CURE: Curriculum-guided Multi-task Training for Reliable Anatomy Grounded Report Generation

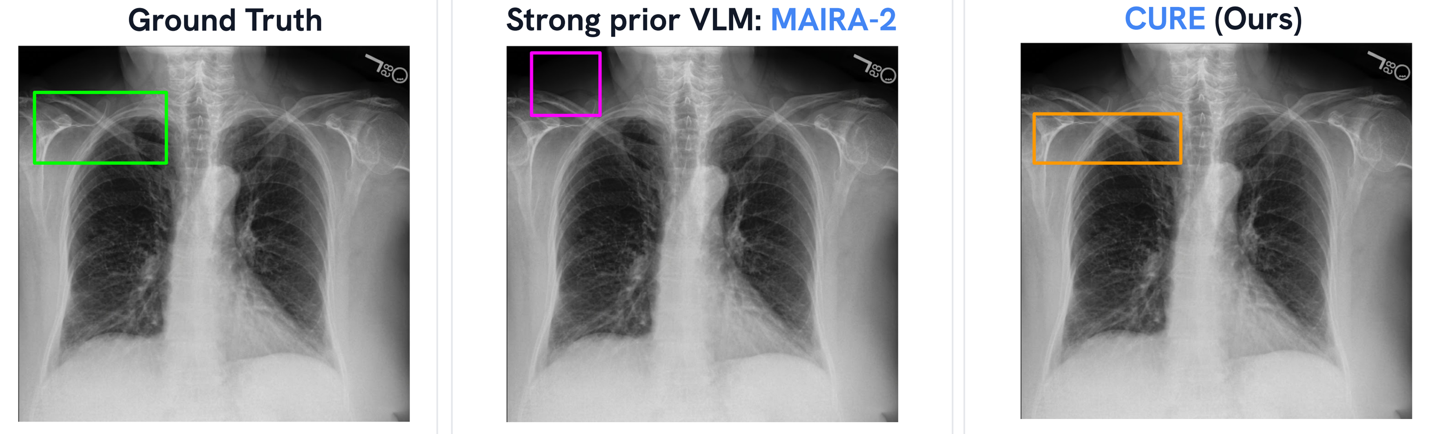
Pablo Messina[1,3,4], Andrés Villa[2], Juan C. Leon[2], Karen Sánchez[2], Carlos Hinojosa[2], Denis Parra[1,3,4], Álvaro Soto[1,3], Bernard Ghanem[2]
1. Pontificia Universidad Católica de Chile 2. KAUST 3.CENIA 4.iHEALTH

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation: VLM Hallucinations & Data Imbalance

Grounded medical VLMs can hallucinate abnormalities

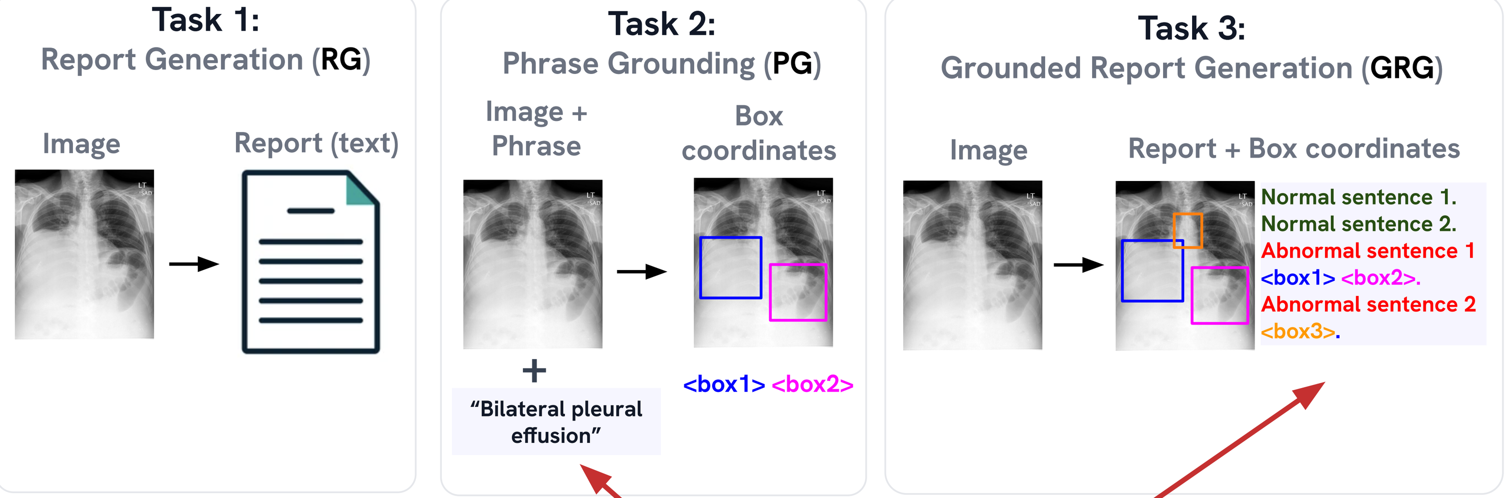
TASK: "Locate and describe the right clavicle"



