The automatic generation of highly clinically accurate radiology reports from Chest X-Ray images could improve clinical outcomes by reducing radiologist workload, prioritizing severe cases, and augmenting existing radiograph processing pipelines.

**Techniques**
- Template matching. It is too restricted method, we did not consider it.
- Retrieval-based. Baseline method. Use K tags from most similar images.
- Encoder-Decoder Generative model; There are a lot of things to try. It is our main method.

**Takeaways**
- Providing image representation and pathological probability outputs to encoder improves the performance
- Joint loss helps significantly

**Literature**
- WCL: Cluster reports with labels for contrastive loss.
- IFCC: Combine factual metric loss with a language model loss and an NLG loss.

**Metrics**
- NLG metric BLEU doesn’t show Clinical Efficacy (CE).
- Compare pathology labels from original and generated text for CE metrics.
- Baseline method uses tag Retrieval from corpus of ground truth clinical tags.

**Data & Experiments**
- IU X-Ray Frontal images, reports and pathology labels (1952 for training, and 488 for testing)
- Experiment with/out pathological probability outputs
- Experiment with/out contrastive loss

**Results**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Best Model from the paper</th>
<th>BLU-1</th>
<th>BLU-4</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMIC-CXR</td>
<td>IFCC</td>
<td>11.1</td>
<td>11.1</td>
<td>40.6*</td>
<td>72.9*</td>
<td>56.4*</td>
</tr>
<tr>
<td></td>
<td>WCL</td>
<td>16.7</td>
<td>16.7</td>
<td>38.5**</td>
<td>77.4**</td>
<td>59.4**</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td></td>
<td></td>
<td>in progress</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JUX-ray</td>
<td>Retrieval</td>
<td>8.78</td>
<td></td>
<td>percent correct tags generated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2Gen</td>
<td>16.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>27.2</td>
<td>27.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The micro average of accuracy, precision, recall, and F1 scores are calculated over 5 observations for: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion

**Future Work**
- Train and validate on a full-size MIMIC-CXR dataset. (it is not possible in project time due to limited computational resources)
- Experiment with model architectures for pathological probabilities class predictions.
- Experiment with different architectures for generation based approach.
- Experiment more with joint loss functions, including contrastive loss functions.