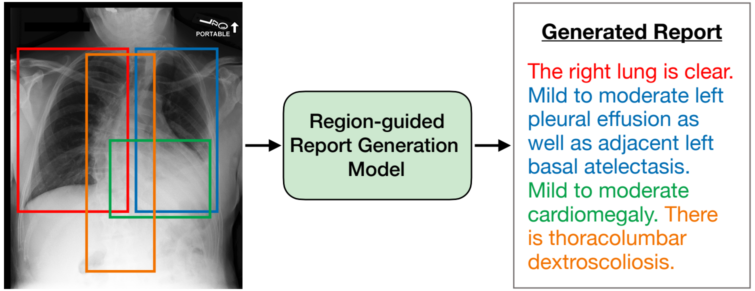
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# Interactive and Explainable Region-guided Radiology Report Generation

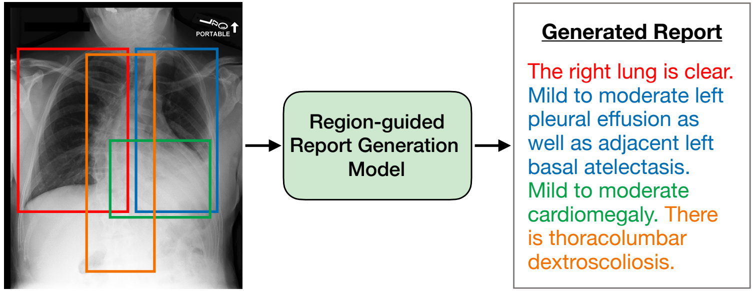
Tim Tanida[1,*], Philip Müller[1,*], Georgios Kaissis[1,2], Daniel Rueckert[1,3]

[1]Technical University of Munich, [2]Helmholtz Zentrum Munich, [3]Imperial College London

Paper Tag: **TUE-PM-316**



# Motivation




## Radiology reports

- Written by radiologists in clinical practice
- Interpratation of medical images
- Consisting of several sentences

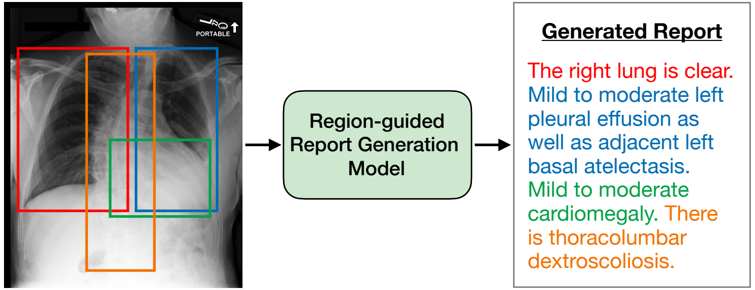
**=> Goal: Automation of report writing**

## Problems of current methods

- No explicit focus on salient regions
- Factual inconsistencies and incompleteness
- Limited explainability
- No human interaction in generation process

