><td>Ind</td><td>PI</td><td>PR</td></tr><tr><td>0.486/0.000</td><td>0.788/0.162</td><td>×</td><td>×</td><td>×</td></tr></table>	B-4↑	RG↑	Ind	PI	PR	0.486/0.000	0.788/0.162	×	×	×	<p><b>Reference report:</b> in comparison with the study of <u>    </u>, the monitoring and support devices remain in place. continued substantial enlargement of the cardiac silhouette with bilateral pleural effusions, compressive basilar atelectasis, and <u>moderate pulmonary edema.</u></p>	<p><b>Generated report by SEI:</b> comparison is made to previous study from <u>    </u>. there is a right ij central line with the distal lead tip in the proximal right atrium. there is a left - sided central venous line with the distal lead tip in the proximal right atrium. this could be pulled back 2 to 3 cm for more optimal placement. there is cardiomegaly. there are bilateral pleural effusions, left greater than right. there is a left retrocardiac opacity. there are no pneumothoraces.</p>	<p><b>Generated report by MLRG (Ours):</b> in comparison with the study of <u>    </u>, <u>there is continued enlargement of the cardiac silhouette</u> with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases. monitoring and support devices remain in place.</p>
B-4↑	RG↑	Ind	PI	PR									
0.486/0.000	0.788/0.162	×	×	×									

Figure A1. Examples of generated the “FINDINGS” section on the MIMIC-CXR test set. Each “A/B” cell corresponds to “MLRG/SEI”. Sentences from the reference report are highlighted in unique colors to clarify alignment with descriptions in the generated reports. Matching content in generated reports is shown in the same color. Correct temporal descriptions and failure descriptions of our MLRG are in **bold** and underlined. “Ind”, “PI”, and “PR” represent patient-specific indications, previous images, and previous reports, respectively.

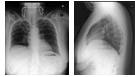
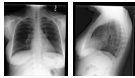

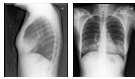
X-rays	Reference Report	Generated Report by Our MLRG										
<div><table><tr><td>B-4↑</td><td>RG↑</td><td>Ind</td><td>PI</td><td>PR</td></tr><tr><td>0.728</td><td>0.829</td><td>✓</td><td>✓</td><td>×</td></tr></table></div>	B-4↑	RG↑	Ind	PI	PR	0.728	0.829	✓	✓	×	<p>[FINDINGS] pa and lateral views of the chest provided. there is no focal consolidation, effusion, or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen. elevation of the right hemidiaphragm is unchanged from chest radiograph</p> <p>[IMPRESSION] no acute intrathoracic process.</p>	<p>[FINDINGS] pa and lateral views of the chest provided. lung volumes are somewhat low though allowing for this, there is no focal consolidation, effusion, or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.</p> <p>[IMPRESSION] no acute intrathoracic process.</p>
B-4↑	RG↑	Ind	PI	PR								
0.728	0.829	✓	✓	×								
<div><table><tr><td>B-4↑</td><td>RG↑</td><td>Ind</td><td>PI</td><td>PR</td></tr><tr><td>0.512</td><td>0.878</td><td>✓</td><td>×</td><td>×</td></tr></table></div>	B-4↑	RG↑	Ind	PI	PR	0.512	0.878	✓	×	×	<p>[FINDINGS] well expanded and clear lungs. no pleural effusion or pneumothorax. heart size, mediastinal contour, and hila are within normal limits. visualized upper abdomen is unremarkable.</p> <p>[IMPRESSION] normal chest radiograph. no pleural effusion or pneumonia</p>	<p>[FINDINGS] the lungs are well-expanded and clear. no pleural effusion or pneumothorax. heart size, mediastinal contour, and hila are unremarkable. limited assessment of the upper abdomen is within normal limits.</p> <p>[IMPRESSION] normal chest radiograph.</p>
B-4↑	RG↑	Ind	PI	PR								
0.512	0.878	✓	×	×								
<div><table><tr><td>B-4↑</td><td>RG↑</td><td>Ind</td><td>PI</td><td>PR</td></tr><tr><td>0.760</td><td>0.933</td><td>×</td><td>×</td><td>×</td></tr></table></div>	B-4↑	RG↑	Ind	PI	PR	0.760	0.933	×	×	×	<p>[FINDINGS] the heart is normal in size. the mediastinal and hilar contours appear within normal limits. there is no pleural effusion or pneumothorax. the lungs appear clear.</p> <p>[IMPRESSION] no evidence of acute disease.</p>	<p>[FINDINGS] the heart is normal in size. the mediastinal and hilar contours appear within normal limits. the lungs appear clear. there are no pleural effusions or pneumothorax.</p> <p>[IMPRESSION] no evidence of acute disease.</p>
B-4↑	RG↑	Ind	PI	PR								
0.760	0.933	×	×	×								
<div><table><tr><td>B-4↑</td><td>RG↑</td><td>Ind</td><td>PI</td><td>PR</td></tr><tr><td>0.178</td><td>0.300</td><td>×</td><td>✓</td><td>×</td></tr></table></div>	B-4↑	RG↑	Ind	PI	PR	0.178	0.300	×	✓	×	<p>[FINDINGS] pa and lateral views of the chest. there are new bibasilar opacities compatible with right middle lobe and lingular pneumonia. elsewhere, the lungs are clear and there is no effusion. cardiomeastinal silhouette is within normal limits. no acute osseous abnormality.</p> <p>[IMPRESSION] right middle lobe and lingular pneumonia. recommend repeat after treatment to document resolution.</p>	<p>[FINDINGS] pa and lateral views of the chest were provided. there is airspace consolidation in the right lower lobe concerning for pneumonia. no large effusion or pneumothorax. the heart and mediastinal contours are normal. the imaged osseous structures are intact. no free air below the right hemidiaphragm.</p> <p>[IMPRESSION] right lower lobe pneumonia.</p>
B-4↑	RG↑	Ind	PI	PR								
0.178	0.300	×	✓	×								

