MedVLM-Report: A Unified Vision–Language Framework for Automated Generation of X-Ray and Ultrasound Reports

***Abstract*—Manual radiology reporting is time-intensive and inconsistent across clinicians, and most automated report-generation systems address only a single imaging modality while treating explainability as optional. Our central contribution is a novel *detect-then-describe* conditioning mechanism that explicitly injects Stage-1 detector findings into the language decoder's input and constrains generation through a faithfulness-aware objective, rather than relying on implicit visual grounding alone. We instantiate this mechanism in MedVLM-Report, a unified vision–language system that generates reports for both chest X-ray and breast ultrasound from one shared decoder and attaches a four-part explainability package to every report, normal or abnormal: a modality-specific Stage-1 detector (a five-network DenseNet ensemble for X-ray; a weakly-supervised ResNet-18 classifier for ultrasound) runs first, and its findings are injected as an explicit text clause into a LoRA-adapted BioGPT decoder's prompt, together with a faithfulness-weighted loss that forces the decoder to verbalise them. Across a 14-version ablation on 1,703 held-out studies (drawn from 141,747 X-ray and 2,458 ultrasound training studies), detect-then-describe conditioning and faithfulness weighting are the dominant drivers of clinical accuracy; a five-network detector ensemble (mean AUC 0.704) reduces missed diagnoses to 39.8% at a clinical F1 of 0.451; and – as a notable negative result – neither precision-tuned detector thresholds nor a stronger, domain-pretrained vision backbone (RAD-DINO) reduce hallucination or improve clinical accuracy, indicating that the Stage-1 detector's recall, not the language decoder or the vision encoder, is the binding constraint on this system. At this accuracy level, we position the system as a research-stage step toward reliable automated reporting – a decision-support signal to be reviewed by a clinician, not yet a deployment-ready replacement for radiologist review.**

***Index Terms*—radiology report generation, vision-language models, explainable artificial intelligence, chest X-ray, breast ultrasound, multimodal learning, Q-Former, low-rank adaptation, detect-then-describe**

I. Introduction

X-ray and ultrasound are the two most common diagnostic imaging modalities in clinical practice. X-ray is standardised and supported by large paired image–report datasets, while ultrasound is real-time but operator-, device-, and viewpoint-dependent, with far fewer paired datasets available [1], [2]. As imaging volumes grow, manual reporting remains a bottleneck: time-intensive, cognitively demanding, and inconsistent across clinicians. Vision–language models (VLMs) that link image evidence to clinical language have been proposed to ease this workload [1], [7], but three limitations recur in the literature: (i) *modality separation* – most systems target a single modality and transfer poorly to the noisier ultrasound domain [2], [9]; (ii) *unreliable transparency* – explainability is often applied only to abnormal cases or as a single global heatmap [3]–[5]; and (iii) *evaluation that rewards fluency over correctness*, since BLEU/ROUGE/CIDEr do not guarantee that named findings are actually present [10].

This work originated from a proposal for a single end-to-end system in which a lesion-detection front end would crop regions of interest and feed one shared pipeline for both modalities. Across fourteen development iterations, this design plateaued in clinical accuracy: region-restricted training measurably *hurt* performance, since many clinically relevant findings (support devices, cardiomediastinal, osseous) lie outside any single detected region. The system reported here therefore separates detection and description into two explicit stages – *detect-then-describe* – with a loss term that keeps the decoder faithful to what the detector found; we report this as a deliberate, ablation-justified redesign rather than a deviation to be hidden.

Contributions: (1) a unified X-ray/ultrasound VLM – a Swin backbone, a BLIP-2-style Q-Former, and a LoRA-adapted BioGPT decoder sharing one language model across modalities; (2) our central contribution, *detect-then-describe* conditioning: a novel mechanism in which a Stage-1 finding detector's output is injected as an explicit text clause into the decoder prompt – rather than left as an implicit, unsupervised visual signal – together with a faithfulness-weighted language-modelling loss that constrains the decoder to verbalise what the detector found; (3) a mandatory four-technique explainability package (attention rollout, per-finding Grad-CAM, sentence-to-region grounding, token/sentence rationale) computed for every report; and (4) a systematic 14-version ablation reporting negative results alongside positive ones – region restriction, threshold-based hallucination control, and a stronger vision backbone all fail to improve on the best configuration.

