Segmentation-Guided Radiology Report Generation for Pneumothorax Detection in Chest X-Rays

Yiming Jia[1], **Ahmed T. Elboardy**[1], **Essam A. Rashed**[1,2]

[1]Graduate School of Information Science, University of Hyogo, Kobe 650-0047, Japan
[2]Advanced Medical Engineering Research Institute, University of Hyogo, Himeji 670-0836, Japan
ad24w007@guh.u-hyogo.ac.jp, af25s001@guh.u-hyogo.ac.jp, rashed@gsis.u-hyogo.ac.jp

Abstract

Recent developments on chest radiographs has primarily focused on developing multi-disease frameworks that aim to diagnose a wide range of thoracic abnormalities from the Chest X-ray datasets. In contrast, this study specifically targets pneumothorax, a life-threatening condition commonly referred to as a collapsed lung, which requires timely detection and accurate clinical reporting. Existing automated report generation Vision-Language Models (VLMs) mainly rely on image-level features and often fail to fully leverage the rich structural information embedded in medical image segmentation. To address this limitation, we propose a distinct strategy to incorporate pneumothorax segmentation masks, which delineate affected regions and provide precise localization guidance to enhance the accuracy of medical image interpretation. Experimental results demonstrate that the proposed segmentation-guided approach integrates visual and textual understanding more effectively for pneumothorax diagnosis from chest radiographs. By employing segmentation masks as guidance, VLMs can accurately localize pathological regions while preserving anatomical context, thereby improving both interpretability and diagnostic precision. Quantitative evaluations across multiple metrics further confirm the effectiveness of the proposed methods in bridging the gap between image-level localization and report-level reasoning.

Introduction

Accurate and interpretable diagnostic reporting is crucial for clinical decision-making, yet current automated systems often fail to reflect the reasoning process of radiologists and concentrate on multi-disease diagnostic system. Motivated by the need to bridge this gap, we focus on pneumothorax, a critical and time-sensitive condition where precise localization and interpretation directly affect patient outcomes. Pneumothorax, clinically characterized by partially or completely collapsed lung, occurs when air or gas enters the pleural cavity, the space between the lung and chest wall (Iqbal, Hallifax, and Rahman 2025). With the growing adoption of VLMs, these models have been increasingly applied in the medical domain to generate diagnostic reports by jointly learning textual feature from expert-written radiology reports and visual features from medical images (Li et al. 2025). For pneumothorax, X-ray images are the primary diagnostic modality as it enables rapid, non-invasive visualization of intrapleural air and assessment of lung collapse with high clinical reliability (Singer et al. 2025). Compared with natural image, X-ray images show low contrast, subtle intensity variations, and overlapping anatomical structures, which makes it particularly challenging for VLMs to achieve accurate interpretation and automated analysis.

To address these challenges of accurate and interpretable pneumothorax report generation, we propose a segmentation-guided approach that leverages pneumothorax segmentation masks into chest radiographs. This study aims to mimic the real-world clinical diagnostic workflow for pneumothorax, which typically involves three sequential stages: first medical imaging acquisition, then radiologist evaluation with lesion annotation, and finally diagnostic report generation. By integrating this process into our modeling framework, we enable VLMs to better align with the actual reasoning pipeline of radiologists.

In this work, we utilized the CANDID-PTX dataset, a comprehensive collection of chest radiographs with expert-annotated pneumothorax segmentation masks and experted-generated pneumothorax report. Specifically, we employed both the chest X-ray images and their corresponding segmentation masks as visual input. Furthermore, to investigate the robustness of segmentation-guided report generation, we conduct experiments by using both ground-truth masks and imperfect segmentation results with lower accuracy, analyzing how segmentation quality influences the generated reports. These modalities are used to fine-tune state-of-the-art open-source VLMs under a supervised fine-tuning setup. Model performance is rigorously evaluated across six established metrics to ensure both diagnostic accuracy and linguistic quality.The key contribution of this study is as following:

- We propose a segmentation-guided approach for pneumothorax diagnosis generation that emulates the clinical pneumothorax diagnosis workflow, then analyze the generalized report with zero-shot, with/without prior segmentation knowledge to validate segmentation guidance enhances both accuracy and interpretability of report generation.
- We further analyze the effect of segmentation accuracy by comparing ground-truth masks with lower-quality

segmentation outputs, revealing how different segmentation qualities impact report generation performance.

- We utilize the activation heatmap to show how segmentation guidance influences the visual representation of the VLMs.

Related Work

Medical Image Segmentation

In recent years, medical image segmentation has become a cornerstone of AI-assisted Computer-Aided Diagnosis (CAD) and medical image analysis (Kabil et al. 2025), enabling precise delineation of anatomical structures for diagnosis. Recent advances in medical image segmentation have been driven by deep learning architectures that effectively integrate both local and global contextual information. Transformer-based hybrid networks such as TransUNet (Chen et al. 2024), Swin-UNet (Cao et al. 2021), and foundation-style models like MedSAM (Ma et al. 2024) have demonstrated impressive generalization across diverse imaging modalities. These models leverage large-scale pretraining and attention mechanisms to enhances spatial reasoning and robustness. Despite the rapid emergence of increasingly complex architectures, the nnUNet framework remains one of the most influential and widely adopted paradigms in medical image segmentation due to its self-configuration and dataset-adaptive design (Isensee et al. 2021).

