Vision-Language Models for Chest X-Ray Radiology

Francesco Dalla Serra

Submitted in fulfilment of the requirements for the

Degree of Doctor of Engineering (EngD)

School of Computing Science

College of Science and Engineering

University of Glasgow

September 2024

Contents

Abstract

Chest X-ray (CXR) is a widely requested imaging test used as a quick and non-invasive procedure to examine various pathologies in the chest cavity. When radiologists interpret CXR scans, they typically consult additional clinical information about the patient under examination and document the relevant findings visualised in the CXR into free-text radiology reports. Therefore, in clinical practice, CXRs are often accompanied by supplementary textual documents that provide important context for accurate diagnosis.

This thesis explores the potential of Visual-Language Models (VLMs)—AI systems designed to process and integrate both visual and textual information—to develop flexible autonomous decision support tools for CXR analysis. We investigate several multimodal tasks, including medical finding classification, Automated Radiology Reporting (ARR), and medical Visual Question Answering (VQA). Medical finding classification involves identifying and categorising specific pathologies or abnormalities present in the CXR images. ARR corresponds to the task of generating free-text radiology reports for each scan, providing comprehensive descriptions and diagnoses. Medical VQA focuses on answering questions about the visual content of medical scans, facilitating deeper interaction with the imaging data.

We address these tasks by improving the visual representation of the CXR scans and providing the VLM with additional relevant textual information. This includes utilising patients' medical history and the reasons for the scans, as detailed in the indication field of the radiology report, available at the time of imaging. We leverage expert-written radiology reports to supervise ARR models and guide VQA model responses through specific textual queries. By integrating both textual and visual data, we aim to improve the models' ability to

accurately interpret and interact with the imaging data. This thesis is organised as follows.

We start by addressing the medical finding classification task. In particular, we investigate how different pre-training strategies of the image encoder impact the performances of a multimodal model and how these degrade in the scenario of limited labelled data. We demonstrate the impact of self-supervised pre-training strategies on this task.

Second, we focus on the ARR task. We start by exploring the effect of incorporating structured information extraction from each scan – expressed in the form of triples (entity1, relation, entity2). Triples extraction is used as the intermediate task in a two-step pipeline for ARR, showing improved results. Additionally, we propose the extraction of more fine-grained visual representations, each specific to an anatomical region of the CXR, which are used as the visual input representation to perform ARR. Our approach offers an effective solution for encoding detailed information about abnormalities within each anatomical region. Following, we demonstrate how to manipulate these region-specific representations to model the evolution of findings over time (*e.g.*, by examining longitudinal scans) and enable controllable partial reporting – the task of generating the radiology report for a selected set of anatomical regions. We then integrate all the proposed solutions for ARR into a single model and provide a human evaluation of its performance to assess its accuracy and clinical utility.

Finally, we focus on the medical VQA task, by exploring how ARR and VQA can be integrated into a unified pipeline. We show that grounding the VQA model on the predicted radiology reports improves its ability to answer queries related to the CXR images.

This work demonstrates how to effectively tackle visio-linguistic tasks specific to CXR scans, addressing the unique challenges of each task. The research presented here offers valuable insights that may guide future studies, ultimately contributing to the successful integration of VLMs into radiologists' workflows for improved clinical outcomes.

List of Tables

List of Figures

List of Abbreviations

AE AutoEncoder

AI Artificial Intelligence

AG Answer Generator

AP Anteroposterior

ARR Automated Radiology Reporting

AUROC Area Under the Receiver Operating Characteristic Curve

A_{tok} Anatomical Token

BERT Bidirectional Encoder Representations from Transformers

BLEU BiLingual Evaluation Understudy

CE Clinical Efficiency

CIDEr Consensus-based Image Description Evaluation

CNN Convolutional Neural Network

CoT Chain-of-Thought

CT Computed Tomography

CV Computer Vision

CXR Chest X-Ray

EHR Electronic Health Record

FC Fully-Connected

FPN Feature Pyramid Network

GPT Generative Pre-trained Transformer

GPU Graphichs Processing Unit

Grad-CAM Gradient-weighted Class Activation Mapping

GRU Gated Recurrent Unit

GT Ground Truth

Ind Indication Field

InfoNCE Info Noise Contrastive Estimation

IoU Intersection over Union

ITM Image Text Matching

LLM Large Language Model

LM Language Model

LPM Longitudinal Projection Module

LSTM Long Short-Term Memory

mAP mean Average Precision

METEOR Metric for Evaluation of Translation with Explicit ORdering

MLLM Multimodal Large Language Model

MLM Masked Language Modeling

MLP Multi-Layer Perceptron

MMBT MultiModal BiTransformer

MOC Masked Object Classification

MoCo Momentum Contrast

MRI Magnetic Resonance Imaging

NLG Natural Language Generation

NLP Natural Language Processing

PA Posteroanterior

PHR Personal Health Record

R-CNN Region-based Convolutional Neural Networks

RG Report Generator

RNN Recurrent Neural Network

ROUGE Recall-Oriented Understudy for Gisting Evaluation

RPN Region Proposal Network

ReLU Rectified Linear Unit

RoI Region of Interest

TE Triples Extractor

TD-IDF Term Frequency - Inverse Document Frequency

Trp Triple

t-SNE t-distributed Stochastic Neighbour Embedding

VLM Vision-Language Model

VQA Visual Question Answering

ViT Vision Transformer

List of Publications and Patents

Publications

The list below is arranged in chronological order:

- Francesco Dalla Serra, Grzegorz Jacenków, Fani Deligianni, Jeff Dalton, and Alison Q. O'Neil. **"Improving Image Representations via MoCo Pre-training for Multimodal CXR Classification"**. In *Annual Conference on Medical Image Understanding and Analysis*, pp. 623-635. (MIUA 2022)

- Francesco Dalla Serra, Fani Deligianni, Jeff Dalton, and Alison Q. O'Neil. **"CMRE-UoG team at ImageCLEFmedical Caption 2022: Concept Detection and Image Captioning"**. In *CLEF 2022 Working Notes, CEUR Workshop Proceedings*. (CLEF 2022)

- Francesco Dalla Serra, William Clackett, Hamish MacKinnon, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q. O'Neil. 2022. **"Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations"**. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 615-624. 2022. (AACL-IJCNLP 2022).

- Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q.

O'Neil. **"Finding-Aware Anatomical Tokens for Chest X-Ray Automated Reporting"**. In *International Workshop on Machine Learning in Medical Imaging*, pp. 413-423. 2023. (MLMI 2023)

- Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q. O'Neil. **''Controllable Chest X-Ray Report Generation from Longitudinal Representations"**. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing*, pp. 4891-4904. 2023. (EMNLP 2023)

Patents

- Francesco Dalla Serra, Alison Q. O'Neil, and Chaoyang Wang. **"Image data processing apparatus and method"** U.S. Patent Application No. 18/581,231. Patent filed in 2024 relating to contributions in Chapter 5 and Chapter 6.

Acknowledgements

I would like to express my gratitude to my supervisors: Dr Alison Q. O'Neil (Canon Medical Research Europe), Dr Fani Deligianni (University of Glasgow) and Dr Zaiqiao Meng (University of Glasgow), for their advice, guidance, and time and effort invested into this project. I would also like to thank my previous supervisors: Dr Jeffrey Dalton (University of Edinburgh, formerly at University of Glasgow) and Dr Maciej Pajak (Bloomberg LP, formerly at Canon Medical Research Europe).

Thanks are also due to my colleagues at Canon Medical Research Europe: Dr Patrick Schrempf, Dr William Clackett, Dr Giovana Klefti, Dr Chaoyang Wang, Dr Noon Altijani, Grzegorz Jacenków and Hamish MacKinnon for their collaboration; Dr William Clackett, Dr Giovana Klefti, Dr Chaoyang Wang, and Dr Noon Altijani for helping to design and run the clinical evaluations; Dr Stuart Thomson and Dickon Fell for their support using the evaluation platforms; Dr William Clackett, Dr Giovana Klefti, Dr Chaoyang Wang, Dr Noon Altijani, and Dr Hannah Watson for their many clinical advice; Dr Patrick Schrempf for the support and guidance on how to organise the thesis; Dr Patrick Schrempf, Dr Antanas Kascenas, Dr Chaoyang Wang, and Hamish MacKinnon for reviewing my work.

Finally, I want to say a big thank you to my partner for her constant support, encouragement, and understanding, which have meant so much to me during this time. I'm also really grateful to my family and friends for always being there and cheering me on throughout this journey.

This work was supported by Canon Medical Research Europe Limited and the UKRI EPSRC Centre for Doctoral Training in Applied Photonics [EP/S022821/1].

Chapter 1

Introduction

1.1 Motivation

The field of radiology plays a crucial role in modern healthcare, enabling the diagnosis and monitoring of numerous medical conditions through advanced imaging techniques. Among these, *Chest X-Rays* (CXRs) are some of the most widely requested imaging tests [141] due to their efficiency, accessibility, and diagnostic value. CXRs are quick, non-invasive procedures that provide a 2D visual representation of the chest cavity using X-ray radiation, providing critical visualisation of the lungs, heart, and chest wall. These scans are essential diagnostic tools used by radiologists to identify a variety of conditions, ranging from acute diseases like trauma, pneumothorax, and pneumonia to chronic diseases such as cancer and chronic obstructive pulmonary disease. [1]

CXRs are often accompanied by a scan request, which includes details about the patient's medical history and the reason for the scan [145]. This information is used by the triaging radiologist to determine the necessity and type of scan required. For instance, the indication field may describe that the patient was previously intubated or experienced dizziness, outlining symptoms or specific diagnostic questions the referring physician wants the radiologist to consider while interpreting the scan. Radiologists then interpret the scan by examining lung fields, heart size and shape, bone structures, and other anatomical features for abnormalities.[2] After this detailed analysis, the radiologist compiles a report that describes any abnormalities and their potential implications for the patient's health, which is then made available to the referring physician to guide further care.

generating detailed radiology reports is time-consuming and highly dependent on radiologist expertise. With an ever-increasing volume of imaging studies [185], driven by factors such as the ageing population, radiologists are often overwhelmed, leading to potential delays in diagnosis and treatment. These delays, caused by institutional constraints and a global shortage of radiologists [171, 28], can negatively impact patient outcomes [29]. Backlogs in

[1] Jones J, Hacking C, Walizai T, et al. Chest radiograph. Reference article, Radiopaedia.org (Accessed on 22 Jul 2024) https://doi.org/10.53347/rID-14511

[2] Ryu Y, Bell D, Knipe H, et al. Chest radiograph assessment using ABCDEFGHI. Reference article, Radiopaedia.org (Accessed on 22 Jul 2024) https://doi.org/10.53347/rID-7267

reporting are particularly severe in remote and underserved regions, where access to expert radiologists is limited.