# Our Approach:

## Region-Guided Radiology Report Generation (RGRG)



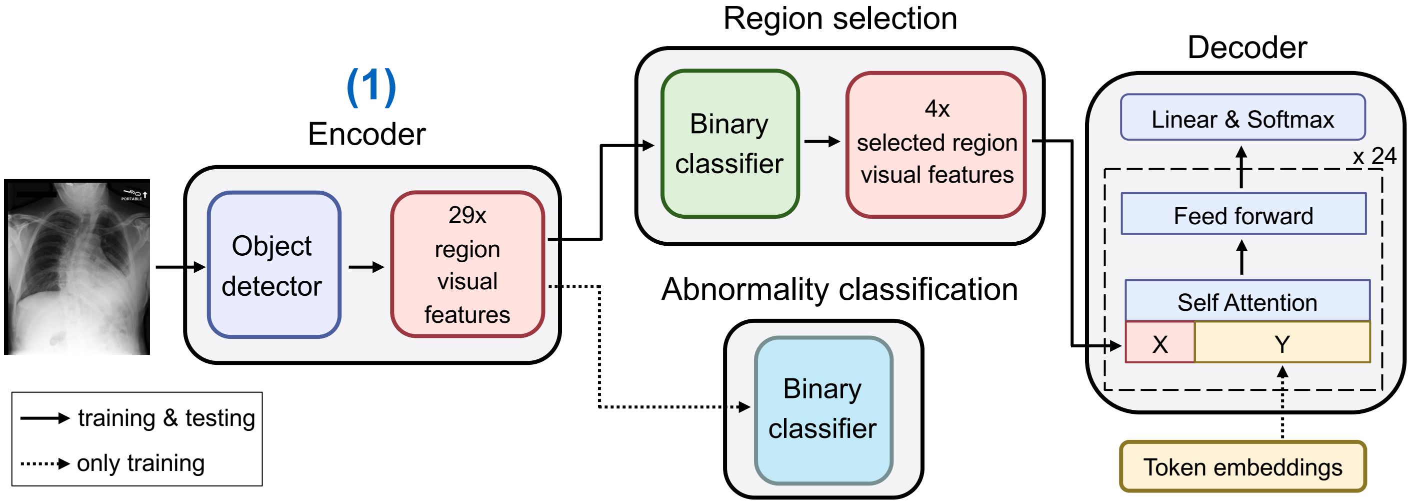

### Our Approach: RGRG

- Explicit detection of anatomical regions
- Explicit description of each (salient) region
- Option to manually specify regions

### Benefits

- Factual completeness and consistency
- High degree of explainability
- Interactive generation process with radiologist