Originally proposed by Isensee et al., nnUNet established a standard methodology that transcends architecture novelty by automating the full segmentation pipeline, including the data preprocessing, network configuration etc. Instead of handcrafting hyperparameters or architectures, nnUNet performs dataset fingerprinting to infer task characteristics and automatically determines optimal configurations, enabling its robust performance across a wide range across datasets without manual tuning.

VLMs For Radiology Report Generation

In recent years, VLMs have shown remarkable capability in bringing visual and textual modelities, particularly in medical application such as automated medical radiology report generation (Mamdouh et al. 2025; Elboardy et al. 2025). These models typically build upon domain-adapted vision encoders and language decoders, integrating visual features with contextual linguistic knowledge to produce clinically coherent and detailed reports.

Despite the rapid evolution of vision encoders in natural image domains, there remains not sufficient dedicated architecture explicitly designed for medical imaging. To bridge this gap, recent studies have sought to adapt vision encoders to the medical domain. For instance, MedCLIP enhances the vision encoder component of VLMs for radiology generation by leveraging a decoupled image-text contrastive learning framework (Wang et al. 2022), effectively enriching visual-semantic alignment. By combining VLMs with radiologists, Flamingo-CXR represents a diagnostic system for automated radiology report generation that can produce reports that are clinically comparable to those written by human physicians (Tanno et al. 2024), while revealing common error patterns and opportunities for physician-AI collaborative workflows. In addition, some recent models enhance VLM-based medical report generation by incorporating organ mask to guide the model's attention toward clinically revelant regions. For example, the Complex Organ Mask Guided (COMG) model leverages multi-organ masks and prior disease knowledge during feature fusion (Gu et al. 2024), then improving cross modal consistency and generating more detailed and accurate radiology reports.

Methods

Dataset

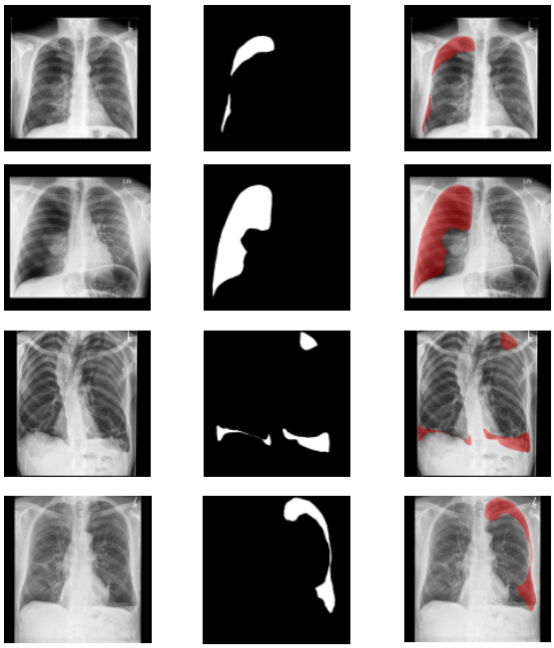
In this study, our experiments are implemented on the Chest X-ray Anonymized Dataset in Dunedin - Pneumothorax (CANDID-PTX), a private dataset for pneumothorax representing 19,237 chest radiographs (Feng et al. 2021). The age range is 16-101 years (mean=60.1 and STDEV=20.1). Among the whole dataset, 3,196 chest radiography represent pneumothorax cases, while the remaining images are normal cases. The data are acquired from three imaging devices manufactured by Philips, GE Healthcare and Kodak. The images are annotated using MD.ai platform with free form line marking. The image size is 1024×1024 pixels in DICOM format, and the segmentation ground truth is provided in Run-Length Encoding (RLE) format. In this study, we adopt controlled experiments setup to evaluate the fine-tuned VLM. Specifically, we randomly select 1000 samples from the dataset, 950 samples of them are served as training set, ensuring a diverse representation of image findings and report variations. The remaining 50 samples are served as independent test set to evaluate the model's generalization capability.

Data Pre-Processing

For the textual data, radiology reports were cleaned to remove sentences referring to previous examinations and follow-up comparisons, as such longitudinal information is irrelevant to generating new diagnostic reports. Non-diagnostic elements such as template headers or administrative remarks were also excluded, leaving only the essential clinical observations and impressions. For the imaging data, due to VLMs are mainly pretrained on natural images, all X-ray images were converted from DICOM to PNG format, while the segmentation masks were decoded from their Run-Length Encoding (RLE) representations into binary mask images in PNG format. Both the medical images and corresponding masks were included into structured Hugging Face dataset with initial resolution (e.g., 1024×1024) to provide more detailed information and ensure spatial consistency and pixel-wise alignment across modalities.

Image-Mask Fusion

In the image-mask fusion strategy, the segmentation mask is visually integrated with the original X-ray image through a transparent overlay with the transparency is set as 0.4.



(a) Original Image (b) Segmentation mask (c) Image-Mask fusion

Figure 1: Example of image-mask fusion.

The segmented lesion area are highlighted using semi-transparent color-coding, preserving the underlying anatomical structure while emphazing region of clinical interest. A corresponding prompt is designed to explicitly draw the model's attention to highlighted regions, encouraging localized reasoning and more precise diagnosis generation. Example are shown in Figure 1.

Prompt

The prompt we used to generate the report with/without segmentation guidance is shown correspondingly in Prompt 1 and 2, and the differences are marked by **bold**.