These challenges highlight the need for effective decision-support tools, to improve radiology workflows and facilitate the timely delivery of accurate reports. However, the 2D nature of CXRs, which causes different anatomical structures to overlap, combined with the diverse reasons for their prescription and the variety of detectable findings, makes interpreting CXRs highly challenging. As a consequence, correctly assessing the content of CXRs, as done by radiologists, is a very difficult task to automate.

Artificial intelligence (AI) has emerged as a transformative technology in various domains, including healthcare. Consequently, in radiology, AI systems have demonstrated the ability to accurately analyse medical images [162]. These AI diagnostic tools can assist in detecting abnormalities, prioritising urgent cases, and providing second opinions, thereby enhancing the overall efficiency and effectiveness of medical diagnosis. The integration of AI into radiology holds promise for addressing the increasing demand for imaging studies and improving patient outcomes through timely and accurate diagnoses.

To address these issues, the integration of advanced technologies, particularly *Vision-Language Models* (VLMs) [24], offers promising solutions. VLMs combine visual and textual information, enabling them to understand and interpret complex medical images in conjunction with contextual information such as patients' histories. This dual capability makes VLMs well-suited for tasks like automated radiology reporting and visual question answering, where understanding both the image and related textual data is crucial.

In this thesis, we explore the development and application of VLMs in the context of CXR radiology. The aim is to develop decision-support tools that assist radiologists by automating specific tasks, such as generating accurate and coherent radiology reports and answering clinically relevant questions about CXR images. These tools are intended to support radiologists in their workflows, rather than replace their expertise, since the critical role of human judgment in interpreting findings remains central.

While VLMs and, more broadly, AI systems are not intended to replace radiologists, it is crucial for professionals to understand how these systems function and explore ways to integrate them into their practice. This integration can significantly enhance their ability to deliver accurate, efficient, and timely care [3]. The primary objective of these technologies is to reduce the burden of routine and repetitive tasks in radiology, allowing radiologists to dedicate more time and focus to complex and challenging cases.

By integrating VLMs into the diagnostic process, these tools can enhance the speed and consistency of report generation while providing reliable supplementary insights. Ultimately, this research seeks to create assistive technologies that improve radiologists' efficiency and accuracy, contributing to better patient outcomes and overall healthcare quality.

1.2 Thesis Summary

This thesis investigates the use of Vision-Language Models as an effective and flexible solution for creating decision-support tools in CXR radiology. We aim to determine whether a multimodal approach—incorporating additional non-imaging inputs, generating text sequences, or responding to specific textual queries—can enhance the performance and flexibility of automated systems in tasks such as performing medical findings classification, radiology report generation or visual question answering. We hypothesise that by providing automated systems with the same comprehensive information available to human experts, we can bring their performance closer to that of expert radiologists. Our goal is to develop systems that perform close to expert radiologists and help address the radiologist capacity shortfall.

We explore various forms of automation based on how the task is framed:

- **Medical Findings Classification**: Predicting a predefined set of medical findings in a medical scan.
- **Automated Radiology Reporting** (**ARR**): Generating a radiology report that provides an exhaustive textual description of the visual content observed in a medical scan.

Figure 1.1: Summary of the contributions of each technical chapter (3-8).

- **Visual Question Answering** (**VQA**): An interactive approach where the system answers specific questions related to the visual aspects of a medical scan.

We address these tasks by combining both visual and textual modalities as input and/or output in a VLM, thus categorising our solutions under multimodal learning. This approach supports the model with additional patient information, such as the indication field, which contains the patient's medical history and the reason for the scan. Moreover, the intrinsic multimodal nature of the ARR and VQA tasks further justifies this method. By leveraging both visual and textual data, we aim to develop robust and accurate decision-support tools that can significantly improve the efficiency and effectiveness of CXR interpretation.

The technical chapters of this thesis are structured as follows. In Chapter 3, we investigate

how various image representations influence the performance of a VLM in detecting radiological findings in CXRs. Our study focuses on identifying conditions such as atelectasis, pleural effusion, and surgical devices using an image-text classification approach. We compare how different image pretraining methods for the image encoder impact VLM performance after fine-tuning. This includes fully-supervised training, where models use labeled datasets, and self-supervised approaches, which leverage unsupervised pretraining to autonomously label data and learn meaningful feature representations. In Chapter 4, we discuss how structured information can be extracted from CXRs to generate more clinically relevant radiology reports. We break down automated reporting into two separate sub-tasks where we first extract meaningful factual information in a structured format, expressed as triples (entity1, relation, entity2), used to condition the generation of the radiology report in the second step. In Chapter 5, we show how to effectively extract from CXRs local representations specific to the anatomical regions, as opposed to image-level representations. These local representations are then used as the input visual representation to an automated radiology reporting system, improving the predicted textual description. In Chapter 6, we demonstrate how to manipulate these anatomy-specific representations for different purposes. Firstly, we show how to combine the local representations of two subsequent scans to model the temporal evolution of findings. Secondly, we introduce a training strategy for the ARR model, enabling it to predict partial reports corresponding to a selected subset of regions. In Chapter 7, we integrate all the solutions proposed in previous chapters related to ARR into a unified model. We then conduct a human evaluation of its performance by implementing a comprehensive evaluation protocol. This protocol highlights errors and correct mentions by comparing the predicted reports with the original reports written by radiologists. In Chapter 8, we adapt our ARR solutions to enable CXR VQA, tackling both questions about individual CXRs and those regarding differences between two CXRs taken at different times. We then demonstrate how grounding the VQA with the predicted reports enhances the accuracy and quality of the generated answers.

A schematic summary of the contributions of each technical chapter is shown in Figure

1.1, and more details are provided in Chapter 1.7.

1.3 Chest X-Rays

A chest X-ray, also known as a chest radiograph, is a medical imaging test that uses a small amount of ionizing radiation to create 2D images of the chest, including the lungs, heart, ribs, and surrounding structures. It is one of the most commonly performed imaging tests [141] and is used to diagnose and monitor various medical conditions affecting the chest area.

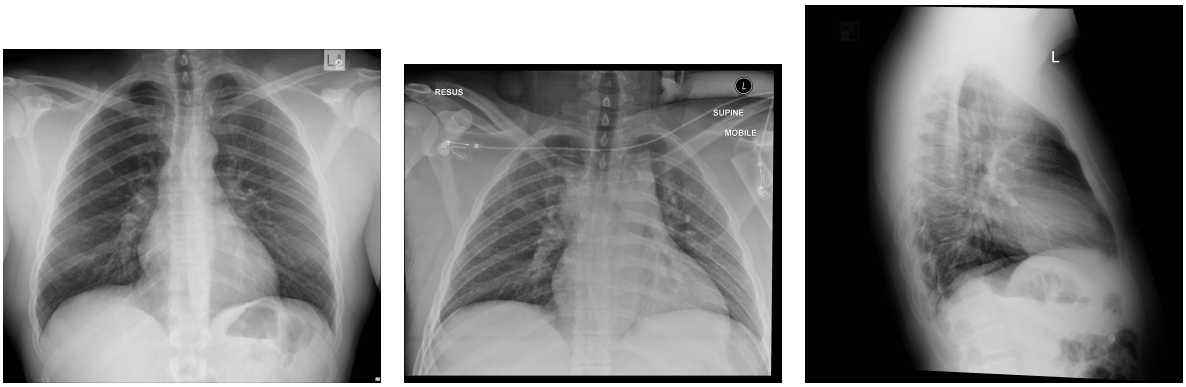
Different tissues in the body absorb X-rays differently due to variations in their density and composition. Tissues with higher density absorb more X-rays, appearing whiter on the resulting image, while tissues with lower density absorb fewer X-rays, appearing darker. For example, the dense tissues of bones absorb a significant portion of X-rays, resulting in white or light grey areas on the image. Soft tissues like muscles and organs have intermediate densities, leading to varying shades of grey on the image, while air has a very low density and absorbs minimal X-rays, resulting in black areas on the image. [3]

1.3.1 Medical Uses

Chest X-rays are often used as one of the initial investigations when suspecting thoracic disease, or to monitor their evolution. They are used to diagnose a wide range of conditions including, but not limited to:

- **Heart conditions**: enlargement of the heart and fluid accumulation around the heart.
- **Pneumonia**: inflammation or infection of the lungs.
- **Lung cancer**: abnormal growth of cells in the lungs.
- **Pleural effusion**: accumulation of fluid in the space between the lungs and chest wall.
- **Postoperative changes**: to monitor the recovery of a patient after surgery.

[3] https://www.nibib.nih.gov/science-education/science-topics/x-rays



(a) PA view. (b) AP view. (c) Lateral view.

Figure 1.2: Chest X-ray views of the same patient. Case courtesy of Yi-Jin Kuok, Radiopaedia.org, rID: 17910, Radiopaedia.org. [4]

- **Rib fractures**: breaks or cracks in the ribs.
- **Foreign bodies**: objects lodged in the airways, windpipe or chest cavity.
- **Placement of medical devices**: to assess that medical devices, such as pacemakers or catheters, are positioned correctly.

A CXR can sometimes detect an abnormality in the lung but does not show sufficient detail to accurately determine the exact diagnosis. For instance, an abnormal density may be due to simple pneumonia but diseases such as an underlying cancer cannot be excluded with absolute certainty on the basis of a CXR alone. When conditions are suspected on CXRs, additional imaging exams (*e.g.,* computed tomography scan) may be requested to provide further evidence.

1.3.2 Views

Chest X-rays can be taken from various views to provide different perspectives of the chest anatomy. [5] CXRs, as 2D projections of 3D anatomical structures, encounter limitations like

[4] Kuok Y, Differences in PA versus AP projection on a chest radiograph. Case study, Radiopaedia.org (Accessed on 12 Aug 2024) https://doi.org/10.53347/rID-17910.

