SusufDoctor: Vision-Language Transformers for Chest X-Ray Report Generation with Longitudinal Comparison

Abstract

Chest X-ray interpretation is a critical diagnostic task, yet radiologists in low-resource settings often face high workloads and long reporting times due to severe workforce shortages. Current automated report generation systems primarily rely on single-image analysis and cannot incorporate longitudinal comparisons or patient metadata, limiting their clinical usefulness. This paper presents SusufDoctor, an intelligent multimodal chest X-ray report generation system powered by a fine-tuned SmolVLM-500M vision-language transformer. The system integrates current and prior chest X-ray images, radiology reports, and patient metadata to produce comprehensive, structured reports covering Findings and Impression sections. A longitudinal multimodal dataset was constructed from the CheXpert-Plus dataset to train and evaluate the model. Through LoRA-based fine-tuning and quantization, the model achieved BLEU of 61.53%, ROUGE-L of 66.08%, and BERTScore (F1) of 93.92%, up from baseline values of 1.29%, 8.26%, and 80.48% respectively. The system was deployed as a web application enabling real-time inference and practical integration into radiologist workflows, demonstrating strong potential to reduce reporting burden and improve access to quality radiology services in resource-limited environments such as Ghana and Rwanda.

Keywords: chest X-ray, radiology report generation, vision-language model, longitudinal analysis, LoRA fine-tuning

Introduction

Medical imaging generates approximately 90% of all healthcare data, with an estimated 2 billion chest X-rays performed annually worldwide, accounting for 30–40% of all imaging modalities. Chest X-rays are essential for diagnosing and monitoring thoracic conditions such as pneumonia, pulmonary edema, and pneumothorax.[1] Despite this volume, over two-thirds of the global population lacks access to quality medical imaging diagnosis, with the radiologist-to-patient ratio at roughly 1 per 500,000 people in parts of Africa. Ghana has fewer than 150 practising radiologists serving over 30 million people, while Rwanda has fewer than 15 nationwide. Manual report writing takes 20–30 minutes per study, and radiology reports contain clinically important errors in up to 30% of cases, with radiologists disagreeing on findings in 20–30% of instances.[3]

A major limitation in current automated systems is over-reliance on single-image interpretation, which neglects sequential comparison with prior exams, a practice central to real radiologist workflows.[5] Existing longitudinal approaches have focused only on pre-filling the Findings section without generating complete structured reports [6], while region-guided models improve interpretability but are limited to single-timepoint analysis. Most models also ignore patient metadata despite its diagnostic value. This work addresses these gaps by developing SusufDoctor: a multimodal, longitudinal-aware system that automates full chest X-ray report generation using a fine-tuned vision-language model.

Methods

Dataset and preprocessing

Training data was sourced from CheXpert-Plus [2], containing over 200,000 frontal and lateral chest X-rays paired with full radiology reports and patient metadata, including study dates, demographics, and BMI. Since the data was not provided in longitudinal format, a custom pipeline was developed to organise patient studies chronologically and pair each prior study with its corresponding current study. A 97 GB subset was processed into a cleaned 46 GB dataset comprising 1,356 training pairs and 147 test pairs. Identical prior-current report pairs were removed to prevent the model from learning non-progressive patterns.

Model architecture and fine-tuning

SusufDoctor is built on the SmolVLM-500M-Instruct vision-language model, running inference on a single image with just 1.23 GB of GPU RAM. The architecture integrates three input streams: (1) the current chest X-ray, encoded into visual embeddings; (2) the prior radiology report, encoded as temporal memory for longitudinal comparison; and (3) patient metadata (age, sex, BMI). These are fused via cross-attention, and a transformer decoder generates fully structured reports, including the Findings and Impression sections.

To adapt the model under computational constraints, 4-bit quantisation (BitsAndBytes NF4) was applied alongside Low-Rank Adaptation (LoRA) with rank $r = 16$, scaling $\alpha = 32$, and dropout 0.05, targeting attention projection and feed-forward layers. Only 9,568,256 parameters (1.85% of total) were trained. Performance was evaluated using BLEU, ROUGE-L, and BERTScore, capturing lexical precision, structural coherence, and

semantic similarity respectively. Training converged within three epochs using early stopping, achieving a final validation loss of approximately 0.50.

Results

Table 1: Model performance before and after fine-tuning

Model	BLEU	ROUGE-L	BERTScore (F1)
Base SmolVLM-500M	1.29%	8.26%	80.48%
SusufDoctor (fine-tuned)	61.53%	66.08%	93.92%

The BLEU score increased from 1.29% to 61.53% reflects the model's successful learning of domain-specific radiological vocabulary and phrasing. The ROUGE-L improvement from 8.26% to 66.08% reflects better structural correspondence with ground-truth reports, capturing clinically meaningful phrases such as "pleural effusion," "bibasilar opacity," and "interval change." BERTScore (F1) rose from 80.48% to 93.92%, confirming enhanced semantic alignment. Integration testing achieved a 100% success rate across 50 test scenarios with API response times below 500 ms. The model generated full reports in 3–30 seconds with a 98% success rate. Acceptance testing with five radiologists showed over 90% task completion, with particular appreciation for the longitudinal comparison language and the report editing feature.

Discussion and Conclusion

SusufDoctor demonstrates that a lightweight, parameter-efficient vision-language model can be adapted to generate clinically aligned, longitudinally aware radiology reports with limited data and computational resources. The BLEU improvement confirms that domain-specific fine-tuning is essential for radiology-specific language generation. Qualitative analysis confirmed the model correctly incorporates temporal comparison language such as "interval improvement" and "stable," mirroring real radiologist practice of comparing current findings against prior studies.

Key limitations include the relatively small training set, which may restrict generalisation to rare pathologies, occasional model hallucinations, and the absence of multi-view imaging or richer clinical inputs such as vital signs and laboratory data. This research is also exploring different multimodal architectures to identify better-performing models for report generation. Future directions include expanding the dataset, integrating additional clinical modalities, and deploying SusufDoctor within hospital PACS to enable automatic retrieval of prior studies. This work contributes a new longitudinal multimodal dataset, a practical fine-tuning methodology for resource-constrained settings, and a deployed clinical tool with measurable impact in contexts like Ghana and Rwanda, where radiologist scarcity makes AI-assisted reporting especially valuable.

Further work is also in progress to train different model architectures to determine which would give the best results compared to this SmolVLM.

References

[1] Akhter, Y., Singh, R., & Vatsa, M. AI-based radiodiagnosis using chest X-rays: A review. Frontiers in Big Data, 6, 1120989, 2023.

[2] Chambon, P., et al. CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats. arXiv:2405.19538, 2024.

[3] Sarkodie, B., et al. Density and Regional Distribution of Radiologists in a Low-Income Country: The Ghana Situation. Chinese Journal of Academic Radiology, 6(4), 188–195, 2023.

[4] Islam, S. K. M. S., et al. Introduction of Medical Imaging Modalities. arXiv:2306.01022, 2023.

[5] Gefter, W. B., Post, B. A., & Hatabu, H. Commonly Missed Findings on Chest Radiographs: Causes and Consequences. Chest, 163(3), 650–658, 2022.

[6] Zhu, Q., et al. Utilizing Longitudinal Chest X-Rays and Reports to Pre-Fill Radiology Reports. arXiv:2306.08749, 2023.