• **Prompt 1:** *"Consider that you are a professional radiologist with extensive experience and now you are treating a patient.* ***In the image, the overlapped area is pneumothorax.*** *Analyze the provided chest X-ray image and generate a comprehensive pneumothorax diagnostic report including findings and conclusions."*

• **Prompt 2:** *"Consider that you are a professional radiologist with extensive experience and now you are treating a patient. Analyze the provided chest X-ray image and generate a comprehensive pneumothorax diagnostic report including findings and conclusions."*

Data Pipeline

The data pipeline of this study is shown in Figure 2. After pre-processing, both the medical images and corresponding radiology reports are fused into one image and then prepared as structured Hugging Face dataset for VLMs along with the pneumothorax reports and prompts. The VLMs are subsequently fine-tuned to adapt to the domain-specific characteristic of pneumothorax diagnosis by leveraging Quantized Low-Rank Adaption (QLoRA) (Dettmers et al. 2023), while enables efficient parameter updates while reducing computational overhead. The fine-tuning process is orchestrated using the Supervised Fine-Tuning (SFT Trainer), ensuring that the models learn to align visual features, segmentation masks with textual patterns from expert handwriting reports. Finally, the generated reports will be evaluated across four established metrics to compare the quality of generated reports with radiologist-generated reports.

Fine-tuned Models

• **Gemma series** Gemma is a family of lightweight, state-of-the-art open models from Google DeepMind. Gemma 3 models are multimodal, handling text and image input and generating text output, with open weights for both pre-trained variants and instruction-tuned variants, well-suited for a variety of text generation and image understanding tasks, including question answering, summarization, and reasoning. (GemmaTeam 2025). MedGemma is a suite of open-source VLM based on Gemma 3 variants (Sellergren et al. 2025). Powered by MedSigLIP, a medically-tuned SigLIP encoder, these models are specialized for medical text and image comprehension through fine-tuning on diverse healthcare datasets, including chest X-rays and histopathology slides.

PaliGemma integrates the SigLIP vision encoder with the Gemma 2B language model, forming a unified framework for multimodal understanding and generation (Beyer et al. 2024). In its architecture, the SigLIP encoder, a pre-trained Vision Transformer optimized for visual representation, transforms input images into sequences of visual tokens. These visual tokens are subsequently combined with textual prompts, applying full attention across both image and text tokens.

• **Qwen-VL series** Qwen-VL is a series of VLMs designed by Alibaba Cloud (Wang et al. 2024) (QwenTeam 2025a) (QwenTeam 2025b). They introduce Naive Dynamic Resolution for adaptive visual tokenization, which enables the model to flexibly map the image of different resolutions into a dynamic number of visual tokens, and Multimodal Rotray Position Enbedding (M-ROPE), which decomposses positional ecnoding into textual, visual and video representation.

• **Phi-Vision series** Phi-Vision is a series of VLMs designed by Microsoft (Marah Abdin 2024). Both the Phi-3-Vision model and the Phi-3.5-Vision model are 4.2-billion-parameter multimodal language models capable of jointly processing visual and textual inputs. The Phi-3.5-Vision model integrates a CLIP ViT-L/14 image encoder with a phi-3.5-mini transformer decoder, enabling coherent text generation grounded in visual context. The model employs a dynamic cropping mechanism to effectively handle high-resolution and variable-aspect-ratio images by dividing them into spatial blocks whose token representations are concatenated. For multi-image tasks, it aggregates visual information by concatenating tokens across images, allowing flexible and efficient multi-modal understanding.

• **Llama-Vision series** Llama-Vision is a collection of multimodal LLMs developed by Meta, comprising pretrained and

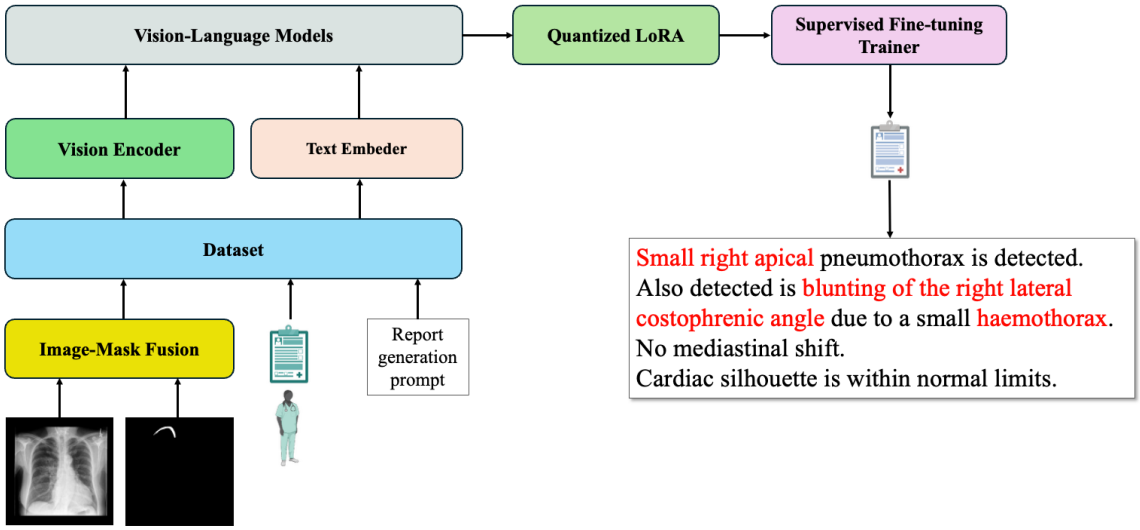


Figure 2: Pipeline for fine-tuning Vision Language Models.

instruction-tuned variants at 11B and 90B scale. Architecturally, Llama 3.2-Vision extends the Llama 3.1 text-only foundation and further aligns its behavior with human expectations of usefulness and safety through supervised fine-tuning and reinforcement learning (RL) with human's feedback (LlamaTeam 2024). To enable visual understanding, the model incorporates a dedicated, separately trained vision adapter composed of multiple cross-attention layers, which inject image encoder representations directly into the core language model, thereby enabling seamless integration of visual information into text generation.