[5] Jones J, Hacking C, Walizai T, et al. Chest radiograph. Reference article, Radiopaedia.org (Accessed on 22 Jul 2024) https://doi.org/10.53347/rID-14511

the superimposition of structures, lack of depth perception, and challenges in detecting small lesions. These issues can obscure critical details and complicate accurate assessments of abnormalities. Thus, each view offers unique advantages and is chosen based on the clinical scenario and the information needed. Common views for CXR are described below.

Posteroanterior (PA) View In a PA view, the X-ray beam enters the patient's back (posterior aspect) and exits through the front (anterior aspect), striking the detector placed against the patient's chest. The patient is typically requested to raise their arms in order to move the scapular bones outwards (laterally), thus minimising overlap with the thorax. The PA view is the standard view for obtaining a chest X-ray. It allows for better visualization of the heart and mediastinal structures with minimal overlap, providing a clearer image for diagnostic purposes.

Anteroposterior (AP) View In an AP view, the X-ray beam enters the patient's front (anterior aspect) and exits through the back (posterior aspect), striking the detector placed behind the patient's back. The AP view is typically used when the patient cannot stand or sit upright, such as in emergencies or when the patient is bedridden. However, in the AP view, the heart and mediastinal structures may appear larger and the lungs may appear more compressed than in the PA view, making interpretation more challenging. Further, the AP view causes increased radiation exposure to the anterior radiosensitive organs.

Lateral View In a lateral view, the X-ray beam enters from one side of the chest and exits through the opposite side, providing a side-to-side perspective of the chest anatomy. Lateral views are useful for evaluating the depth and position of structures within the chest, such as any abnormalities located behind the heart, which would be obscured by overlying structures on an AP or PA projection. There are two main types of lateral views: right and left lateral views. Lateral views are often used in conjunction with PA or AP views to provide a more comprehensive assessment of the chest.

Date: DD/MM/YY

View: Posteroanterior (PA)

Indication Field: 35-year-old male with a productive cough and fever.

Findings: There are patchy opacities noted in the right lower lung field suggestive of airspace consolidation. No evidence of pleural effusion. The heart size is within normal limits. The mediastinal contours appear unremarkable. No acute bony abnormalities identified.

Impression: Appearances in keeping with right lower zone pneumonia.

Figure 1.3: Example of synthetic CXR radiology report, using ChatGPT 3.5 [147] and verified by a junior radiology trainee. The main body of the report is divided into three sections: *indication field* (available at imaging time), *findings*, and *impression*.

Overall, the choice of CXR view depends on the clinical situation, the patient's condition, and the information needed for diagnosis and management. Each view provides different information about the chest anatomy and pathology, and they can be used in combination to obtain a comprehensive evaluation. An illustration of the different views is shown in Figure 1.2.

Throughout this thesis, we focus on AP and PA views, as lateral CXRs are less commonly requested and tend to be more challenging to interpret compared to frontal CXRs. Additionally, the annotations required for some of the proposed approaches in this work are available only for AP and PA views.

1.3.3 Radiology Report

After the acquisition and the interpretation of a CXR, radiologists document the observed findings in a radiology report. In the report, they typically describe the abnormalities and their locations, state their clinical impression (may be a differential diagnosis), and if necessary

make recommendations for further imaging.

The report generally includes three main sections:

- **Indication Field**: This section outlines the reason for the imaging study, including the patient's clinical history, symptoms, and any relevant medical information. It provides context for the radiologist's interpretation and helps guide the focus of the examination.
- **Findings**: In this section, radiologists detail the observed abnormalities and their precise locations. They describe the appearance of the lungs, heart, bones, and soft tissues, noting any deviations from normal anatomy or signs of disease. This objective account serves as the primary documentation of what is seen in the imaging study.
- **Impression**: The impression section synthesises the findings into a coherent clinical interpretation. Radiologists state their diagnostic conclusions, which may include a differential diagnosis, and assess the clinical significance of the findings. If necessary, they make recommendations for further imaging or follow-up to aid in patient management.

An example of a radiology report is shown in Figure 1.3.

1.4 What is Automated Radiology Reporting?

Automated reporting refers to the process of generating reports automatically through the use of AI, without requiring significant human intervention. In the context of radiology, automated reporting involves the use of AI to predict from medical images free-text descriptions of the findings, their anatomical location, and appearance. Auxiliary data could also be considered (*e.g.,* electronic health records (EHRs), laboratory results, etc.) to support the AI with more relevant information about the patients. By facilitating the timely delivery of accurate reports, automated reporting has the potential to improve the radiology workflow, often subject to delays, which can result in adverse patient outcomes. Therefore, some benefits of automating the reporting workflow in healthcare are:

- **Efficiency**: it saves time for healthcare professionals, enabling faster decision-making and more proactive management of patient care [17, 13].
- **Risk reduction**: it reduces the risk of human errors, caused by extensive working hours or distractions [194].
- **Standardisation**: it may improve consistency and standardisation in the format and content of reports, adhering to predefined templates and guidelines [160, 94].
- **Scalability**: automated reporting systems can handle large volumes of data efficiently, allowing healthcare organisations to scale their reporting capabilities to meet growing demands [194].

Overall, automated reporting has the potential to play a crucial role in modern healthcare by enabling efficient, accurate, and timely reporting of clinical data. By leveraging technology to automate repetitive tasks and streamline reporting workflows, healthcare organisations can enhance their efficiency, productivity, and quality of care.

1.5 What is Visual Question Answering?

Visual Question Answering (VQA) is a task that involves answering questions about images or visual data. While humans have an innate ability to perform VQA, automating this task requires complex steps of visual-linguistic reasoning which requires understanding both the content of the image and the semantics of the question posed about that image. First, the semantics of a question constrains the attention to focus on some specific details in the image. Second, these details and their relations with each other and the question are used to formulate the answer. This process applies naturally to humans whenever the answer requires generic knowledge. On the other hand, whenever the questions and the images involve some specialised expertise, only trained professionals can properly answer. For example, in the medical domain, specialised doctors are trained to perform this task accurately. Radiologists,

for instance, are often required to diagnose diseases or recommend treatments based on medical images such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI).

In the context of healthcare, VQA has various applications that leverage medical imaging data and clinical questions, including:

- **Clinical Decision Support** [58]: VQA systems can be integrated into clinical decision support systems to help healthcare professionals make informed decisions. By asking questions related to diagnostic findings (*e.g.,* "Is there evidence of pneumonia in this chest X-ray?"), treatment options (*e.g.,* "Based on the fracture location, what is the best course of treatment?"), or patient conditions (*e.g.,* "What is the severity of this patient's lung disease?"), clinicians can receive immediate answers or recommendations based on visual data and medical knowledge. For instance, it could help radiologists to speed up the diagnosis and treatment workflow, by asking multiple questions relevant to a set of medical images and collecting the answers to form the radiology report.

- **Patient Assistant** [27]: VQA can enhance patient education by providing explanations to patients' questions about their Personal Health Records (PHRs). It is unlikely that patients have the formation to interpret what abnormalities can be detected from a medical image, and the medical terminology is often too technical to be understood. This system could help explain (*e.g.*, in the form of an app) any doubt about the disease or treatments based on visual and textual information contained in the PHR and some other sources (*e.g.* medical journals). This could prevent patients from looking at misleading online sources, which are not necessarily wrong, but an inexperienced person may easily misinterpret their content. This system would need to relate information from PHRs and formulate the answer in a comprehensible manner.

- **Medical Training** [105]: VQA systems can be used for medical training and education purposes. Medical students, residents, and healthcare professionals can ask questions about medical images or clinical scenarios, and receive detailed explanations or annotations to aid in their learning process.

Overall, VQA holds significant promise for improving healthcare delivery, patient outcomes, and medical research by enabling intelligent interaction with visual data and supporting decision-making in clinical settings. A medical VQA system can alleviate the burden on healthcare professionals by answering routine questions, thus enhancing efficiency and reducing their workload. Unlike traditional methods such as medical finding classification, which predicts a limited set of predefined labels, or ARR, which describes findings in an image, VQA offers a more interactive approach. However, its complexity lies in its multimodal nature, as it requires interpreting visual information in the context of textual queries, making it a powerful but challenging tool for advancing medical diagnostics and care.

1.6 Thesis Statement

This thesis argues that effectively integrating VLMs into the radiological workflow can enhance the effectiveness of decision-support tools in CXR radiology. By processing and generating semantically rich textual information, VLMs can better align with radiologists' workflows. Furthermore, employing multiple VLMs—each dedicated to a specific subtask—can improve overall task performance, including Automated Radiology Reporting, and Visual Question Answering. Additionally, the strategic extraction of meaningful image representations enhances VLM performance in CXR interpretation. Finally, the flexibility of VLMs in handling diverse inputs and outputs allows for greater control over generated outputs and deeper interaction with imaging data.

1.7 Contributions

This section overviews the main contributions specific to each technical chapter.

Chapter 3 – Multimodal CXR Classification from Self-Supervised Image Encoders In this chapter, we investigate how to extract meaningful image representations for the task

of CXR finding classification. We focus on comparing the impact of different pre-training strategies for the image encoder when fine-tuned on the target task. We frame this as a multimodal task, by incorporating both visual inputs from CXR images and textual inputs from the indication field, aiming to leverage the complementary information provided by these two modalities. We explore how these pre-training strategies perform when the multimodal model is fine-tuned on a reduced subset of the training data, a scenario which is critical in real-world applications, where labelled data can often be scarce. We identify some pre-training strategies that prove effective in such constrained conditions, offering insights into their robustness and adaptability. To validate our approach, we conduct a series of experiments using publicly available datasets. We report comprehensive quantitative results that highlight the comparative performance of each pre-training strategy. Additionally, we examine the activation of image features across different strategies, providing a deeper understanding of how pre-training influences the model's ability to capture and utilise relevant information for CXR finding classification.

Chapter 4 – CXR Automated Reporting using Intermediate Triples Representations In this chapter, we study the effect of employing multiple VLMs—each dedicated to a specific subtask— to tackle ARR. We propose to use a clinically informed schema, to express the information in CXR radiology reports in a structured form, using triples (entity1, relation, entity2). Building on this, we propose a two-step pipeline for automated report generation that leverages this structured information, enhancing the quality of the generated reports. We define this as *TE-RG*, which performs *Triples Extractor* (TE) followed by *Report Generator* (RG). Our methodology is evaluated through extensive experiments on a large, publicly available dataset, where we demonstrate state-of-the-art performance across different metrics. To further validate the effectiveness of our approach, we conduct a human evaluation, categorising and quantifying different types of errors in the generated reports.