II. Related Work

Early report generators paired CNN encoders with sequential decoders and struggled to keep long clinical reports coherent [6]. Mohsan *et al.* combine a transformer image encoder with clinical text generation but evaluate only on X-ray, with no guaranteed explanation coverage [1]; Wang *et*

al. (R2GenGPT) show that a frozen LLM with a lightweight visual-alignment module produces fluent reports, while noting that fluency alone does not guarantee grounding [7] – a concern this paper addresses directly via the faithfulness loss. Goswami *et al.* propose a unified vision–language architecture that learns shared representations across modalities [8], while Yang *et al.* and Yan *et al.* improve ultrasound-specific reporting with adaptive attention or CLIP-style encoders [2], [9], but neither delivers one system spanning both modalities with consistent per-sentence grounding. On explainability, Chaddad *et al.* and Ras *et al.* argue that explanations must be faithful to model behaviour, not an afterthought [3], [4]; Suara *et al.* show that Grad-CAM heatmaps do not always localise clinically meaningful evidence [5]; and Reyes *et al.* recommend pairing visual maps with semantic explanation [10]. Region-of-interest emphasis has separately helped ultrasound segmentation [11] and lightweight lesion detection [12], but neither targets report generation – and, as Section VII shows, region restriction that helps segmentation-style tasks actively *hurts* report generation, since reportable findings are not confined to one region.

Beyond these radiology-specific systems, several widely-cited biomedical VLMs are frequently raised as comparison points, though none target the same task/modality combination evaluated here. LLaVA-Med [25] and Med-Flamingo [26] are instruction-tuned and few-shot conversational assistants evaluated on biomedical visual question answering, not free-text report generation. BioViL [28] is a contrastive vision–language pretraining method for chest X-ray representation learning and phrase grounding, not a report generator. M3D [29] targets 3D volumetric modalities (CT/MRI), a different imaging setting entirely. MiniGPT-Med [27] is the closest in task – radiology report generation – but is evaluated on a single-modality X-ray benchmark using BERT-Sim/CheXbert-Sim, metrics not computed by any of the other systems above or by this work. Because no shared test set, modality pairing, or metric protocol exists across these five systems, R2GenGPT, and ours, we do not report a numeric head-to-head comparison table – doing so would imply a controlled comparison the current evaluation landscape does not support. We treat this absence of a common unified X-ray-plus-ultrasound report-generation benchmark as a gap for the field, not one we attempt to paper over post hoc (Section VII).

III. System Architecture

Fig. 1 shows the pipeline: a visual-representation branch (backbone $\rightarrow$ modality adapter $\rightarrow$ Q-Former) and a detect-then-describe branch (Stage-1 detector $\rightarrow$ text clause) are fused before decoding, while a frozen text encoder and contrastive grounding head support both training and post-hoc explainability.

A. Vision–Language Backbone

The default backbone is a frozen Swin-Base transformer [15] (224px, 1024-d), optionally LoRA-adapted [14], with a lightweight per-modality bottleneck adapter (dim 128) so X-ray and ultrasound share most visual parameters. Two alternative backbones (BiomedCLIP, RAD-DINO [22]) are supported behind the same interface; RAD-DINO is evaluated as the v14 ablation (Section VI). Up to three views per study are encoded independently and mean-pooled. A BLIP-2-style Q-Former [13] (16 learnable queries, 4 layers of self-attention plus cross-attention to image tokens – the HPO-selected size, Section III-D) converts the image-token sequence into a fixed-length representation; earlier iterations drove this cross-attention with ground-truth text, which silently became a no-op at inference, so the final design conditions on image tokens only. Query outputs are linearly projected into BioGPT's [16] hidden size, rescaled by its embedding scale, and BioGPT itself is LoRA-adapted (rank 8) to keep the trainable footprint small enough for bf16 fine-tuning.