To provide a clear overview of the models utilized in this study, Table 1 summarizes all fine-tuned models along with their corresponding configurations and training settings.

Evaluation Metrics Comparison between Different Segmentation Accuracy

In this study, we also employ nnUNet to obtain pneumothorax segmentation masks from chest X-ray images. To systematically investigate the influence of segmentation accuracy on report generation, we extracted 20 nnUNet's segmentation result masks for each Dice Score Coefficient (DSC) level ranging from 0.9 to 0.5, representing varying levels of mask precision. For each range, considering the insignificant difference between segmentation masks with little amount of difference in Dice score, in critical point, we select samples with an error range within 0.03 for its range, such as for critical point 0.8, we select samples with Dice score ranging from [0.77, 0.83] in its range. By comparing the reports generated from these predicted masks against the corresponding ground-truth radiology reports, we analyzed how progressive deviation in mask quality affects the diagnostic accuracy and clinical relevance of the generated reports. The example of the masks used for this comparative analysis are illustrated in Table 3.

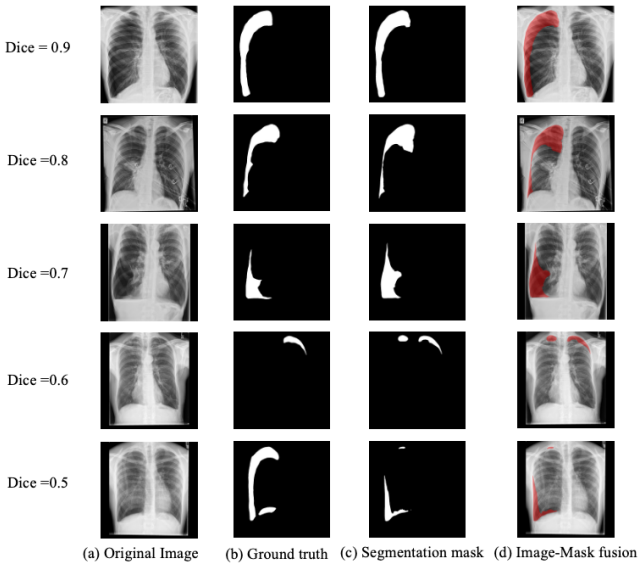


Figure 3: Examples of masks used for comparative analysis.

Patch-Level Activation Heatmap Derived From Visual Hidden States

Activation heatmap is a visualization technique that highlights spatial regions in an input image that contributes most to a model's internal activations, thereby providing interpretability of a VLM's decision-making process (Selvaraju et al. 2017). To investigate how segmentation guidance influences the internal visual representations of the model, we construct a patch-level activation heatmap derived from visual hidden states by using the 4B MedGemma model with-/without segmentation-guided fine-tuning.

To get the activation heatmap, on input image I, the visual encoder partitions it into N non-overlapping patches. Each

Developer	Model	Size (B)	Abbreviation
Google DeepMind	Gemma-3-4B-PT	4	GE04
	MedGemma-4B-IT	4	MG04
	MedGemma-27B-IT	27	MG27
Microsoft	Phi-3.5-vision-Instruct	4	PV3.5-04
	Phi-3-vision-128k-instruct	4	PV3-04
Alibaba Cloud	Qwen2-VL-7B-Instruct	7	QW2-07
	Qwen2.5-VL-7B-Instruct	7	QW2.5-07
	Qwen2.5-VL-32B-Instruct	32	QW2.5-32
	Qwen3-VL-8B-Instruct	8	QW3-08
	Qwen3-VL-32B-Instruct	32	QW3-32
Meta	Llama-3.2-11B-Vision-Instruct	11	LM11

Table 1: Vision-Language Models for fine-tuning classified by developer.

patch is encoded into a hidden representation

$$\mathbf{h}_i \in \mathbb{R}^d, \quad i = 1, \ldots, N, \tag{1}$$

where d denotes the hidden dimension of the visual encoder.

To quantify the representational strength of each patch, we compute the ℓ_2 norm of its hidden representation as a scalar activation score:

$$a_i = \|\mathbf{h}_i\|_2 . \tag{2}$$

For numerical stability and comparability across samples, the activation scores are normalized as:

$$\hat{a}_i = \frac{a_i}{\max_j a_j + \epsilon}, \tag{3}$$

where ϵ is a small constant.

The normalized patch-wise activation scores are then mapped back to the spatial layout of the image. Let $\mathrm{reshape}(\cdot)$ denote the operation that rearranges the patch sequence into a two-dimensional grid corresponding to the original patch layout:

$$\mathbf{A} = \mathrm{reshape}\left(\left\{ \frac{\|\mathbf{h}_i\|_2}{\max_j \|\mathbf{h}_j\|_2 + \epsilon} \right\}_{i=1}^{N} \right). \tag{4}$$

The final activation heatmap $\mathbf{A}$ reflects the spatial distribution of representational capacity within the visual encoder, indicating which image regions are encoded with stronger activation responses.