Chapter 5 – CXR Automated Reporting using Finding-Aware Anatomical Tokens In this chapter, we introduce a novel approach to extracting meaningful image representations of the CXRs, each corresponding to a different anatomical region contained in the image. Our method leverages a multi-task learning framework to localise anatomical regions, as well as detect relevant clinical findings within these regions. The extracted features, which we term *finding-aware anatomical tokens*, capture both anatomical and diagnostic information, providing a more fine-grained and informative visual representation. These tokens are then integrated into the previously introduced TE-RG pipeline for radiology report generation, enhancing the quality of the generated reports. We validate the effectiveness of this approach through comprehensive experiments on a widely used chest X-ray dataset, demonstrating improvements in the quality of the generated reports. To further assess the impact of the finding detection task on anatomical representation extraction, we employ a popular statistical method for visualising high-dimensional data. The results suggest that integrating finding detection as an auxiliary task improves the ability to extract meaningful anatomical features, leading to a more effective differentiation between normal and abnormal anatomical regions.

Chapter 6 – Longitudinal and Controllable CXR Automated Reporting In this chapter, we introduce a novel solution to enhance the generation of radiology reports by modelling the temporal evolution of CXRs from the same patient. Our approach involves aligning and integrating representations of equivalent anatomical regions from prior and current scans, allowing for a comprehensive joint representation that captures changes over time. This method provides a more nuanced understanding of patient progress, thereby improving the quality of the generated reports. Additionally, we propose a novel training strategy, called *sentence-anatomy dropout*, which trains the model to generate partial reports based on a sampled subset of anatomical regions. This approach enables the model to learn associations between specific anatomical regions and corresponding report content, offering greater control over which anatomical regions are reported on in the predicted report. We validate the effectiveness of our proposed methods through extensive experiments, demonstrating state-of-the-art performance

in report generation on a large-scale chest X-ray dataset.

Chapter 7 – Assessing Integrated Automated Reporting Solutions: A Human Evaluation In this chapter, we consolidate the CXR automated radiology reporting techniques proposed in earlier chapters into a unified method. This integrated approach combines the strengths of the previously proposed solutions, aiming to enhance the overall quality and accuracy of the generated reports. We then introduce a detailed protocol for expert human evaluation, designed to assess the clinical relevance and quality of the generated reports. We apply this evaluation protocol to a subset of radiology reports predicted by our unified method, providing valuable insights into the effectiveness and reliability of the proposed approach from a clinical perspective.

Chapter 8 – Grounding CXR Visual Question Answering with Radiology Reports In this chapter, we explore the use of sequential VLMs, each tailored to specific subtasks, to tackle VQA in the context of CXR. Building on our previously developed ARR solutions, our approach is designed to handle both individual CXR images and comparisons between successive images, allowing the system to provide accurate and relevant answers to clinical questions. Our solution is structured into two sequential stages: *Report Generation* and *Answer Generation*. This two-stage process leverages predicted radiology reports to inform and enhance the answer generation, ensuring that the VQA system benefits from a comprehensive clinical context. This integration improves the relevance and accuracy of the responses. We validate the effectiveness of our method by demonstrating state-of-the-art performance on a CXR VQA dataset.

Chapter 2

Technical Background

2.1 Introduction

The rapid advancement of machine learning has considerably influenced and shaped many research fields, including the healthcare sector. In recent years, numerous medical applications have arisen based on machine learning solutions, including assisting patients in retrieving health-related information from reliable online sources [45]; supporting the diagnosis of diseases through medical imaging analysis [205], or automating some clinicians' duties such as radiology report generation [138]. Many of these data-driven applications have relied on different biomedical data modalities: medical images, textual findings (*e.g.*, radiology reports) or collections of different records (*e.g.*, Electronic Health Records (EHRs)).

This chapter reviews some of the recent technical advancements in Computer Vision (CV) and Natural Language Processing (NLP), and how these have shaped the medical field. We discuss the similarities and differences that characterised the general and medical domains, showing some of the key challenges that affect the second. We start by giving a general overview of the methods in CV, NLP, and multimodal learning, and we introduce some of the metrics that have been widely used to measure the quality of text generative tasks (*e.g.*, image captioning, machine translation), highlighting some of the limitations. We then deepen into the literature of the main tasks covered in this thesis: Automated Radiology Reporting (ARR) and medical Visual Question Answering (VQA).

2.2 Computer Vision

Computer vision is the branch of AI that focuses on interpreting and understanding the world through imaging sensors (*e.g.*, images and videos). The recent trend in CV research is based on data-driven approaches and, in particular, deep learning models. CV tasks include *image classification* – where each image has to be assigned to its correct semantic category; *image object detection* – which requires detecting the coordinates of the bounding box for specific objects; *image segmentation* – similar to object detection, this time it is drawn a pixel-wise

mask for each object.

2.2.1 CNNs

In 2012, the presentation of AlexNet [104] during the ImageNet challenge [175] determined the passage to Convolutional Neural Networks (CNNs) as the standard approach in most CV tasks. CNN is a class of deep neural networks; it was inspired by the organization of the animal visual cortex, where neurons have receptive fields and respond to overlapping regions in the visual field. The main idea behind CNNs is to automatically and adaptively learn spatial hierarchies of features from the input data through the application of convolutions. The fundamental building block of a CNN are convolutional layers. Each consists of n 3D weight matrices, called filters or kernels, that convolve along the width and height of an input tensor (*e.g.*, an RGB image) to detect specific patterns or features. CNNs have shown to be particularly powerful for tasks involving images and spatial data, and numerous other CNNs have been proposed, each introducing new architectural designs. For instance, ResNet [65] introduced residual skip connections to overcome the problem of vanishing gradients during back-propagation. This class of CNNs is widely used as the backbone to numerous applications in image classification, object detection models (*e.g.*, Faster R-CNN [169]), and image segmentation models (*e.g.*, Mask R-CNN [64]). Numerous other CNNs architectures [73, 192] have been proposed to tackle CV tasks.

2.2.2 Faster R-CNN

Faster R-CNN [169] is a widely-used deep learning architecture for object detection, designed to improve the efficiency of region-based approaches by integrating a Region Proposal Network (RPN) directly into the detection pipeline. Faster R-CNN builds upon earlier models like R-CNN [56] and Fast R-CNN [55], by increasing both speed and accuracy in object detection tasks.

The architecture of Faster R-CNN usually consists of a CNN as the backbone network,

to extract feature maps from the input image. These feature maps serve as the input to the RPN. The RPN slides a small network over the feature map, predicting two outputs for each anchor: an objectness score and bounding box refinements. The objectness score indicates whether an anchor contains an object or belongs to the background, while the bounding box refinements adjust the anchor box coordinates to better fit the potential object. The RPN uses predefined anchor boxes of different scales and aspect ratios, allowing it to detect objects of various shapes and sizes. These anchors are matched with ground-truth boxes during training using the Intersection over Union (IoU) criterion, where anchors with high IoU (*i.e.*, greater than a threshold value) are labeled as positive, and those with low IoU (*i.e.*, lower than a threshold value) are labeled as negative.

The output of the RPN consists of a set of region proposals, which are then passed to the RoI Align [64] or RoI Pooling layer [56], to extract fixed-size feature maps for each proposed region from the feature map of the backbone. Unlike the RoI Pooling method, RoI Align is designed to avoid quantisation errors by using bilinear interpolation, ensuring that the spatial information is preserved. The fixed-size feature maps are fed into fully connected layers, which output the final predictions through two branches: one for classification to determine the object class and another for bounding box regression to refine the proposals.

Bounding box regression involves predicting offsets for the center coordinates, width, and height of the boxes. These offsets are calculated as follows:

$$\Delta x = \frac{x_{\text{gt}} - x_{\text{a}}}{w_{\text{a}}}, \quad \Delta y = \frac{y_{\text{gt}} - y_{\text{a}}}{h_{\text{a}}}, \quad \Delta w = \log\left(\frac{w_{\text{gt}}}{w_{\text{a}}}\right), \quad \Delta h = \log\left(\frac{h_{\text{gt}}}{h_{\text{a}}}\right) \tag{2.1}$$

where $x_{\text{gt}}, y_{\text{gt}}, w_{\text{gt}}, h_{\text{gt}}$ represent the center coordinates, width, and height of the ground-truth box, and $x_{\text{a}}, y_{\text{a}}, w_{\text{a}}, h_{\text{a}}$ represent the same for the anchor box. The predicted offsets are then used to adjust the anchor box parameters to better align with the ground-truth boxes.

Faster R-CNN employs a multi-task loss function to train the RPN and the final detection heads. The total loss is given by:

$$\mathcal{L} = \mathcal{L}_{\text{RPN}} + \mathcal{L}_{\text{Detector}} \tag{2.2}$$

The RPN loss consists of two components: the classification loss for objectness scores and the regression loss for bounding box refinements. This can be written as:

$$\mathcal{L}_{\text{RPN}} = \frac{1}{N_{\text{cls}}} \sum_i \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \frac{\lambda}{N_{\text{reg}}} \sum_i p_i^* \mathcal{L}_{\text{reg}}(t_i, t_i^*) \tag{2.3}$$

Here, $\mathcal{L}_{\text{cls}}$ is the binary cross-entropy loss for objectness scores, $\mathcal{L}_{\text{reg}}$ is the Smooth L1 loss for bounding box regression, p_i is the predicted probability of the anchor being foreground, p_i^* is the ground-truth label (1 for foreground, 0 for background), and t_i, t_i^* are the predicted and ground-truth regression parameters for the anchors. The terms N_{cls} and N_{reg} are normalisation factors for classification and regression, respectively, and λ is a weighting factor to balance the two losses.

The detector loss follows a similar form, combining classification and regression losses for the RoI-based predictions:

$$\mathcal{L}_{\text{Detector}} = \frac{1}{N_{\text{cls}}} \sum_j \mathcal{L}_{\text{cls}}(q_j, q_j^*) + \frac{\lambda}{N_{\text{reg}}} \sum_j q_j^* \mathcal{L}_{\text{reg}}(v_j, v_j^*) \tag{2.4}$$

Here, q_j represents the predicted class probabilities for a region, q_j^* is the ground-truth class label, and v_j, v_j^* are the predicted and ground-truth bounding box parameters for the regions.

