Benchmarking and Adapting DeepSeek-OCR and Vision-Language Models for Chest X-ray Report Generation

Emily Xie EMXIE@UCSD.EDU
University of California, San Diego

Ryan Khalloqi KHALLOQI@WISC.EDU
University of Wisconsin-Madison

Rui Qiu QIUR14@MCMASTER.CA
McMaster University

Jun Ma JUNMA.MA@MAIL.UTORONTO.CA
AI Hub, University Health Network

Editors: Under Review for MIDL 2026

Abstract

We investigate whether vision-language models designed for document understanding can be repurposed for radiology report generation. Surprisingly, DeepSeek-OCR, an optical character recognition-centric model with no medical pretraining, achieves state-of-the-art performance on chest X-ray report generation (GREEN = 0.846) after supervised fine-tuning, outperforming medical-domain models including MedGemma-4B. We attribute this to aggressive visual token compression, which proves effective for encoding radiographic detail. Component analysis reveals that location accuracy and entity matching are the main bottlenecks in zero-shot models, with DeepSeek-OCR showing +262% and +230% improvements, respectively, after fine-tuning. We further show that reinforcement learning with RadGraph-based clinical rewards yields gains beyond supervised fine-tuning saturation, improving entity matching by 6% on Qwen3-VL-4B. Our results suggest that document-understanding architectures offer an underexplored pathway for medical image interpretation.

Keywords: Radiology report generation, vision-language models, chest X-ray, supervised fine-tuning, reinforcement learning, visual token compression.

1. Introduction

The domain of medical artificial intelligence (AI) stands at a juncture, specifically within the sub-discipline of automated radiology report generation. The ambition to create computational systems capable of interpreting radiographic images has been driven by the escalating global shortage of radiologists and the increasing complexity of diagnostic imaging data (Sloan et al., 2025)—imaging volumes increased 1,091% between 2009–2022 while the radiology workforce grew only 19% (Troupis et al., 2024; van Leeuwen et al., 2022). Although Large Language Models (LLMs) and Vision-Language Models (VLMs) have catalysed a shift from template-based systems to generative narratives, significant architectural bottlenecks remain (Mamdouh et al., 2025). Foremost among these is the "resolution-context trade-off": the computationally prohibitive cost of processing high-resolution medical images within token constraints, leading to visual hallucinations, lack of spatial grounding, and loss of fine-grained pathological detail (Lu et al., 2024).