Results

Zero-shot Diagnosis Generation

In the zero-shot diagnosis generation setting, VLMs were directly evaluated without task-specific fine-tuning, in order to rigorously assess their inherent capability to generate diagnostic reports from chest radiographs. In our study, we do zero-shot pneumothorax report generation for two times by using both the original X-ray image and X-ray image after image-mask fusion as input Each model was provided with the same input image and prompt. The results are shown in Table 2. For X-ray images with/without segmentation anotation, because most of the evaluated VLMs were primarily pre-trained on natural image-text dataset and are not specifically optimized for medical domain, their zero-shot performance on radiologist task is inherently limited. In the meantime, for inference tasks MedGemma models use its predefined template, generating large amount of conclusions irrelevant with pneumothorax which leads to a significant degradation in accuracy.

Impact of Segmentation Guidance on VLMs

To rigorously evaluate the effect of segmentation guidance on biomedical report generation, we conducted a systematic and quantitative metric comparison across multiple VLMs, examining their performance both with and without segmentation guidance. The comparative results are summarized in Table 3. Incorporating the segmentation guidance consistently enhance the model performance across almost all evaluation metrics. Specifically, VLMs equipped with segmentation guidance achieved higher scores in all evaluation metrics, reflecting improved lexical fidelity and sentence-level coherence with the reference radiology reports. This improvement suggests that segmentation masks provide an explicit spatial prior, enabling the model to attend more precisely to diagnostically relevant regions and better localization of pathologies. Thereby, VLMs are able to facilitate more accurate alignment between textual and visual presentations. Our experimental results demonstrates that integrating segmentation information is an effective strategy for enhancing the clinical reliability and interpretability of VLMs in pneumothorax report generation.

Influence of Segmentation Accuracy

While segmentation information substantially enhances the quality of pneumothorax diagnostic-oriented report generation, the extent to which how segmentation accuracy affects report accuracy and quality remains an open question. To address this, in this section, we conducted a detailed analysis cross different segmentation levels (DSC), ranging from 0.9 to 0.5, and evaluated the corresponding report metrics, including ROUGE-1, ROUGE-2 and ROUGE-L, as illustrated Figures 4 and 5. Based on our experiments, the fine-tuned MedGemma model achieved the best performance overall metrics among the evaluated VLMs, therefore, it was used as the model for report generation in this part.

Model	Guidance	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-P
GE04	N/A	0.0940	0.0071	0.1541	0.0368	0.0929	0.7957
	+Seg	0.1323	0.0170	0.2079	0.0555	0.1176	0.8128
MG04	N/A	0.1431	0.0148	0.2094	0.0505	0.1112	0.8166
	+Seg	0.1792	0.0256	0.2464	**0.0699**	0.1489	0.8302
MG27	N/A	0.1497	0.0112	0.2233	0.0532	0.1145	0.8128
	+Seg	0.1559	0.0097	0.2280	0.0437	0.1129	0.8158
PV3.5-04	N/A	**0.2411**	**0.0325**	**0.2941**	**0.0743**	**0.1945**	**0.8594**
	+Seg	**0.2605**	**0.0337**	**0.2885**	0.0644	**0.1803**	**0.8545**
QW2-07	N/A	0.1378	0.0070	0.1874	0.0318	0.1234	0.8195
	+Seg	0.1517	0.0057	0.2246	0.0496	0.1298	0.8082
QW2.5-07	N/A	0.1207	0.0110	0.1879	0.0415	0.1089	0.8108
	+Seg	0.1420	0.0102	0.2137	0.0535	0.1263	0.8148
QW2.5-32	N/A	0.0943	0.0054	0.1636	0.0399	0.0982	0.8075
	+Seg	0.0980	0.0069	0.1642	0.0392	0.0958	0.8059
QW3-08	N/A	0.1376	0.0059	0.2162	0.0465	0.1203	0.8085
	+Seg	0.1142	0.0054	0.1945	0.0415	0.1091	0.8056
QW3-32	N/A	0.0898	0.0045	0.1601	0.0402	0.0938	0.8066
	+Seg	0.1049	0.0093	0.1860	0.0440	0.0973	0.8092
LM11	N/A	0.0830	0.0036	0.1032	0.0129	0.0735	0.7731
	+Seg	0.0863	0.0043	0.1082	0.0139	0.0832	0.7789

Table 2: Evaluation metrics comparison between VLMs for zero-shot generation task, **Bold** indicates top score, and underline indicates second rank score both with and without segmentation guidance.

The boxplots reveal that the influence of segmentation quality on report generation is consistent yet highly non-linear. When the segmentation accuracy remains relatively high (e.g., DSC$\geq$0.8), the ROUGE scores exhibit only marginal declines, suggesting that model can effectively leverage segmentation guidance to maintain coherent descriptions. In this regime, the VLMs can still accurately localize and interpret the pneumothorax regions, effectively capturing critical spatial and morphological cues. Consequently, the generated descriptions remain largely consistent with the reference reports in terms of diagnostic accuracy, clinical completeness, and linguistic coherence.