After the RPN generates proposals, post-processing steps are applied. These include filtering out low-scoring proposals, clipping proposals that extend beyond image boundaries, and performing Non-Maximum Suppression to remove redundant proposals that have a high degree of overlap. The remaining proposals are refined further by the detection heads to produce the final object detections.

Faster R-CNN integrates region proposal generation directly into the detection pipeline, improving efficiency compared to earlier approaches that relied on external proposal methods. By sharing convolutional features between the RPN and detection heads, it achieves a balance

between computational speed and detection accuracy. However, Faster R-CNN can be computationally intensive compared to single-stage detectors like YOLO [167] and SSD [133], which unify region proposal and detection.

2.2.3 Transfer Learning to Medical Images

CNNs and other CV architectures have proven to be successful in many general-purpose tasks due to large-scale open-source datasets. In the general domain, specific tasks with fewer available data have generally benefited from information learned from larger and more generic datasets, given that natural image datasets follow quite similar distributions. Differently, medical images are perceptually very dissimilar from natural images and their distribution varies considerably depending on the modalities used to capture them. Therefore, a key challenge consists of how to adapt CV solutions from the general domain to the medical domain. Yet, the use of CNN – and deep learning approaches in general – for medical applications has strongly relied on transferring the knowledge learnt from large-scale general domain collections, as the size of the medical image databases is remarkably smaller. This technique is referred to as *transfer learning*. More specifically, transfer learning is generally based on a two steps training procedure: (1) *pre-training* – the model is first trained on large-scale general domain datasets (*e.g.*, Imagenet [175]); (2) *fine-tuning* – the pre-trained model is then trained on the task-specific dataset, which domain may be different. Raghu et al. [161] explored the impact that transfer learning has on medical imaging, analysing the results on two datasets – Retina dataset [59] and CheXpert [83] – of different CNNs and different initialisation techniques – pre-training and random initialisation. Their results suggest that transfer learning does not significantly affect the performances of CNN models, especially when these have fewer parameters; while for larger models the benefit of transfer learning may be due to over-parametrisation rather than reusing relevant features. Raghu et al. suggest using a hybrid approach, where the weights at the lowest layers are reused, which contains the most meaningful features; redesigning the top layers with fewer parameters, and initialising

them randomly.

2.3 Natural Language Processing

NLP refers to the branch of AI dealing with natural language, where a computer is asked to understand, interpret, and generate human language. NLP covers a wide range of different tasks such as summarisation, question-answering, machine translation, and sentiment analysis. In the era of big data, NLP methods have gradually shifted from sophisticated rule-based models to data-driven approaches. The large collection of available data has inspired researchers to adopt machine-learning approaches: from multi-layer perceptrons to more complicated deep neural networks.

2.3.1 RNNs

Nowadays, the use of neural networks in NLP has become a standard in almost all state-of-the-art solutions. Recurrent Neural Networks (RNNs) have been used as the backbone of many NLP models – due to their ability to process sequential data – such as the human language. RNNs have the advantage of handling input text without any restriction on the length and with no increase in the model size. However, vanilla RNNs lack of accessing information from many steps back. In the case of human language, this means that RNNs lack linking the representations of words whose position is far apart in the text. To address this problem, together with the need to increase the representation capacity of these models, RNN architectures have been further improved with the introduction of Long Short-Term Memory (LSTM) [68] and Gated Recurrent Unit (GRU) [34].

2.3.2 Transformers

The increment of large open-source text corpora (*e.g.*, Wikipedia) has encouraged researchers to design highly parallelisable models. This led to the Transformer [198] architecture, in

2017, a neural network based on the attention mechanism. Compared to RNNs that process input data sequentially, Transformers have been designed to process input data in parallel, reducing the training time. Based on this architecture, numerous pre-trained models have been introduced – Bidirectional Encoder Representations from Transformers (BERT) [47] and Generative Pre-trained Transformer (GPT) [25], to name a few. These models are characterizsd by unsupervised pre-training objectives, which enable the model to learn bidirectional representations from large corpora of unlabelled text. For instance, Masked Language Modeling (MLM) corresponds to the task of predicting masked tokens, which normally correspond to 15% of the input tokens.

2.3.3 Medical Transformers

Inspired by the success of general-purpose models, similar approaches have been adopted for biomedical and clinical applications. In recent years, the use of Transformer-based models and their pre-training strategies (*e.g.*, MLM) have shown state-of-the-art results in biomedical text mining. Indeed, the main difference between generic and medical text consists in the lexicon adopted, where medical texts are generally quite technical and require specific knowledge. To overcome this, the success of biomedical NLP solutions can be attributed to the large open-source biomedical corpora (*e.g.*, PubMed) adopted in the pre-training step, such that models can learn the domain-specific vocabulary. One of the early works by Lee et al. [106] proposed BioBERT. This model follows a two steps pre-training approach: (1) BioBERT is initialised using the weights of BERT, pre-trained on general domain corpora (English Wikipedia and BooksCorpus [239]); (2) BioBERT is trained on domain-specific datasets (PubMed abstracts and PMC full-text articles). BioBERT is then fine-tuned on the task-specific dataset. Similar works in domain adaptation based on Transformer architectures are: SciBERT [18] – pre-trained on 1.14M papers from Semantic Scholar [36]; ClinicalBERT [74] – pre-trained on MIMIC-III database [90]; BioELMo [88] – pre-trained on 10M PubMed abstracts.

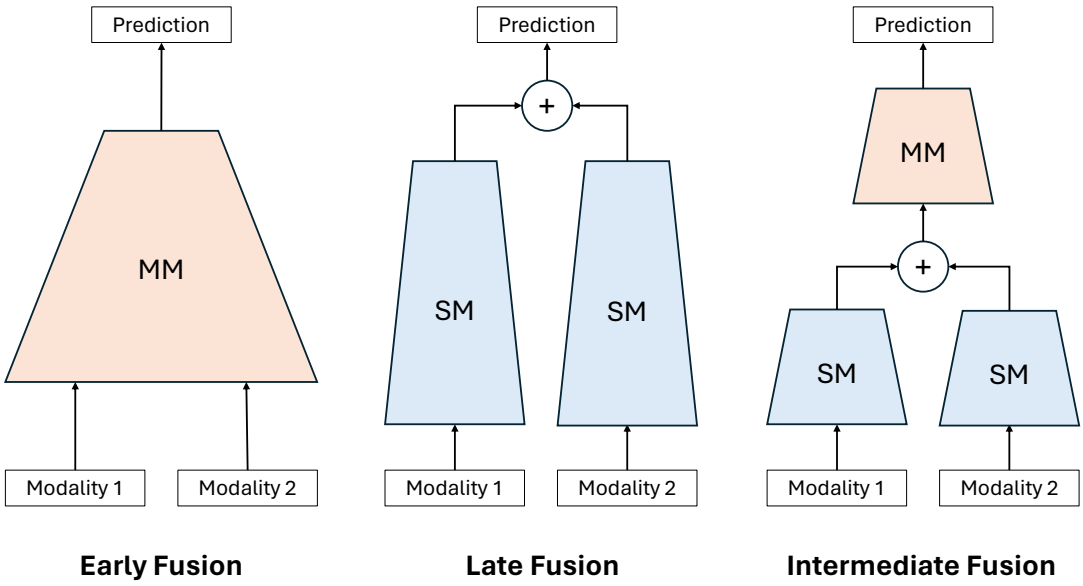


Figure 2.1: Examples of the different fusion techniques show how to combine singlemodal (SM) and multimodal (MM) components. In these examples, we only show the fusion techniques for two modalities, but similar approaches apply when we have more.

2.4 Multimodal Learning

Multimodal learning refers to the process of training models using data from multiple modalities, such as text, images, audio, video, sensor data, etc. For example, in a task like sentiment analysis, a multimodal model might analyse both the text content of a review and the accompanying images to make a more accurate prediction of the sentiment expressed. By combining information from different sources, multimodal learning aims to exploit the complementary nature of different modalities to build more robust and versatile models capable of handling diverse types of data, which can lead to better performance on various tasks such as classification, regression, generation, and more. This is particularly relevant for biomedical applications [101] where predictions are made based on multiple data sources: images (*e.g.*, CT, X-Rays), text (*e.g.*, radiology reports, EHR), time series, structured data, etc. As we describe in more detail in later sections, most of the tasks tackled in this thesis have an intrinsic multimodal nature: *automated radiology reporting* – an image-to-text task of generating a textual description

based on a medical image; and *visual question answering* – the task of answering questions (expressed in free-text) referred to the visual features of an image.

2.4.1 Fusion Techniques

One of the main challenges of multimodal learning is the effective fusing of different modalities in a common space. The fusion techniques are generally characterised into 3 main groups depending at which stage in the architecture of a model the different data sources are merged: *early*, *late*, and *intermediate fusion* (Figure 2.1).

Early Fusion

In early fusions, different data sources are joined together at the data level, before the learning process. These are typically vectorised into a common input space and passed to a multimodal model. Different methods of extracting features may depend on the nature of the modality [60], and multiple methods have been adopted to project all data in a common feature space (*e.g.* canonical correlation analysis [23] and non-negative matrix factorisation [7]). These methods have the advantage of capturing correlations between different data modalities but do not take advantage of more advanced learnable feature extractors such as CNNs for images or Transformers for text.

Late Fusion

In late fusion, the model consists of specialised components one for each modality, and, similar to ensemble methods, the predictions are based on combining the output of the individual components through some fusion operations at the decision level of the architecture. For instance, by performing some fusion operations of the output features of each component (*e.g.* concatenation [233, 84], element-wise addition, [177] or element-wise multiplication [9]), or combining the different predictions (*e.g.*, through a voting mechanism [178]). This results in very simple approaches but, since each modality is processed separately, they cannot learn

multimodal representations, often needed when making predictions.

Intermediate Fusion

Most recent approaches adopt intermediate fusion, as it has the advantage of both alternatives: better feature extraction using trainable encoders, and multimodal modules to fuse the different data sources into joint representations. These methods have been very successful, especially for image-text multimodal models where images are encoded through CNNs [98, 77, 85] or by extracting the region of interests from an object detector [8, 134, 111, 188, 154, 191], and use RNNs or Transformers to learn joint representations. More recently, similar architectures have been proposed to combine different data modalities into LLMs [148, 195], to form what is called Large Multimodal Models (LMMs) or Multimodal Large Language Models (MLLMs) [109, 216, 132, 108].

