

# MAIRA-Seg: Enhancing Radiology Report Generation with Segmentation-Aware Multimodal Large Language Models

**Harshita Sharma** *Microsoft Health Futures, Cambridge, UK*

**Valentina Salvatelli** *Microsoft Health Futures, Cambridge, UK*

**Shaury Srivastav** *Microsoft Research India*

**Kenza Bouzid** *Microsoft Health Futures, Cambridge, UK*

**Shruthi Bannur** *Microsoft Health Futures, Cambridge, UK*

**Daniel C. Castro** *Microsoft Health Futures, Cambridge, UK*

**Maximilian Ilse** *Microsoft Health Futures, Cambridge, UK*

**Sam Bond-Taylor** *Microsoft Health Futures, Cambridge, UK*

**Mercy Prasanna Ranjit** *Microsoft Research India*

**Fabian Falck** *Microsoft Health Futures, Cambridge, UK*

**Fernando Pérez-García** *Microsoft Health Futures, Cambridge, UK*

**Anton Schwaighofer** *Microsoft Health Futures, Cambridge, UK*

**Hannah Richardson** *Microsoft Health Futures, Cambridge, UK*

**Maria Teodora Wetscherek**

*Microsoft Health Futures, Cambridge, UK; Department of Radiology, University of Cambridge and Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK*

**Stephanie L. Hyland** *Microsoft Health Futures, Cambridge, UK*

**Javier Alvarez-Valle** *Microsoft Health Futures, Cambridge, UK*

HARSHITA.SHARMA@MICROSOFT.COM

V.SALVATELLI@MICROSOFT.COM

T.SSRIVASTAV@MICROSOFT.COM

KENZA.BOUZID@MICROSOFT.COM

SHRUTHI.BANNUR@MICROSOFT.COM

DACOEELH@MICROSOFT.COM

MAXILSE@MICROSOFT.COM

SBOND TAYLOR@MICROSOFT.COM

MERCY.RANJIT@MICROSOFT.COM

FABIAN.FALCK@MICROSOFT.COM

FERNANDO.PEREZGARCIA@MICROSOFT.COM

ANTONSC@MICROSOFT.COM

HANNAH.MURFET@MICROSOFT.COM

A-MWETSCHERE@MICROSOFT.COM

STEPHANIE.HYLAND@MICROSOFT.COM

JAALVARE@MICROSOFT.COM

## Abstract

There is growing interest in applying AI to radiology report generation, particularly for chest X-rays (CXRs). This paper investigates whether incorporating pixel-level information through segmentation masks can improve fine-grained image interpretation of multimodal large language models (MLLMs) for radiology report generation. We introduce *MAIRA-Seg*, a segmentation-aware MLLM framework designed to utilize semantic segmentation masks alongside CXRs for generating radiology reports. We train expert segmentation models to obtain mask pseudolabels for radiology-specific structures in CXRs. Subsequently, building on the architectures of MAIRA, a CXR-specialised model for report generation, we integrate a trainable segmentation tokens extractor that leverages these mask pseudolabels, and employ mask-aware prompting to generate draft radiology reports. Our experiments on the publicly available MIMIC-CXR dataset show that MAIRA-Seg outperforms non-segmentation baselines. We also investigate set-of-marks prompting with MAIRA and find that MAIRA-Seg consistently demonstrates comparable or superior performance. The results confirm that using segmentation masks enhances

the nuanced reasoning of MLLMs, potentially contributing to better clinical outcomes.

**Keywords:** Semantic segmentation, multimodal large language models, chest x-rays, radiology report generation

**Data and Code Availability** We use the MIMIC-CXR dataset (Johnson et al., 2019a,b) and CXR segmentation datasets which are all publicly available, referring to Section 3.1 and Section 4.1 for further details. At present, we do not release the code.

**Institutional Review Board (IRB)** Proposed use of public datasets was reviewed by home institution. Under policy, use of de-identified public datasets is classified as Not Human Subjects Research [per 45§46.102(e)(1)(ii), 45§46.102(e)(5)]. Guidance and data reflection questions are provided to researchers including considerations to support representativeness, transparency and intended use.

## 1. Introduction

Radiology report generation involves the automated creation of free-text draft reports from medical images (Liu et al., 2019). Research indicates that ap-

plying Artificial Intelligence (AI) in this area could significantly enhance radiology workflows and clinical outcomes (Huang et al., 2023; Yildirim et al., 2024). There is increasing interest in using multi-modal large language models (MLLMs) for chest X-ray (CXR) report generation, demonstrating impressive performance (Tu et al., 2024; Chaves et al., 2024; Zhou et al., 2024; Bannur et al., 2024; Hyland et al., 2024). However, current MLLMs often neglect the integration of pixel-level inputs alongside images for generating radiology reports, limiting their region-based and fine-grained image interpretation capabilities. This is particularly significant in the biomedical domain, where a single medical image can contain multiple subtle findings, nuanced structures and relevant context representing the regions of interest (ROI). This gap presents an opportunity to enhance MLLM outputs by incorporating segmentation masks, which we aim to explore in this paper. We hypothesize that providing localized pixel-level details alongside images can enhance MLLM’s perceptual and reasoning abilities for biomedical applications like radiology report generation.

We propose *MAIRA-Seg*, a segmentation-aware MLLM that utilizes fine-grained mask features from semantic medical image segmentation alongside CXR input images to generate draft radiology reports. We build upon the model architectures and training method from the Multimodal AI for Radiology Applications (MAIRA) series of MLLMs (Bannur et al., 2024; Hyland et al., 2024). To the best of our knowledge, ours is the first work leveraging semantic image segmentation for instruction tuning MLLMs for CXR report generation. By integrating pixel-level knowledge in the form of segmentation and mask-aware information into the prompt instructions of the MLLM, we aim to improve the pixel-wise visual understanding and enhance the quality and accuracy of draft radiology reports generated from CXRs.

**Contributions** Our contributions are: 1) We propose MAIRA-Seg, a segmentation aware framework for radiology report generation. Semantic segmentation masks generated from expert models are integrated into the MLLM input using a segmentation tokens extractor, enabling fine-grained supervision along with CXR images. 2) We train MAIRA-Seg with segmentation masks of multiple anatomical structures, support devices, and pathological regions. By incorporating semantic segmentation as additional visual inputs along with single or multiple CXR views,

we achieve superior performance in radiology report generation. We demonstrate notable quantitative and qualitative improvements of MAIRA-Seg over the non-segmentation baseline models. 3) We investigate an additional general-domain method for leveraging segmentation as MLLM inputs: set-of-marks (SoM) prompting (Yang et al., 2023). Here, visual marks (e.g., contours) are directly superimposed on the image for visual instruction tuning. For SoM, we observe improvements over the non-segmentation baselines and comparable performance to MAIRA-Seg.

## 2. Related Work

**Multimodal LLMs in Radiology** Recent advances in AI for generating free-text radiology reports suggest improvements in operational efficiency, reduction in radiologist workloads, and enhancement of patient care quality (Huang et al., 2023; Yildirim et al., 2024; Liu et al., 2019). As a result, there has been a growing research interest in the generation of free-text, narrative-style reports from radiology images (Sloan et al., 2024). Specifically, MLLMs have been increasingly explored and have demonstrated a promising performance for radiology report generation. Recent MLLMs in the literature encompass generalist biomedical models for multiple imaging modalities (Zhou et al., 2024; Tu et al., 2024; Yang et al., 2024) and specialist radiology domain models (Hyland et al., 2024; Bannur et al., 2024; Li et al., 2024; Bai et al., 2024; Chaves et al., 2024). For this work, we build on the CXR-specific MAIRA framework (Hyland et al., 2024; Bannur et al., 2024), as it has demonstrated competitive CXR report generation performance over prior works such as LLaVA-Rad (Chaves et al., 2024), MedPalm-M (Tu et al., 2024) and MedVersa (Zhou et al., 2024).

### Segmentation to Prompt Multimodal LLMs

Recent work in the general domain shows that using visual prompts (e.g. bounding boxes, markers, segmentation masks) with images for MLLM visual instruction tuning can enhance their visual perception capabilities (Wu et al., 2024). Specifically, among methods using segmentation masks, the set-of-marks (SoM) prompting method (Yang et al., 2023) uses off-the-shelf interactive segmentation models to partition the input image into semantically meaningful regions using “marks” (e.g., bounding boxes, contours, alphanumeric marks). They query GPT-4V using such marked images and observe that it can

provide visually grounded zero-shot responses. Additionally, [Yan et al. \(2024\)](#) presents a method to enhance SoM prompting ability of existing MLLMs by using augmented prompts that incorporate mark information. Datasets of SoM-augmented images and prompts are curated for fine-tuning MLLMs. We conduct ablations and comparative analysis on SoM prompting with MAIRA-Seg and report our key findings. Another approach, the Osprey method ([Yuan et al., 2024](#)), uses a mask-aware visual extractor to obtain segmentation tokens interleaved with image tokens to prompt the MLLM, which is fine-tuned using a curated visual instruction tuning dataset. We adapt the Osprey architecture for the radiology domain (details in Section 3). In contrast to Osprey and SoM, we generate online mask-aware prompts without the need to generate new instruction tuning datasets for generating the radiology reports.

For radiology images, recent work ([Denner et al., 2024](#)) shows that incorporating visual prompts like arrows, circles, and contours for BiomedCLIP models significantly improves lung nodule malignancy classification metrics in CXRs. Another related study ([Zhao et al., 2023](#)) uses the Segment Anything Model (SAM) ([Kirillov et al., 2023](#)) to segment meaningful ROI of the image and a supervised contrastive learning method showing promising performance for the report generation task. In our knowledge, ours is the first paper leveraging segmentation masks in MLLMs to improve their nuanced visual understanding for radiology report generation.