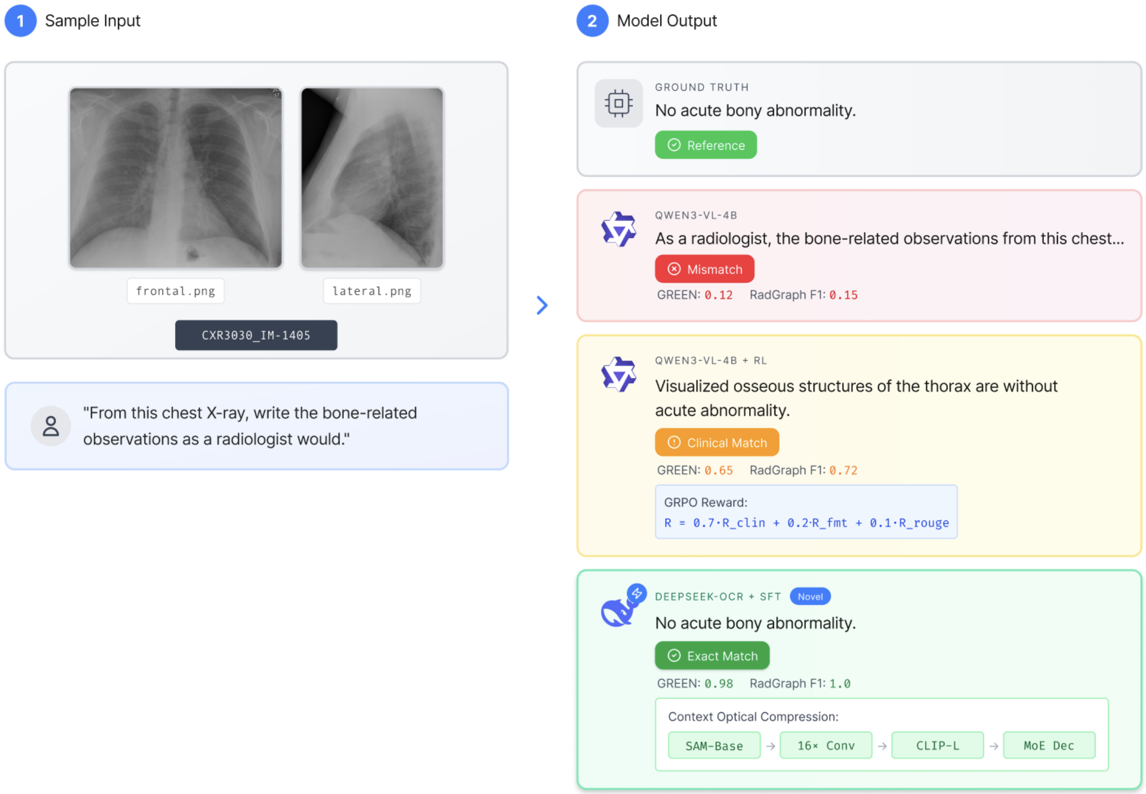

1 Sample Input
frontal.png
lateral.png
CXR3030_IM-1405
"From this chest X-ray, write the bone-related observations as a radiologist would."
2 Model Output
GROUND TRUTH
No acute bony abnormality.
Reference
QWEN3-VL-4B
As a radiologist, the bone-related observations from this chest...
Mismatch
GREEN: 0.12 RadGraph F1: 0.15
QWEN3-VL-4B + RL
Visualized osseous structures of the thorax are without acute abnormality.
Clinical Match
GREEN: 0.65 RadGraph F1: 0.72
GRPO Reward:
R = 0.7·R_clin + 0.2·R_fmt + 0.1·R_rouge
DEEPSEEK-OCR + SFT
Novel
No acute bony abnormality.
Exact Match
GREEN: 0.98 RadGraph F1: 1.0
Context Optical Compression:
SAM-Base
16× Conv
CLIP-L
MoE Dec

Figure 1: Overview of our approach: DeepSeek-OCR, an OCR-centric model with aggressive visual token compression, achieves near-exact clinical matches after supervised fine-tuning (SFT), outperforming general-purpose VLMs. Reinforcement learning (RL) with clinical rewards further improves Qwen3-VL-4B beyond SFT alone.

Early approaches adapted CNN-RNN architectures from image captioning (Vinyals et al., 2015; Shin et al., 2016), though fixed-length visual representations limited pathology localisation. Attention mechanisms improved text-pathology correspondence (Xue et al., 2018), and transformers subsequently became dominant: R2Gen introduced memory-driven transformers (Chen et al., 2020), R2GenCMN enhanced cross-modal alignment (Chen et al., 2022), and RATCHET demonstrated joint diagnosis and reporting (Hou et al., 2021). A central tension exists between domain-specific models like MedGemma, which incorporate medical pretraining but may lack sophisticated reasoning (Sellergren et al., 2025), and generalist models such as Qwen2.5-VL and Qwen3-VL with native dynamic resolution and extended context windows (Wang et al., 2024; Bai et al., 2025b,a).

This research investigates a novel cross-domain adaptation: applying DeepSeek-OCR, a specialised VL architecture designed for optical character recognition (OCR), to X-ray report generation (Wei et al., 2025). Although seemingly counter-intuitive, DeepSeek-OCR functions not merely as a text extractor but as an engine for Context Optical Compression, a mechanism that separates high-resolution visual information (1024×1024 pixels and beyond) into compressed semantic tokens (Dutt, 2025). We hypothesise its tri-stage encoder (SAM-Base → convolutional compressor → CLIP-Large) transfers effectively to radiographic interpretation, as chest radiographs share characteristics with structured documents: predictable layout, hierarchical information, and dense extraction requirements (Xie et al., 2025).

Parameter-efficient fine-tuning via LoRA and QLoRA enables adaptation with minimal parameters (Hu et al., 2021; Dettmers et al., 2023). While SFT optimises cross-entropy loss, it may not correlate with clinical utility. RL enables direct optimisation of clinically meaningful objectives; GRPO provides stable training for optimising rewards based on clinical entity extraction (RadGraph F1 (Delbrouck et al., 2024)) (Shao et al., 2024). The objectives of this research are the following.