These methods are often characterised by a self-supervised joint pre-training step. For instance, Lu et al. [134] presented Vision & Language BERT (ViLBERT), a BERT architecture composed of two streams – one for each modality – that communicate to each other through co-attention transformer blocks. The input of the visual stream corresponds to the region-based features extracted from the Faster R-CNN and the spatial location of the bounding box. This model is pre-trained using two self-supervised tasks: (1) *Masked Multi-Modal Modelling* (4M) task – the 15% of the input tokens and bounding-boxes are masked and the model is trained to reconstruct them; (2) *Image-Text Matching* (ITM) – the model is trained to recognise if the text is relevant to the image and vice-versa. After pre-training, ViLBERT was then transferred to different vision-and-language tasks including VQA.

Concurrently, a similar model named VisualBERT was proposed [111]. This model differs from [134] since only a single BERT-like model is trained for both visual and image features, reducing the number of parameters. Furthermore, it is also pre-trained on the same objectives with the only difference that instead of 4M it performs Masked Language Modeling (MLM), where only the input tokens are masked.

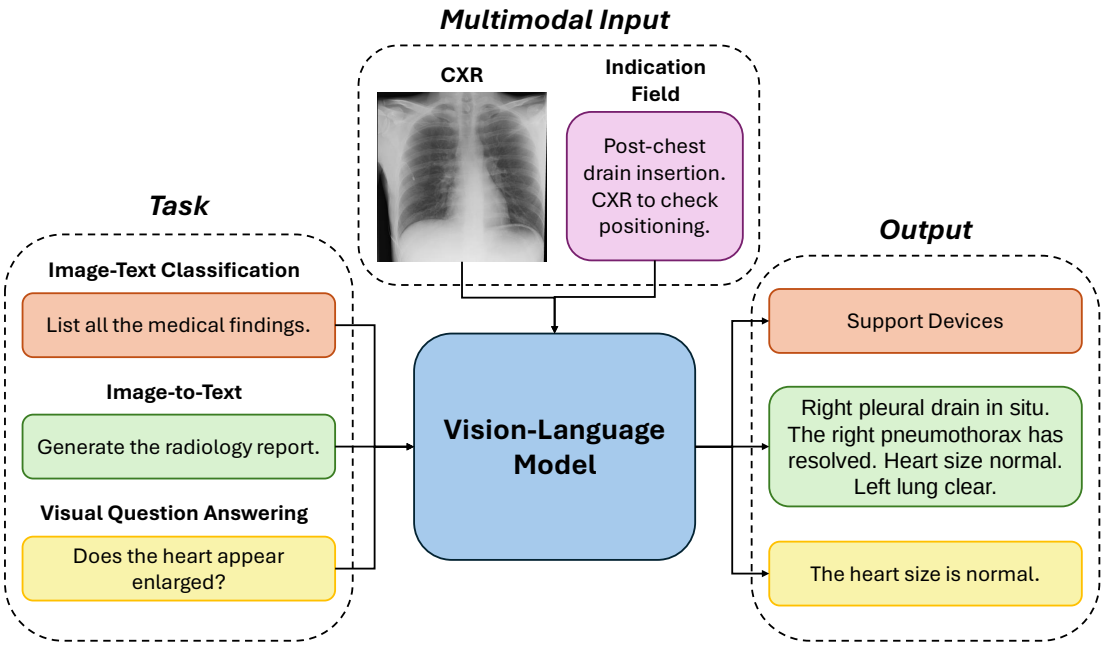


Figure 2.2: Overview of Vision-Language model applications for CXR interpretation.

Other similar models have been proposed [188, 154, 191, 235], each introducing new pre-trained objectives or architecture designs. More precisely, ImageBERT [154] proposed a multi-stage pre-training using four different objectives for each stage: (1) MLM; (2) Masked Object Classification (MOC) – only the bounding-boxes are masked; (3) Masked Region Feature Regression – similar to MOC but aiming at reconstructing the feature embedding by adding a fully connected layer on top; (4) ITM. ImageBERT is first pre-trained on a large-scale weak-supervised image-text data collection, also proposed in [154], and further pre-trained on the less noisy open-source collections.

Throughout our thesis, we adopt intermediate fusion to merge textual and visual modalities.

2.4.2 Vision-Language Models

VLMs [24] are a class of AI designed to interpret, generate, and reason with both visual data – such as images or videos that depict visual representations – and textual data – which includes linguistic descriptions or information related to the visual content.

VLMs are employed across a diverse range of tasks, each leveraging the interplay between

visual and textual information. Depending on the specific application, VLMs can process and/or generate images and text, performing tasks such as:

- **Multimodal Classification**: classification tasks that rely on the combined interpretation of both textual and visual information.
- **Image-To-Text (Image Caption)**: the generation of descriptive text that accurately reflects the objects, actions, or scenes depicted in an image.
- **Visual Question Answering**: the ability to answer questions posed about a given image, leveraging both visual and textual understanding
- **Text-to-Image**: the creation of images based on textual descriptions, synthesising visuals that align with the provided context..[1]

Notable vision-language models include CLIP [158], DALL·E [164, 163], LXMERT [191], BLIP [110, 109], and LLaVa [131]. CLIP uses contrastive learning to associate images with descriptions, enabling tasks like image classification and zero-shot learning. DALL·E generates creative images from textual descriptions by understanding and synthesising visual context. LXMERT extends the BERT architecture to handle both visual and textual data, excelling in tasks requiring deep multimodal understanding. BLIP, or Bootstrapping Language-Image Pre-training, enhances performance across various tasks by generating rich, semantically meaningful image-text pairs. LLaVa, a recent advancement, improves multimodal integration by leveraging large-scale pre-training and fine-tuning strategies to enhance performance in complex vision-language tasks.

In the context of this thesis, we will explore the increasing significance of VLMs in the medical field, with a particular focus on CXR interpretation (see Figure 2.2 for an overview of different applications). VLMs can effectively utilise the textual information often associated with medical scans, such as radiology reports, and enable advanced interaction capabilities, including responding to specific queries about the scans. This thesis highlights the potential

[1]Throughout this thesis, we will focus solely on the first three tasks and will not focus on image generation.

of VLMs to enhance medical diagnostics and patient care through improved multimodal understanding.

2.5 End-to-end vs Multi-Stage

The *end-to-end* [65, 198, 47] and *multi-stage* [232, 207] approaches represent two distinct strategies in machine learning, each with its own set of strengths and weaknesses.

In an *end-to-end* approach, the model is designed to learn a direct mapping from input to output without any intermediate steps. The system is trained as a single unified model that takes raw data as input and produces the final result in one step. This method simplifies the design by eliminating the need for separate modules or stages, allowing the system to optimise the entire pipeline jointly. Additionally, end-to-end systems can be highly efficient in optimising performance across the entire pipeline, as they leverage large datasets to learn complex mappings directly. End-to-end models also offer greater adaptability, as they are easier to apply to new domains or tasks by fine-tuning the model with appropriate data. However, end-to-end approaches can suffer from limited transparency, as the system functions as a black box, making it difficult to incorporate domain-specific knowledge or make changes to individual components.

In contrast, a multi-stage approach divides the process into distinct, modular components, with each stage focusing on a specific task or subtask. The output from one stage serves as input to the next, and the system often requires explicit intermediate representations. This modularity allows for more flexibility and interpretability, as each stage can be tailored or refined independently. Multi-stage approaches are particularly well-suited for tasks where domain-specific expertise must be incorporated. The flexibility of multi-stage systems also supports easier scalability, as new stages can be added to the pipeline without significantly disrupting the rest of the system. Furthermore, because each stage is processed independently, errors can be isolated for each stage and corrected without affecting the entire process. However, multi-stage approaches suffer error propagation, as errors in the earlier stages can

propagate through the pipeline, potentially affecting the final output.

2.6 Automated Radiology Reporting

ARR is a special case of the broader image captioning task, applied to radiology. Image captioning corresponds to generating a textual description based on the visual content of the image. In the case of ARR, the image corresponds to a medical scan and the textual description to the associated radiology report. Recent methods have adopted encoder-decoder architectures, in which the image embeddings are normally computed using CNNs (*e.g.,* [65]) or Vision Transformers (ViT) [49] and the text is generated using Recurrent Neural Networks (RNNs) (*e.g.,* [68] and [34]), or, more recently, using Transformer-based architectures [198].

Advancements in the general domain have often inspired ARR approaches [8, 37, 118, 229]. However, despite image captioning and ARR being *image-to-text* tasks, they present some differences and unique challenges. The former usually aims to generate a short description of the visual scene shown in a natural image, where subjects are usually placed in the foreground and the context (*e.g.*, the location of the scene) is defined from the background. The appearance of natural images can vary enormously – due to the range of possible objects, locations, lights, etc. – and the corresponding captions can be phrased in multiple ways. Whereas, for each type of medical scan (*e.g.* CT, MRI, X-Rays, etc.) and the same region of the body (*e.g.* head, chest, leg, etc.) the high-level appearance may look very similar for non-experts, and the description of similar findings in different radiology reports are often phrased in similar ways (*e.g.,* "The cardiac silhouette is normal.", "The cardiac silhouette is unremarkable."). The target of ARR is not to generate a generic and "superficial" description of what the image is showing (*e.g.,* body region, organs, etc.), but rather a long textual description of subtle details emerging in the scan, describing both normal and abnormal features. Due to these differences, simply transferring image captioning approaches to ARR is often insufficient but requires some ad-hoc solutions.