B. Detect-then-Describe Conditioning

A Stage-1 detector runs *before* generation and its output is injected as an explicit clause into the decoder's prompt (e.g. "An automated detector flags: cardiomegaly, atelectasis"), tokenised through the decoder's normal embedding path rather than fused as a hidden vision feature; a weaker secondary channel also prepends one soft-conditioning token built from the detector's multi-hot vector. For X-ray, the detector is an ensemble of five off-the-shelf DenseNet-121 checkpoints (`torchxrayvision` [17]), independently pretrained on five chest-radiograph corpora (an all-sources aggregate, MIMIC-CXR [19], CheXpert [18], NIH ChestX-ray14, PadChest), combined by simple averaging across models that report each pathology; per-pathology thresholds are tuned by F-β grid search against a rule-based reference labelling (β=1, i.e. F1, for the best configuration). This five-network composition was itself chosen empirically, not by default availability: a sixth candidate, a higher-resolution ResNet-50 checkpoint trained on non-MIMIC sources, was validated and *excluded* after it lowered ensemble mean AUC (0.65 vs. 0.704 for the retained five). Spot-checked per-pathology AUCs for the retained ensemble on the MIMIC validation split include edema (0.81), effusion (0.81), pneumothorax (0.82), and cardiomegaly (0.73). For ultrasound, the detector is a ResNet-18 lesion classifier trained under weak, regex-derived supervision from report text, with a recall-favouring threshold (β=1.5), appropriate for a screening use case where missing a lesion is costlier than a reviewable false alarm.

C. Faithfulness Loss and Contrastive Grounding

Injecting findings into the prompt does not guarantee the decoder will mention them, so the language-modelling loss up-weights the cross-entropy of tokens belonging to a fixed finding-naming lexicon (e.g. *cardiomegaly, effusion, opacity*) by a faithfulness weight λ_f, increased from 4.0 (first introduced with detect-then-describe) to 5.0 across the best-performing versions. A grounding head separately projects pooled image and text (frozen Bio_ClinicalBERT [21]) representations into a shared 512-d space with a symmetric InfoNCE objective, added to training with a small weight (≈ 0.2–0.28); at inference, the same space is reused, without