- DeepSeek-OCR, an OCR model with no medical pretraining, achieves state-of-the-art GREEN (0.846) after SFT, outperforming medical-domain models, suggesting visual token compression transfers effectively to radiographic interpretation.
- RL with clinical rewards (RadGraph F1) yields +6% entity matching gains beyond SFT on Qwen3-VL-4B.
- Component analysis identifies location accuracy and entity matching as primary zero-shot bottlenecks, with DeepSeek-OCR showing +262% and +230% improvements after SFT.

2. Method

2.1. Models

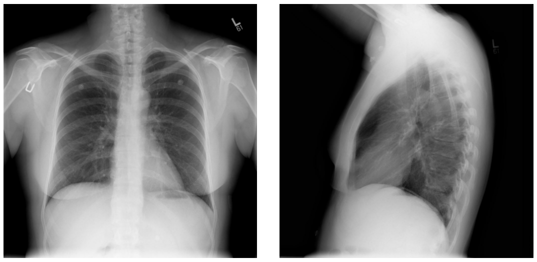
We evaluated four VLMs spanning different architectural paradigms. DeepSeek-OCR is a document-understanding model with a tri-stage visual encoder (SAM-Base → 16× compressor → CLIP-Large) and 3B-parameter MoE decoder (570M active) (Wei et al., 2025). Qwen2.5-VL-7B is a 7B-parameter VLM with ViT encoder and dynamic resolution (Bai et al., 2025b), while Qwen3-VL-4B is a 4B-parameter variant selected for RL experiments due to its efficiency (Bai et al., 2025a). MedGemma-4B is a medical foundation model with MedSigLIP encoder trained on 33M+ medical image-text pairs (Sellergren et al., 2025). All models used Unsloth for memory-efficient fine-tuning (Han et al., 2023).

2.2. Experimental Setup

Both training and validation sets follow a unified multi-turn conversational format where each sample pairs chest radiograph(s) with a clinical question and ground-truth response. Figure 2 illustrates a representative example.

2.2.1. Zero-Shot Baseline

For zero-shot evaluation, each model was prompted with the radiograph and associated clinical question without any domain-specific fine-tuning. Inference was performed on NVIDIA H100 GPUs. For DeepSeek-OCR, images were processed at 1024×1024 resolution for the



User: `<frontal.png>` `<lateral.png>` Write the part of the radiology report that describes the heart.
Assistant: Heart size within normal limits.

Figure 2: Example training/evaluation sample: paired frontal and lateral chest radiographs with clinical question and ground-truth response.

global view with dynamic tiling disabled to establish a controlled baseline. Generated outputs were truncated to 500 characters to standardise comparison across models with varying verbosity.

2.2.2. Supervised Fine-Tuning

DeepSeek-OCR. Fine-tuning was performed using Low-Rank Adaptation (LoRA) to enable parameter-efficient training while preserving pretrained representations (Hu et al., 2021). LoRA adapters were applied to the query, key, value, output, gate, up, and down projection matrices of the language model component with rank $r = 16$, scaling factor $\alpha = 32$, and dropout probability $p = 0.05$. We developed a custom data collator to handle VL alignment, computing per-image token allocations based on the encoder's compression ratio (16× downsampling with patch size 16) and interleaving image tokens with text according to `<image>` placeholder positions. Images were processed at 1024×1024 resolution using the `BasicImageTransform` with mean and standard deviation normalisation of $(0.5, 0.5, 0.5)$.

Qwen models. For Qwen3-VL-4B and Qwen2.5-VL-7B, we applied LoRA to all linear layers in both the vision encoder and language model with $r = 16$, $\alpha = 16$, and no dropout. We additionally saved the output head and embedding modules to preserve vocabulary alignment. Training used Unsloth's vision data collator for automatic image processing and chat template application.