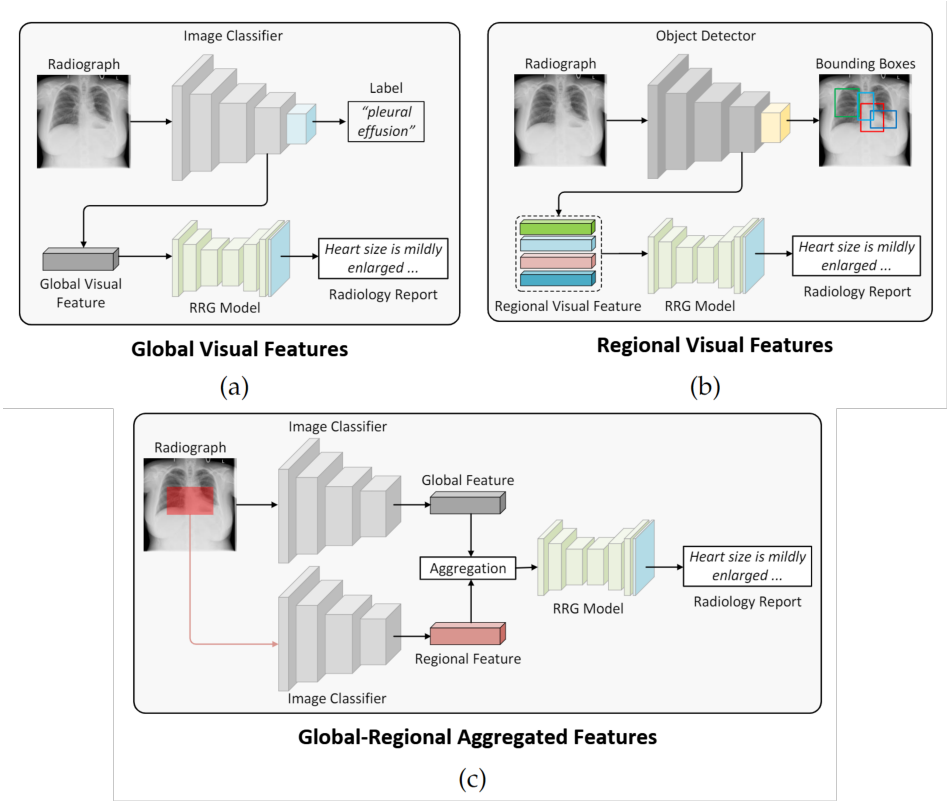


Figure 2.3: Different architectural designs of the image encoding of ARR approaches. They vary depending on the granularity of the visual features extracted from the medical image. Image taken from [126].

2.6.1 Taxonomy of ARR approaches

In this section, we follow the categorisations between different ARR approaches as proposed in [126], where they compare the different methods based on the architectural designs and training strategies specific to each modality involved: *vision*, *text*, and *cross-modal*.

Vision

We first categorise different approaches depending on how the input image is encoded into visual features. This is inspired by the different ways that visual-linguistic models in the general domain encode images. Common solutions are using CNNs (*e.g.* VGG [182], ResNet [65], DenseNet [73], etc.) [77] or using the region features extracted by the Region Proposal Network (RPN) of an object detector (*e.g* Faster R-CNN [169]) [8].

Depending on the granularity of these features, ARR methods can be classified into three main categories: global, regional, and global-regional aggregation (Figure 2.3). Approaches focusing on *global visual features* are the majority and they aim to extract a global image representation of the whole image using a CNN [32, 225, 31, 78, 128, 155] or ViT [211, 114]. Differently, *regional visual features* aims at providing fine-grained representations of single regions within the image. These are extracted using object detection algorithms [193] (*e.g.*, Faster R-CNN [169]) or through selective search algorithm to unsupervisely generate region proposals [212]. Finally, approaches using *global-regional aggregated features* aim at extracting both types of image representations followed by an aggregation step, performed through a self-adaptive fusion module [116, 210].

The use of global features has the advantage of having a simpler model design and a faster model during inference. Region-based solutions, on the contrary, are more label-demanding – since the object detector is usually trained using bounding box annotations, which are expensive to collect – but have a higher degree of interpretability, given that each input embedding is associated with a corresponding anatomical region.

Text

Most existing works have handled the text generation process by connecting the extracted visual features to an autoregressive model (*e.g.,* Transformer [198] or LSTM [68]), to generate the output text sequentially. Compared to image captions in the general domain, radiology reports are often very long textual descriptions, dense in medical terminology, requiring very specialised expertise, and highly patterned. ARR methods have considered these aspects by proposing ad-hoc solutions. For instance, they have considered extracting medical terms from the image and using them to enhance report generation. These terms are often extracted as a fixed set of radiographic observations [228, 6], or using external knowledge graphs [107, 221, 115].

Other solutions have taken advantage of the highly patterned nature of radiology reports

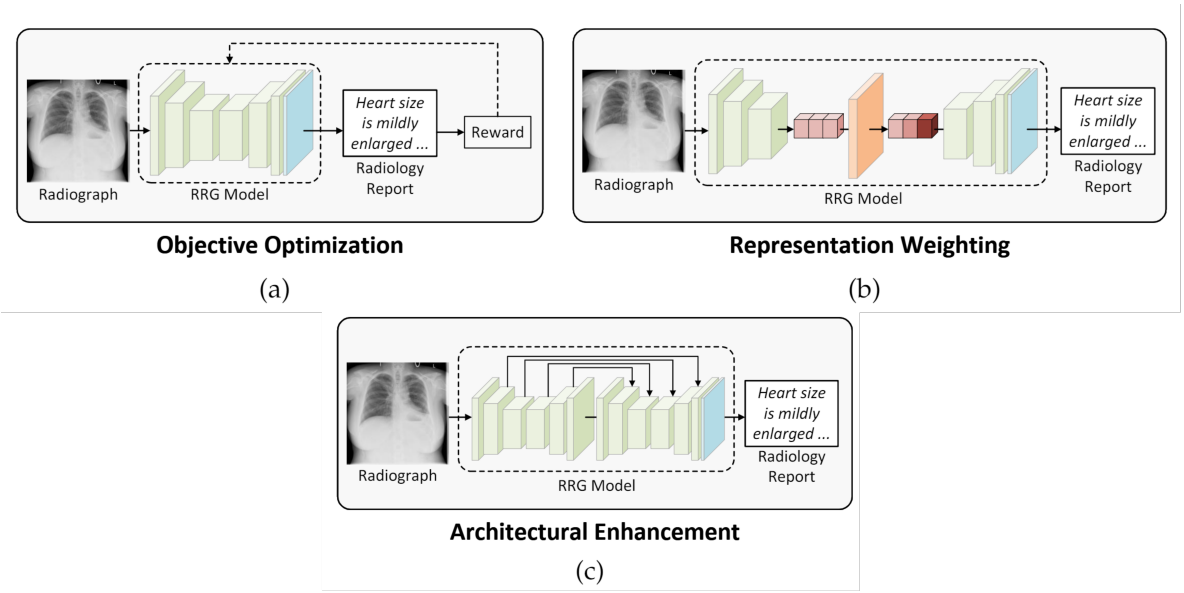


Figure 2.4: Different ARR approaches to enhance the cross-modal alignment between the visual (radiograph) and textual (report) components. Image taken from [126].

by considering report templates or report clusters. In [107, 96], based on the detected abnormalities from the image, they retrieve sentence templates specific to the abnormalities to form the final report. In [218], reports in the training set which are semantically similar to one another are clustered together and propose a weakly supervised contrastive learning approach to encourage the generated reports to be semantically close to the corresponding cluster.

Other works have considered dividing the challenging one-step report generation approach into multiple steps, either by first extracting intermediate high-level summaries of the report [144] or by generating the report in a sentence-to-sentence paradigm [217].

Cross-modal

Many works have leveraged the multimodal nature of ARR, to enhance the alignment between the visual and textual components. This has been done through *objective optimisation*, *representation weighting* and *architecture enhancement* (Figure 2.4). Objective optimisation has been achieved through reinforcement learning [143, 42, 137, 155], as opposed to standard cross-entropy optimisation, to better align the loss function with the evaluation metrics – by considering these metrics as rewards – or to improve consistency in the generated reports

[137]. Alternatively, a multimodal curriculum learning approach has been proposed in [127], to gradually train from easy to more complex samples, based on the number of abnormalities. Other objective optimisation techniques include self-boosting [212] – where two separate branches are considered to perform radiograph-report matching and report generation – and pre-training [142] – where they study how different pre-trained models affect ARR.

Representation weighting leverages the attention mechanisms and the memory network to better align the cross-modal representations. For instance, in [89], they propose a co-attention module to weight visual representation with medical terms, while in [128, 129] they improve the attention modules in the Transformer. Another work has introduced relational memory modules [32] to allow the model to memorise information from previous generations, and cross-modal memory modules [31] to encourage better alignment between visual and textual information.

Other works have focused on improving the architecture of ARR models. In [211], multiple learnable "expert" tokens are introduced into the encoder and decoder of a transformer and are encouraged to attend to different regions of the input image and to capture complementary information. In [224], they propose using a hierarchical transformer to align the input regions of the radiograph and disease tags.

2.7 Medical Visual Question Answering

VQA correspond to the task of asking questions that can only be answered by examining a specific image. For example, given an image containing some people, a question could be about the action performed by one of them, about their clothing, and so on. VQA can be considered as a cognitive task, requiring multi-modal reasoning as well as understanding the relationship between objects and the context of the scene.

VQA systems typically combine CV techniques to understand the visual content and NLP techniques to comprehend and answer questions in human language. Most of the proposed solutions are based on deep neural networks, given the great results obtained in the fields

of CV and NLP. The image feature extractor backbone is generally a CNN architecture (*e.g* ResNet [65], VGG [182], Faster R-CNN [169], U-Net [173], etc.). The question understanding and answer generation are generally achieved using RNNs (*e.g.* LSTM [68], GRU [34]) or Transfomer-based architectures [198] (e.g. BERT [47], GPT [25], etc.).

Due to its multi-modal nature, VQA inherits most of the issues described in NLP and CV when applied to the medical domain, as well as sharing some of the challenges already discussed for ARR, and multimodal learning in general, such as thoroughly fusing image and textual representations. Another challenge is posed by the scarcity of labelled VQA medical data – in general, multimodal medical collections are rarely publicly released, due to privacy concerns. Besides containing fewer samples, questions are often trivial (*e.g.* "What plane is this?", "What modality is this?"), which makes it harder to establish the level of multi-modal reasoning learnt by models. For this reason, top-scoring models in some medical VQA challenges [62, 21, 19] approach VQA as an image classification problem. Due to such limitations, medical VQA is still at an early stage. Figure 2.5 showcases examples of medical VQA samples drawn from various datasets.

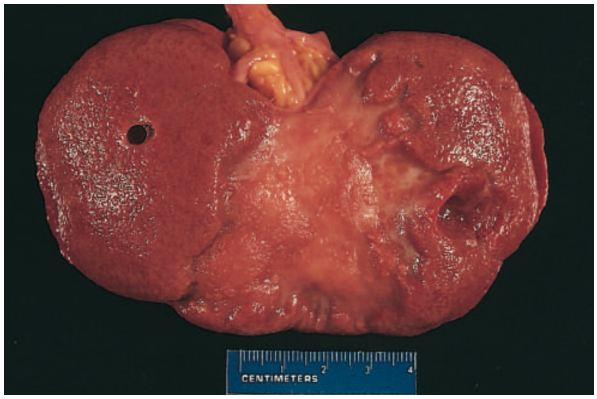
2.7.1 Taxonomy of Medical VQA approaches

We now describe the most common approaches to tackle medical VQA introduced in recent years. VQA can be formulated as a classification task, for multiple-choice answers or when the dataset contains a fixed set of possible answers; or more broadly as a text-generation task, for open-ended questions.

