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

Anonymous Authors[1]

Abstract

Clinical vision–language models (VLMs) for chest X-ray report generation are typically evaluated on generated text quality, but strong generation performance does not necessarily imply fine-grained visual grounding. In this work, we empirically evaluate how much spatial visual information a clinical VLM uses for radiology report generation. Using our own implementation of LLaVARad, a state-of-the-art VLM for radiology report generation, we apply a simple intervention framework that selects, removes, or randomly samples visual tokens before projection into the language model. We find that compressing or removing large portions of the visual token set produces performance very close to the full-token baseline, with the largest degradation appearing in CheXbert-based clinical metrics. Notably, retaining only $\sim 1\%$ (i.e., T=14) of randomly sampled visual tokens before mean-pooling, nearly matches the full-token setting. These results suggest that the model uses the image primarily through a low-dimensional visual conditioning signal rather than strong fine-grained spatial grounding, raising concerns about the limited use of visual inputs by current clinical VLMs.

1. Introduction

Clinical vision–language models (VLMs) have shown strong performance in chest X-ray (CXR) report generation, producing fluent and clinically plausible findings on benchmarks such as MIMIC-CXR, IU X-Ray, and PadChest (Johnson et al., 2019; Liu et al., 2023; Bustos et al., 2020; Zambrano Chaves et al., 2025; Bannur et al., 2024; Chen et al., 2024). These systems are typically evaluated using language and radiology-based metrics such as BLEU, ROUGE, CheXbert, and RadGraph. However, in clinical settings, plausible language is not sufficient: generated findings must be supported by visual evidence in the image. Strong generation performance does not necessarily imply strong visual grounding (Park et al., 2025; Chen et al., 2025). VLMs may rely on language priors, dataset statistics, and prompt context rather than the input image, leading to hallucinated findings or brittle behavior under image perturbations (Ceron et al., 2024; Cai et al., 2025; Meddeb et al., 2025).

This is especially difficult to detect in radiology report generation, where reports are highly structured, common findings are repeated frequently, and prompts may include clinically meaningful context such as the indication or reason for the exam (Zhang et al., 2025). As a result, it remains unclear whether improvements in report metrics reflect better image understanding or better modeling of report distributions (Liu et al., 2025) (Figure 1). A common misconception is that dense visual tokens provide fine-grained spatial evidence, but access to spatial features does not imply that the LLM really uses them (Fu et al., 2025). This raises a simple question: *how much spatial visual information does a clinical VLM actually need for radiology report generation?*

Our starting point is an unexpected observation: compressing the full 1369-token visual representation into a single mean-pooled token performs similarly to the baseline LLaVARad (Zambrano Chaves et al., 2025). To understand this behavior, we apply a simple causal probing framework that selects, removes, or randomly samples visual patch tokens before projection into the LLM. Unlike attention-based analyses, our approach directly intervenes on the visual information available to the model and does not require attention extraction or external segmentation models. Our results suggest that although the model is not image-independent, dense spatial visual structure is less necessary than expected. Based on these findings, we make the following contributions:

1. An empirical analysis of spatial visual information in VLM-based radiology report generation, examining if and how the image is used for findings generation.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

