Do Clinical VLMs Need Dense Visual Tokens? Probing Spatial Grounding in Radiology Report Generation

L. Julián Lechuga López [1,2] Tim G. J. Rudner[3] Farah E. Shamout [1,2]

[1] New York University, [2] New York University Abu Dhabi, [3] University of Toronto

Is your chest X-ray needed to generate a radiology report?

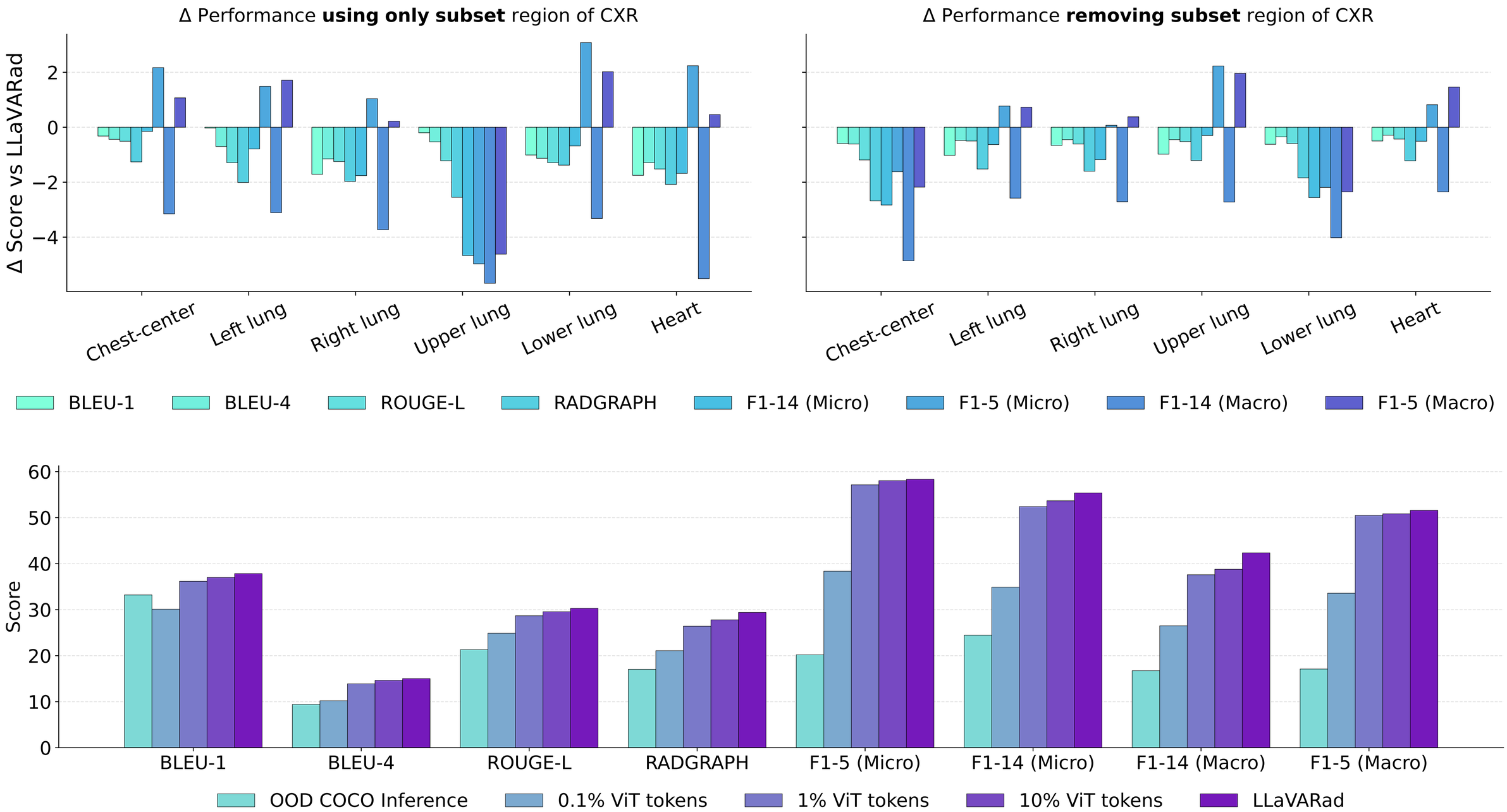

Figure 1: **Visual token interventions reveal weak fine-grained visual dependence. Top:** Removing or isolating anatomical regions degrades performance, especially for CheXbert-based clinical metrics, suggesting at first glance that the model depends on organ-level visual information. **Bottom:** However, this apparent dependence weakens under random token sampling: performance drops sharply when only 0.1% ($T = 1$) of visual tokens are retained, but rapidly recovers as more tokens are sampled, with only 1% ($T = 14$) already approaching the full-token baseline. These results suggest that while the image influences the model, much of this influence may come from a weak global conditioning signal rather than robust, localized clinical evidence.

Benchmarking VLMs for Radiology Report Generation

- State-of-the-art clinical VLMs **can generate plausible clinical reports** on out-of-distribution and corrupted visual inputs.
- This suggests that strong report-generation performance **does not necessarily imply fine-grained visual grounding**.
- The chest X-ray appears to **weakly condition the report**, rather than serving as the primary source of clinical evidence.

Figure 2: **Generation remains clinically plausible under visual and prompt perturbations.** Comparison of LLaVARad generation under i) original CXR and prompt, ii) removal of the `REASON` field, iii) strong CXR corruptions, and iv) replacement of the CXR with a non-clinical image. Although perturbations degrade clinical metrics, generation remain fluent and does not fully collapse, motivating our analysis of how much visual information the model actually uses. ■ Correct ■ Wrong ■ Irrelevant/ambiguous. *(Radiologist annotation.)*