# RGRG: Architecture



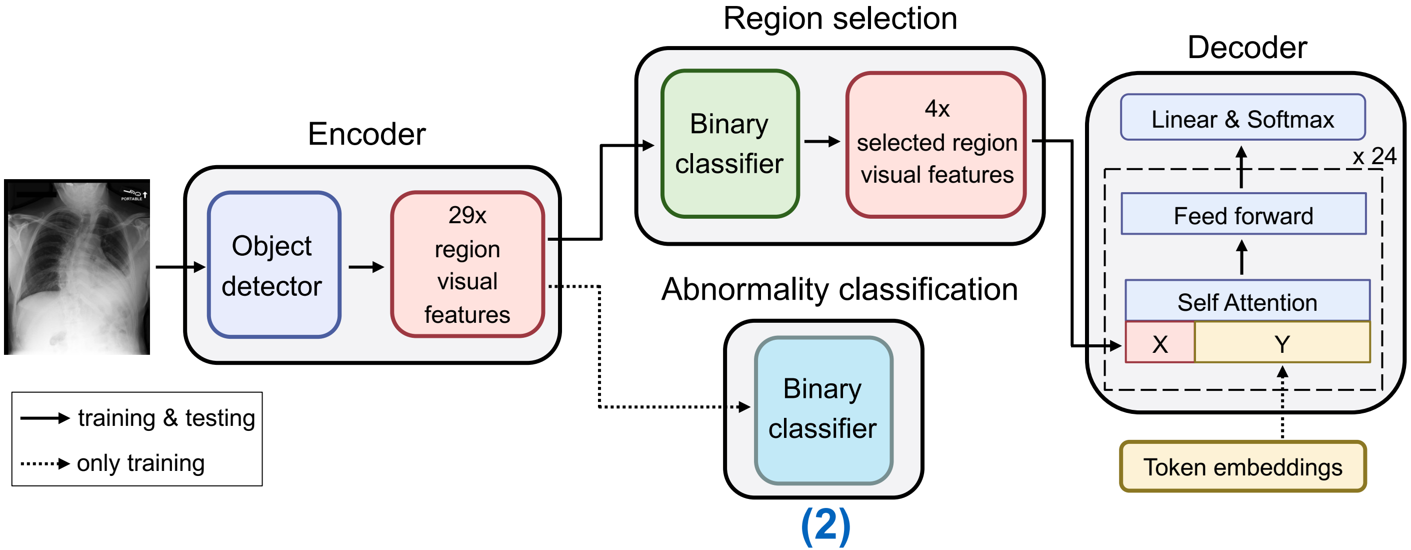

## (1) Encoder with Object Detector

- ResNet50 + Faster R-CNN detecting 29 anatomical regions

→ Top-1 object proposal for each class

→ Extract region features using RoI pooling

# RGRG: Architecture

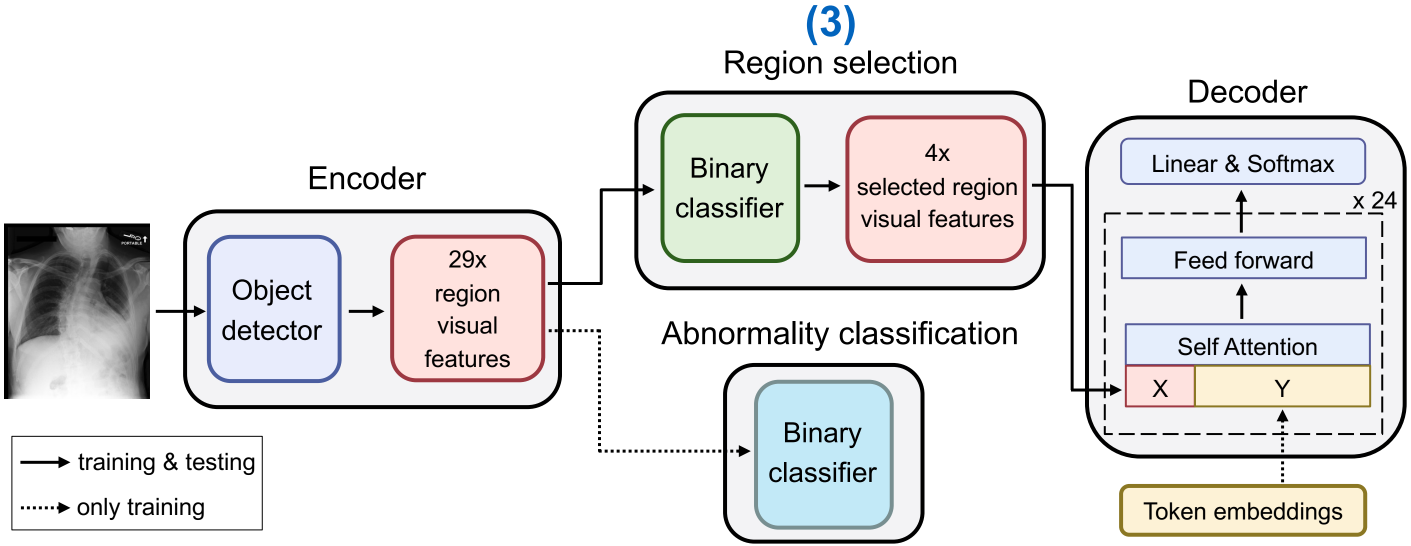



## (2) Abnormality Classification

- Binary classifier on region features: **Normal** (healthy) / **Abnormal** (pathology present)