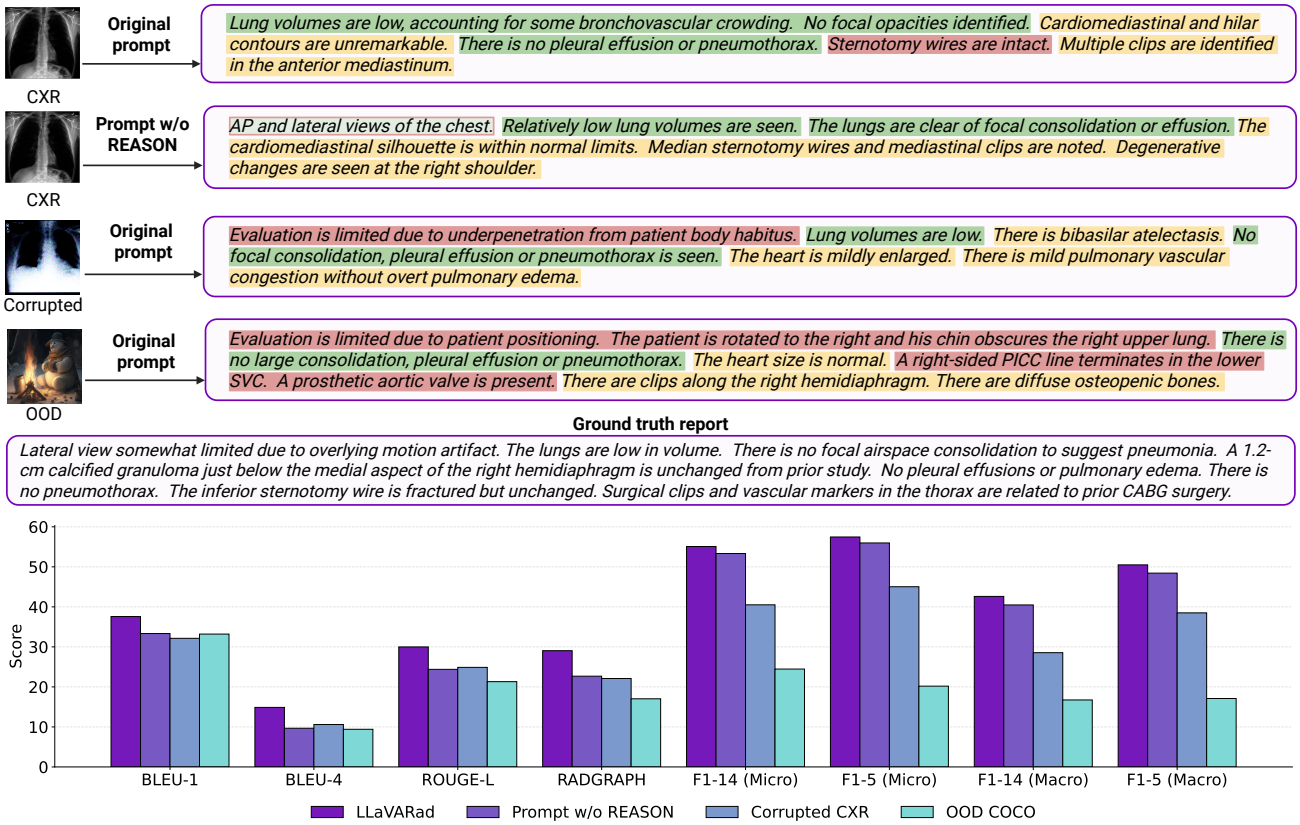

Figure 1. **Generation remains clinically plausible under visual and prompt perturbations.** We compare LLaVARad generation under i) original CXR and prompt, ii) removal of the `REASON` field, iii) strong CXR corruptions, and iv) replacement of the CXR with a non-clinical image. Although perturbations degrade clinical metrics, generation remain fluent and does not fully collapse, motivating a direct analysis of how much visual information the model actually uses. ■ Correct ■ Wrong ■ Irrelevant/ambiguous.

2. We evaluate a LLaVARad implementation across controlled input ablations that modify visual tokens, anatomical regions, random token subsets, image corruptions, and clinical prompt context.

3. We show that region-level interventions cause measurable but modest degradation, while $\sim 1\%$ (T=14) of randomly sampled visual tokens mostly matches the full-token setting, suggesting weak spatial grounding.

2. Related Work

Clinical VLMs for Radiology Report Generation. LLaVARad (Zambrano Chaves et al., 2025), MAIRA-2 (Bannur et al., 2024), and CheXagent (Chen et al., 2024) adapt general-purpose multimodal architectures to radiology by coupling a CXR vision encoder with a large language model. In this setup, patch-level visual features are projected into the LLM embedding space before generation. Although these models achieve strong generation performance, standard metrics do not reval how much spatial visual information is actually used.

Grounding and Hallucination. Recent work has shown that fluent radiology reports can still contain hallucinated or weakly grounded findings. Grounded report generation methods, such as MAIRA-2, address this by pairing generated text with spatial localizations and evaluating sentence-level factuality and grounding (Bannur et al., 2024). Other approaches use external verification or clinically informed evaluation (Yu et al., 2024; Meddeb et al., 2025). However, these methods do not quantify how much visual information is used by the LLM to produce grounded findings.

Relevance of Visual Inputs. Studies show that VLMs can rely on coarse image summaries, overlook accessible visual representations, or underperform vision encoders on vision-centric tasks (Zhang et al., 2024; Kaduri et al., 2025; Fu et al., 2025). In clinical VLMs, this question remains underexplored: existing work has mainly evaluated robustness to medical image artifacts, quantitative medical image analysis, or broad clinical validation, revealing brittleness, weak measurement and detection ability, and limited evaluation beyond text-based metrics (Cheng et al., 2025; Yao et al., 2025b; Dutta et al., 2025).