However, when the segmentation accuracy further deteriorates (e.g., DSC$\leq$0.7), all the ROUGE variants, particularly ROUGE-2 and ROUGE-L show a noticeable decline, indicating reduced semantic consistency and contextual integrity in the generated text. The degradation becomes more pronounced at DSC$\leq$0.6, where reports tend to contain incomplete lesion description. Under these conditions, the model tends to produce descriptions that exhibit partial lesion identification, underestimation of pneumothorax extent, or even misinterpretation of lesion laterality. Such discrepancies suggest that poor segmentation not only disrupts the spatial grounding between image and text but also weakens the model's ability to align visual abnormalities with appropriate clinical terminology. This observation highlights a critical dependency that while moderate segmentation errors can be tolerated without major loss of performance, maintaining high-quality and anatomically precise segmentation is essential for achieving clinically reliable, contextually coherent, and trustworthy automated radiology report generation.

Activation Heatmap

Samples of attention map for MedGemma is shown in Figure 6. By comparing activation heatmaps before and after segmentation-guided fine-tuning, we observe a systematic redistribution of visual representations. Specifically, the fine-tuned model exhibits stronger and more spatially coherent activations over the target anatomical regions, while activations in irrelevant background areas are relatively suppressed. This suggests that segmentation guidance changes the focus of the model toward encoding clinically meaningful regions more prominently, thereby improving visual grounding for downstream medical report generation.

Conclusion

This study proposed a segmentation-guided approach for pneumothorax report generation from chest radiographs, aiming to bridge the gap between visual understanding and clinical report generation. By explicitly integrating pixel-level segmentation masks into the report generation process, the model is guided to focus on pathologically meaningful regions. This targeted attention not only facilitates more precise localization and spatial localization of pneumothorax but also promotes the generation of of reports that exhibit improved consistency with expert radiological descriptions. The proposed approach effectively combines the fine-grained visual interpretability offered by segmentation with linguistic fluency and contextual reasoning of VLMs, thereby enabling a synergistic multimodal learning framework that enhances both descriptive precision and clinical reliability.

First, our experiments shows that the segmentation-guided fine-tuning helps the VLMs to learn the way to

Model	Guidance	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-P
GE04	N/A	0.3329	0.1615	0.4210	0.2288	0.3397	**0.8929**
	+Seg	**0.3835**	**0.1924**	**0.4661**	**0.2580**	**0.3621**	0.8912
MG04	N/A	0.3755	0.1646	0.4420	0.2187	0.3449	**0.8867**
	+Seg	**0.4167**	**0.1876**	**0.4586**	**0.2310**	**0.3525**	0.8833
MG27	N/A	0.4147	0.2039	0.4842	0.2738	0.3866	0.8866
	+Seg	**0.5023**	**0.2647**	**0.5369**	**0.3081**	**0.4430**	**0.8990**
PV3.5-04	N/A	0.3714	0.1339	0.4188	0.1694	0.3250	0.8953
	+Seg	**0.3915**	**0.1602**	**0.4372**	**0.1907**	**0.3458**	**0.8997**
PV3-04	N/A	0.3627	0.1368	0.4121	0.1666	0.3132	0.8857
	+Seg	**0.3740**	**0.1512**	**0.4332**	**0.1932**	**0.3443**	**0.8992**
QW2-07	N/A	0.3274	0.1255	0.3877	**0.1658**	**0.3008**	0.8726
	+Seg	**0.3533**	**0.1308**	**0.3920**	0.1599	0.2958	**0.8731**
QW2.5-07	N/A	0.3704	0.1250	0.3868	0.1399	0.2837	0.8637
	+Seg	**0.3754**	**0.1321**	**0.3878**	**0.1420**	**0.2916**	**0.8657**
QW2.5-32	N/A	0.3974	0.1794	0.4304	0.1858	0.3227	0.8861
	+Seg	**0.4278**	**0.1986**	**0.4473**	**0.1970**	**0.3529**	**0.8879**
QW3-08	N/A	0.3899	0.1778	0.4290	0.1916	0.3326	0.8830
	+Seg	**0.4195**	**0.1895**	**0.4423**	**0.1978**	**0.3432**	**0.8858**
QW3-32	N/A	0.4026	0.1513	0.4120	0.1662	0.3155	0.8739
	+Seg	**0.4482**	**0.2182**	**0.4691**	**0.2121**	**0.3749**	**0.8947**
LM11	N/A	0.3536	0.1245	0.3749	**0.1464**	0.2750	0.8629
	+Seg	**0.3872**	**0.1377**	**0.3856**	0.1451	**0.2772**	**0.8652**

Table 3: Evaluation metrics comparison between VLMs with/without segmentation guidance.

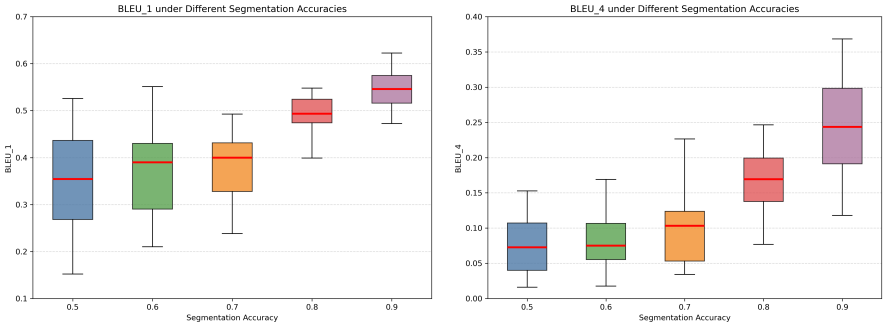


Figure 4: Comparison of BLEU metrics under varying mask qualities, demonstrating. From left to right, the plots represent BLEU-1, BLEU-4 comparisons across segmentation accuracies (DSC = 0.5–0.9).