→ Encourages meaningfull region features

# RGRG: Architecture

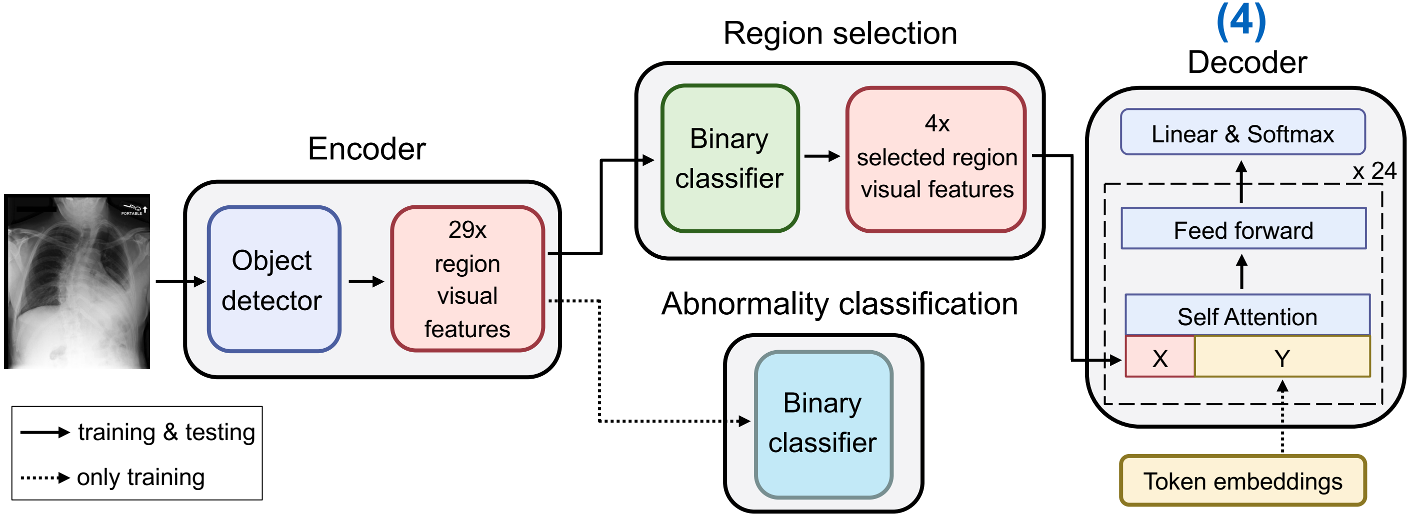



## (3) Region Selection

- Binary classifier on region features: **Select** / **Ignore** region for report generation

→ Pre-selection to only generate sentences for relevant (salient) regions

# RGRG: Architecture

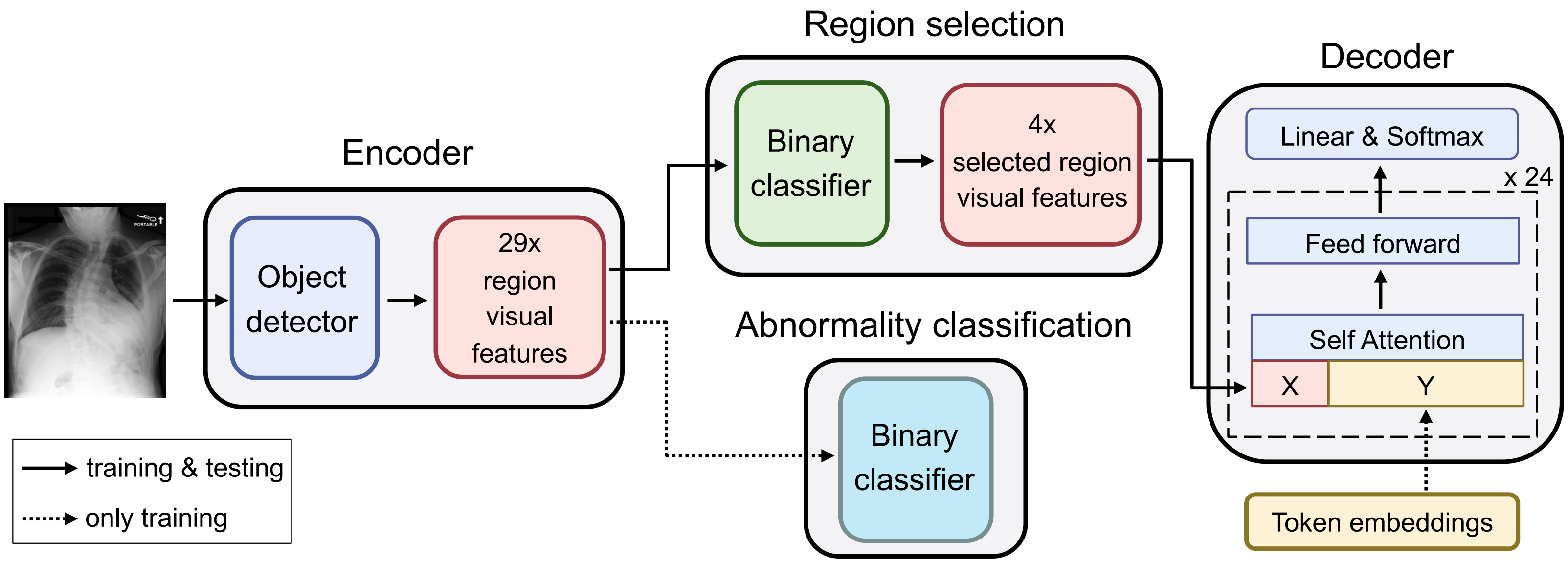



## (4) Decoder for Sentence Generation

- Generated sentences independently per region using region features
- 355M-parameter GPT-2 Medium pre-trained on PubMed abstracts