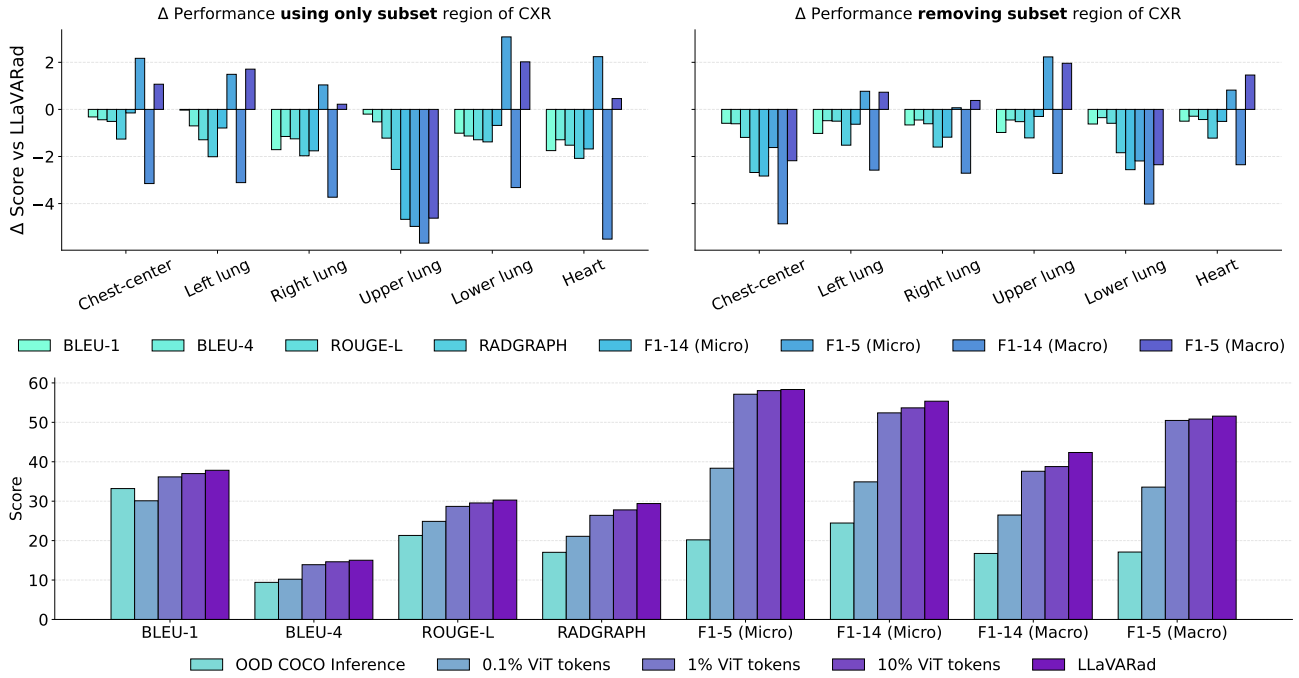

Figure 2. **Visual token interventions reveal weak fine-grained visual dependence. Top:** Removing or isolating specific regions affects performance, especially for CheXbert-based clinical metrics, but the overall drops remain modest relative to the amount of spatial information removed. **Bottom:** Performance drops sharply when only 0.1% (T=1) of visual tokens are retained, but rapidly recovers as more tokens are sampled; retaining 1% (T=14) of tokens already approaches the full-token baseline. Together, these results suggest that the model uses visual information, but primarily through a coarse visual conditioning signal rather than strongly localized spatial evidence.

3. Experimental Setup

Following LLaVARad (Zambrano Chaves et al., 2025), we train our own implementation on the same preprocessed MIMIC-CXR (Johnson et al., 2019) reports, vision encoder (i.e., BiomedCLIP-CXR) and hyperparameters. Full implementation details, including the training configuration and preprocessing steps, are provided in Appendix D.

Mean-pooled Visual Representation. Let $V = \{v_i\}_{i=1}^{T}$ denote the visual patch tokens produced by the encoder, with $T = 1369$ in the full-token representation. Our implementation compresses the image using a single mean-pooled visual token:

$$v_{\text{global}} = \frac{1}{T}\sum_{i=1}^{T} v_i.$$

We compare this setting with the full-token LLaVARad baseline to test whether dense visual token sequences are necessary for maintaining performance.

Spatial Region Interventions. To probe region-level visual relevance, we reshape the visual tokens into their 37×37 spatial grid and define coarse CXR regions in the preprocessed image space (Figure D.2). For a region R, we test *sufficiency* by mean-pooling only tokens inside the region, and *necessity* by removing the region before pooling:

$$v_R = \frac{1}{|R|}\sum_{i \in R} v_i, \qquad v_{\neg R} = \frac{1}{T - |R|}\sum_{i \notin R} v_i.$$

These interventions ask whether a region alone is enough to preserve performance, and whether removing that region causes measurable degradation.

Random Token Sampling. As a spatial control, we uniformly sample retained patches from the full set of 1369 visual tokens and mean-pool only the selected tokens. This tests whether anatomically coherent regions are required, or whether a sparse random sample of visual features is sufficient to approximate the full-image representation. We currently evaluate each sampling rate with a single random seed, which limits our ability to estimate variance across random subsets.

Evaluation. We report standard language metrics, including BLEU-1, BLEU-4, and ROUGE-L, and clinical metrics, including CheXbert F1 variants and RadGraph. Because high report-generation scores do not necessarily imply visual grounding, we focus primarily on performance changes relative to LLaVARad (i.e., original CXR and prompt).

4. Results

4.1. One visual token matches the full-token baseline

Across language and clinical metrics, our compressed model performs similarly to the full-token LLaVARad baseline, with the largest observed drop in CheXbert F1-14 Macro, where performance decreases by 2.11 points (Table A.1). Importantly, this model is trained with a single mean-pooled visual token rather than the full 1369-token visual sequence. Thus, the result does not show that a full-token model is invariant to compression; instead, it shows that when trained on the same preprocessed reports and given the same language inputs as LLaVARad, a single global visual token is sufficient to reach comparable generation performance.

> If a model trained with one pooled visual token can match a dense-token baseline, strong report-generation scores alone cannot be taken as evidence of fine-grained visual grounding.

4.2. Region interventions show weak spatial grounding

We next test whether the compressed model remains sensitive to localized visual evidence. Removing clinically relevant regions such as the chest center or lower lung, or using only limited regions such as the upper lung, produces the largest degradation (Figure 2, Table B.2, Table B.3). This shows that the model is not image-independent: changing the available visual content can hurt performance. However, the drops are modest and do not follow a clear anatomical trends, suggesting that the model does not rely strongly on fine-grained spatial structure.

> Region removal affects performance, but only modestly; the model uses visual information, yet its successful generation is not directly tied to localized anatomical evidence.