Ground Truth	Strong prior VLM: MAIRA-2	CURE (Ours)
The right clavicle is normal	"Right clavicle is fractured" • Visual Grounding ✅ (roughly) • Hallucinated finding ❌	"No acute osseous abnormalities" • Visual Grounding ✅ • Healthy description ✅

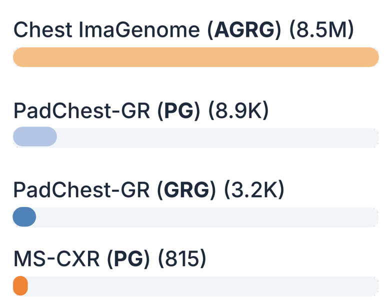
Why do grounded medical VLMs hallucinate? Because grounding labels are abnormality-biased

PRIOR WORK



Abnormality bias: Prior work datasets only provide grounding supervision for abnormal sentences

Another problem: Dataset & Class Imbalance



Public datasets, very different sizes

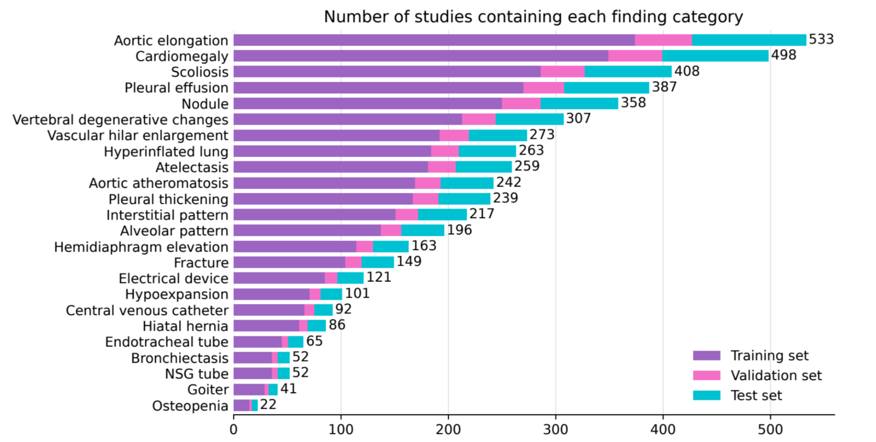
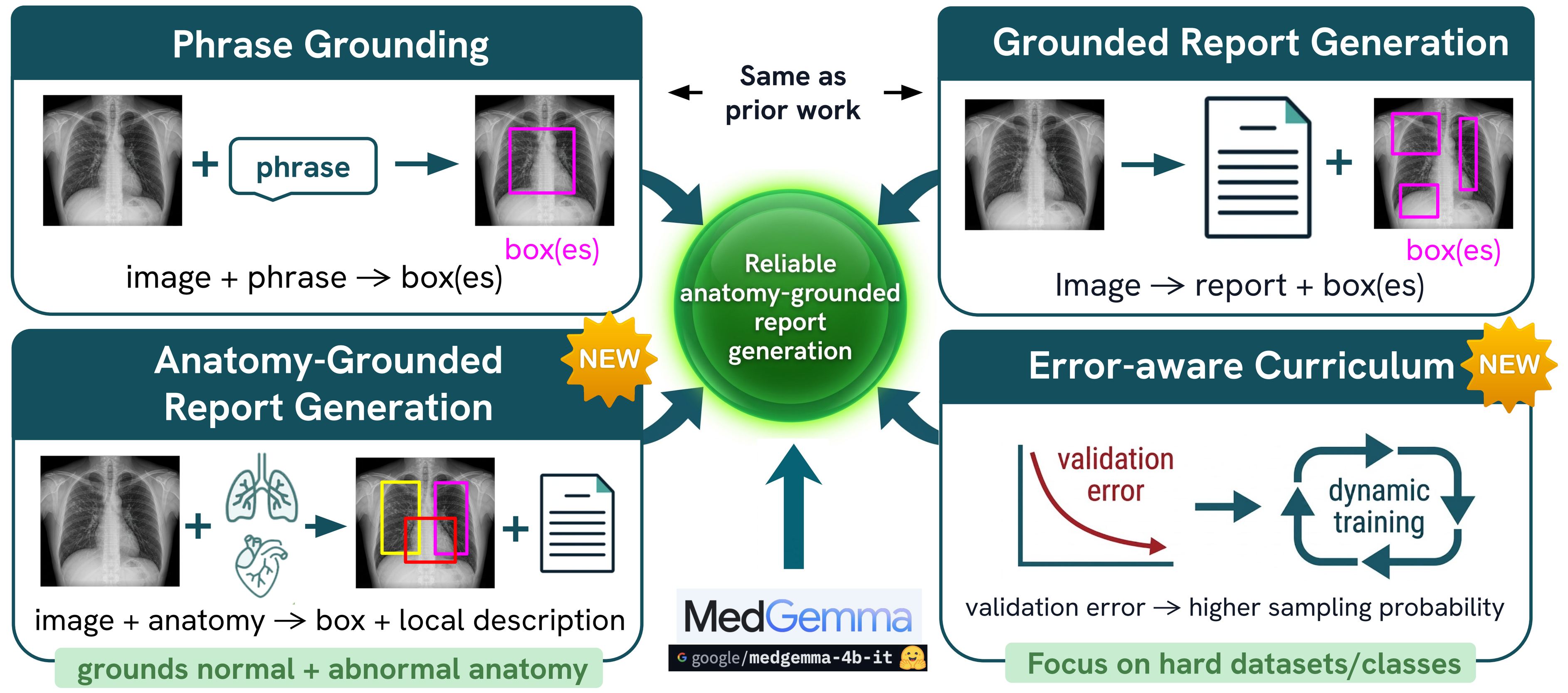