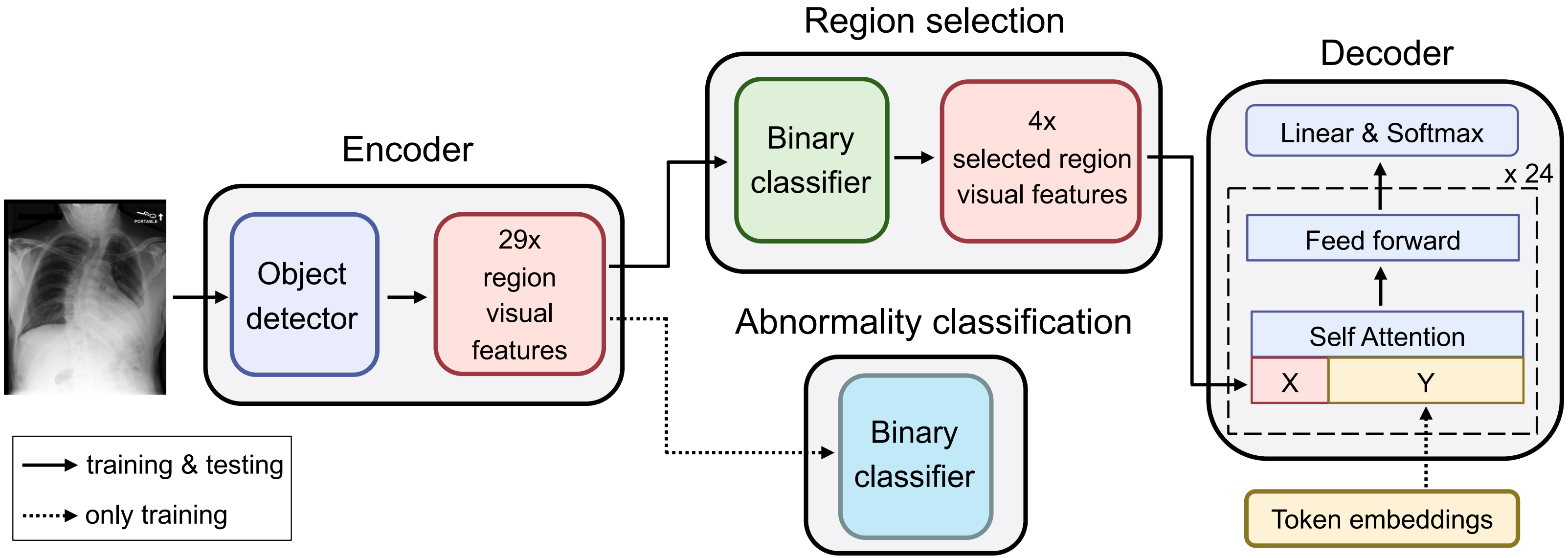
→ Conditioning using pseudo self-attention: extend key/value sequences in attention

# RGRG: Inference




1. **Full radiology report generation**
   - Concatenation of generated sentences
2. **Anatomy-based sentence generation (interactive)**
   - Region selection module exclusively chooses radiologist's selection
3. **Selection-based sentence generation (interactive)**
   - Manually drawn bounding box → RoI pooling → Decoder

# RGRG: Training




$$\mathcal{L} = \lambda_{obj}\mathcal{L}_{obj} + \lambda_{select}\mathcal{L}_{select} + \lambda_{abnormal}\mathcal{L}_{abnormal} + \lambda_{language}\mathcal{L}_{language}$$

- $\mathcal{L}_{obj}$ : Faster R-CNN loss
- $\mathcal{L}_{select}$ : Binary cross-entropy loss
- $\mathcal{L}_{abnormal}$ : Binary cross-entropy loss
- $\mathcal{L}_{language}$ : Cross-entropy loss

# Chest ImaGenome Dataset [1]

- Automatically constructed from the MIMIC-CXR [2] dataset
- 242,072 frontal chest X-ray images
- Scene graph data structure (inspired by Visual Genome)
- Each image has:
  - Bounding box coordinates for 29 anatomical regions
  - Reference sentences describing regions (if exist in reference report)

[1] Wu, J., Agu, N., et al. Chest ImaGenome Dataset for Clinical Reasoning. In PhysioNet, 2021.
[2] A. E. W. Johnson, et al. Mimic-cxr database (version 2.0.0). PhysioNet, 2019.