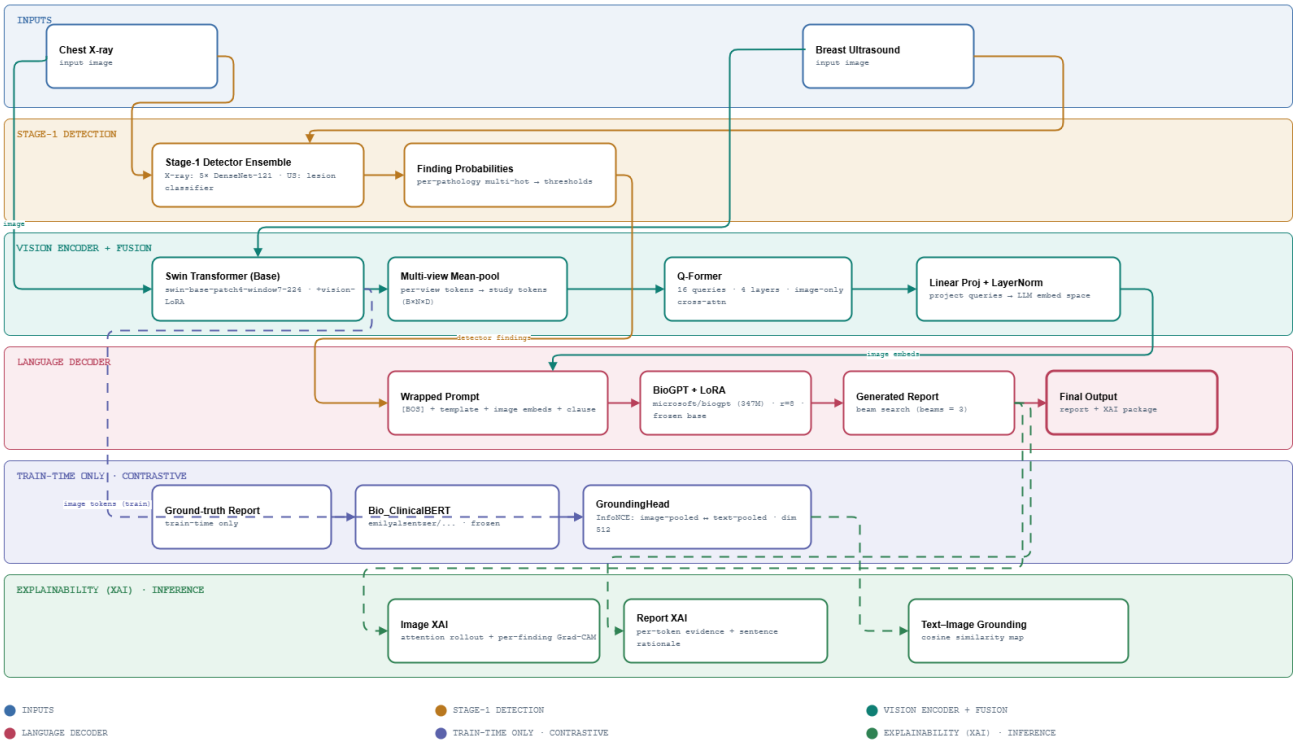

Fig. 1. MedVLM-Report v12 pipeline, by lane: *Inputs* (chest X-ray, breast ultrasound); *Stage-1 Detection*, a modality-specific detector ensemble producing per-pathology finding probabilities; *Vision Encoder + Fusion*, a shared Swin-Base backbone with a per-modality adapter, mean-pooled across views and fed through a 16-query, 4-layer Q-Former; *Language Decoder*, where detector findings are wrapped into the prompt as an explicit text clause alongside the projected image embeddings and decoded by LoRA-adapted BioGPT; *Train-time-only Contrastive* alignment between pooled image and text (Bio_ClinicalBERT) representations via a InfoNCE grounding head; and *Explainability (XAI)*, computed at inference from the same Q-Former attention, detector, and grounding-head signals. No lesion-cropping or region-of-interest stage precedes the vision encoder – the detector conditions the decoder through text, not through the image.

ever re-running cross-attention on generated text, to compute sentence-to-image-region similarity for explainability.

D. Training and Explainability

The model is trained with AdamW, a cosine schedule with warmup, bf16 mixed precision, and early stopping. Fourteen hyperparameters (learning rate, LoRA rank/α, Q-Former size, loss weights, etc.) are searched with Optuna's Tree-structured Parzen Estimator sampler [23] on proxy subsets; the best configuration is then retrained on the full corpus, with checkpoint selection weighting clinical micro-F1 alongside BLEU-4. Four explainability techniques are computed for *every* report: (1) Q-Former attention rollout, averaged across layers and summed over queries into an image-level saliency map; (2) per-finding Grad-CAM [20] on an intermediate backbone feature map, one map per detected finding; (3) sentence-to-region grounding, matching each generated sentence to image regions via the contrastive embedding space; and (4) token/sentence rationale scoring, surfacing the top-3 highest-scoring sentences. All four, plus a lung ROI overlay, are served through a FastAPI web application and exportable as a PDF carrying a fixed clinical-use disclaimer.

IV. Datasets

X-ray data comes from MIMIC-CXR (141,747 training studies, native splits); ultrasound from a breast (mammary) corpus of 2,458 studies, capped to 1,697 training studies after duplicate-template capping to curb memorisation. The ultrasound corpus's size is constrained less by patient volume than by annotation throughput: high daily ultrasound caseloads left sonographers and radiologists little time to save images and record findings during the clinical workflow, and, unlike the X-ray acquisition pipeline, most of the source ultrasound machines had no automatic image-save mode, so retrospective collection depended on manual, per-case capture – a site-infrastructure constraint rather than a modelling choice. All test-set results are computed on a shared, held-out split of 1,703 studies across both modalities (1,000 X-ray, 703 ultrasound); the ultrasound test split is proportionally large relative to its training set even though the training set itself is small. The two modalities are combined with a modality-balanced weighted sampler (weight $\propto (1/N_{\text{modality}})^{0.5}$), so the two-orders-of-magnitude size gap between corpora does not starve ultrasound of training signal. X-ray images receive CLAHE normalisation and small affine augmentation with no flip (laterality-preserving); ultrasound receives despeckling and log compression and allows flips but no rotation (fixed probe orientation).

TABLE I
DEPLOYED MODEL (V12) EFFICIENCY, MEASURED ON AN IDLE RTX 5090; SINGLE IMAGE, BEAM SIZE 3, MAX 160 NEW TOKENS.

Metric	Value
Total parameters	584.2 M
Trainable parameters	42.4 M (7.3%)
GPU memory (weights)	2.38 GB
GPU memory (peak, inference)	2.47 GB
Latency / report	0.27 s
Throughput	3.66 reports/s

V. EXPERIMENTAL SETUP

Text-generation quality is measured with BLEU-1–4, METEOR, ROUGE-L, and CIDEr [24] (COCO caption-evaluation toolkit). Clinical correctness is measured with a negation-aware, rule-based CheXpert-style labeller over 13 findings, reporting micro precision/recall/F1 (Clin-P/R/F1) plus two case-level safety rates: *Missed%* (abnormal cases naming none of the true findings) and *Halluc%* (normal cases naming a finding that is not present). We treat this labeller as an approximation to a trained clinical labeller (e.g. CheXbert); it is applied consistently across all versions, so relative comparisons remain valid. Training and evaluation used a single NVIDIA RTX 5090 GPU with bf16 mixed precision under PyTorch Lightning.