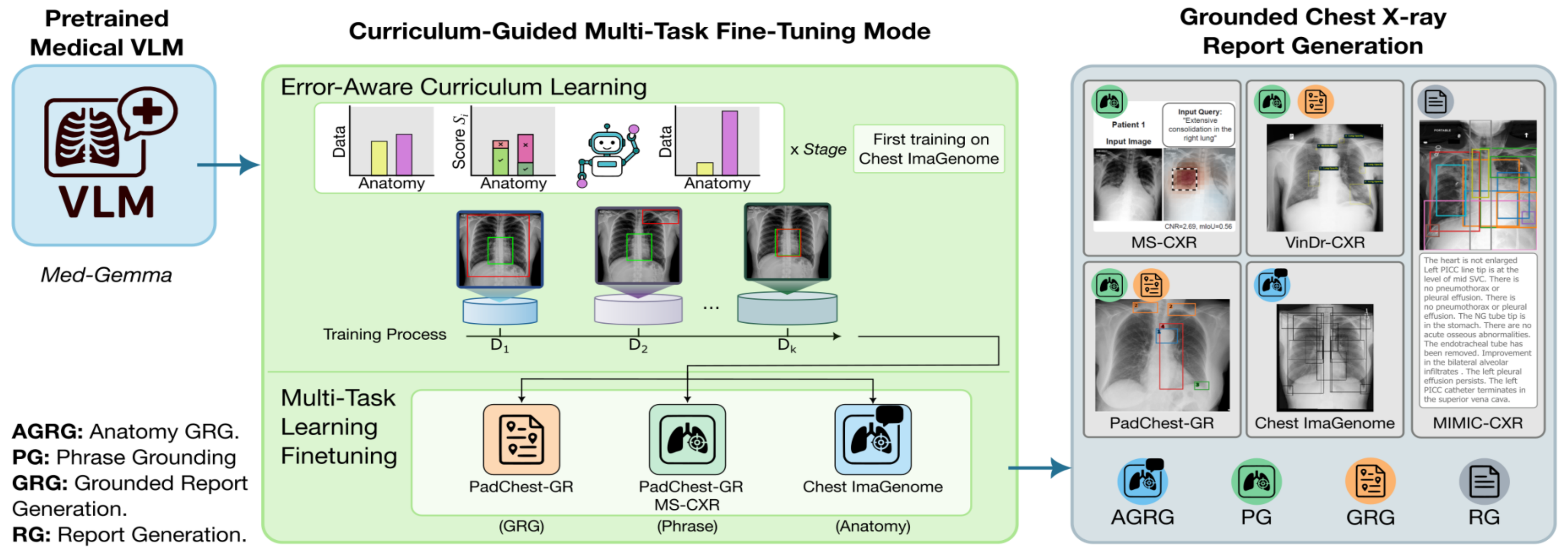
Figure taken from PadChest-GR paper

Solution: CURE