4.3. A sparse random sample of visual tokens recovers most performance

Finally test whether spatially coherent regions are required by randomly retaining a fraction of the 1369 visual tokens, mean-pooling the retained tokens, and generating findings from the resulting single visual token. Retaining only 0.1% of tokens (i.e., 1 randomly sampled patch), causes a large drop, especially in CheXbert metrics. However, with only $\sim 1\%$ of tokens (i.e., 14 randomly sampled patches), the model nearly matches the full-token baseline (Figure 2, Table C.4).

The pooled representation is also highly redundant under random subsampling. Retaining 10% of visual tokens yields a cosine similarity of 0.91 with the full-token mean embedding, while retaining 25% or more yields near-identical embeddings (≥ 0.99).

> A small random sample can approximate the pooled visual signal, further indicating that performance is driven by coarse image-level conditioning rather than spatially organized evidence.

5. Discussion

Our results suggest a gap between findings-generation performance and fine-grained visual reliance. The model is not image-independent: region removal, image corruption, and severe token dropout degrade performance, especially on clinical metrics. However, these drops are modest relative to the amount of visual information removed, suggesting that the image may function primarily as a simple conditioning signal while textual context and learned report priors carry much of the generation. This is a safety-relevant failure mode for clinical VLMs, since fluent findings may appear visually grounded even when the model is weakly sensitive to actual clinical evidence from the patient.

Limitations & Future Work Our regions are manually defined in the preprocessed image space and are not anatomical segmentations. Our compressed model is trained with a single visual token, which encourages adaptation to a global visual interface. We evaluate only MIMIC-CXR findings generation and do not test downstream clinical decisions. Future work should develop methods that make clinical VLMs heavily dependent on visual evidence rather than weak image conditioning. Extending this analysis to sentence-level and abnormality-specific generation and training models across multiple visual-token budgets. Clinical VLMs should be optimized not for plausible reports, but for findings that are demonstrably grounded in the image.

6. Conclusion

We studied how much fine-grained spatial visual information a clinical VLM needs for MIMIC-CXR findings generation. Using mean-pooled visual-token compression, region-level interventions, input ablations, and random token sampling, we found that dense visual token grids are not necessary to achieve the same performance as the LLaVARad baseline. The model is not image-independent, but its performance appears to rely mainly on a coarse visual conditioning signal rather than strong fine-grained spatial grounding. These results suggest that clinical VLM evaluation should explicitly test visual reliance, since plausible findings and strong aggregate metrics may persist even when visual evidence is heavily compressed or spatially disrupted.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bannur, S., Bouzid, K., Castro, D. C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Bustos, A., Pertusa, A., Salinas, J.-M., and De La Iglesia-Vaya, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Cai, L., Kang, J., Li, S., Ma, W., Xie, B., Qin, Z., and Liang, J. From local details to global context: Advancing vision-language models with attention-based selection. *arXiv preprint arXiv:2505.13233*, 2025.

Ceron, T., Falk, N., Barić, A., Nikolaev, D., and Padó, S. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms. *Transactions of the Association for Computational Linguistics*, 12:1378–1400, 2024.

Chen, S., Zhang, J., Zhu, T., Liu, W., Gao, S., Xiong, M., Li, M., and He, J. Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*, 2025.

Chen, Z., Varma, M., Xu, J., Paschali, M., Van Veen, D., Johnston, A., Youssef, A., Blankemeier, L., Bluethgen, C., Altmayer, S., et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv e-prints*, pp. arXiv–2401, 2024.

Cheng, Z., Ong, A. Y., Wagner, S. K., Merle, D. A., Ju, L., Zhang, H., Chen, R., Pang, L., Li, B., He, T., et al. Understanding the robustness of vision-language models to medical image artefacts. *NPJ Digital Medicine*, 8(1): 727, 2025.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Dutta, N., Bose, K., Syailendra, E., Chu, L., and Gupta, P. Vision-language models in diagnostic imaging: review of technical advances, clinical validation, and practical deployment. *International Journal of Medical Informatics*, pp. 106227, 2025.

Fu, S., Bonnen, T., Guillory, D., and Darrell, T. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Kaduri, O., Bagon, S., and Dekel, T. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14549–14558, 2025.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

Liu, B., Zou, K., Zhan, L.-M., Lu, Z., Dong, X., Chen, Y., Xie, C., Cao, J., Wu, X.-M., and Fu, H. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21310–21320, 2025.

Liu, W., Xue, Y., Lin, C., and Boumaraf, S. Dynamic multi-domain knowledge networks for chest x-ray report generation. *arXiv preprint arXiv:2310.05119*, 2023.

Meddeb, A., Rangus, I., Pagano, P., Dkhil, I., Jelassi, S., Bressem, K., Scheel, M., Wattjes, M. P., Nagi, S., Pierot, L., et al. Evaluating the diagnostic accuracy of vision language models for neuroradiological image interpretation. *npj Digital Medicine*, 8(1):666, 2025.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Park, S.-J., Heo, K.-S., Shin, D.-H., Son, Y.-H., Oh, J.-H., and Kam, T.-E. Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15580–15589, 2025.

Pellegrini, C., Özsoy, E., Busam, B., Navab, N., and Keicher, M. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023.