Efficiency. Table I reports the deployed model's (v12) resource footprint, measured on an idle RTX 5090 (single image, beam size 3, max 160 new tokens). Of the 584.2 M total parameters, 88.0 M belong to the vision encoder (Swin-Base + adapter), 108.3 M to the frozen text encoder (Bio_ClinicalBERT), 38.6 M to the Q-Former, 0.9 M to the grounding head, 347.5 M to the BioGPT decoder, and 0.8 M to the projection/conditioning heads; only 42.4 M (7.3%) are trainable, since LoRA adapts just 0.8 M of the decoder's 347.5 M parameters. At this size the model fits on a single consumer-grade GPU (2.47 GB peak) and generates a report in 0.27 s. For context, swapping in a larger decoder – BioGPT-Large (1.5 B) or a 4-bit-quantised BioMistral-7B – raises peak memory to 7.63 GB and 30.11 GB and latency to 0.64 s and 1.69 s respectively, at a similarly small trainable fraction; this cost is reported for scale only; it is not an accuracy comparison, since decoder scaling does not target the detector-recall bottleneck identified in Section VII.

VI. RESULTS

Table II reports all 14 comparable versions (v1–v2 used an earlier, non-comparable test-cleaning and are excluded). The best configuration (v12) combines the backbone above with the 5-network detector ensemble (mean AUC 0.704) and F1-tuned thresholds, reaching Clin-F1 0.451, Clin-R 0.411 (best of all versions), and Missed% 39.8% (tied-lowest), at BLEU-4 0.277 and CIDEr 1.088. Two trade-offs are visible beyond the headline clinical metrics. First, **the best text-fluency model is not the best clinically-accurate one**: v11 has the highest BLEU-1–4/METEOR/ROUGE-L, yet trades a little BLEU for v12's better clinical recall and lower missed-diagnosis rate. Second, **CIDEr peaks at v10** (1.255), whose recall-objective checkpoint selection favours more varied phrasing despite a worse missed-diagnosis rate (42.7% vs. 39.8%) – reinforcing why this system anchors model selection on the clinical labeller rather than NLG scores. For reference, the excluded v1 reached a deceptively high CIDEr of 1.507 by reproducing generic prior-study boilerplate, at a 67.1% missed-diagnosis rate.

Fig. 2 shows the full explainability output for two representative v12 test cases, with generated and reference report text shown side by side (shared phrasing highlighted). In the X-ray case, the model correctly reports an enlarged cardiac silhouette and correctly states no pleural effusion, matching the reference's "heart is moderately enlarged" and "no focal infiltrate or effusion," though it omits the reference's mild pulmonary vascular redistribution; the region-of-interest overlay, Q-Former attention map, and four per-finding Grad-CAM maps (with detector confidence and % overlap with the lung mask) show the model attending to the lung fields for every candidate finding regardless of whether it is ultimately asserted in the text. In the ultrasound case, the generated report reproduces the reference's breast-gland description almost verbatim but under-reports the reference's axillary lymph-node finding, illustrating that recall gaps persist even on an otherwise well-matched report; the attention map concentrates on the central glandular tissue. This is one of two qualitative cases; XAI is computed for every report, normal or abnormal.

VII. DISCUSSION

Detect-then-describe dominates. The largest single jump in clinical recall occurs at v8, when detector-conditioning and the faithfulness loss are introduced together (Clin-R 0.283→0.382), far exceeding gains from architecture search (v7) or contrastive grounding (v6) alone – the binding constraint on earlier versions was the absence of an explicit signal forcing the decoder to commit to findings, not representational capacity.

Missed-diagnosis is detector-recall-bound. Ensembling one to five detector networks (v8→v11→v12) steadily lowers Missed% (46.0→41.4→39.8) while the ensemble's own mean AUC stays modest (0.704); because the decoder is trained to faithfully relay the detector, the ∼40% missed-diagnosis floor across v10–v14 reflects detector coverage limits, not a language-generation failure.

Hallucination does not respond to threshold tuning. Raising detector thresholds to favour precision (v13, β=0.85) only cost recall (Missed% 39.8→46.6) without reducing Halluc%, suggesting hallucination originates in the decoder's language prior rather than over-triggered detections.