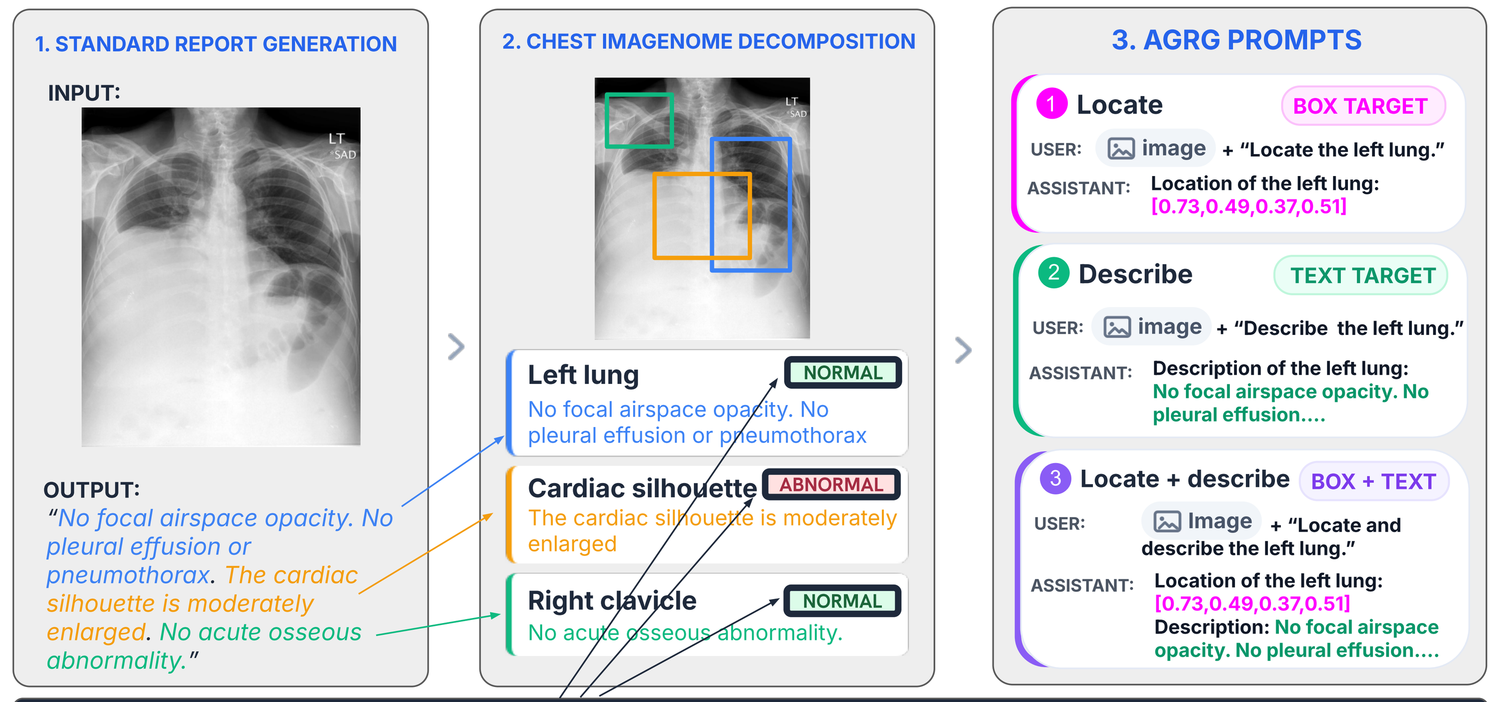
CURE: teaching grounded VLMs to locate and describe both normal and abnormal anatomy