MedGemma. For MedGemma-4B, we employed Quantized LoRA (QLoRA) with 4-bit quantisation to enable training within GPU memory constraints (Dettmers et al., 2023). The model was fine-tuned on the complete set of 19 FLARE 2025 datasets to leverage cross-task transfer learning (Yin et al., 2025). Additional memory optimisations included dynamic batch sizing, gradient checkpointing, and fast image loading with OpenCV caching.

Training configuration. To focus learning on clinical text generation, we masked all user-turn tokens when computing the cross-entropy loss (response-only training). Common hyperparameters included: batch size 1 per device, gradient accumulation over 4 to 8 steps, learning rate 1×10^{-4} to 2×10^{-4} with linear warmup (5 steps) and linear decay, AdamW optimiser with 8-bit quantisation (Loshchilov and Hutter, 2019), and mixed-precision training (bfloat16 where supported, fp16 otherwise). Early stopping with patience of 3 evaluation rounds was applied based on validation loss.

2.2.3. Reinforcement Learning

For Qwen3-VL-4B, we additionally explored reinforcement learning fine-tuning using Group Relative Policy Optimisation (GRPO) (Shao et al., 2024) to directly optimise for clinical accuracy, initialising from the SFT checkpoint. GRPO training used learning rate 5×10^{-6}, batch size 1, gradient accumulation over 8 steps, and 2 generations per prompt for advantage estimation. We employed sequence-level importance sampling with DR-GRPO loss. Training proceeded for 300 steps with checkpoints saved every 100 steps. Final rewards were clipped to $[-2, 2]$ for stability.

Reward function. We implemented a composite reward function combining clinical accuracy, format quality, and lexical overlap:

$$R(y, y^*) = 0.7 \cdot R_{\text{clin}} + 0.2 \cdot R_{\text{fmt}} + 0.1 \cdot R_{\text{rouge}} \tag{1}$$

where y is the generated report and y^* is the reference.

The clinical reward R_{clin} was computed using RadGraph-XL (Delbrouck et al., 2024), which extracts clinical entities and relations from radiology reports and computes F1 overlap between generated and reference graphs:

$$R_{\text{clin}} = F1_{\text{RadGraph}}(y, y^*) \tag{2}$$

The format reward R_{fmt} penalised malformed outputs containing thinking tags, code blocks, or excessive length:

$$R_{\text{fmt}} = 1.0 - \sum_{v \in \mathcal{V}} \nVdash [v \in y] \tag{3}$$

where $\mathcal{V}$ is the set of format violations. The ROUGE reward R_{rouge} used ROUGE-L F1 score for lexical overlap (Lin, 2004).

2.3. Evaluation Metrics

We evaluate generated reports using three complementary metrics. BLEU-4 (Papineni et al., 2002) measures corpus-level n-gram precision with brevity penalty, capturing lexical overlap with reference reports. The GREEN score (Grounding Radiology Evaluation with Expert Notations) (Ostmeier et al., 2024) provides comprehensive clinical evaluation through seven weighted components: entity matching (0.30), location accuracy (0.20), negation handling (0.15), temporal accuracy (0.10), measurement accuracy (0.10), clinical significance (0.10), and structure completeness (0.05). Clinical Efficacy (CE) evaluates coarse diagnostic alignment by categorising reports into critical, abnormal, normal, or uncertain classes, with partial credit for severity confusion within abnormal categories.

3. Experiments

3.1. Dataset

We employ the IU_XRay subset of the FLARE Task5 benchmark, a medical visual question answering (VQA) dataset for radiology report generation (Yin et al., 2025). The dataset consists of 5,908 chest radiographs with 9,742 associated question-answer pairs focused on free-text report generation. The dataset follows a standardised directory structure with separate image folders for training and validation splits, accompanied by JSON annotation files that contain image paths, questions, and ground-truth reports.

3.2. Experimental Results

Model	GREEN	BLEU-4	CE
Zero-Shot			
DeepSeek-OCR	0.498	0.002	0.851
Qwen2.5-VL-7B	0.505	0.000	0.936
Qwen3-VL-4B	0.695	0.003	0.327
MedGemma-4B	0.691	0.000	0.850
Supervised Fine-Tuning (LoRA/QLoRA)			
DeepSeek-OCR	0.846	0.176	0.957
Qwen2.5-VL-7B	0.824	0.117	0.953
Qwen3-VL-4B	0.769	0.050	0.843
MedGemma-4B	0.791	0.050	0.900
Reinforcement Learning (GRPO)			
Qwen3-VL-4B	0.793	0.066	0.836

Table 1: Performance on IU_XRay validation set ($n = 1945$). Best results underlined.