### 3. Methodology

The MAIRA-Seg architecture and method are demonstrated in Figure 1. We first train structure-specific expert models for segmenting multiple CXR structures. We use these models to generate segmentation masks for the CXR images and feed these as pseudolabel inputs to the MLLM for training or inference. The masks along with image encoder features are then used to train a segmentation tokens extractor based on the Osprey architecture ([Yuan et al., 2024](#)) that generates two additional segmentation tokens (mask token, spatial token) for each individual mask. We investigate methods to integrate these tokens into the large language model (LLM)’s input, and use interleaved segmentation tokens with text and image tokens. We augment the input prompts on the fly using the available mask information, without the overhead of curating new instruction tuning datasets to train

the MLLM. We describe the individual components of the proposed method in the following sections.

#### 3.1. Expert Models for Semantic Medical Image Segmentation

We leverage expert semantic segmentation models trained for segmenting multiple structures in CXRs, and predict the corresponding semantic segmentation masks in the input radiology image. [Pérez-García et al. \(2024\)](#) empirically evaluate multiple segmentation models for CXR structures; the best performance is achieved by EfficientNet-UNet, a U-Net based on the EfficientNet backbone ([Tan and Le, 2020](#)), as such, we use this architecture for our expert models.

For semantic segmentation, we select structures that are relevant to common pathological observations in CXRs (e.g. CheXpert pathological findings ([Irvin et al., 2019](#))), and can improve the pixel-level understanding and reasoning capabilities of the trained MLLM for CXR report generation. We group these structures into three main categories: anatomical, pathological and support devices (Section A.2, Table 4). We note that the trained expert semantic segmentation models perform adequately for the respective structures (Section A.2, Table 5).

There has been an influx of general-purpose, prompt-based (bounding box, points, text) segmentation models in the biomedical domain, for instance, MedSAM ([Ma et al., 2024](#)) and BiomedPARSE ([Zhao et al., 2024](#)). Compared to a U-Net, these prompt-based models require additional input prompts to segment the structures of interest in the CXR. Since we strive for a fully automated system for CXR report generation without relying on input prompts, we used the ‘segment everything’ approach of these models, and during our initial experiments, we found that these models were sub-optimal compared to domain- and structure-specific U-Nets. Hence, general-purpose, prompt-based segmentation models may not yield precise segmentation for clinically critical tasks like draft radiology report generation.

#### 3.2. Leveraging Semantic Image Segmentation in MLLMs

**MAIRA Architecture** We leverage the MAIRA framework ([Hyland et al., 2024](#); [Bannur et al., 2024](#)) to experiment with semantic segmentation as additional visual inputs along with CXR images for generating draft reports. MAIRA MLLMs use a pretrained CXR-specific image encoder (RAD-DINO) based on

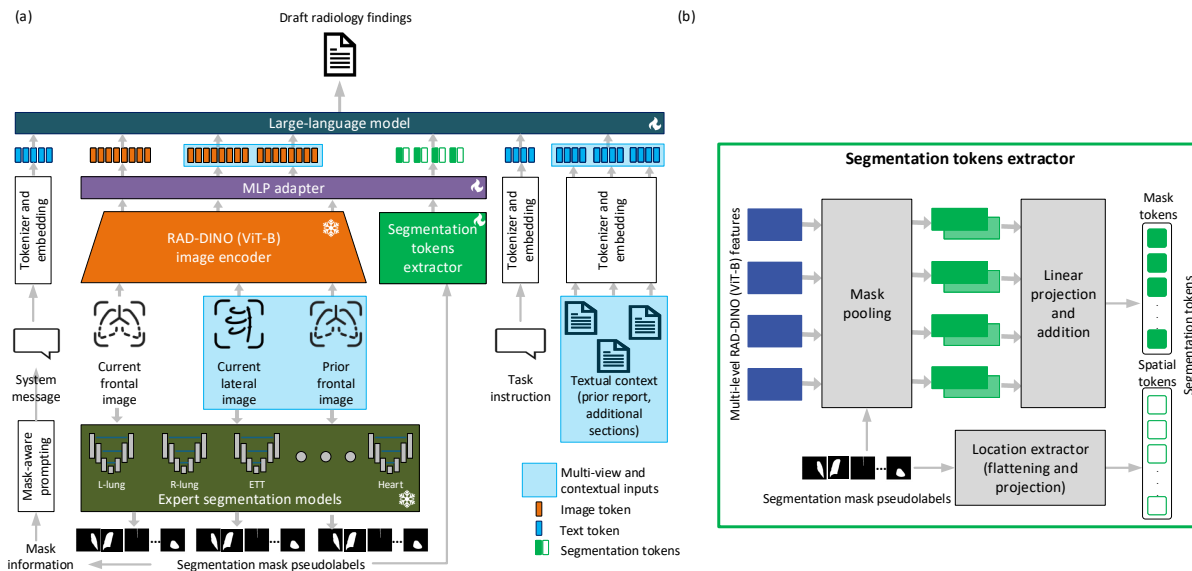


Figure 1: (a) MAIRA-Seg model architecture. Multi-view and textual context inputs are shown in blue boxes (Bannur et al., 2024). (b) Segmentation tokens extractor architecture based on Yuan et al. (2024).

a ViT-B architecture (Pérez-García et al., 2024) and a pretrained LLM Vicuna-7B v1.5 (Chiang et al., 2023), and align these using a 4-layer multi-layer perceptron (MLP) adapter based on the LLaVA framework (Liu et al., 2023b,a). The CXR-specific RAD-DINO (Pérez-García et al., 2024) image encoder is pretrained following the self-supervised DINOv2 approach (Oquab et al., 2024); it is kept frozen during the MLLM training. We use the image encoder checkpoint from Bannur et al. (2024). The LLM and adapter are fine-tuned in a single stage to generate the draft radiology report.

**MAIRA-Seg Architecture** MAIRA-Seg follows the MAIRA architecture for leveraging image and text inputs (Figure 1(a)), with some simplifications aimed at reducing the computational needs. We replace the LLM with Phi-3-mini consisting of 3.8B parameters (Abdin et al., 2024), which demonstrated performance close to Vicuna-7B in recent work (Srivastav et al., 2024). In addition, we do not apply GPT augmentations at training time. Also, we don’t perform grounding in our experiments as in Bannur et al. (2024). As detailed in Section 4.1, the amount of data we train on differs.

We experiment with two flavours of MAIRA-Seg. A single-view model *MAIRA-Seg-Frontal* that takes as input only the frontal chest x-ray and their corresponding mask and reports; and a multi-view model *MAIRA-Seg-Multi* that takes as input also priors and laterals, if available, and the corresponding masks. In the MAIRA-Seg-Multi model we also incorporate additional ‘textual context’ (prior report, indication, technique and comparison sections of the current report) in the text prompts as in (Bannur et al., 2024).

In order to leverage the mask inputs, we incorporate into the MAIRA framework, a trainable segmentation tokens extractor module based on the Osprey method that uses a visually-aware mask extractor (Yuan et al., 2024). The segmentation tokens extractor utilizes the output image features from RAD-DINO (ViT-B architecture) Pérez-García et al. (2024) and individual segmentation masks to generate visually-aware mask and spatial tokens (Yuan et al., 2024). A mask token (for one structure) is obtained as a result of mask pooling operation between image and mask features followed by linear projections, addition and MLP, to capture mask-level image features. A spatial token is obtained via flattening and linear projection of the



segmentation mask only which extracts the spatial location of the structure of interest. We collectively refer to the mask and spatial tokens as ‘segmentation tokens’. Since this approach involves the extraction of only two additional segmentation tokens per structure, it is scalable to extend to multiple structures and hence masks in the image – a common scenario found in biomedical images with multiple regions of interest, such as CXRs. This flexibility makes the approach suitable for fixed computational budgets and context lengths of LLMs.

A major difference from the Osprey method (Yuan et al., 2024) is that we use a transformer-based image encoder RAD-DINO (ViT-B backbone) in contrast to a convolutional neural network (CNN)-based CLIP vision encoder used in Yuan et al. (2024). We empirically investigate four design considerations to integrate the the segmentation tokens extractor with MAIRA MLLMs: 1) the interchangeability of linear and MLP layers within the original mask extractor, 2) embedding dimension sizes, 3) layer output indices from the ViT-B based image encoder, and 4) the number of segmentation tokens used to prompt the LLM. We select linear layers for both segmentation tokens, smaller embedding dimension of 768 (opposed to 1024 in the Osprey method), and layer indices [2,4,6,8] for ViT-B (in contrast to first four layers of the convolutional encoder in Osprey method). We use two segmentation tokens, namely one token for the mask information and one for spatial information (similar to Osprey method), opposed to fewer (1 token) or more (4 or 8 tokens). We propose the architecture of the segmentation tokens extractor as shown in Figure 1(b).

The segmentation tokens extractor, the MLP adapter and LLM are all finetuned, while the RAD-DINO image encoder is frozen during the MLLM training. More implementation details for MAIRA-Seg models are in Section C.