Region restriction and a stronger encoder both fail. Lung-only training (v9) regressed, since over 40% of findings lie outside the lungs – unlike segmentation, where region emphasis helps [11]. Replacing Swin with a chest-X-ray-pretrained RAD-DINO backbone (v14, with its own HPO) tied v12 on Clin-F1 (0.447 vs. 0.451) but worsened hallucination

TABLE II

FULL METRIC COMPARISON ACROSS THE VERSION ABLATION, ON THE SHARED 1,703-STUDY HELD-OUT TEST SET (V3–V14; V1–V2 USED AN EARLIER, NON-COMPARABLE TEST-CLEANING AND ARE EXCLUDED, SEE TEXT). B1–B4 = BLEU-1–4; MET = METEOR; R-L = ROUGE-L; CLIN-P/R/F1 = CLINICAL PRECISION/RECALL/F1 FROM THE RULE-BASED LABELLER. **BOLD** = BEST IN COLUMN.

Ver.	Key change	B1	B2	B3	B4	MET	R-L	CIDEr	Clin-P	Clin-R	Clin-F1	Miss.%	Hall.%
v3	US duplicate cap	0.413	0.328	0.279	0.246	0.210	0.368	0.750	0.577	0.198	0.295	63.4	19.9
v4	+ modality-balanced sampling	0.421	0.343	0.297	0.267	0.219	0.392	1.014	**0.614**	0.277	0.382	53.7	15.4
v5	+ internal finding-head conditioning	0.422	0.343	0.297	0.267	0.218	0.390	0.954	0.590	0.288	0.387	53.2	16.1
v6	+ contrastive grounding tuning	0.432	0.352	0.305	0.274	0.223	0.394	1.070	0.608	0.300	0.402	51.6	17.0
v7	+ HPO-tuned architecture	0.410	0.334	0.290	0.262	0.216	0.394	1.088	0.574	0.283	0.379	53.5	13.8
v8	+ detect-then-describe, faithfulness 4.0	0.429	0.348	0.302	0.271	0.220	0.392	0.969	0.446	0.382	0.412	46.0	**13.1**
v9	lung-focus (masked) – regressed	0.434	0.345	0.295	0.262	0.217	0.378	0.879	0.497	0.341	0.405	46.1	19.1
v10	+ faithfulness 5.0, recall objective	0.437	0.356	0.310	0.279	0.223	0.400	**1.255**	0.497	0.386	0.435	42.7	21.4
v11	+ 3-net detector ensemble	**0.446**	**0.364**	**0.317**	**0.286**	**0.226**	**0.401**	1.183	0.486	0.407	0.443	41.4	18.4
v12	+ 5-net ensemble, F1 thresholds (best)	0.435	0.355	0.308	0.277	0.225	0.399	1.088	0.499	**0.411**	**0.451**	**39.8**	20.0
v13	precision-tuned thresholds (β=0.85) – rejected	0.433	0.354	0.307	0.276	0.223	0.398	1.029	0.512	0.361	0.424	46.6	19.0
v14	RAD-DINO vision encoder – rejected	0.420	0.332	0.281	0.248	0.212	0.364	0.801	0.503	0.402	0.447	**39.8**	24.3