Zero-shot. Qwen3-VL-4B attains the highest zero-shot GREEN (0.695) but low CE (0.327), indicating verbose yet diagnostically misaligned outputs. DeepSeek-OCR starts with lower GREEN (0.498) yet strong CE (0.851), suggesting its document-centric pre-training captures coarse clinical categories despite limited radiology exposure. BLEU-4 is near-zero for all models, reflecting stylistic divergence from reference reports (Table 1).

Supervised fine-tuning. SFT produces substantial improvements across all models (Table 1). Figure 3 shows GREEN scores before and after SFT: DeepSeek-OCR improves from 0.498 to 0.846 (+70%), Qwen2.5-VL-7B from 0.505 to 0.824 (+63%), Qwen3-VL-4B from 0.695 to 0.769 (+11%), and MedGemma-4B from 0.691 to 0.791 (+14%). Models with lower zero-shot performance exhibit larger improvements. BLEU-4 rises from near-zero to 0.05 to 0.18, indicating stylistic alignment; CE converges to 0.84 to 0.96. After SFT, DeepSeek-OCR moves from an OCR-heavy baseline to the Pareto front, attaining the top GREEN (0.846) and BLEU-4 (0.176) while matching the highest CE (0.957). Its gains show that aggressive visual token compression can transfer to radiology once guided by domain supervision.

Reinforcement learning. GRPO on Qwen3-VL-4B yields consistent gains beyond SFT across all metrics: GREEN improves by +3.1% (0.769 → 0.793), BLEU-4 by +32% (0.050 → 0.066), while CE remains stable (0.843 → 0.836) (Table 1). These results demonstrate that RL with clinical reward signals (RadGraph F1) can extract additional performance gains after SFT has plateaued, particularly refining entity-level precision. The RL bar in Figure 3 shows Qwen3-VL-4B approaching the clinical threshold after GRPO training, suggesting RL as a promising direction for further improving DeepSeek-OCR and other models in future work.