Sheng, H., Ma, L., Samson, J.-F., and Liu, D. Barlowtwins-cxr: enhancing chest x-ray abnormality localization in heterogeneous data with cross-domain self-supervised learning. *BMC Medical Informatics and Decision Making*, 24(1):126, 2024.

Thawakar, O. C., Shaker, A. M., Mullappilly, S. S., Cholakkal, H., Anwer, R. M., Khan, S., Laaksonen, J., and Khan, F. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pp. 440–448, 2024.

Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., and Rajpurkar, P. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. In *International Conference on*, 2021.

Yao, J., Wang, X., Song, Y., Zhao, H., Ma, J., Chen, Y., Liu, W., and Wang, B. Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning. *npj Digital Medicine*, 8(1):678, 2025a.

Yao, Y., Zong, Y., Dutt, R., Yang, Y., Tsaftaris, S. A., and Hospedales, T. Medvision: Dataset and benchmark for quantitative medical image analysis. *arXiv preprint arXiv:2511.18676*, 2025b.

You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E. K., Baek, W., and Roh, B. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 101–111. Springer, 2023.

Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

Zambrano Chaves, J. M., Huang, S.-C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

Zhang, M., Shen, Y., Li, Z., Sha, H., Hu, B., Wang, Y., Huang, C., Liu, S., Tong, J., Jiang, C., et al. Llmeval-med: A real-world clinical benchmark for medical llms with physician validation. *arXiv preprint arXiv:2506.04078*, 2025.

Zhang, Y., Unell, A., Wang, X., Ghosh, D., Su, Y., Schmidt, L., and Yeung-Levy, S. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024.

Appendix

Table of Contents

A Benchmarking Foundation Models and Vision Encoder Variants 8

B Detailed Region Sufficiency and Necessity Results 9

C Extended Random Token Sampling Results 10

D Experimental Details 11

Reproducibility

We support reproducibility by providing detailed descriptions of the experimental setup, model configuration, preprocessing steps, and sampling parameters in Appendix D. All datasets used in our experiments are publicly available, and our implementation follows the LLaVARad preprocessing and training setup as closely as possible. We do not release code or checkpoints with this workshop version, but plan to do so in a future full-paper release..

A. Benchmarking Foundation Models and Vision Encoder Variants

We include additional benchmark results to contextualize the LLaVARad baseline used throughout the main paper. First, we compare LLaVARad against several radiology VLMs for report generation. Second, we evaluate our LLaVARad implementation with different vision encoders while keeping the downstream pipeline fixed. The results show that LLaVARad is a strong reference model, and that our implementation remains close to the original baseline across several encoder choices.

Table A.1. **Benchmarking radiology foundation models and vision encoder variants.** We compare LLaVARad against several radiology VLM baselines, and compare our LLaVARad-style implementation using different vision encoders with Vicuna-7B. The bottom row reports the difference between our BiomedCLIP-CXR implementation and the original LLaVARad baseline.

Model	BLEU-1	BLEU-4	ROUGE-L	RADGRAPH	F1-14 (Micro)	F1-5 (Micro)	F1-14 (Macro)	F1-5 (Macro)
Foundation Models								
CheXAgent (Chen et al., 2024)	24.07	5.97	23.40	26.84	**57.79**	**62.07**	**<u>42.53</u>**	**54.34**
MAIRA-2 (Bannur et al., 2024)	31.04	8.89	24.15	21.33	50.14	54.07	35.33	45.69
LLaVA-Med (Li et al., 2023)	21.72	0.92	13.30	5.51	26.09	25.04	18.21	20.41
Xray-GPT (Thawakar et al., 2024)	8.88	1.58	16.11	20.53	36.66	42.85	22.54	35.81
RaDialog (Pellegrini et al., 2023)	29.37	7.70	21.58	19.15	22.88	19.21	17.39	17.35
LLaVARad (Zambrano Chaves et al., 2025)	**37.58**	**14.88**	**29.99**	**29.05**	**<u>55.08</u>**	57.46	**42.61**	50.49
Vision Encoder + Vicuna 7b								
DINOv2 (OOD vision encoder) (Oquab et al., 2023)	33.67	13.23	28.30	25.11	47.41	53.61	32.46	46.03
CXR-CLIP (You et al., 2023)	34.02	13.33	28.29	25.42	47.66	53.87	32.05	45.62
BarlowTwins-CXR (Sheng et al., 2024)	34.77	13.50	28.55	25.81	50.55	56.39	35.52	49.38
ChexZero (Tiu et al., 2022)	33.46	12.95	28.18	25.78	50.26	56.87	34.91	49.38
EVA-X (Yao et al., 2025a)	35.77	13.90	29.04	27.01	52.95	57.99	37.42	50.57
BiomedCLIP-CXR (Zambrano Chaves et al., 2025)	**<u>36.99</u>**	**<u>14.59</u>**	**<u>29.64</u>**	**<u>27.94</u>**	54.96	**<u>59.12</u>**	40.50	52.24
Δ (BiomedCLIP-CXR (T=1) vs LLaVARad (T=1369))	-0.59	-0.29	-0.35	-1.11	-0.11	+1.66	-2.11	+1.75