Table 1: **Benchmarking radiology foundation models and vision encoder variants.** Comparison of LLaVARad against different radiology VLMs, and our own LLaVA-style implementation using different vision encoders with Vicuna-7B. Δ row reports the difference between BiomedCLIP-CXR ($T = 1$) and the original LLaVARad baseline ($T = 1{,}369$).

Model	BLEU-1 ↑	ROUGE-L ↑	RadGraph ↑	F1-14 (Macro) ↑	F1-5 (Macro) ↑
Foundation Models					
CheXAgent (Chen et al., 2024)	24.07	23.40	26.84	42.53	**54.34**
MAIRA-2 (Bannur et al., 2024)	31.04	24.15	21.33	35.33	45.69
LLaVA-Med (Li et al., 2023)	21.72	13.30	5.51	18.21	20.41
Xray-GPT (Thawakar et al., 2024)	8.88	16.11	20.53	22.54	35.81
RaDialog (Pellegrini et al., 2023)	29.37	21.58	19.15	17.39	17.35
LLaVARad (Zambrano et al., 2025)	**37.58**	**29.99**	**29.05**	**42.61**	50.49
Vision Encoder + Vicuna 7B					
DINOv2 (OOD) (Oquab et al., 2023)	33.67	28.30	25.11	32.46	46.03
CXR-CLIP ((You et al., 2023)	34.02	28.29	25.42	32.05	45.62
BarlowTwins-CXR (Sheng et al., 2024)	34.77	28.55	25.81	35.52	49.38
ChexZero (Tiu et al., 2022)	33.46	28.18	25.78	34.91	49.38
EVA-X (Yao et al., 2025a)	35.77	29.04	27.01	37.42	50.57
BiomedCLIP-CXR (Zambrano et al., 2025)	36.99	29.64	27.94	40.50	52.24
Δ (BiomedCLIP-CXR vs LLaVARad)	-0.59	-0.35	-1.11	-2.11	+1.75

- Notably, compressing LLaVARad's full visual representation ($T = 1{,}369$) into a single mean-pooled token ($T = 1$) performs comparably to the full-token state-of-the-art baseline.
- Furthermore, under this training setting, retaining only ∼1% of randomly sampled visual tokens ($T = 14$) before mean-pooling nearly matches the full-token model.
- We note that the strong baseline performance under image ablations is driven by **extensive preprocessing of the text inputs** and the dominant role of the LLM in the model pipeline.
- The findings of our study raise concerns that current clinical VLMs **rely only weakly on visual evidence**, despite achieving strong report-generation scores.