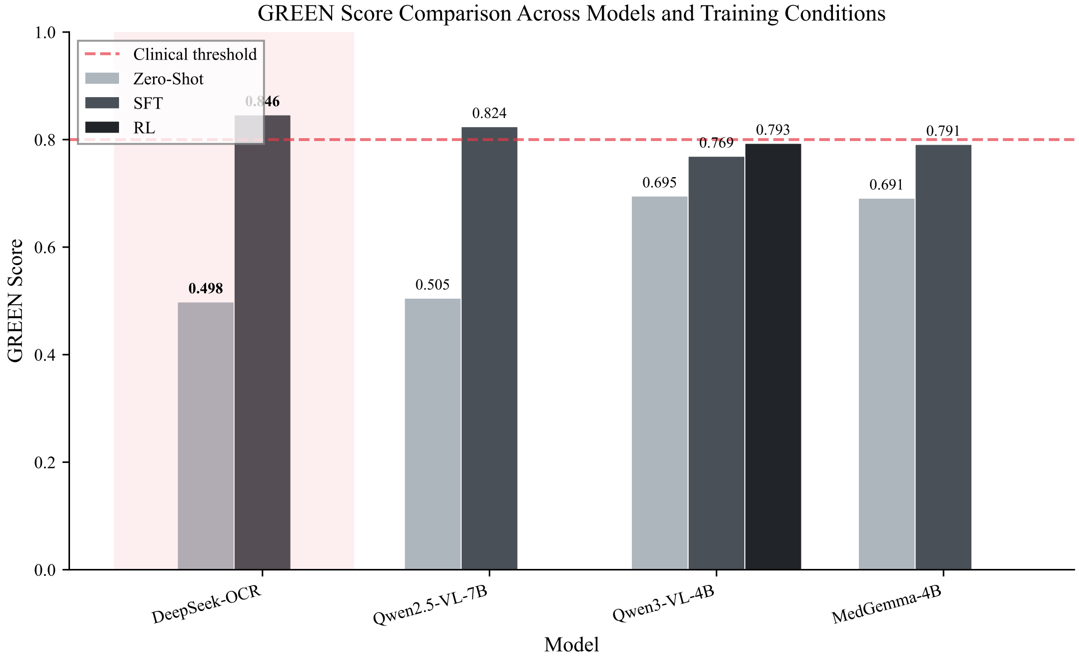


Figure 3: GREEN score comparison across four models under zero-shot and SFT conditions ($n = 1945$ validation samples). The dashed line at 0.8 indicates a clinical acceptability threshold. DeepSeek-OCR exhibits the largest absolute gain (0.498 to 0.846, $\Delta = +0.348$) despite having the lowest zero-shot baseline.

3.3. Component-Wise Analysis

To understand why DeepSeek-OCR achieves state-of-the-art performance after SFT, we analyse the seven GREEN components. Figure 4 shows the distribution of component scores across all models and training conditions: zero-shot models exhibit high variance (particularly in location accuracy and clinical significance), while SFT compresses the distribution toward higher scores. Figure 5a visualises DeepSeek-OCR's component-wise transformation: the zero-shot model (grey) shows severe deficits in entity matching (0.248) and location accuracy (0.218), while SFT (red) dramatically expands the performance envelope across all dimensions. Table 2 provides exact scores for DeepSeek-OCR and Qwen3-VL-4B.

Location. The largest SFT improvements occur in location accuracy: DeepSeek-OCR improves from 0.218 to 0.789 (+262%), and Qwen3-VL-4B from 0.236 to 0.538 (+128%). This indicates that zero-shot models struggle to ground findings to anatomical structures

Component	DeepSeek-OCR Base	 SFT	Qwen3-VL-4B Base	 SFT	 RL
Entity Matching	0.248	0.819	0.746	0.767	0.813
Location Accuracy	0.218	0.789	0.236	0.538	0.564
Negation Handling	0.964	0.986	0.974	0.980	0.977
Temporal Accuracy	0.899	0.932	0.763	0.894	0.898
Measurement Accuracy	0.998	1.000	0.996	0.997	0.996
Clinical Significance	0.083	0.577	0.677	0.566	0.625
Structure Completeness	0.742	0.875	0.694	0.770	0.767

Table 2: GREEN component scores across models and training conditions.

(e.g., "left lower lobe," "right costophrenic angle"), a skill that requires domain-specific supervision to acquire.

Entity matching. DeepSeek-OCR shows dramatic entity matching improvement (0.248 → 0.819, +230%), the largest gain among all models, suggesting its document-centric pretraining provides a strong foundation that transfers effectively to clinical entity recognition once fine-tuned. Qwen3-VL-4B, by contrast, has strong zero-shot entity matching (0.746) that improves modestly with SFT (0.767). Figure 5b visualises the full training paradigm progression for Qwen3-VL-4B: RL (dark) expands beyond SFT (teal) in entity matching (0.767 → 0.813, +6%), location accuracy (0.538 → 0.564, +5%), and clinical significance (0.566 → 0.625, +10%). This demonstrates that clinical reward signals can drive targeted improvements beyond cross-entropy training, suggesting RL could yield similar gains for DeepSeek-OCR.

Stable components. Negation handling (0.96 to 0.99) and measurement accuracy (0.99 to 1.00) remain consistently high across all conditions, suggesting these capabilities transfer well from general pretraining. Models rarely confuse present and absent findings or misreport numerical measurements.

3.4. Qualitative Analysis

The quantitative gains reported above manifest as qualitatively distinct output behaviours. Table 3 presents representative model outputs for a normal chest radiograph across two clinical questions: cardiac and pulmonary findings.

Zero-shot models exhibit two failure modes. First, *verbose hallucination*: Qwen3-VL-4B fabricates pathologies absent from the image, reporting "cardiomegaly," "pleural effusion," "pneumothorax," and "consolidation" for a study with no abnormalities. Second, *format mismatch*: DeepSeek-OCR produces irrelevant meta-commentary ("this image is an example of a false positive case") rather than clinical findings, while Qwen2.5-VL-7B generates incomplete preambles.