**Tokenization** The segmentation tokens retrieved from the segmentation tokens extractor are used to prompt the LLM along with the image and text tokens. We explore multiple approaches to incorporate segmentation tokens as input of the LLM, including directly concatenating all image and segmentation tokens, concatenating segmentation tokens in addition to the image tokens, and using separate segmentation tokens for individual structures in the image. We report these in Table 2. Best outcomes are achieved using separate segmentation tokens for individual structures in the image where the structure-specific tokens

interleave with the image and text tokens in the input prompt. Every segmentation token is not always used in the text prompt, as the input tokens are only added when a positive segmentation mask is available. Here, positive masks are defined as binary masks with at least one ‘1’ pixel. The rest of the tokenization process follows the method of MAIRA framework for MAIRA-Seg-Frontal (Hyland et al., 2024) and MAIRA-Seg-Multi (Bannur et al., 2024), respectively.

**Mask-aware Prompting** We perform online mask-aware prompting using input mask information, i.e. the structure names when a positive mask is available, followed by the corresponding segmentation tokens. This strategy helps us to quickly prototype without the need to curate new instruction tuning datasets to train the MLLM. The prompt format and ordering is the following: system message, current frontal image and tokens, positive mask names and tokens, current lateral image tokens, positive mask names and tokens, prior frontal image, positive mask names and tokens, instruction, textual context (prior report, other sections). Lateral image, prior image and textual context are added in the MAIRA-Seg-Multi prompt, similar to Bannur et al. (2024). An example of a text prompt of the MAIRA-Seg-Multi model for a study with current and prior frontal images and corresponding positive masks is: ‘You are an expert radiology assistant tasked with interpreting a chest X-ray study. Given the current frontal image <Image>, left lung mask <LLseg>, right lung mask <RLseg>, endotracheal tube mask <ETTseg>, heart mask <Heartseg> and the prior frontal image <PriorImage>, prior left lung mask <priorLLseg>, prior right lung mask <PriorRLseg>, prior heart mask <PriorHeartseg>, provide a description of the findings in the radiology study in comparison to the prior frontal image. Where segmentation masks are provided to highlight specific image regions...<textual context>’.

## 4. Experimental Setup

We address the task of generating the main body section *Findings* of the text report accompanying a chest X-ray. The task is identical to the radiology report generation task reported in Hyland et al. (2024). We perform experiments for single-and multi-view MAIRA-Seg and SoM, and compare with their respective baselines.

#### 4.1. Dataset and Evaluation

We perform the MLLM report generation experiments using the MIMIC-CXR dataset (Johnson et al., 2019a,b) hosted on PhysioNet (Goldberger et al., 2000). This dataset from the Beth Israel Deaconess Medical Center in Boston comprises a total of 377,110 DICOM images across 227,835 studies. Each imaging study is accompanied by a report. For each report, we extract the *Findings* section using the official MIMIC-CXR codebase<sup>1</sup>.

We train all models on the training split of MIMIC-CXR and report results on the official MIMIC-CXR test split using standard lexical and clinical metrics (Banmur et al., 2024). We also report additional Mask-Relevant (MR) clinical metrics (macro and micro F1-MR) on the CheXpert pathological findings relevant to the segmented structures, namely ‘Lung Opacity’, ‘Cardiomegaly’, ‘Pneumothorax’, ‘Support Devices’, ‘Pleural Effusion’. These pathological findings are directly correlated with the input segmentation masks (Section A.2, Table 4). It is worth noting that our macro and micro F1-MR are different from the standard macro and micro F1-5 (Miura et al., 2021) due to the difference in the selected pathological findings for analysis. We select BLEU-4, RadCliQ, Macro F<sub>1</sub>-MR, Micro F<sub>1</sub>-MR and Radfact/logical\_F1 for further analysis in the paper, with additional metrics in Appendix tables. We present the median along with 95% confidence intervals calculated from 500 bootstrap samples. Bold indicates best performance for that metric, or overlapping CIs with best, compared to the baselines. All metrics are higher is better except where ‘↓’ indicates lower is better. CheXpert F1 metrics are computed based on CheXbert labeller outputs (Hyland et al., 2024). We also present qualitative results in Figures 3, 6, 7 and 8, which were reviewed by a board-certified radiologist.

#### 4.2. Baselines and SoM Prompting

We compare MAIRA-Seg against baseline models, trained solely on input CXR images without using semantic segmentation masks, referred to as *MAIRA-Frontal* and *MAIRA-Multi*.

Moreover, we explore the set-of-marks (SoM) prompting method (Yang et al., 2023, 2024) that uses “marks” derived from segmentation overlaid on input images to visually prompt the MLLMs and has shown promising results in the general domain.

1. [https://github.com/MIT-LCP/mimic-cxr/blob/master/txt/section\\_parser.py](https://github.com/MIT-LCP/mimic-cxr/blob/master/txt/section_parser.py)

For this purpose, we create grayscale contours and alphanumeric marks using the segmentation mask pseudolabels and overlay these on the CXR images (example in Section A.1, Figure 4). We demonstrate the usability of the SoM prompting for the radiology report generation task through ablations on contours and alphanumeric marks, and the plain prompt and augmented prompt approaches (Section A.1, Table 3). For the augmented prompts, we use the online mask-aware augmentations as explained in Section 3.2. We also list all masks sequentially with the corresponding mark numbers at the end of the prompt (Yang et al., 2024). We find that using contours and alphanumeric marks with augmented prompts performed the best in our use-case (Section A.1, Table 3).

## 5. Results

We report the results for the single view and multi-view inputs in Table 1, and extended tables with additional lexical and clinical metrics in Section A.3, Tables 6 and 7 to supplement the key results. We also present the stratified F1-score for the CheXpert MR pathological findings, along with support for each multi-label finding in Figure 2 (extended 14 CheXpert pathologies in Section A.3, Figure 5). Results for ablations are presented in Table 2. We present one qualitative example in Figure 3, with more examples in Section B.

### 5.1. MAIRA-Seg-Frontal

**Quantitative Analysis** From Table 1, We observe that for the single view experiments, leveraging segmentation along with input images to prompt the MLLM improves report generation performance. MAIRA-Frontal is outperformed by MAIRA-Seg-Frontal in all the clinical metrics. We observe comparable results for MAIRA-SoM-Frontal, with significant improvements over baseline in 3 out of 4 clinical metrics, but MAIRA-Seg-Frontal is superior for extended clinical metrics as shown in Table 6. For the BLEU-4 lexical metric, we do not find significant differences, which aligns with our expectation since we aim at enhancing visual understanding capabilities of the MLLM, indicated by the clinical metrics. When further investigating the F1-MR stratified scores, in Figure 2(a), we find that MAIRA-Seg-Frontal outperforms MAIRA-Frontal on all five mask-relevant pathological findings, with significant gains in support devices, lung opacity and cardiomegaly.

Table 1: Experimental results for single view and multi-view setup on the official MIMIC-CXR test split. We compare the SoM and MAIRA-Seg methods against the non-segmentation baselines. Bold means superior to baselines (i.e., medians do not fall into mutual CIs). F1-MR\* are F1 scores on the mask-relevant CheXpert pathological findings, namely, ‘Lung Opacity’, ‘Cardiomegaly’, ‘Pneumothorax’, ‘Support Devices’, ‘Pleural Effusion’.

Method	BLEU-4	RadCliQ(↓)	Macro F <sub>1</sub> -MR *	Micro F <sub>1</sub> -MR *	RadFact/logical_fl
MAIRA-Frontal	14.2 [13.7, 14.8]	3.19 [3.15, 3.22]	54.1 [51.2, 56.2]	61.4 [60.0, 62.7]	42.4 [41.5, 43.6]
MAIRA-SoM-Frontal	14.6 [14.1, 15.2]	<b>3.14</b> [3.10, 3.18]	55.3 [53.1, 57.8]	<b>63.1</b> [61.9, 64.4]	<b>44.0</b> [43.0, 45.1]
MAIRA-Seg-Frontal	14.5 [14.0, 15.1]	<b>3.11</b> [3.08, 3.15]	<b>59.2</b> [56.7, 61.7]	<b>65.4</b> [64.1, 66.7]	<b>44.7</b> [43.8, 45.8]
MAIRA-Multi	19.5 [19.0, 20.1]	2.90 [2.86, 2.94]	52.5 [50.0, 55.0]	55.6 [54.2, 57.0]	45.4 [44.2, 46.5]
MAIRA-SoM-Multi	20.3 [19.8, 20.8]	<b>2.81</b> [2.77, 2.85]	53.9 [51.5, 56.3]	<b>59.9</b> [58.6, 61.1]	<b>47.4</b> [46.3, 48.6]
MAIRA-Seg-Multi	19.6 [19.1, 20.1]	<b>2.82</b> [2.78, 2.86]	55.9 [53.6, 58.7]	<b>60.9</b> [59.6, 62.5]	<b>47.1</b> [46.1, 48.3]

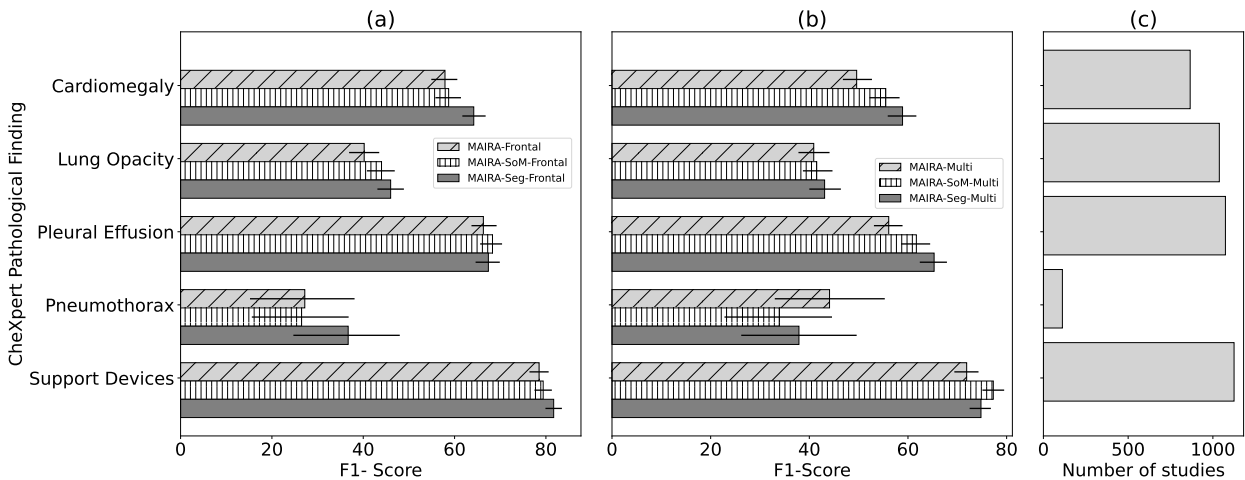


Figure 2: Stratified F1-scores for the five mask-relevant findings comparing the respective baselines, SoM prompting and MAIRA-Seg models for (a) single view (b) multi-view experiments (c) Support for each pathological finding in the MIMIC-CXR test set.