# Results: Full Report Generation

Natural language generation (NLG) metrics

| Dataset | Method | Year | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|---|---|---|
| MIMIC-CXR | R2Gen [7] | 2020 | 0.353 | 0.218 | 0.145 | 0.103 | 0.142 | 0.277 | 0.406† |
| | CMN [6] | 2021 | 0.353 | 0.218 | 0.148 | 0.106 | 0.142 | 0.278 | - |
| | PPKED [24] | 2021 | 0.360 | 0.224 | 0.149 | 0.106 | 0.149 | 0.284 | 0.237 |
| | $\mathcal{M}^2$ TR. PROGRESSIVE [30] | 2021 | 0.378 | 0.232 | 0.154 | 0.107 | 0.145 | 0.272 | - |
| | Contrastive Attention [25] | 2021 | 0.350 | 0.219 | 0.152 | 0.109 | 0.151 | 0.283 | - |
| | AlignTransformer [53] | 2021 | 0.378 | 0.235 | 0.156 | 0.112 | 0.158 | 0.283 | - |
| | $\mathcal{M}^2$ Trans w/ NLL [28] | 2021 | - | - | - | 0.105 | - | - | 0.445 |
| | $\mathcal{M}^2$ Trans w/ NLL+BS+$f_{C_E}$ [28] | 2021 | - | - | - | 0.111 | - | - | 0.492 |
| | $\mathcal{M}^2$ Trans w/ NLL+BS+$f_{C_{EN}}$ [28] | 2021 | - | - | - | 0.114 | - | - | **0.509** |
| | ITA [47] | 2022 | **0.395** | **0.253** | 0.170 | 0.121 | 0.147 | 0.284 | - |
| | CvT-212DistilGPT2 [29] | 2022 | 0.392 | 0.245 | 0.169 | 0.124 | 0.153 | **0.285** | 0.361 |
| | RGRG | Ours | 0.373 | 0.249 | **0.175** | **0.126** | **0.168** | 0.264 | 0.495 |

*Δ+10.5%* *Δ+6.3%*

NLG metrics: count matching n-grams ("word overlap")

→ Domain-agnostic

BLEU score boosted by lowercasing

Best without lowercasing

➔ **Competitive/outperforms** methods on NLG metrics
➔ New **SOTA** on METEOR
➔ Lower ROUGE-L score due to low precision of region selection (very subjective)

# Results: Full Report Generation

Clinical efficacy (CE) metrics

| Dataset | Method | RL | Year | $P_{mic-5}$ | $R_{mic-5}$ | $F_{1, mic-5}$ | $P_{ex-14}$ | $R_{ex-14}$ | $F_{1, ex-14}$ |
|---|---|---|---|---|---|---|---|---|---|
| MIMIC-CXR | R2Gen [7] | ✗ | 2020 | 0.412 | 0.298 | 0.346 | 0.331 | 0.224 | 0.228 |
| | $\mathcal{M}^2$ Trans w/ NLL [28] | ✗ | 2021 | 0.489 | 0.411 | 0.447 | - | - | - |
| | $\mathcal{M}^2$ Trans w/ NLL+BS+$f_{C_E}$ [28] | ✓ | 2021 | 0.463 | **0.732** | **0.567** | - | - | - |
| | $\mathcal{M}^2$ Trans w/ NLL+BS+$f_{C_{EN}}$ [28] | ✓ | 2021 | **0.503** | 0.651 | **0.567** | - | - | - |
| | CMN [6] | ✗ | 2021 | - | - | - | 0.334 | 0.275 | 0.278 |
| | Contrastive Attention [25] | ✗ | 2021 | - | - | - | 0.352 | 0.298 | 0.303 |
| | $\mathcal{M}^2$ TR. PROGRESSIVE [30] | ✗ | 2021 | - | - | - | 0.240 | 0.428 | 0.308 |
| | CvT-212DistilGPT2 [29] | ✗ | 2022 | - | - | - | 0.359 | 0.412 | 0.384 |
| | RGRG | ✗ | Ours | 0.491 | 0.617 | 0.547 | **0.461** | **0.475** | **0.447** |