SFT eliminates both failure modes. All models produce concise, clinically accurate responses that align with the ground truth ("Heart size is normal"; "The lungs are clear. No pleural effusion or pneumothorax"). For Qwen3-VL-4B, RL maintains the format disci-

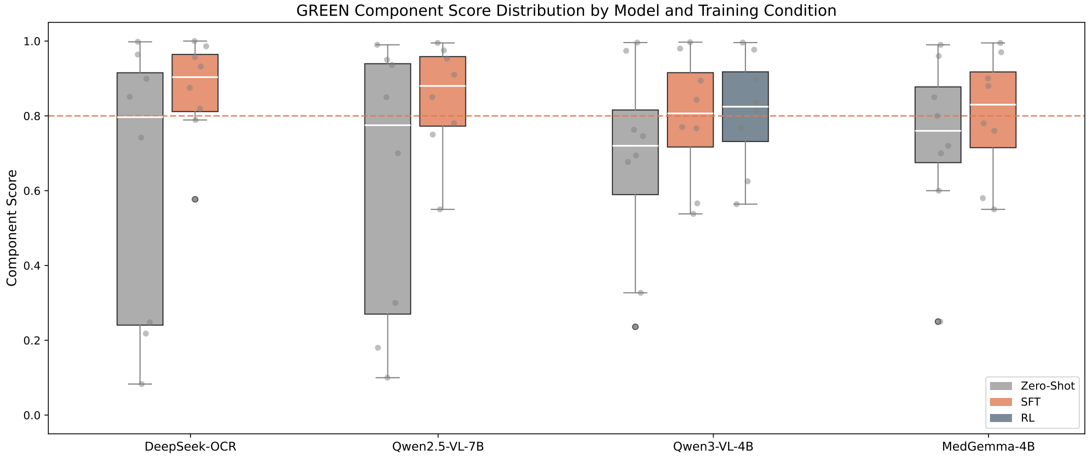
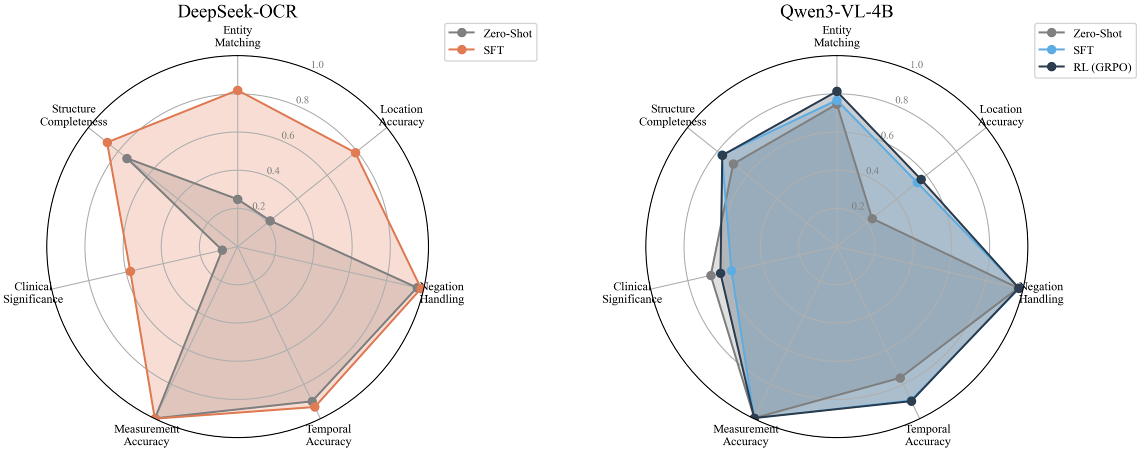


Figure 4: Box plot of GREEN component score distributions across models and training conditions. Each box shows the spread of the seven component scores (EM, LA, NH, TA, MA, CS, SC). DeepSeek-OCR exhibits the largest median shift from zero-shot (grey) to SFT (red), with reduced variance indicating more consistent performance across components.



(a) DeepSeek-OCR

(b) Qwen3-VL-4B

Figure 5: Radar plots of GREEN component scores. (a) DeepSeek-OCR: zero-shot (grey) vs SFT (red). Entity matching improves from 0.248 to 0.819 (+230%) and location accuracy from 0.218 to 0.789 (+262%). (b) Qwen3-VL-4B: zero-shot (grey), SFT (teal), and RL (dark). RL provides incremental gains over SFT in entity matching (+6%), location accuracy (+5%), and clinical significance (+10%).

pline acquired during SFT while preserving diagnostic accuracy, demonstrating that clinical reward signals do not degrade output quality.