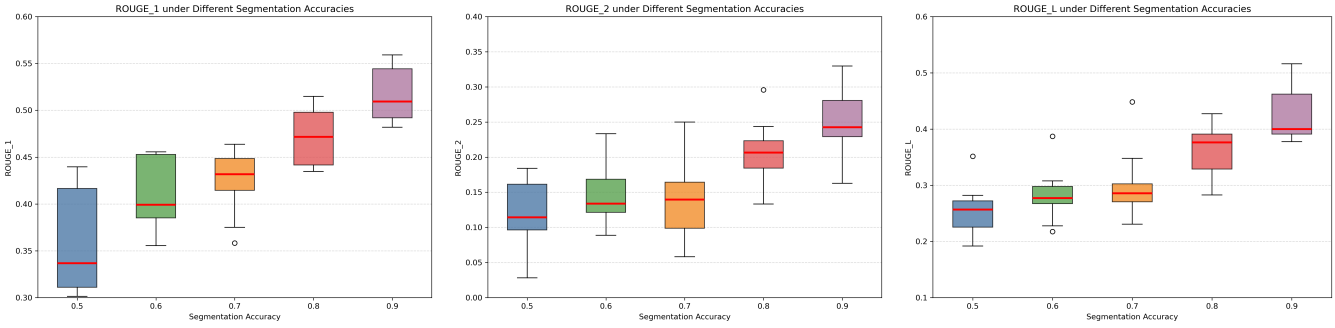


Figure 5: Comparison of ROUGE metrics under varying mask qualities, demonstrating. From left to right, the plots represent ROUGE-1, ROUGE-2, and ROUGE-L comparisons across segmentation accuracies (DSC = 0.5–0.9).

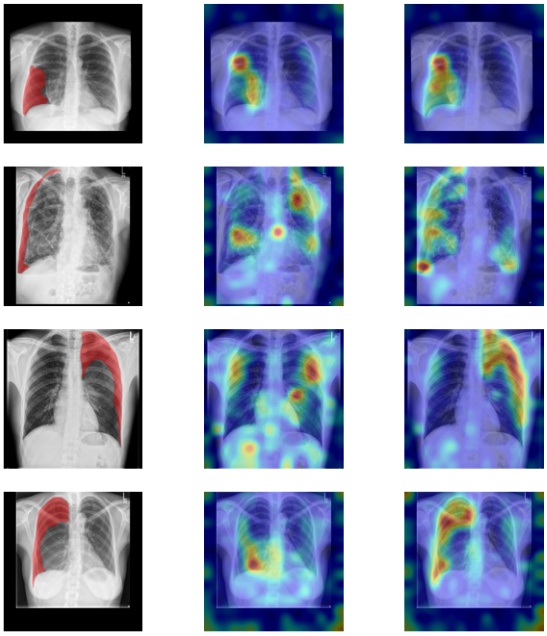


Figure 6: Comparison of activation heatmaps between VLMs with/without segmentation guidance.

write the pneumothorax report following the radiologist style. Extensive experiments demonstrated that the proposed segmentation-guided model outperformed the baseline VLMs across a range of quantitative metrics, including BLEU-1/4, ROUGE-1/2/L and BERTScore. Beyond numerical improvements, qualitative analyses revealed that segmentation guidance leads to more accurate lesion localization and a notable reduction in hallucianted findings, underscoring its role in improving factual alignment and diagnostic completeness. Importantly, this study highlights the broader potential of structured visual supervision inmedical report generation. By leveraging segmentation as an interpretable intermediate representation, our proposed approach not only improves the interpretability of visual–text alignment but also offers clinicians transparent insight into the decision process of the model. Such interpretability is essential for fostering clinical trust and regulatory acceptance for AI-assisted diagnostic system.

Furthermore, our proposed approach provides a scalable foundation for future research on explainable multimodal AI systems in medicine. Integrating segmentation-derived spatial priors with advanced generative architectures may support fine-grained reasoning about disease extent, severity and progression.

Limitation and Future Direction

Limitation

Although this study demonstrates that incorporating segmentation information can substantially enhance both the accuracy and clinical relevance of pneumothorax report generation, several limitations remain. First, the current approach primarily employs segmentation masks as auxiliary visual guidance to direct the model's attention toward pathological regions. Nevertheless, the alignment between visual features and textual representations is still established in a relatively implicit manner, without explicitly modeling the fine-grained correspondence between localized visual semantics and domain-specific linguistic components. Consequently, while the generated reports exhibit improved diagnostic precision, the alignment between image features and clinical terminology remains suboptimal. Furthermore, although quantitative evaluation metrics suggest enhanced performance, the clinical validity of these automatically generated reports must still be verified by experienced radiologists. In addition, the relatively limited size of the annotated dataset may constrain the generalizability of the model, underscoring the need for larger and more diverse datasets to robustly assess and further refine the proposed approach.

Moreover, in present study, only a subset of the dataset was utilized, and the experiments were conducted with relatively small-scale and middle-scale VLMs. In the future work, we plan to expand the number of training and evaluation samples to better represent the full spectrum of clinical representations, and to investigate the performance of larger, more capable VLM architectures, then improve the robustness and generalization ability of the model and provider deeper insights into the interactions between segmentation accuracy and automated report generation at scale.