**Qualitative Analysis** Figure 3 reveals that MAIRA-Frontal omits the detail of lungs being “hyperinflated” in the draft report, however, this is correctly predicted by MAIRA-Seg-Frontal, probably due to enhanced visual understanding using the existing lung masks. Interestingly, in Section B, Figure 6(b)-(c), we find that the tip locations of tubes are correctly predicted by MAIRA-Seg-Frontal in contrast to MAIRA-Frontal, which shows the importance of adding segmentation masks for fine-grained tubular structures. Moreover, we find that the accuracy of masks affects the generated report outcomes, reflected in Section B, Figure 7(a) where overlapping devices lead to over-segmentation and predicting “distal SVC”

in place of “mid SVC” as the tip location, suggesting that more accurate segmentation masks can lead to more precise report generation.

**Ablations** We explore multiple approaches to incorporate segmentation tokens as input of the LLM. These include 1) concatenating image and segmentation features in unified image tokens, where the image and text tokens are interleaved (DC: direct concatenation); 2) concatenating segmentation tokens from all the structures in the image, where the combined segmentation tokens are interleaved with the image and text tokens (CS: Combined segmentation tokens); 3) using separate segmentation tokens for individual

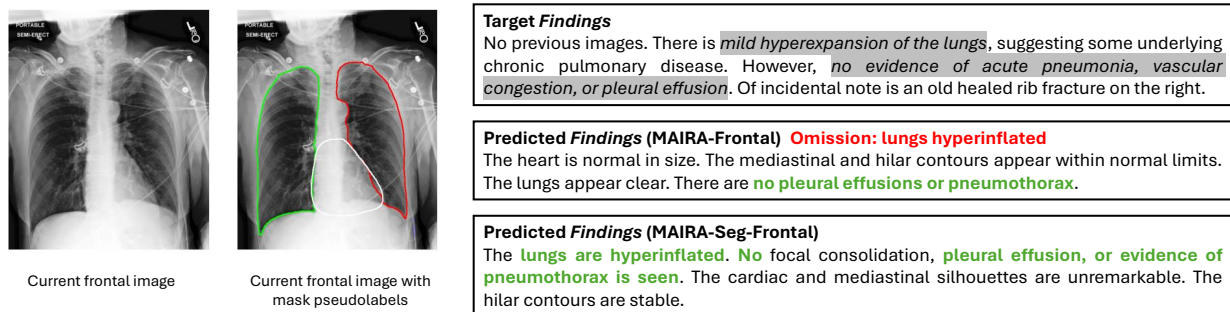


Figure 3: Qualitative result for an example in the MIMIC-CXR test set, showing target and predicted *Findings* using MAIRA-Frontal and MAIRA-Seg-Frontal. Mask pseudolabels are shown overlaid on the CXR image for illustrative purposes (corresponding masks are used to obtain segmentation tokens).

Table 2: Ablation of strategies for incorporating segmentation tokens on MIMIC-CXR validation split. (DC: direct concatenation, CS: combined segmentation tokens, SS: structure-specific segmentation tokens, NS: no segmentation tokens).

Method	MAIRA-Frontal	Ablation 1	Ablation 2	Ablation 3	Ablation 4	Ablation 5
<b>Concatenation strategy</b>	N/A	SS	NS	DC	CS	SS
<b>Augmented mask-aware prompts?</b>	No	No	Yes	Yes	Yes	Yes
BLEU-4	17.5	18.2	18.4	17.9	18.4	<b>18.7</b>
RadCliQ (↓)	2.72 [2.70, 2.75]	2.75 [2.73, 2.78]	2.72 [2.70, 2.75]	2.75 [2.72, 2.78]	2.73 [2.70, 2.75]	<b>2.71</b> [2.68, 2.73]
Macro F <sub>1</sub> -14	33.8 [32.8, 34.8]	34.1 [33.1, 35.3]	33.6 [32.7, 34.5]	33.4 [32.3, 34.4]	35.2 [34.3, 36.2]	<b>37.0</b> [35.8, 38.0]
Micro F <sub>1</sub> -14	51.9 [51.0, 52.7]	51.6 [50.8, 52.4]	52.2 [51.4, 53.0]	50.8 [50.0, 51.6]	53.1 [52.3, 53.9]	<b>54.0</b> [53.2, 54.8]

structures in the image, where the structure-specific tokens interleave with the image and text tokens (SS: structure-specific tokens). Results are depicted in Table 2 for the MIMIC-CXR validation split. We observe that the third strategy (SS) gives the best results and we select this strategy for introducing segmentation tokens at the input of the LLM. We also perform ablations on the impact of segmentation tokens and mask-aware prompting (Table 2, third and fourth columns), and observe improvements from the baseline for both cases, highlighting the importance of both sources of information to train the MLLM.

## 5.2. MAIRA-Seg-Multi

**Quantitative Analysis** In Table 1, we observe significant improvements over the MAIRA-Multi for 3 out of 4 clinical metrics using both MAIRA-SoM-

Multi or MAIRA-Seg-Multi to leverage segmentation in MLLMs. Macro F1-MR improvements are not significant in both cases, which could be attributed to data imbalance.

We find MAIRA-SoM-Multi is closer in performance to MAIRA-Seg-Multi, in contrast to the single-view experiments. This can be due to the fact that our architecture choices are based on ablation and tuning experiments using the single view counterparts (fewer input tokens to the LLM). Also, using a larger number of input tokens in MAIRA-Seg-Multi with the same-sized LLM could lead to more complex interactions, unlike MAIRA-SoM-Multi, which maintains a constant number of image tokens. Thus, we report MAIRA-Seg-Multi results as proof-of-concept with significant gains over MAIRA-Multi and consistent findings with MAIRA-Seg-Frontal, however, there is potential for enhancing the MAIRA-Seg-Multi archi-



ture for optimal performance which we leave for future work. For the pathology-stratified F1-MR scores in Figure 2(b), we observe significant improvements for MAIRA-Seg-Multi from the non-segmentation baseline for most relevant pathological findings such as support devices, cardiomegaly and pleural effusion.

**Qualitative Analysis** Observing the qualitative results, in Section B, Figure 8(a), MAIRA-Seg-Multi and MAIRA-Multi correctly mention stable cardiomegaly, bilateral pleural effusions and left retrocardiac opacity. However, MAIRA-Multi hallucinates a right-sided PICC line, which is not mentioned by MAIRA-Seg-Multi, also not detected as a mask by the expert segmentation model in current or prior images. Moreover, MAIRA-Multi hallucinates edema. Improvement of aeration is omitted in both predicted draft reports.

### 5.3. Set-of-Marks Prompting

Although SoM prompting shows competitive results and also outperform the non-segmentation baselines in several instances, we noted several challenges with such a prompting method for biomedical images. Firstly, as multiple masks are overlaid on the image in the form of alphanumeric characters and/or contours, these may lead to undesirable occlusions in the image – while this may not be problematic in natural scenes, it could be consequential for biomedical images containing small structures (e.g. rib fractures or lung lesions in CXR images). Moreover, as we scale-up to multiple structures in a single image, it may increase the number of occlusions and lead to confusions in the attended image features by the MLLM (e.g. multiple overlapping lines, overlapping anatomical regions with tubes). In contrast, it is easier to scale up MAIRA-Seg to more structures, which is only limited by the context length of the LLM. Further, as CXR images are grayscale, there are limited options available for contour intensities to appear with reasonable contrast (Section A.1, Figure 4) compared to RGB contours frequently used in the general domain. Lastly, it is unclear if the frozen RAD-DINO image encoder (Pérez-García et al., 2024) trained on CXRs can interpret out-of-domain images with the set of marks.

## 6. Conclusions and Future Work

We presented *MAIRA-Seg*, a proof-of-concept MLLM framework to leverage semantic segmentation masks for CXR radiology report generation. Our experi-

ments demonstrate improvements in clinical metrics (with additional mask-relevant metrics) using the proposed approach on MAIRA architectures. We observe encouraging improvements in quantitative results for MAIRA-Seg when introducing segmentation masks with CXR images to visually prompt the MLLM. This confirms that segmentation can enhance the nuanced fine-granular reasoning and understanding of MAIRA-Seg to interpret complex biomedical images. We compare our method for single- and multi-view baseline models, where we outperform the latter in most cases. However, set-of-marks also achieves competitive metrics over the non-segmentation baselines, reconfirming our hypothesis that semantic segmentation can potentially lead to better MLLM understanding of radiological images and generation of superior draft reports.