Model	Condition	Response (truncated)
Cardiac findings — Ground truth: *Normal cardiac contours.*		
Qwen3-VL-4B	Zero-Shot	*Cardiomegaly*: The heart appears enlarged... [500+ words]
Qwen3-VL-4B	SFT	Heart size is normal.
Qwen3-VL-4B	RL	Heart size is normal.
Qwen2.5-VL-7B	Zero-Shot	The provided chest X-ray image shows... [incomplete]
Qwen2.5-VL-7B	SFT	Cardiac contours are within normal limits.
DeepSeek-OCR	Zero-Shot	Heart shadow is slightly enlarged... [verbose]
DeepSeek-OCR	SFT	Heart size is normal.
Pulmonary findings — Ground truth: *No consolidation. No effusion. No pneumothorax.*		
Qwen3-VL-4B	Zero-Shot	*Pleural effusion... pneumothorax... consolidation...*
Qwen3-VL-4B	SFT	No focal airspace disease. No pneumothorax or effusion.
Qwen3-VL-4B	RL	Lungs clear. No pneumothorax or pleural effusion.
Qwen2.5-VL-7B	Zero-Shot	Standard AP view of the thorax... [incomplete]
Qwen2.5-VL-7B	SFT	No consolidation, pneumothorax or large pleural effusion.
DeepSeek-OCR	Zero-Shot	This is an example of a false positive case... [irrelevant]
DeepSeek-OCR	SFT	Lungs clear. No pleural effusion or pneumothorax.

Table 3: Model outputs for two CXR report generation tasks.

4. Conclusion

We show that DeepSeek-OCR, an OCR model without medical pretraining, achieves state-of-the-art chest X-ray report generation (GREEN = 0.846) after SFT, outperforming medical-domain models. Location accuracy (+262%) and entity matching (+230%) emerge as key zero-shot bottlenecks addressed by fine-tuning. RL with RadGraph rewards yields further gains on Qwen3-VL-4B (+6% entity matching), demonstrating that clinical reward signals can refine precision beyond cross-entropy training. These findings suggest visual token compression transfers effectively to radiographic interpretation, and RL-based optimisation offers a promising direction for medical VLM adaptation.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1439–1449, 2020.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12902–12915, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Aashi Dutt. Deepseek-ocr: A hands-on guide with 7 practical examples, 2025.

Daniel Han, Michael Han, and Unsloth team. Unsloth, 2023.

Benjamin Hou, Georgios Kaissis, Ronald M Summers, and Bernhard Kainz. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 293–303, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Dima Mamdouh, Mariam Attia, Mohamed Osama, Nesma Mohamed, Abdelrahman Lotfy, Tamer Arafa, Essam A Rashed, and Ghada Khoriba. Advancements in radiology report generation: A comprehensive analysis. *Bioengineering*, 12(7):693, 2025.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y K Li, Y Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. *arXiv preprint arXiv:1603.08486*, 2016.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387, 2025.

Christopher John Troupis, Richard Alexander Hyde Knight, and Kenneth Kwok-Pan Lau. What is the appropriate measure of radiology workload: Study or image numbers? *Journal of Medical Imaging and Radiation Oncology*, 68(5):530–539, 2024.

Kicky G van Leeuwen, Maarten de Rooij, Steven Schalekamp, Bram van Ginneken, and Matthieu J C M Rutten. How does artificial intelligence in radiology improve efficiency and health outcomes? *Pediatric Radiology*, 52(11):2087–2093, 2022.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2015.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

Bin Xie, Hao Tang, Dawen Cai, Yan Yan, and Gady Agam. Self-prompt sam: Medical image segmentation via automatic prompt sam adaptation. *arXiv preprint arXiv:2502.00630*, 2025.

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 457–466, 2018.

Shuolin Yin, Beatrice Chen, Yeonwoo Seo, and Jun Ma. Flare-task5-mllm-2d, 2025.