Unified instruction tuning: **image + instruction → response**



AGRG: learn to locate and describe anatomical regions



💡 AGRG grounds both normal and abnormal regions → reducing abnormality bias

Error-aware curriculum: train more on what is currently hard

Periodic validation + Reweighting

1. Hybrid Score (si)

$$s_i = \alpha \cdot \mathrm{IoU}_i + (1-\alpha) \cdot \mathrm{CXRFEScore}_i$$

Combines Grounding and Text Quality metrics:

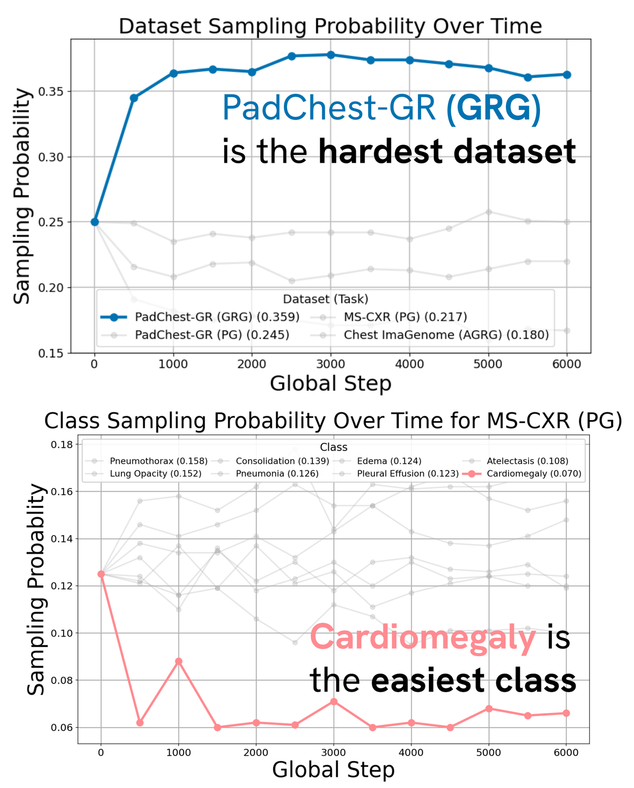
- IoU: Measures grounding accuracy
- CXRFEScore: Measures text generation quality
- α : interpolation weight