The proposed MAIRA-Seg MLLM framework is demonstrated on CXR report generation, however, due to the flexibility in its architecture, it can be extended to other imaging modalities such as CT and MRI, given a modality-specific image encoder, expert segmentation models and modality-specific training report generation data. In future, we intend to explore ways to improve the segmentation quality of the generated mask pseudo-labels (e.g. there may be under or over-segmentation), especially for the fine-grained structures. Moreover, we will explore more generalist segmentation models in our framework in order to scale up to multiple unseen structures without sacrificing segmentation performance. Through model optimization and tuning methods, specifically for the MAIRA-Seg-Multi model, we will aim to further improve model performance (for instance, by using more training datasets and extending to grounded reporting) and optimize computational requirements at inference time (for instance, quantization and pruning), with the end goal of clinical deployment. We intend to perform more elaborate assessment of results through human-centric evaluation, similar to the study described in Bannur et al. (2024). Lastly, we aim to continue exploring the interactions between the visual (images, segmentation) and text tokens in MLLMs.

## References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL <https://arxiv.org/abs/2404.14219>.
- Fan Bai, Yuxin Du, Tiejun Huang, Max Q. H. Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models, 2024. URL <https://arxiv.org/abs/2404.00578>.
- Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. URL <https://arxiv.org/abs/2406.04449>.
- Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *arXiv preprint arXiv:2403.08002*, 2024.
- Li-Ching Chen, Po-Chih Kuo, Ryan Wang, Judy Gichoya, and Leo Anthony Celi. Chest x-ray segmentation images based on mimic-cxr, 2022. URL <https://physionet.org/content/lung-segment-mimic-cxr/1.0.0/>.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%\* ChatGPT quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Stefan Denner, Markus Bujotzek, Dimitrios Bounias, David Zimmerer, Raphael Stock, Paul F. Jäger, and Klaus Maier-Hein. Visual prompt engineering for medical vision language models in radiology, 2024. URL <https://arxiv.org/abs/2408.15802>.
- Sijing Feng, Damian Azzollini, Ji Soo Kim, Cheng-Kai Jin, Simon P Gordon, Jason Yeoh, Eve Kim, Mina Han, Andrew Lee, Aakash Patel, et al. Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence*, 3(6):e210136, 2021.
- Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Julia Mariel Saidman, Martina Aineseder, Diego H. Milone, and Enzo Ferrante. Chexmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *Scientific Data*, 11(1), May 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03358-1. URL <http://dx.doi.org/10.1038/s41597-024-03358-1>.
- Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a

- new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.
- Jonathan Huang, Luke Neill, Matthew Wittbrodt, David Melnick, Matthew Klug, Michael Thompson, John Bailitz, Timothy Loftus, Sanjeev Malik, Amit Phull, et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. *JAMA network open*, 6(10):e2336100–e2336100, 2023.
- Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. Maira-1: A specialised large multimodal model for radiology report generation, 2024.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019)*, volume 33, pages 590–597. AAAI Press, July 2019. doi: 10.1609/aaai.v33i01.3301590.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019a. doi: 10.1038/s41597-019-0322-0.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Roger G. Mark, and Steven Horng. MIMIC-CXR database (version 2.0.0). PhysioNet, 2019b.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL <https://arxiv.org/abs/2304.02643>.
- Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, Jiing-Feng Lirng, Kai-Wei Chang, and Shih-Hwa Chiou. Towards a holistic framework for multimodal large language models in three-dimensional brain ct report generation, 2024. URL <https://arxiv.org/abs/2407.02235>.
- Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest X-ray report generation. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269. PMLR, 09–10 Aug 2019. URL <https://proceedings.mlr.press/v106/liu19a.html>.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, December 2023b. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf).
- Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL <http://dx.doi.org/10.1038/s41467-024-44824-z>.
- Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation, 2021. URL <https://arxiv.org/abs/2010.10042>.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, et al. DINOv2: Learning robust visual features without supervision.

- Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=a68SUt6zFt>.
- Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Rad-dino: Exploring scalable medical image encoders beyond text supervision, 2024. URL <https://arxiv.org/abs/2401.10815>.
- Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Björn H. Menze. cldice - a novel topology-preserving loss function for tubular structure segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.01629. URL <http://dx.doi.org/10.1109/CVPR46437.2021.01629>.
- Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *arXiv preprint arXiv:2405.10842*, 2024.
- Shaury Srivastav, Mercy Ranjit, Fernando Pérez-García, Kenza Bouzid, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Harshita Sharma, Maximilian Ilse, Valentina Salvatelli, Sam Bond-Taylor, Fabian Falck, Anja Thieme, Hannah Richardson, Matthew P. Lungren, Stephanie L. Hyland, and Javier Alvarez-Valle. MAIRA at RRG24: A specialised large multimodal model for radiology report generation. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 597–602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.bionlp-1.50>.
- Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2020. URL <https://arxiv.org/pdf/1905.11946.pdf>.
- Jennifer SN Tang, Jarrel CY Seah, Adil Zia, Jay Gajera, Richard N Schlegel, Aaron JN Wong, Dayu Gai, Shu Su, Tony Bose, Marcus L Kok, et al. Clip, catheter and line position dataset. *Scientific Data*, 8(1):285, 2021.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, et al. Towards generalist biomedical AI. *NEJM AI*, 1(3):AIoa2300138, February 2024. doi: 10.1056/AIoa2300138.
- Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A. Rossi, Ruiyi Zhang, Subrata Mitra, Dimitris N. Metaxas, Lina Yao, Jingbo Shang, and Julian McAuley. Visual prompting in multimodal large language models: A survey, 2024. URL <https://arxiv.org/abs/2409.15310>.
- An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, and Lijuan Wang. List items one by one: A new data source and learning paradigm for multimodal llms, 2024. URL <https://arxiv.org/abs/2404.16375>.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023.
- Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024.
- Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. Multimodal healthcare ai: Identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for



Computing Machinery. ISBN 9798400703300.  
doi: 10.1145/3613904.3642013. URL  
<https://doi.org/10.1145/3613904.3642013>.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning, 2024. URL <https://arxiv.org/abs/2312.10032>.

Ruoqing Zhao, Xi Wang, Hongliang Dai, Pan Gao, and Piji Li. Medical report generation based on segment-enhanced contrastive representation learning, 2023. URL <https://arxiv.org/abs/2312.15869>.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Mounq-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once, 2024. URL <https://arxiv.org/abs/2405.12971>.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.

## Appendix A. More Experimental Details

To utilize the available computational requirements economically for fixed computational budgets, we perform ablations and tuning experiments using the MAIRA-Seg-Frontal architecture which uses only frontal images at the input. We then transfer key findings to MAIRA-Seg-Multi with multi-view inputs including current lateral image, prior image, additional report sections (see Figure 1(a)). We report our ablation results on the MIMIC-CXR validation split (Bannur et al., 2024).

### A.1. Ablations: Set-of-marks Prompting Method

Example of a CXR image overlaid with set-of-marks is presented in Figure 4. Results of ablations on set-of-marks prompting are depicted in Table 3 for the MIMIC-CXR validation split. We observe that mask-aware augmented prompts help compared to settings

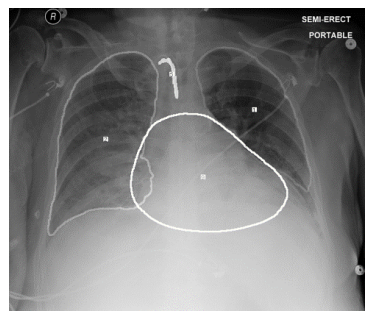


Figure 4: Example of a CXR image overlaid with set-of-marks for the SoM prompting method.

without prompt augmentation. Also, presence of contours and alphanumeric marks are more beneficial than just contours, consistent with the observations in (Yang et al., 2023). We also performed preliminary sanity checks where we used the same grayscale intensity for contours, or the same alphanumeric mark, for different masks, however these settings lead to worse performance compared to the classic SoM method.

### A.2. Expert Semantic Segmentation Models

We present the datasets used for training expert EfficientNet-UNet semantic segmentation models in Table 4 and report segmentation metrics in Table 5. For the latter, we aggregate the scores only over the positives and provide mean and standard deviation across the test set (70/15/15 split over each dataset in Table 4). CL-Dice metric (Shit et al., 2021) is additionally reported to measure segmentation performance on tubular structures in the support devices category. Implementation details including segmentation preprocessing are provided in Section C.

### A.3. Detailed Results

We present the experimental results for MAIRA-Seg-Frontal and MAIRA-Seg-Multi with additional standard lexical and clinical metrics for radiology report generation in Table 6 and Table 7. We illustrate stratified F1-scores for 14 CheXpert pathological findings in Figure 5.

## Appendix B. Qualitative Results

We present qualitative results in Figures 6 and 7 and Figure 8 for MAIRA-Seg-Frontal and MAIRA-Seg-

Table 3: Ablations of set-of-marks prompting method on MIMIC-CXR validation split.