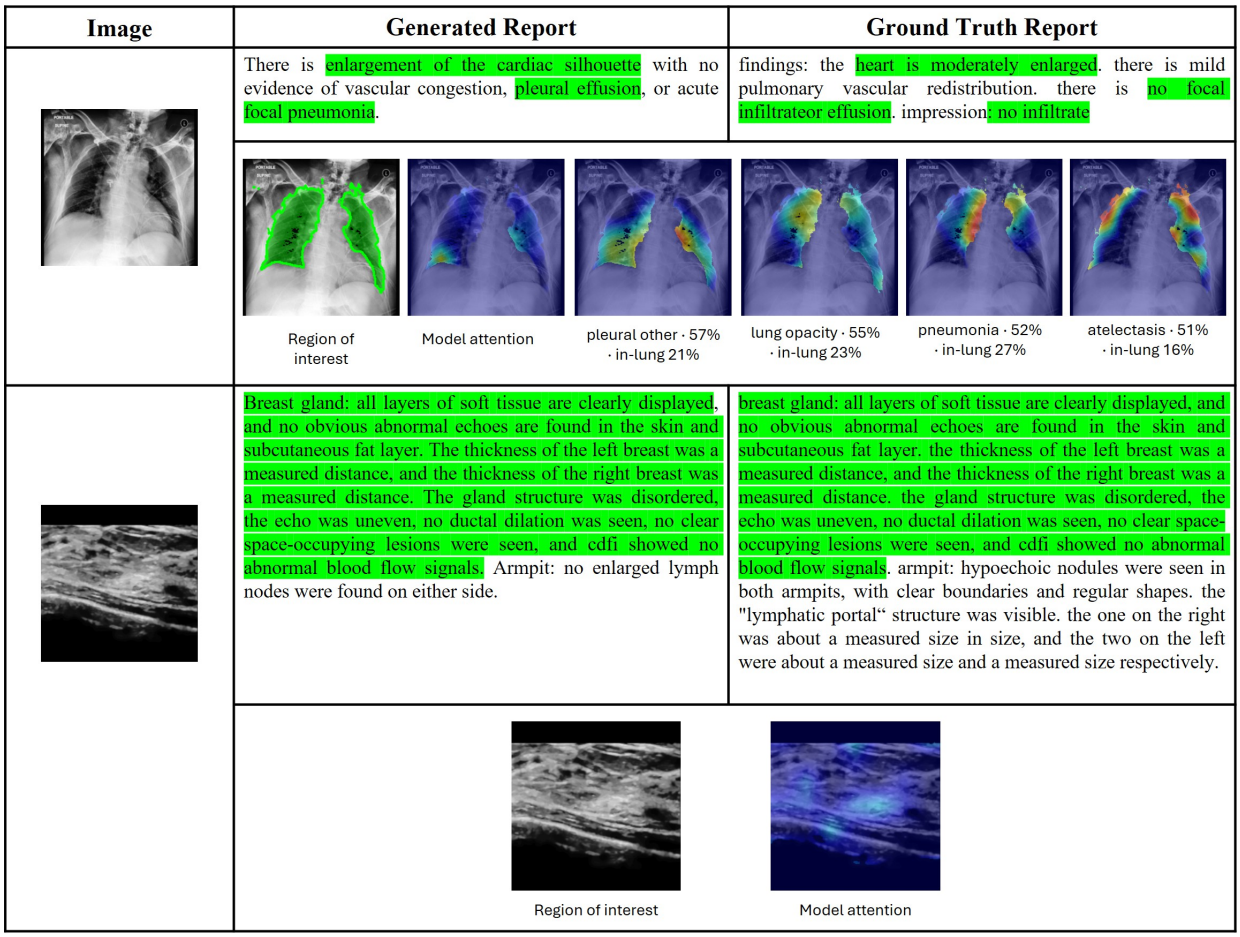

Fig. 2. Generated vs. reference report text (shared phrasing highlighted) with the full XAI package for two v12 test cases. Top: chest X-ray, with region-of-interest overlay, Q-Former attention map, and per-finding Grad-CAM (confidence % and % overlap with the lung mask) for four candidate findings. Bottom: breast ultrasound, with region-of-interest overlay and attention map.

(24.3% vs. 20.0%), confirming the Stage-1 detector, not the vision encoder, gates further accuracy gains.

Overall, v12 trades ∼1.6 points of extra hallucination for the lowest missed-diagnosis rate – a reasonable choice for screening, where a missed finding is costlier than a reviewable false alarm. At a clinical F1 of 0.451, though, roughly two in five abnormal studies still have their true finding(s) entirely unmentioned (Missed% 39.8): this is evidence for where the system's bottleneck lies, not a claim of clinical readiness. We see MedVLM-Report as a research testbed for the detect-then-describe mechanism and a diagnosis-support signal for clinician review, not a standalone report-writer.

Limitations: the clinical labeller is rule-based rather than a trained model (e.g. CheXbert); the detector ensemble is used

off-the-shelf, capping recall; the ultrasound training corpus is two orders of magnitude smaller than the X-ray corpus, a consequence of the manual, throughput-constrained collection process described in Section IV rather than a modelling choice, and likely limits generalisation on less-common ultrasound findings even though the ultrasound test split itself (703 studies) is proportionally large; the explainability package has not yet been validated against radiologist judgement; and, as discussed in Section II, no shared benchmark exists across this system and the closest related unified or biomedical VLMs, so our evaluation is an internal ablation rather than an external state-of-the-art comparison.