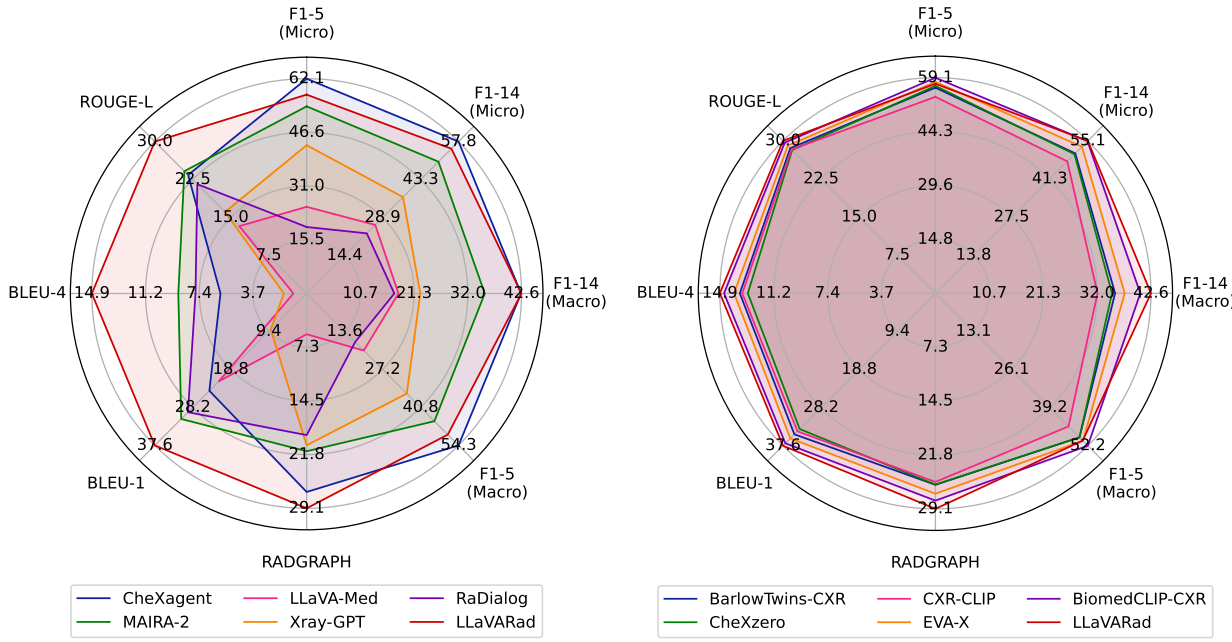

Figure A.1. **Benchmarking LLaVARad and vision encoder variants. Left:** Comparison of LLaVARad against radiology VLM baselines for report generation. LLaVARad provides a strong reference point across the evaluated metrics. **Right:** Comparison between LLaVARad and our implementation using different vision encoders. Despite changing the vision encoder, performance remains close to LLaVARad across metrics, suggesting that downstream performance is relatively insensitive to the specific encoder choice in this setup.

B. Detailed Region Sufficiency and Necessity Results

This section provides the full region-intervention results corresponding to the analysis in Section 4.2. We report the change in performance relative to the LLaVARad baseline for both region sufficiency and region necessity settings. Cells highlighted in red indicate drops larger than 2 points, while green indicates improvements. Overall, most changes remain small despite removing or isolating large portions of the visual token grid, suggesting that the model's performance is only weakly sensitive to spatially localized visual information.

One consistent pattern is that CheXbert F1-14 Macro is the most affected metric across interventions. While this may reflect loss of clinically relevant visual information, it could also arise from an interaction between rare-label report language, CheXbert label extraction, and the coarse visual conditioning signal. Future work should therefore analyze abnormality-level outputs and label-specific errors to determine whether these drops correspond to genuine visual failures or metric-specific sensitivity.

Table B.2. **Region sufficiency results.** Performance change when findings are generated using only tokens from the selected region. Delta values are computed relative to the original full-token LLaVARad baseline. Green cells indicate improvements, and red cells indicate drops larger than 2 points.

Region	Vision Tokens	Δ BLEU-1	Δ BLEU-4	Δ ROUGE-L	Δ RADGRAPH	Δ F1-14 (Micro)	Δ F1-5 (Micro)	Δ F1-14 (Macro)	Δ F1-5 (Macro)
Chest-center	702	-0.32	-0.44	-0.51	-1.26	-0.15	+2.17	-3.15	+1.07
Left-lung	338	-0.03	-0.70	-1.29	-2.01	-0.79	+1.49	-3.11	+1.71
Right-lung	364	-1.71	-1.15	-1.25	-1.97	-1.76	+1.04	-3.73	+0.22
Upper-lung	351	-0.20	-0.53	-1.22	-2.55	-4.67	-4.97	-5.68	-4.62
Lower-lung	351	-1.01	-1.13	-1.29	-1.38	-0.68	+3.08	-3.32	+2.02
Heart	225	-1.75	-1.29	-1.52	-2.08	-1.68	+2.24	-5.51	+0.46
Upper	666	-0.70	-0.52	-0.95	-2.56	-3.00	-2.05	-5.02	-2.74
Lower	703	-1.16	-0.76	-1.04	-1.88	-1.16	+2.06	-3.70	+1.39
Left	666	-0.36	-0.47	-0.86	-1.92	-1.04	+0.46	-3.13	+0.98
Right	703	-1.46	-0.71	-0.56	-1.79	-1.22	+1.05	-3.12	+0.71

Table B.3. **Region necessity results.** Performance change when tokens from the selected region are removed before mean pooling. Delta values are computed relative to the original full-token LLaVARad baseline. Green cells indicate improvements, and red cells indicate drops larger than 2 points.