Figure A2. Generated examples of “FINDINGS” and “IMPRESSION” sections on the MIMIC-CXR test set. Sentences from the reference report are highlighted in unique colors to clarify alignment with descriptions in the generated reports. Matching content in generated reports is shown in the same color. Failure descriptions of our MLRG are underlined. “Ind”, “PI”, and “PR” represent patient-specific indications, previous images, and previous reports, respectively.

**Reference Report:** the lungs are clear bilaterally with no areas of focal consolidation. there is no pleural effusion or pneumothorax. patient is status post cabg. cardiomegaly is stable. mediastinal silhouette is within normal limits.  
**Generated Report by MLRG (Ours):** frontal and lateral views of the chest were obtained. the patient is status post median sternotomy and aortic valve replacement. the cardiac and mediastinal silhouettes are stable. no focal consolidation, pleural effusion, or pneumothorax is seen. there is no pulmonary edema.

[Explanation]:

GREEN Response

The candidate report has a few discrepancies when compared to the reference report. The candidate report mentions an aortic valve replacement, which is not mentioned in the reference report. The reference report mentions cardiomegaly, which is not mentioned in the candidate report. The reference report also mentions that the lungs are clear bilaterally, which is not explicitly stated in the candidate report.

[Clinically Significant Errors]:

- (a) False report of a finding in the candidate: 1. Aortic valve replacement.
- (b) Missing a finding present in the reference: 1. Cardiomegaly.
- (c) Misidentification of a finding's anatomic location/position: 0.
- (d) Misassessment of the severity of a finding: 0.
- (e) Mentioning a comparison that isn't in the reference: 0.
- (f) Omitting a comparison detailing a change from a prior study: 0.

#Clinically Significant Errors↓						#Matched Findings↑	GREEN↑
(a)	(b)	(c)	(d)	(e)	(f)		
1	1	0	0	0	0	3	0.6

[Clinically Insignificant Errors]:

- (a) False report of a finding in the candidate: 0.
- (b) Missing a finding present in the reference: 0.
- (c) Misidentification of a finding's anatomic location/position: 0.
- (d) Misassessment of the severity of a finding: 0.
- (e) Mentioning a comparison that isn't in the reference: 0.
- (f) Omitting a comparison detailing a change from a prior study: 0.

[Matched Findings]:

- 3. Status post surgery; No focal consolidation; No pleural effusion or pneumothorax.

Figure A3. An output result of the GREEN model [10] on the MIMIC-CXR test set. “#Clinically Significant Errors” and “#Matched Findings” represent the number of clinically significant errors and matched findings, respectively.

- viana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597, 2019. 1
- [5] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR, a large publicly available database of labeled chest radiographs, 2019. 1
- [6] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhushi Zhong, Zhicheng Jiao, Grayson Baird, Harrison Bai, and Qiguang Miao. Structural entities extraction and patient indications incorporation for chest x-ray report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 433–443, Cham, 2024. Springer Nature Switzerland. 1, 2, 3
- [7] Qiguang Miao, Kang Liu, Zhuoqi Ma, Yunan Li, Xiaolu Kang, Ruixuan Liu, Tianyi Liu, Kun Xie, and Zhicheng Jiao. Evoke: Elevating chest x-ray report generation via multi-view contrastive learning and patient-specific knowledge, 2025. 1
- [8] Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian J. McAuley. Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1954–1960, 2020. 1
- [9] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023. 1, 3
- [10] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA, 2024. Association for Computational Linguistics. 1, 5
- [11] Honglong Yang, Hui Tang, and Xiaomeng Li. Fita: Fine-grained image-text aligner for radiology report generation, 2024. 1, 3
- [12] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. 1