2. Sampling Probability (pi)

$$p_i = \frac{1 - s_i}{\sum_j (1 - s_j)}$$

- Inversely proportional to performance
- Normalized across datasets or classes

low score → high error → higher sampling probability



Results & Examples

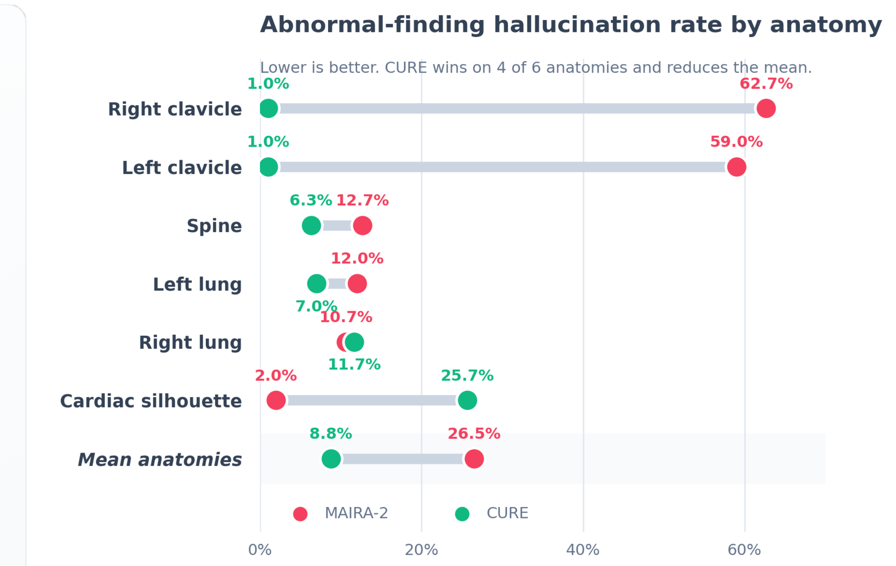
CURE reduces abnormality hallucinations by two-thirds