VIII. Conclusion

We presented MedVLM-Report, a unified chest X-ray and breast-ultrasound report-generation system built around a Swin/Q-Former/BioGPT-LoRA backbone, whose central contribution is a two-stage *detect-then-describe* conditioning mechanism with a faithfulness-weighted loss, and a mandatory four-technique explainability package attached to every report. Across a systematic 14-version ablation, detector conditioning and faithfulness weighting produced the largest clinical-accuracy gains, detector ensembling reduced missed diagnoses to 39.8% at a clinical F1 of 0.451, and two negative results – threshold-based hallucination control and a stronger vision backbone – both failed to improve on this, pointing to Stage-1 detector recall as the system's binding constraint. Future work includes fine-tuning the detector directly on the target reporting corpus, adopting a trained clinical labeller (CheXbert or RadGraph), expanding the ultrasound corpus, and validating the explainability package against radiologist-rated evidence relevance.

References

[1] M. M. Mohsan, M. U. Akram, G. Rasool, N. S. Alghamdi, M. A. A. Baqai, and M. Abbas, "Vision transformer and language model based radiology report generation," *IEEE Access*, 2023.

[2] S. Yang, J. Niu, J. Wu, Y. Wang, X. Liu, and Q. Li, "Automatic ultrasound image report generation with adaptive multimodal attention mechanism," *Neurocomputing*, vol. 427, pp. 40–49, 2021.

[3] G. Ras, N. Xie, M. van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *arXiv preprint arXiv:2004.14545*, 2021.

[4] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of explainable AI techniques in healthcare," *Sensors*, vol. 23, no. 2, p. 634, 2023.

[5] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is Grad-CAM explainable in medical images?," *arXiv preprint arXiv:2307.10506*, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[7] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2GenGPT: Radiology report generation with frozen large language models," *Meta-Radiology*, vol. 1, no. 3, p. 100033, 2023.

[8] D. Goswami, R. Subedi, and S. Chakraborty, "MediVLM: A vision language model for radiology report generation," in *Findings of the Assoc. for Computational Linguistics: EMNLP*, 2025, pp. 10287–10304.

[9] L. Yan, X. Zhou, Y. Wang, X. Chang, Q. Li, and G. Han, "Automated ultrasound diagnosis via CLIP-GPT synergy: A multimodal framework for image classification and report generation," *IEEE Access*, vol. 13, pp. 107950–107960, 2025.

[10] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. von Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, art. e190043, 2020.

[11] A. M. Mostafa, A. S. Alaerjan, B. Aldughayfiq, H. Allahem, A. A. Mahmoud, W. Said, H. Shabana, and M. Ezz, "Optimized YOLOv8 for enhanced breast tumor segmentation in ultrasound imaging," *Discover Oncology*, vol. 16, p. 1152, 2025.

[12] Z. Yu, Q. Guan, J. Yang, Z. Yang, Q. Zhou, Y. Chen, and F. Chen, "LSM-YOLO: A compact and effective ROI detector for medical detection," *arXiv preprint arXiv:2408.14087*, 2024.

[13] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Machine Learning (ICML)*, 2023, pp. 19730–19742.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[16] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, art. bbac409, 2022.

[17] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, and H. Bertrand, "TorchXRayVision: A library of chest X-ray datasets and models," in *Proc. Medical Imaging with Deep Learning (MIDL)*, 2022, pp. 231–249.

[18] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artificial Intelligence*, 2019, pp. 590–597.

[19] A. E. W. Johnson *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 317, 2019.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

[21] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 72–78.

[22] F. Pérez-García *et al.*, "Exploring scalable medical image encoders beyond text supervision," *arXiv preprint arXiv:2401.10815*, 2024.

[23] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.

[24] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[25] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets and Benchmarks Track*, 2023.

[26] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-Flamingo: A multimodal medical few-shot learner," *arXiv preprint arXiv:2307.15189*, 2023.

[27] A. Alkhaldi, R. Alnajim, L. Alabdullatef, R. Alyahya, J. Chen, D. Zhu, A. Alsinan, and M. Elhoseiny, "MiniGPT-Med: Large language model as a general interface for radiology diagnosis," *arXiv preprint arXiv:2407.04106*, 2024.

[28] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the most of text semantics to improve biomedical vision–language processing," in *Proc. European Conf. Computer Vision (ECCV)*, Lecture Notes in Computer Science, vol. 13696, 2022, pp. 1–21.

[29] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, and B. Zhao, "M3D: Advancing 3D medical image analysis with multi-modal large language models," *arXiv preprint arXiv:2404.00578*, 2024.