Method	MAIRA-Frontal	Ablation 1	Ablation 2	Ablation 3	Ablation 4
<b>Contours?</b>	No	Yes	Yes	Yes	Yes
<b>Alphanumeric marks?</b>	No	No	No	Yes	Yes
<b>Augmented mask-aware prompts?</b>	No	No	Yes	No	Yes
<i>Lexical</i>					
ROUGE-L	34.8 [34.4, 35.3]	34.5 [34.0, 34.9]	35.0 [34.6, 35.4]	34.9 [34.4, 35.3]	35.3 [34.9, 35.7]
BLEU-1	38.5	37.6	38.2	37.9	<b>38.9</b>
BLEU-4	17.5	17.1	17.5	17.3	<b>17.8</b>
METEOR	37.5	36.8	37.4	37.2	<b>37.9</b>
<i>Clinical</i>					
RadGraph-F1	30.6 [30.1, 31.2]	30.0 [29.4, 30.5]	30.8 [30.3, 31.3]	30.4 [30.0, 30.9]	31.5 [30.9, 32.0]
RGER	35.0 [34.5, 35.5]	34.3 [33.8, 34.8]	35.1 [34.6, 35.6]	34.8 [34.3, 35.3]	<b>35.8</b> [35.4, 36.3]
CheXbert vector	51.5 [50.9, 52.1]	50.7 [50.1, 51.3]	51.7 [51.1, 52.3]	51.0 [50.5, 51.6]	52.4 [51.7, 52.9]
RadCliQ ( $\downarrow$ )	2.72 [2.70, 2.75]	2.76 [2.73, 2.79]	2.72 [2.69, 2.74]	2.74 [2.71, 2.76]	<b>2.68</b> [2.66, 2.71]
Macro F <sub>1</sub> -14	33.8 [32.8, 34.8]	32.5 [31.6, 33.5]	34.9 [33.9, 36.1]	33.6 [32.7, 34.6]	<b>36.3</b> [35.2, 37.4]
Micro F <sub>1</sub> -14	51.9 [51.0, 52.7]	49.7 [48.8, 50.4]	52.2 [51.4, 53.0]	51.5 [50.7, 52.4]	<b>53.6</b> [52.8, 54.4]

Table 4: Dataset details for training expert segmentation models and associated CheXpert findings classes.

Category	Structure	Training dataset	Number of images	Correlated CheXpert findings (Irvin et al., 2019)
Anatomical	Left lung	MIMIC-CXR subset (Chen et al., 2022)	1,138	Lung opacity
	Right lung			
	Heart	CheXmask (Gaggion et al., 2024)	242,000	Cardiomegaly
Support devices	Central Venous Catheter (CVC)	RANZCR-CLiP (Tang et al., 2021)	30,083	Support devices
	Endo-Tracheal Tube (ETT)			
	Naso-Gastric Tube (NGT)			
	Swan-Ganz Catheter (SGC)			
	Chest tube	CANDID-PTX (Feng et al., 2021)	19,237	Support devices, Pneumothorax, Pleural effusion
Pathological	Pneumothorax	CANDID-PTX (Feng et al., 2021)	19,237	Pneumothorax

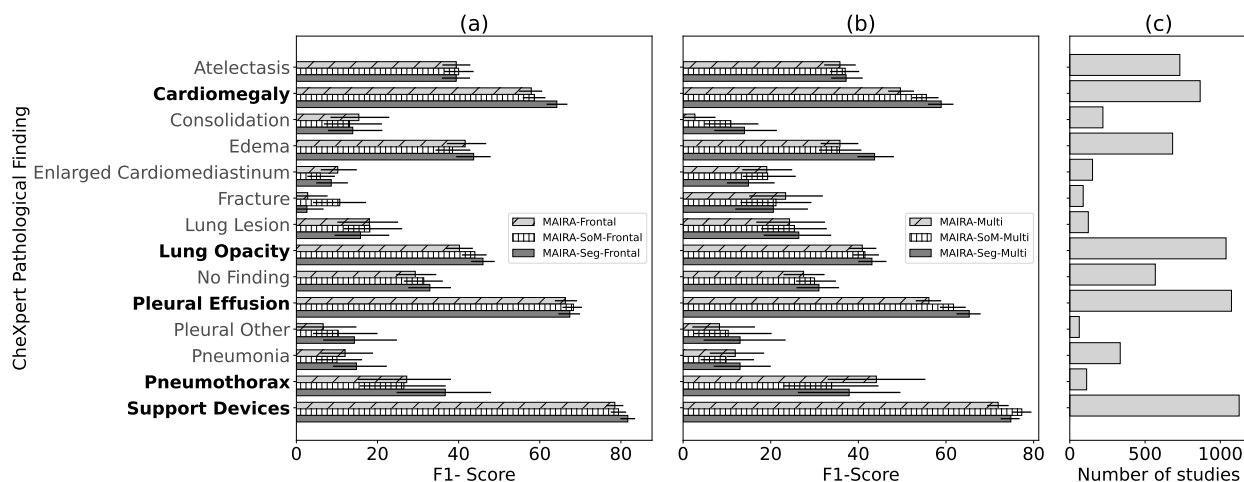


Figure 5: Stratified F1-scores for the 14 CheXpert findings comparing the respective baselines, SoM prompting and MAIRA-Seg models for (a) single view (b) multi-view experiments (c) Support for each pathological finding in the MIMIC-CXR test set.

Table 5: Segmentation metrics for the expert EfficientNetUNet semantic segmentation models. ‘Supp.’ stands for support.

Category	Structure	Dice score	CL-Dice
Anatomical	Left lung	$98.3 \pm 1.1$	N/A
	Right lung	$98.3 \pm 1.4$	N/A
	Heart	$95.3 \pm 2.0$	N/A
Supp.devices	CVC	$77.0 \pm 13.0$	$87.5 \pm 15.0$
	ETT	$58.5 \pm 24.1$	$66.3 \pm 32.8$
	NGT	$69.5 \pm 17.0$	$79.6 \pm 20.9$
	SGC	$73.2 \pm 12.6$	$84.9 \pm 15.5$
	Chest tube	$53.0 \pm 16.8$	$59.5 \pm 20.8$
Pathological	Pneumothorax	$73.5 \pm 26.9$	N/A

Multi, respectively. We present the target report and predicted findings of the baselines and corresponding proposed MAIRA-Seg methods. In the reports, we highlight the phrases selected for further analysis (gray) in the target findings, and the corresponding errors (red) and correct predictions (green). The image labelled as “current image with mask pseudolabels” is only for illustrating the mask pseudolabels used along with the CXR image in the MAIRA-Seg architecture, where the masks in green, red, yellow, blue, light-blue, white correspond to right lung, left lung, NGT, CVC, ETT, and heart respectively.

In Figure 6(a), we find that MAIRA-Frontal hallucinates Cardiomegaly (“cardiac contours are mildly enlarged”) whereas this is correctly reported by MAIRA-Seg-Frontal. Also, bibasilar atelectasis is omitted by MAIRA-Frontal but correctly stated by MAIRA-Seg-Frontal. Both MAIRA-Frontal and MAIRA-Seg-Frontal report an NGT (traced with yellow with MAIRA-Seg-Frontal prediction showing the corresponding NGT mask pseudolabel). In Figure 6(b), we note that MAIRA-Frontal wrongly predicts the tip position of the IJ catheter (a type of CVC), which is correctly predicted by MAIRA-Seg-Frontal (as correctly depicted by the blue CVC mask). Moreover, MAIRA-Frontal hallucinates atelectasis, and mentions ‘bilateral’ pleural effusion – this is correctly predicted as ‘right’ only by MAIRA-Seg-Frontal (target report mentions ‘no left effusion’). In Figure 6(c), we observe that again MAIRA-Frontal erroneously mentions the tip of the IJ line in the right atrium, which is correctly predicted as cavo-atrial junction by MAIRA-Seg-Frontal, also depicted in the corresponding blue CVC mask. We also find that MAIRA-Frontal hallucinates cardiomegaly and omits right pleural effusion – these are correctly predicted by MAIRA-Seg-Frontal. For MAIRA-Seg-Frontal, the extents of pleural effusions is mistaken as ‘moderate’ compared to extensive, and atelectasis is omitted. In Figure 7(a), MAIRA-Frontal omits atelectasis and mentions of any support

Table 6: Extended single view experimental results on the official MIMIC-CXR test split and comparison of the segmentation-aware methods MAIRA-SoM-Frontal and MAIRA-Seg-Frontal against the MAIRA-Frontal baseline.

Method	MAIRA-Frontal	MAIRA-SoM-Frontal	MAIRA-Seg-Frontal
<i>Lexical</i>			
ROUGE-L	29.3 [28.8, 29.9]	30.0 [29.5, 30.5]	29.8 [29.3, 30.3]
BLEU-1	36.0 [35.3, 36.9]	36.7 [35.9, 37.4]	37.1 [36.4, 37.8]
BLEU-4	14.2 [13.7, 14.8]	14.6 [14.1, 15.2]	14.5 [14.0, 15.1]
METEOR	31.2 [30.7, 31.9]	31.9 [31.3, 32.4]	32.0 [31.5, 32.5]
<i>Clinical</i>			
RadGraph-F1	23.4 [22.9, 24.1]	24.4 [23.8, 25.0]	24.5 [23.8, 25.0]
RG <sub>ER</sub>	28.1 [27.4, 28.8]	<b>29.0</b> [28.4, 29.6]	<b>29.3</b> [28.7, 29.9]
CheXbert vector	41.7 [40.8, 42.5]	42.3 [41.5, 43.2]	<b>43.4</b> [42.5, 44.3]
RadCliQ (↓)	3.19 [3.15, 3.22]	<b>3.14</b> [3.10, 3.18]	<b>3.11</b> [3.08, 3.15]
Macro F <sub>1</sub> -14	31.9 [30.3, 33.4]	32.6 [31.0, 33.9]	<b>34.5</b> [32.9, 36.0]
Micro F <sub>1</sub> -14	50.4 [49.2, 51.7]	51.6 [50.5, 52.9]	<b>53.5</b> [52.3, 54.6]
Macro F <sub>1</sub> -MR *	54.1 [51.2, 56.2]	55.3 [53.1, 57.8]	<b>59.2</b> [56.7, 61.7]
Micro F <sub>1</sub> -MR *	61.4 [60.0, 62.7]	<b>63.1</b> [61.9, 64.4]	<b>65.4</b> [64.1, 66.7]
RadFact/logical_precision	46.8 [45.5, 48.0]	<b>48.3</b> [47.0, 49.3]	<b>48.4</b> [47.3, 49.6]
RadFact/logical_recall	38.9 [37.8, 40.1]	<b>40.5</b> [39.4, 41.6]	<b>41.5</b> [40.5, 42.7]
RadFact/logical_f1	42.4 [41.5, 43.6]	<b>44.0</b> [43.0, 45.1]	<b>44.7</b> [43.8, 45.8]

Table 7: Extended multi-view experimental results for MAIRA-Seg-Multi on the official MIMIC-CXR test split and comparison of the segmentation-aware methods MAIRA-SoM-Multi and MAIRA-Seg-Multi against the MAIRA-Multi baseline.