Region	Vision Tokens	Δ BLEU-1	Δ BLEU-4	Δ ROUGE-L	Δ RADGRAPH	Δ F1-14 (Micro)	Δ F1-5 (Micro)	Δ F1-14 (Macro)	Δ F1-5 (Macro)
Chest-center	667	-0.59	-0.61	-1.19	-2.68	-2.83	-1.62	-4.86	-2.18
Left-lung	1031	-1.02	-0.48	-0.50	-1.52	-0.63	+0.77	-2.58	+0.73
Right-lung	1005	-0.66	-0.45	-0.61	-1.60	-1.18	+0.07	-2.71	+0.38
Upper-lung	1018	-0.98	-0.45	-0.52	-1.21	-0.30	+2.23	-2.72	+1.96
Lower-lung	1018	-0.62	-0.35	-0.59	-1.84	-2.56	-2.19	-4.02	-2.35
Heart	1144	-0.50	-0.29	-0.43	-1.22	-0.51	+0.82	-2.35	+1.46
Upper	703	-1.41	-0.92	-1.08	-1.82	-1.22	+2.21	-3.76	+1.44
Lower	666	-0.70	-0.54	-0.97	-2.48	-2.99	-2.22	-5.04	-3.00
Left	703	-1.51	-0.76	-0.68	-1.94	-1.81	+0.50	-3.39	+0.42
Right	666	-0.43	-0.46	-0.82	-1.80	-1.03	+0.42	-3.44	+0.92

C. Extended Random Token Sampling Results

This section provides the full random token sampling results. We randomly retain a subset of visual tokens from the full 1369-token grid, mean-pool the retained tokens, and compare performance against the original full-token LLaVARad baseline. Red cells indicate drops larger than 2 points, while green cells indicate improvements. Overall, performance recovers rapidly as more random tokens are retained, suggesting that the pooled visual representation is highly redundant under random subsampling.

Table C.4. **Random token sampling results.** Performance change when only a random subset of visual tokens is retained before mean pooling. Delta values are computed relative to the original full-token LLaVARad baseline. Green cells indicate improvements, and red cells indicate drops larger than 2 points.

Retained Ratio	Vision Tokens	Δ BLEU-1	Δ BLEU-4	Δ ROUGE-L	Δ RADGRAPH	Δ F1-14 (Micro)	Δ F1-5 (Micro)	Δ F1-14 (Macro)	Δ F1-5 (Macro)
0.10%	1	-7.48	-4.67	-5.11	-7.96	-20.18	-19.09	-16.13	-16.92
0.20%	3	-4.10	-3.19	-4.17	-6.24	-13.29	-11.57	-11.63	-10.83
0.30%	4	-2.24	-2.02	-3.03	-4.42	-6.35	-4.17	-7.58	-4.29
0.40%	5	-1.71	-1.70	-2.54	-3.97	-6.30	-4.13	-8.25	-4.81
0.50%	7	-1.64	-1.45	-2.32	-3.55	-5.32	-3.30	-6.85	-3.67
0.60%	8	-1.19	-1.10	-1.81	-3.16	-4.28	-1.92	-5.73	-2.40
0.70%	10	-1.60	-1.20	-1.74	-3.21	-4.26	-2.56	-5.33	-3.18
0.80%	11	-1.22	-1.01	-1.62	-2.99	-3.81	-2.56	-5.31	-2.27
0.90%	12	-1.46	-0.90	-1.38	-2.76	-3.47	-1.32	-5.21	-1.64
1%	14	-1.42	-0.98	-1.31	-2.65	-2.68	-0.32	-5.03	-0.01
5%	68	-0.91	-0.50	-0.56	-1.55	-1.63	+0.59	-4.59	-0.25
10%	137	-0.58	-0.25	-0.43	-1.27	-1.40	+0.58	-3.83	+0.34
25%	342	-0.58	-0.37	-0.42	-1.18	-0.43	+1.86	-2.60	+2.14
50%	685	-0.63	-0.24	-0.24	-0.93	-0.41	+1.58	-2.74	+1.41
75%	1027	-0.63	-0.38	-0.34	-1.02	-0.24	+1.21	-2.16	+1.47
90%	1232	-0.62	-0.33	-0.29	-1.10	-0.33	+1.23	-1.93	+1.27