Mean hallucination rate

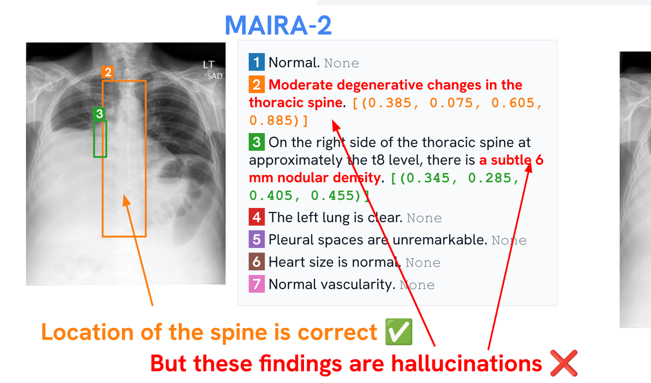
26.5% → 8.8%

−67%

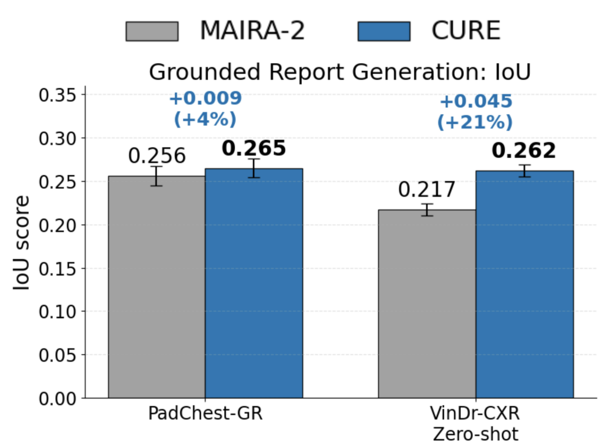
RELATIVE REDUCTION

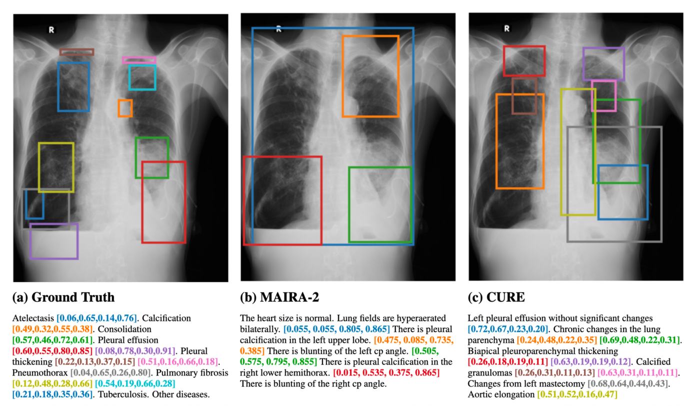


TASK: "Locate and describe the spine"



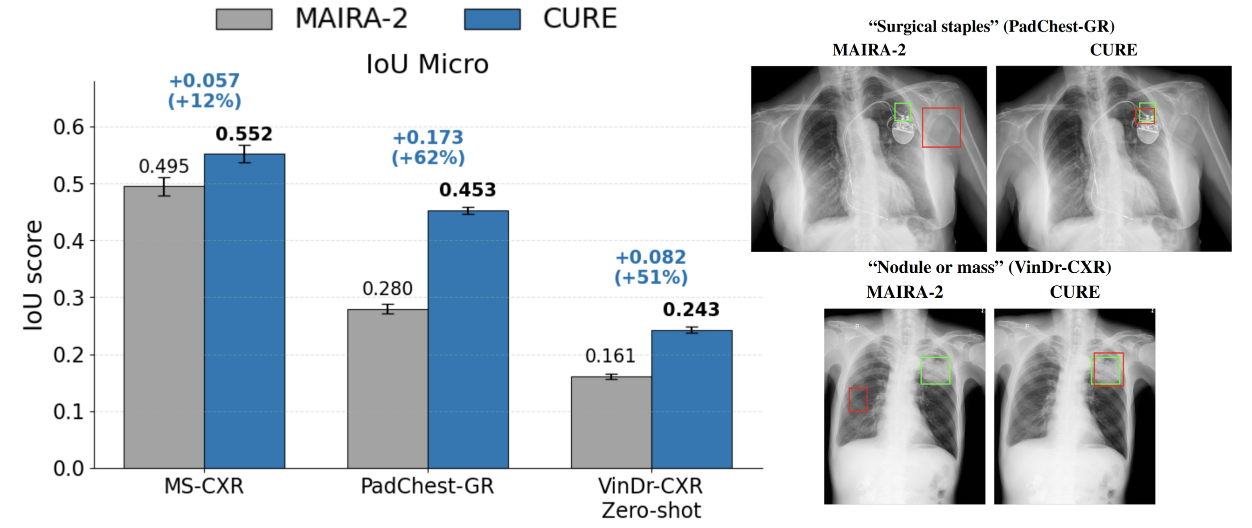
Grounded Report Generation: CURE also achieves higher IoU than MAIRA-2 on PadChest-GR and VinDr-CXR