Method	MAIRA-Multi	MAIRA-SoM-Multi	MAIRA-Seg-Multi
<i>Lexical</i>			
ROUGE-L	35.5 [34.9, 36.1]	<b>36.7</b> [36.2, 37.3]	36.3 [35.8, 36.9]
BLEU-1	39.7 [38.9, 40.5]	40.5 [39.7, 41.3]	39.3 [38.6, 40.2]
BLEU-4	19.5 [19.0, 20.1]	20.3 [19.8, 20.8]	19.6 [19.1, 20.1]
METEOR	37.2 [36.6, 37.8]	<b>38.4</b> [37.8, 38.9]	37.4 [36.9, 38.0]
<i>Clinical</i>			
RadGraph-F1	29.6 [29.0, 30.4]	<b>31.3</b> [30.7, 32.1]	<b>30.9</b> [30.1, 31.6]
RG <sub>ER</sub>	34.6 [34.0, 35.2]	<b>36.3</b> [35.7, 37.0]	<b>35.7</b> [35.1, 36.5]
CheXbert vector	44.1 [43.3, 45.2]	<b>45.9</b> [45.0, 46.8]	<b>46.4</b> [45.4, 47.3]
RadCliQ (↓)	2.90 [2.86, 2.94]	<b>2.81</b> [2.77, 2.85]	<b>2.82</b> [2.78, 2.86]
Macro F <sub>1</sub> -14	32.3 [30.8, 34.0]	33.5 [32.1, 35.1]	<b>35.3</b> [33.5, 37.0]
Micro F <sub>1</sub> -14	46.0 [44.7, 47.3]	<b>49.2</b> [48.0, 50.3]	<b>50.5</b> [49.5, 52.0]
Macro F <sub>1</sub> -MR *	52.5 [50.0, 55.0]	53.9 [51.5, 56.3]	55.9 [53.6, 58.7]
Micro F <sub>1</sub> -MR *	55.6 [54.2, 57.0]	<b>59.9</b> [58.6, 61.1]	<b>60.9</b> [59.6, 62.5]
RadFact/logical_precision	48.3 [47.1, 49.6]	<b>50.7</b> [49.5, 52.0]	<b>51.5</b> [50.2, 52.9]
RadFact/logical_recall	42.8 [41.5, 43.9]	44.4 [43.4, 45.6]	43.4 [42.3, 44.6]
RadFact/logical_f1	45.4 [44.2, 46.5]	<b>47.4</b> [46.3, 48.6]	<b>47.1</b> [46.1, 48.3]

device. In contrast, MAIRA-Seg-Frontal correctly mentions atelectasis (although as ‘bibasilar’ rather than ‘right’). MAIRA-Seg-Frontal mentions the two devices, namely, AICD and PICC line (a type of CVC) correctly, and with the correct position for AICD, but erroneously for PICC line: interestingly, we note that the PICC line mask is over-segmented because of overlap with AICD, and gives the appearance of ‘dis-

tal SVC’ that is predicted by MAIRA-Seg-Frontal rather than mid-SVC. In Figure 7(b), we observe that atelectasis is omitted by MAIRA-Frontal but correctly mentioned by MAIRA-Seg-Frontal. In Figure 7(c), MAIRA-Frontal omits cardiomegaly and pacemaker position, correctly predicted by MAIRA-Seg-Frontal. The former wrongly mentions the side of the costophrenic angle (this detail is not usually a

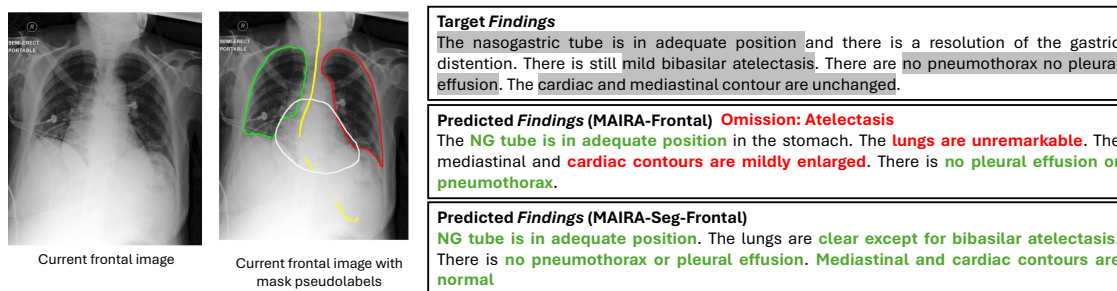


part of the *Findings* section). In Figure 8(b), we note that MAIRA-Seg-Multi correctly predicts the type, side and tip location of the dialysis catheter (corresponding to the blue mask) whereas MAIRA-Multi makes error in the approach. MAIRA-Seg-Multi suggests moderate cardiomegaly and no pleural effusion, which are omitted and hallucinated respectively by MAIRA-Multi. MAIRA-Multi doesn't mention the slightly low lung volumes at all, these are erroneously mentioned as 'well expanded' by MAIRA-Seg-Multi.

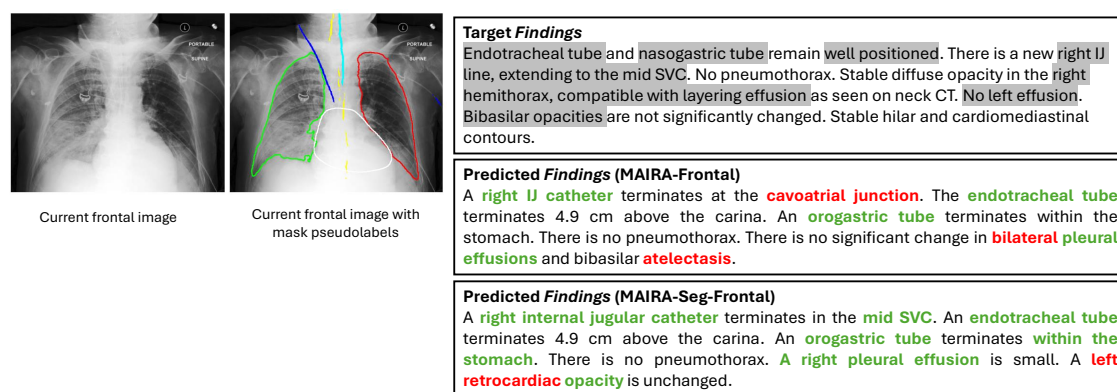
## Appendix C. Implementation Details

We train MAIRA-Seg models (MAIRA-Seg-Frontal and MAIRA-Seg-Multi) with a conventional autoregressive cross-entropy loss on the MIMIC-CXR training set. Following the training method in Hyland et al. (2024), we do a single stage of training with a frozen image encoder and trainable adapter and LLM. We train for 3 epochs and use the final checkpoint in evaluations. We use the AdamW optimiser and a cosine learning rate scheduler. We train MAIRA-Seg-Frontal on 4 NVIDIA A100 GPUs with a global training batch size of 128, training time  $\approx 15$ h, warm-up of 0.03, and a learning rate of  $2 \times 10^{-5}$ . We train MAIRA-Seg-Multi across 16 NVIDIA A100 GPUs with a global training batch size of 256, training time  $\approx 11$ h, warmup ratio 0.03, and a learning rate of  $2 \times 10^{-5}$ . We use a pre-trained RAD-DINO image encoder (Pérez-García et al., 2024) using the same weights as used in Bamur et al. (2024).