D. Experimental Details

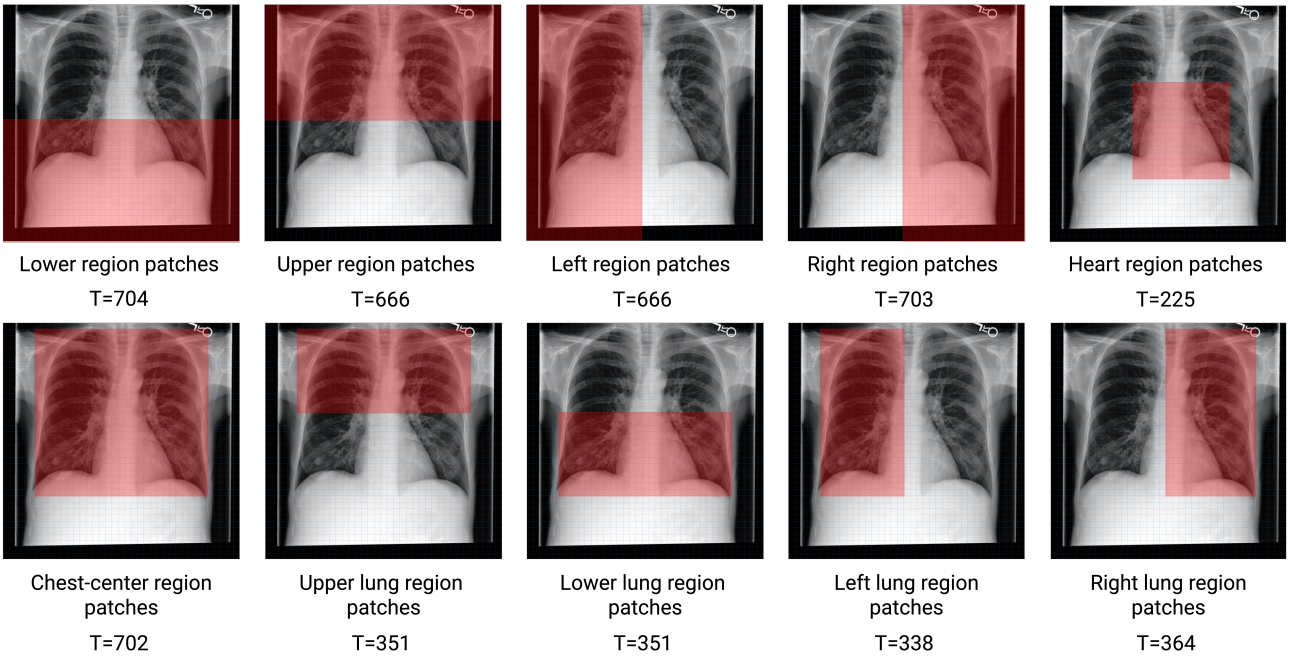

Figure D.2. **CXR region definitions.** Coarse regions are defined over the preprocessed 37×37 visual token grid. T denotes the number of visual tokens contained in each region and used for mean pooling.

LLaVA Training Pipeline. LLaVA-style VLMs are typically trained in stages. First, a vision encoder is initialized from a model pretrained on image–text data. Second, the vision encoder and LLM are frozen while a projection module is trained to map visual features into the LLM embedding space. This stage aligns visual representations with the language model input space. Third, the vision encoder remains frozen while the projector and LLM are fine-tuned for downstream instruction-following or report generation.

For our implementation, we follow the LLaVARad training setup as closely as possible. We use the LLaVARad preprocessed MIMIC-CXR reports, a vanilla Vicuna-7B language model (Chiang et al., 2023), an `mlp2x_gelu` projection layer, and LoRA fine-tuning (Wallis et al., 2021). For the vision encoder, we initialize from the pretrained BiomedCLIP-CXR vision encoder weights used by LLaVARad. Unless otherwise specified, all hyperparameters are the same as LLaVARad, the vision encoder is frozen and all visual token interventions are applied before the projector.

Training Compute. The original LLaVARad pipeline reports training the vision encoder, alignment stage, and fine-tuning stage on a cluster of 8 A100 GPUs, requiring approximately 8 hours, 4 hours, and 16 hours, respectively. In contrast, our experiments train only the alignment and fine-tuning stages with compressed visual tokens. We train in approximately 15 hours using 2 GPUs, or approximately 7.5 hours using 4 A100 GPUs, while achieving the same performance as LLaVARad. This reduction in compute is not the main goal of our work, but it supports our central observation that dense visual token grids are not strictly necessary for this model and task.

Visual Token Compression. In the original full-token setting, the vision encoder produces $T = 1369$ patch embeddings. In our compressed setting, these patch embeddings are mean-pooled before the projection layer, producing a single visual token. Region sufficiency, region necessity, and random token sampling experiments use the same mechanism: visual tokens are first selected, removed, or sampled, and the remaining tokens are then mean-pooled before projection into the LLM.

Image Preprocessing. All CXR images are converted to three-channel grayscale, resized, center-cropped, converted to tensors, and normalized using ImageNet statistics. Region definitions and overlays are computed in this preprocessed image space.

Region Definitions. Visual tokens are reshaped into a 37×37 grid. We manually define approximate CXR regions, including the lungs, chest center, upper lung, lower lung, upper CXR, and lower CXR. These regions are intentionally coarse probes rather than anatomical segmentations.

Random Token Sampling. For random token sampling, we uniformly sample visual tokens from the full set of 1369 patch embeddings before mean-pooling. We evaluate retention ratios of $[0.1\%, 1\%]$, $5\%, 10\%, 25\%, 50\%, 75\%$, and 90%. Results are currently reported for a fixed random seed.

Image Corruption. Strong CXR corruptions combine brightness and contrast shifts, Gaussian blur, sharpness degradation, Gaussian noise, salt-and-pepper noise, motion blur, and random erasing. These perturbations are designed to degrade visual information while preserving the broad appearance of a CXR.