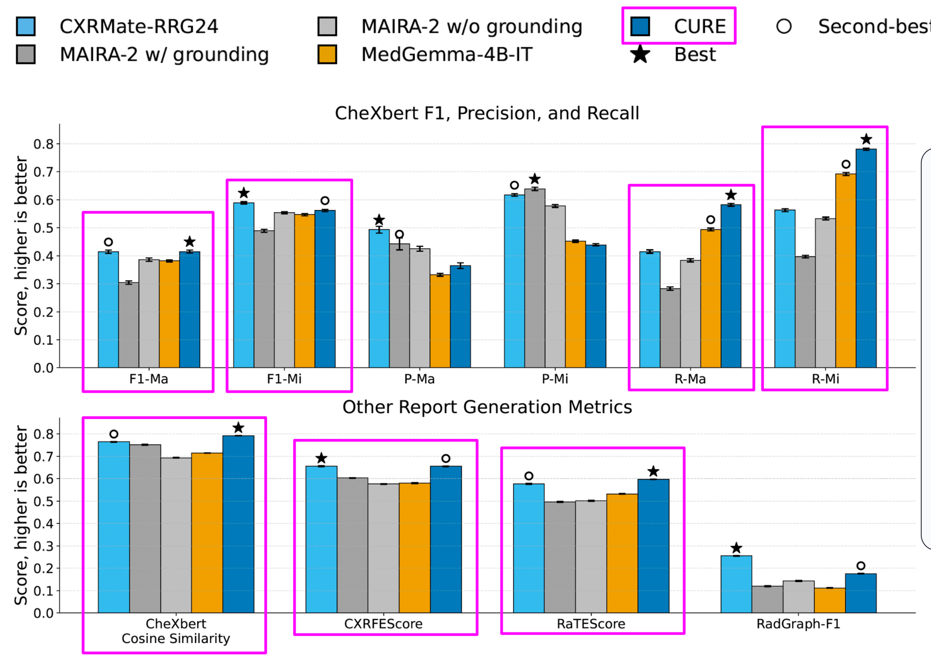


CURE improves visual grounding

Phrase Grounding: CURE achieves higher IoU scores than MAIRA-2 on MS-CXR, PadChest-GR, and VinDr-CXR



CURE shows competitive report generation quality on MIMIC-CXR

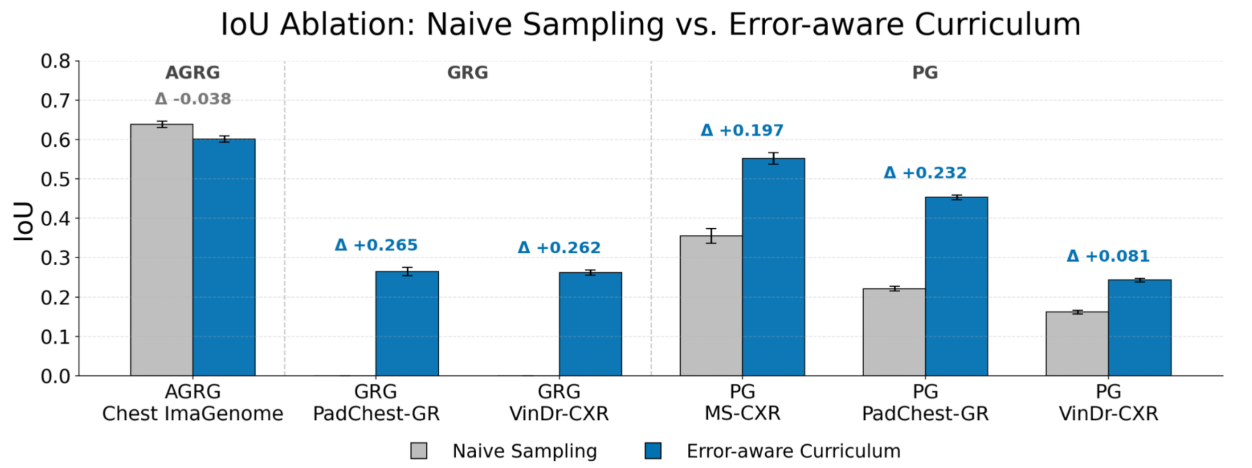


Key Observations

CURE outperforms or is close to major baselines on several metrics.

CURE achieves this by combining AGRG (29 locations) + GRG.

Ablation: Error-aware Curriculum is better than Naive Sampling



Summary & Takeaways

Contributions	Key Results	Takeaways
Anatomy Grounded Report Generation	2/3 fewer abnormality hallucinations	Ground and describe normal anatomy too
Error-aware Curriculum	Better visual grounding	Use adaptive sampling for data imbalance
	Competitive report generation quality	

Project Website