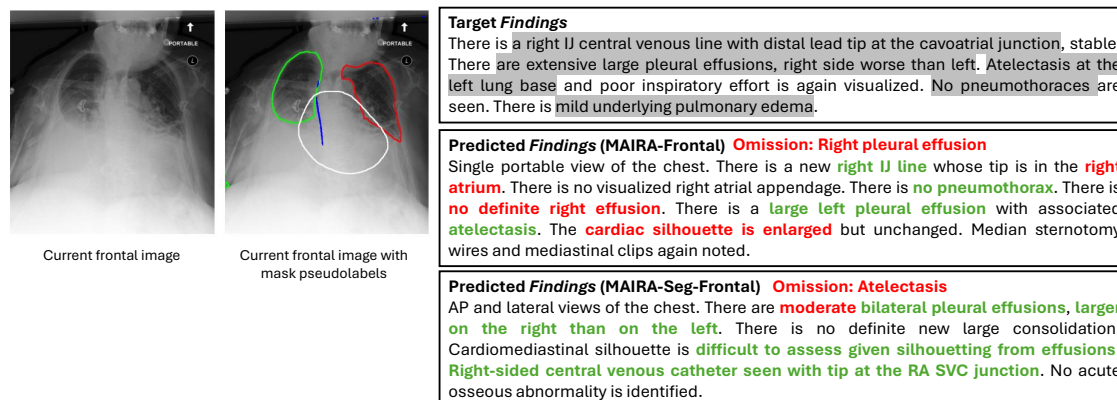
We train expert segmentation models on 8 NVIDIA V100 GPUs. We use a training batch size 80 (10 per GPU), Adam optimizer, base learning rate  $5 \times 10^{-4}$ , and a cosine learning rate scheduler. We use the following preprocessing and augmentations: centre-cropping and resizing ( $512 \times 512$ ), random horizontal flip (except left-right lungs), random affine transform, elastic transform, random brightness and contrast jittering, and random gamma adjustments. The segmentation models are trained for 100 epochs. The checkpoint with minimum loss on the validation set is used for inference on the test set. We use a 70/15/15 split by subjects for train, validation and test sets, respectively, and report metrics on the test set.



(a)


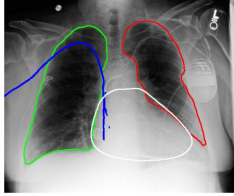


(b)



(c)

Figure 6: Qualitative result for examples in the MIMIC-CXR test set, showing target and predicted Findings using MAIRA-Frontal and MAIRA-Seg-Frontal. Mask pseudolabels are shown overlaid on the CXR image for illustrative purposes (corresponding masks are used to obtain segmentation tokens).


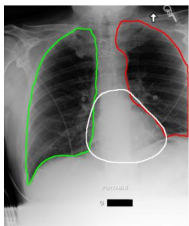



**Target Findings**  
 Right PICC line ends at mid SVC. Left-sided pacer defibrillator with leads through the left transvenous approach is seen to end in the right atrium, right ventricle, and coronary sinus. Minimal right basal atelectasis is unchanged. There are no other lung opacities of concern. Top normal heart size, mediastinal and hilar contours are stable.

**Predicted Findings (MAIRA-Frontal)** **Omission: Atelectasis, support devices**  
 In comparison with the study of \_\_, there is little change in the appearance of the heart and lungs. Specifically, there is no evidence of pneumothorax.

**Predicted Findings (MAIRA-Seg-Frontal)**  
 Frontal and lateral views of the chest were obtained. Dual-lead left-sided AICD is again seen with leads extending to the expected positions of the right atrium and right ventricle. The patient is status post median sternotomy and CABG. There is a right-sided PICC, terminating in the distal SVC. There is minimal bibasilar atelectasis. No focal consolidation, pleural effusion, or evidence of pneumothorax is seen. The cardiac and mediastinal silhouettes are stable.

(a)


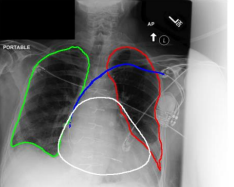



**Target Findings**  
 The patient is status post median sternotomy and CABG. The cardiac, mediastinal, and hilar contours are normal. The pulmonary vascularity is normal. There are streaky opacities in the lung bases, most likely reflective of atelectasis. No focal consolidation, pleural effusion, or pneumothorax is visualized. There are no acute osseous abnormalities.

**Predicted Findings (MAIRA-Frontal)** **Omission: Atelectasis**  
 The patient is status post median sternotomy and CABG. The heart size is normal. The mediastinal and hilar contours are unremarkable. The pulmonary vascularity is normal. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.

**Predicted Findings (MAIRA-Seg-Frontal)**  
 The patient is status post median sternotomy and CABG. The heart size is normal. The mediastinal and hilar contours are unremarkable. The pulmonary vascularity is normal. There is minimal atelectasis in the left lung base. No focal consolidation, pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.

(b)

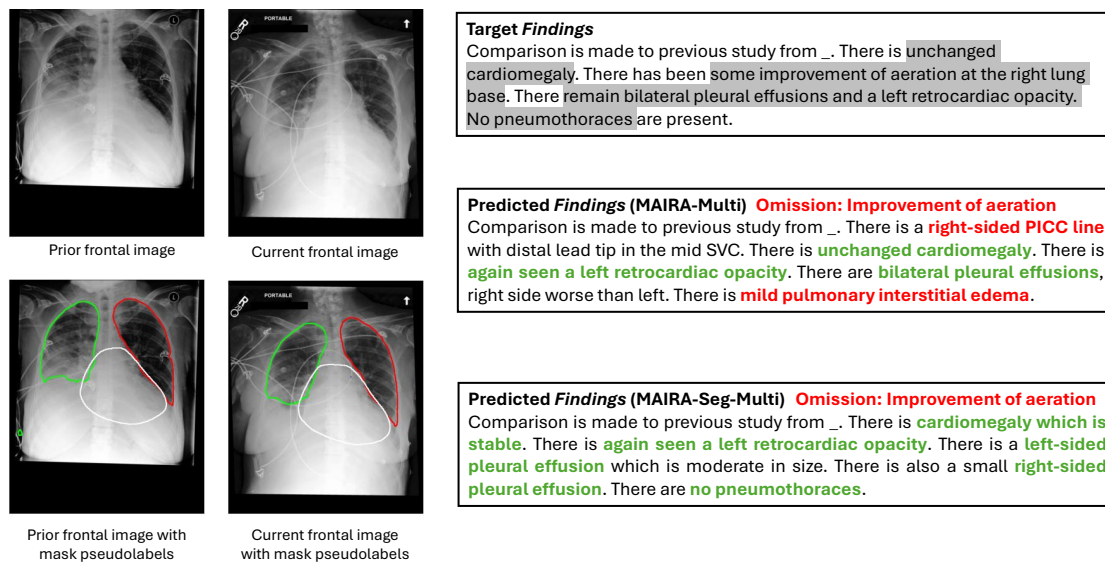
**Target Findings**  
 AP portable upright chest radiograph was provided. Midline sternotomy wires and left chest wall pacer device again noted with pacer lead extending into the region of the right atrium and right ventricle. Multiple mediastinal clips are noted. As seen on prior high res CT, areas of scarring evidenced by subtle linear reticular opacity at the right lung base present. The heart is mildly enlarged. There is no definite effusion, though the left CP angle is excluded. No pneumothorax. No signs of CHF or discrete signs of pneumonia. Bony structures are intact.

**Predicted Findings (MAIRA-Frontal)** **Omission: Pacemaker position, Cardiomegaly**  
 Single portable view of the chest. Left chest wall pacing device is again seen. The lungs are clear of confluent consolidation. There is no definite large effusion noting that the right costophrenic angle is excluded from the field of view. Cardiomeastinal silhouette is stable. Median sternotomy wires again noted.

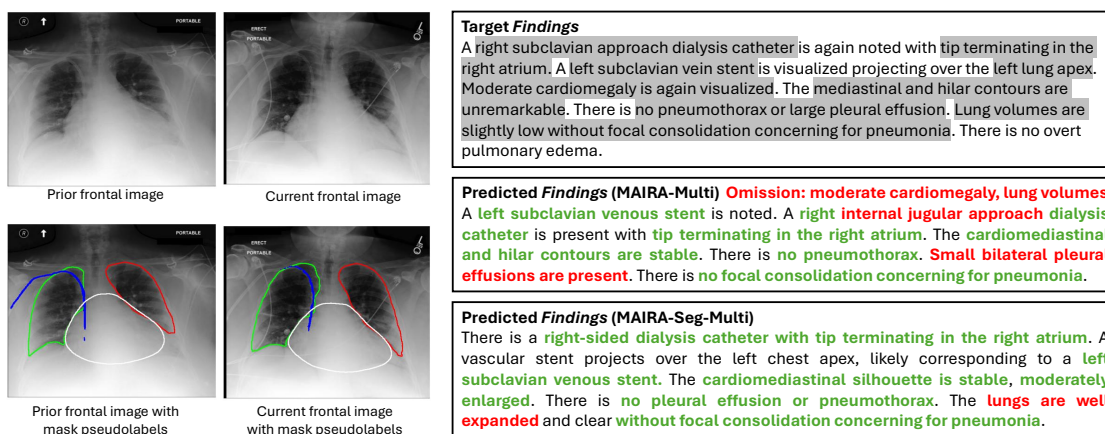
**Predicted Findings (MAIRA-Seg-Frontal)**  
 The patient is status post median sternotomy and CABG. Left-sided AICD/pacemaker device is noted with leads terminating in the right atrium and right ventricle. Heart size is mildly enlarged. The aorta is tortuous and diffusely calcified. There is no pulmonary vascular congestion. Streaky opacities in the lung bases likely reflect atelectasis. No pleural effusion or pneumothorax is identified. No acute osseous abnormalities are seen.

(c)

Figure 7: (Contd.) Qualitative result for examples in the MIMIC-CXR test set, showing target and predicted Findings using MAIRA-Frontal and MAIRA-Seg-Frontal. Mask pseudolabels are shown overlaid on the CXR image for illustrative purposes (corresponding masks are used to obtain segmentation tokens).



(a)



(b)

Figure 8: Qualitative result for examples in the MIMIC-CXR test set, showing target and predicted Findings using MAIRA-Multi and MAIRA-Seg-Multi. Mask pseudolabels are shown overlaid on the CXR image for illustrative purposes (corresponding masks are used to obtain segmentation tokens).