*Δ+22,4%* *Δ+16.4%*

CE metrics: compare generated and reference report w.r.t. clinical observations

→ Evaluate diagnostic accuracy (and factual completeness/consistency)

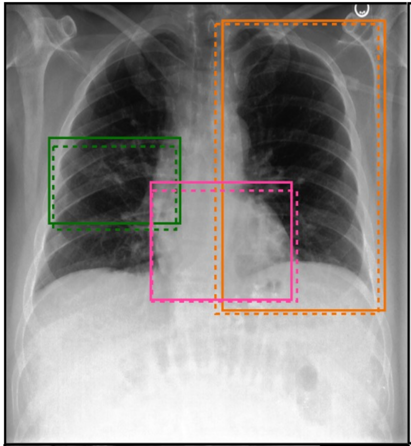
— RL-optimized on CE metrics

➔ **Competitive** with methods directly optimized on CE metrics

➔ **Substantially outperforms** all other methods on CE metrics

➔ Generates **factually complete and consistent** reports

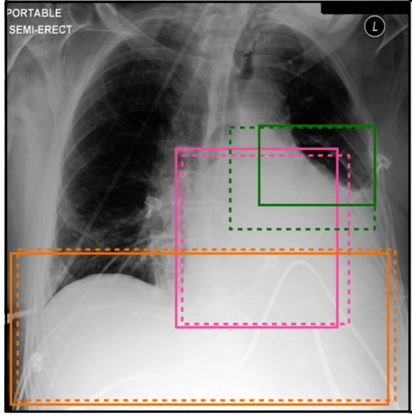
# Results: Anatomy-based Sentence Generation



Right mid lung zone:
Generated: Right lower lobe pneumonia.
Reference: There is persistent subtle airspace opacity in the right mid to lower lung, this may reflect the residua of the patient's known pneumonia.

Left lung:
Generated: There is no pleural effusion or pneumothorax. No acute cardiopulmonary process.
Reference: The left lung is clear. No pleural effusion seen.

Cardiac silhouette:
Generated: The cardiomediastinal silhouette is normal.
Reference: The heart is not enlarged.