Future Direction

In future work, we aim to further enhance cross-modal alignment by integrating medical named entity recognition (NER) and structured representation learning. Medical NER plays a critical role in bridging the gap between visual features and textual descriptions, as medical terminology is highly compositional and often constructed from well-defined roots, prefixes, and suffixes (e.g., "pneumo-," indicating air or lung, and "-thorax," indicating chest). Leveraging morphological analysis, such terms can be decomposed into constituent components, facilitating more precise and granular entity extraction. By combining segmentation-guided localization with linguistically informed NER, the model can explicitly link visual abnormalities observed in imaging data to semantically structured medical entities, thereby promoting a more interpretable and clinically meaningful report generation process. Furthermore, structured representation learning can capture hierarchical relationships among entities and their attributes, enabling the model to reason over both local image features and complex clinical concepts. This integration of visual, linguistic, and structural information is expected to improve the fidelity of automatically generated reports, reduce ambiguities in terminology usage, and provide a foundation for more reliable deployment in real-world clinical settings.

Acknowledgments

This work was supported by JST PRESTO Grant Number JPMJPR23P7, Japan and JST NEXUS Grant Number JP-

MJNX25C4, Japan. The authors would like to thank Sijing Feng from the Department of Radiology, Dunedin Hospital, Dunedin, New Zealand, for providing the CANDID-PTX dataset.

References

Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; Unterthiner, T.; Keysers, D.; Koppula, S.; Liu, F.; Grycner, A.; Gritsenko, A.; Houlsby, N.; Kumar, M.; Rong, K.; Eisenschlos, J.; Kabra, R.; Bauer, M.; Bošnjak, M.; Chen, X.; Minderer, M.; Voigtlaender, P.; Bica, I.; Balazevic, I.; Puigcerver, J.; Papalampidi, P.; Henaff, O.; Xiong, X.; Soricut, R.; Harmsen, J.; and Zhai, X. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2021. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. arXiv:2105.05537.

Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; Lungren, M. P.; Zhang, S.; Xing, L.; Lu, L.; Yuille, A.; and Zhou, Y. 2024. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97: 103280.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314.

Elboardy, A. T.; Khoriba, G.; al Shatouri, M.; Mousa, M.; and Rashed, E. A. 2025. Benchmarking vision-language models for brain cancer diagnosis using multisequence MRI. *Informatics in Medicine Unlocked*, 58: 101692.

Feng, S.; Azzollini, D.; Kim, J. S.; Jin, C.-K.; Gordon, S. P.; Yeoh, J.; Kim, E.; Han, M.; Lee, A.; Patel, A.; et al. 2021. Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence*, 3(6): e210136.

GemmaTeam. 2025. Gemma 3.

Gu, T.; Liu, D.; Li, Z.; and Cai, W. 2024. Complex Organ Mask Guided Radiology Report Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7995–8004.

Iqbal, B.; Hallifax, R.; and Rahman, N. M. 2025. Pneumothorax: An update on clinical spectrum, diagnosis and management. *Clinical Medicine*, 25(3): 100327.

Isensee, F.; Jaeger, P. F.; Kohl, S. A. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2): 203–211.

Kabil, A.; Khoriba, G.; Yousef, M.; and Rashed, E. A. 2025. Advances in medical image segmentation: A comprehensive survey with a focus on lumbar spine applications. *Computers in Biology and Medicine*, 198: 111171.

Li, X.; Li, L.; Jiang, Y.; Wang, H.; Qiao, X.; Feng, T.; Luo, H.; and Zhao, Y. 2025. Vision-Language Models in medical image analysis: From simple fusion to general large models. *Information Fusion*, 118: 102995.

LlamaTeam. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment Anything in Medical Images. *Nature Communications*, 15(1): 654.

Mamdouh, D.; Attia, M.; Osama, M.; Mohamed, N.; Lotfy, A.; Arafa, T.; Rashed, E. A.; and Khoriba, G. 2025. Advancements in Radiology Report Generation: A Comprehensive Analysis. *Bioengineering*, 12(7).

Marah Abdin, e. a. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*.

QwenTeam. 2025a. Qwen2.5-VL.

QwenTeam. 2025b. Qwen3 Technical Report. arXiv:2505.09388.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Singer, D. D.; Scott, H.; Khan, A.; Donnelly, A.; Singer, A. J.; Botwinick, I.; Jawa, R.; Mukhi, A.; Thode, H. C.; and Secko, M. 2025. Emergency Department Accuracy of Point-of-Care Ultrasound in Identifying Clinically Significant Pneumothorax in High-Severity Trauma Patients. *The Journal of Emergency Medicine*, 77: 140–151.

Tanno, R.; Barrett, D. G. T.; Sellergren, A.; Ghaisas, S.; Dathathri, S.; See, A.; Welbl, J.; Lau, C.; Tu, T.; Azizi, S.; Singhal, K.; Schaekermann, M.; May, R.; Lee, R.; Man, S. W.; Mahdavi, S.; Ahmed, Z.; Matias, Y.; Barral, J.; Eslami, S. M. A.; Belgrave, D.; Liu, Y.; Kalidindi, S. R.; Shetty, S.; Natarajan, V.; Kohli, P.; Huang, P.-S.; Karthikesalingam, A.; and Ktena, I. 2024. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2): 599–608.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3876–3887.