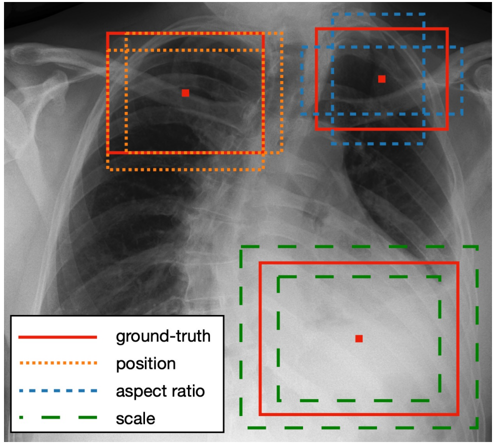


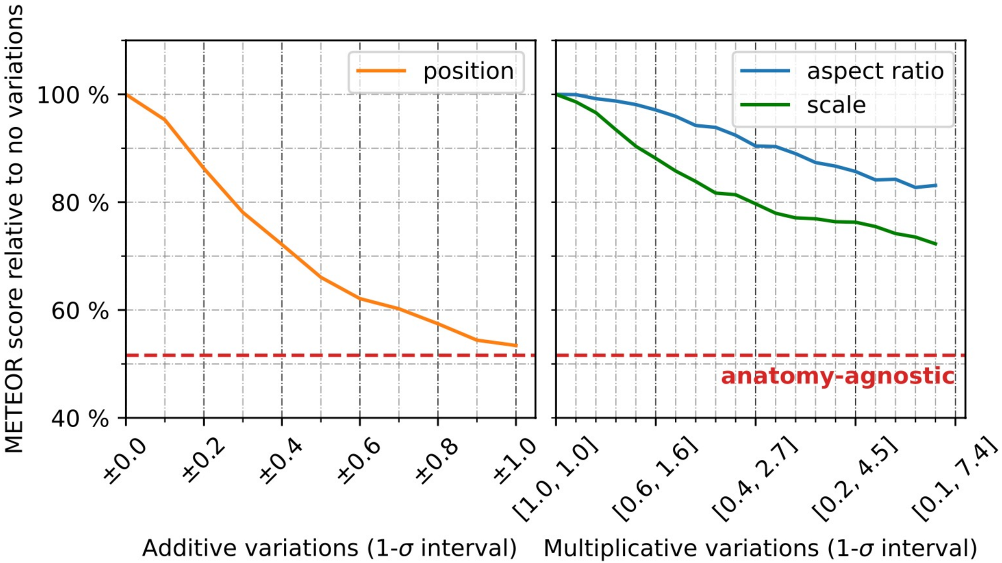

Left mid lung zone:
Generated: Moderate left pleural effusion and moderate left lower lobe atelectasis are unchanged.
Reference: Atelectasis of the left lower lobe with potential accompanying small left pleural effusion.

Abdomen:
Generated: NG tube tip is in the stomach.
Reference: Nasogastric tube is in unchanged position.

Cardiac silhouette:
Generated: Moderate cardiomegaly persists.
Reference: Unchanged moderate cardiomegaly.

# Results: Selection-based Sentence Generation

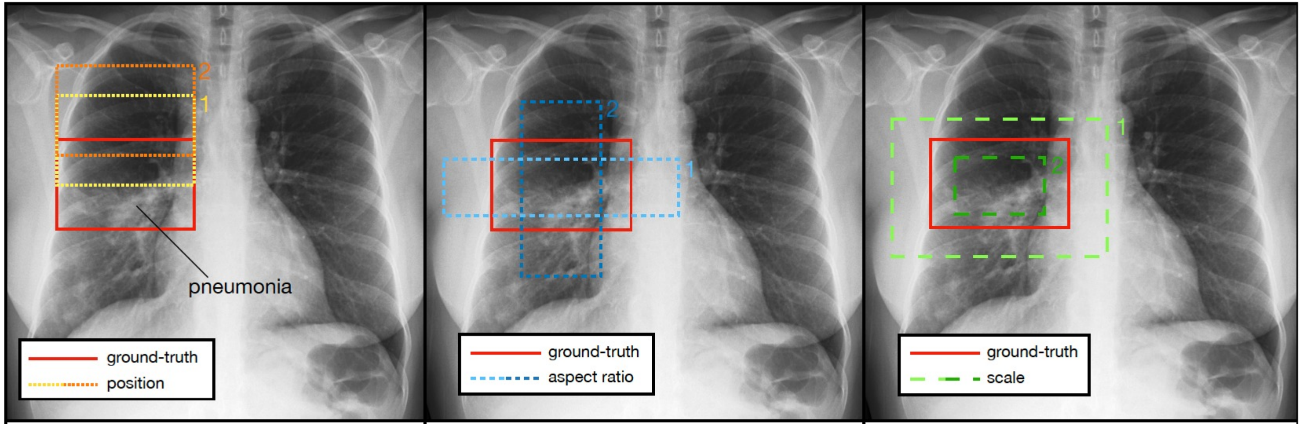



Variation of bounding boxes to simulate manually drawn boxes
→ Evaluate sensitivity to changes

➔ **Location-sensitivity** ➔ **Shape robustness**

# Results: Selection-based Sentence Generation




Ground-truth:
Generated: Right lower lobe pneumonia.
Reference: Interval worsening of right lower lobe pneumonia.

Position 1:
Generated: The right upper lobe opacity has improved since ___.

Position 2:
Generated: Upper lungs are clear.

Ground-truth:
Generated: Right lower lobe pneumonia.
Reference: Interval worsening of right lower lobe pneumonia.

Aspect ratio 1:
Generated: Right lower lobe pneumonia.

Aspect ratio 2:
Generated: Right <u>upper</u> lobe pneumonia.

Ground-truth:
Generated: Right lower lobe pneumonia.
Reference: Interval worsening of right lower lobe pneumonia.

Scale 1:
Generated: Right lower lobe pneumonia.

Scale 2:
Generated: Right <u>upper</u> lobe pneumonia.

# Conclusion

- Simple yet effective approach to radiology report generation
- Focus on salient anatomical regions
- Competitive with/outperforming SOTA methods in full report generation
- Generates region-specific descriptions → High degree of explainability
- Interactive intervention in generation process possible