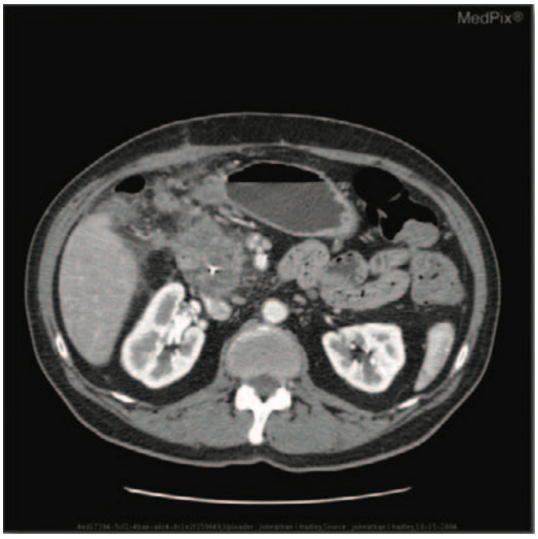
Classification VQA

The task of medical VQA has often been tackled as a classification problem. Methods adopting such an approach have, for a long time, shown empirical better performances, often due to the low variability in the questions and answers. In the case of open-ended VQA, the n possible answers in the dataset are treated as n independent classes. Whereas, for multiple-choice

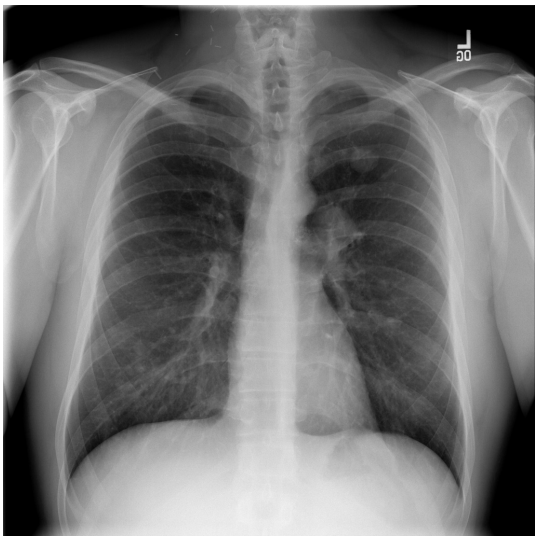
(a) **VQA-Med-2020** [19]
Q: what abnormality is seen in the image?
A: ollier's disease, enchondromatosis



(b) **Path-VQA** [66]
Q: is remote kidney infarct replaced by a large fibrotic scar?
A: yes



(c) **VQA-RAD** [105]
Q: what is the location of the mass?
A: head of the pancreas



(d) **Slake** [125]
Q: where is/are the abnormality located?
A: left lung, upper right

Figure 2.5: Samples taken from the respective medical domain datasets.

VQA, the answer among the *m* provided candidates having the highest likelihood among them is selected.

By taking a look at the VQA-Med challenges that took place in 2018, 2019 and 2020, hosted by ImageCLEF competition [80, 81, 82], we see the leading team approaching VQA as a classification task. In the 2018 VQA-Med challenge [62], the top three leading teams [20, 236, 153] approached the VQA task as a classification problem rather than a text generation task. All three models were based on a CNN-RNN approach, using different network architectures for both image and textual streams, and different fusion and attention techniques. Similarly, in 2019 [21], the top three teams [202, 219, 237] mainly approached the VQA task as a classification problem, except [237] which proposed a two-stage solution: (1) a classifier discriminates between the four question categories – modality, plane, organ system, and abnormality; (2) for the first three categories a classifier is used to determine the answer, while a sequence-to-sequence model is applied for the abnormality type of questions. In the 2020 ImageCLEF competition, the three leading teams of the VQA-Med challenge [19] were [121, 4, 95]. The top two teams formulated the VQA task as an image classification problem, due to the nature of the dataset and its repetitiveness. They consider the visual representation to play a more decisive role in selecting the correct answer. This was achieved by: (1) categorizing the questions into *n* classes; (2) for each of these classes a different image classifier is used to infer the answer. This highlights how this dataset is very limited and the linguistic side of the VQA task can be bypassed through some simple heuristics.

We present other notable methods that have adopted a classification-based approach to perform VQA. In [140], Nguyen et al. explore how to overcome the data limitation problem in medical VQA. They propose the Mixture Enhanced Visual Features model, in which they combine a Convolutional Denoising Auto Encoder and Model-Agnostic Meta-Learning to initialise the model weights for the image feature extraction. Khare et al. [98] introduced Multimodal Medical BERT (MMBERT) where they investigate the use of a joint pre-training strategy for medical VQA. More specifically, MMBERT is pre-trained on the Radiology

Objects in COntext (ROCO) dataset [151], containing over 81,000 radiology images of different modalities and the corresponding captions, using the MLM objective, by masking only medical keywords. More recently, Pellegrini et al. [152], proposed a new benchmark dataset (Rad-ReStruct) and a new method (hi-VQA) to model the structured reporting task as a hierarchical VQA task. To this end, based on the pre-defined question hierarchy, their method considers previously asked questions and answers – higher up in the hierarchy – to answer more specific questions about the same patient.

Other works have focused on VQA in surgery, to answer questions from surgical scenes, which could help medical students and junior surgeons to learn from surgical videos. Again, the limitation posed by the lack of available datasets in this task has leaned in favour of classification-based methods. For instance, Seenivasan et al. [179] proposed SurgicalGPT, a fine-tuned vision-language model based on GPT-2 and a classification head to perform answer classification. In [12], the authors proposed a Transformer with Co-Attention gaTed Vision-Language (CAT-ViL) embedding, a Data-Efficient Image Transformer (DeiT) module to jointly perform answer classification and detection of the region of interest.

Generative VQA

Considering VQA as a classification task highly restricts the problem. This is particularly limiting for open-ended questions, where questions are inquiries that cannot be answered with a simple "yes" or "no" response. Instead, they require more elaborate and thoughtful answers, that can be phrased in multiple ways. This motivates the need for generative VQA systems, typically designed in an encoder-decoder fashion, capable of generating "original" answers.

Ren et al. [168] proposed the CGMVQA model, one of the earliest works in this space, which performs both classification and answer generation. CGMVQA shares the same backbone encoders, based on ResNet and BERT, while only changing the output layers. Similarly, Sharma et al. [181] propose MedFuseNet – an attention-based multimodal model – which tackles the answer prediction as both categorization and generation, by using two

Dataset	Tasks	Annotations
MIMIC-CXR	ARR, Finding Classification	CXRs, radiology reports, medical findings (14 labels).
Chest ImaGenome	Finding Detection, Anatomy Localisation[2]	anatomical region bounding boxes (36 labels) with associated sentences and medical findings (71 labels).
Medical-Diff-VQA	VQA	question-answer pairs related to CXRs

Table 2.1: Datasets used in this thesis, along with corresponding tasks they are used for and annotations they support. Both Chest ImaGenome and Medical-Diff-VQA are derived from MIMIC-CXR and contain CXRs from the same source.

separate heads composed of Fully-Connected and LSTM layers, respectively. More recently, van Sonsbeek et al. [197] proposed a fully generative method, leveraging GPT2-XL [159] as the text generation component, and using Low-Rank Adaptation (LoRA) [69] as the parameter-efficient fine-tuning strategy. Li et al. [117] show the benefit of combining several unimodal and multimodal pre-training strategies to improve the VQA downstream task: Unimodal Contrastive Loss (Text and Image), Multimodal Contrastive Loss (Text-Image), Image Text Matching, Masked Language Modeling.

Hu et al. [72] proposed a novel Chest-X ray Difference VQA task, where the questions require the comparison between current and prior CXRs (*e.g.,* "What has changed compared to the past image?"), and a novel expert knowledge-aware graph representation learning model (EKAID) to address this task. Their method was further superseded by Cho et al. [35], who proposed PLURAL, a pre-trained vision-language model, which shows state-of-the-art performances on the same task.

2.8 Datasets

This section describes the various datasets and their annotations that are utilised throughout the research. A summary of these datasets is presented in Table 2.1, including the tasks they

[2]Finding detection and anatomy localisation are intermediate tasks not directly addressed or improved upon in this thesis. They are used to extract meaningful representations from CXRs, as described in Chapter 5.

Dataset	Image Modality	# Images	# Reports
IU-Xray [44]	CXR	8,121	3,996
PadChest [26]	CXR	160,868	206,222
FFA-IR [112]	FFA	1,048,584	10,790
COV-CTR [26]	CT	728	728
MIMIC-CXR [91, 92, 57]	CXR	377,110	227,835

Table 2.2: Automated radiology reporting datasets. This table contrasts MIMIC-CXR, the dataset used in this thesis, with other open-access datasets.

are used for and the annotations they support.

Research on ARR has predominantly focused on CXRs, largely due to the availability of extensive open datasets that enable the effective training of deep learning models. Among these, MIMIC-CXR [91, 92, 57] has been the primary focus of this thesis. Its extensive collection of image-report pairs, along with derivative datasets that expand its annotations, such as Chest ImaGenome [214] and Medical-Diff-VQA, make it a valuable resource.

In the case of VQA datasets, their creation often requires expert annotation by specialised doctors, making them particularly challenging to curate due to privacy concerns. Consequently, open-access VQA datasets in the medical domain tend to be fewer in number and smaller in size compared to general-domain datasets. Furthermore, they predominantly feature 2D images, often sourced from medical publications. For modalities like CT and MRI, these datasets typically provide only a single slice instead of full volumetric data, thereby limiting the available information.

For this thesis, we focus exclusively on the Medical-Diff-VQA dataset [72, 70], which contains question-answer pairs associated with CXRs from MIMIC-CXR. This dataset was selected for its substantial number of question-answer pairs, its focus on CXRs, and its distinctive feature of including questions about changes in findings between two CXRs of the same patient—a central area of interest in this research. Furthermore, as Medical-Diff-VQA is derived from MIMIC-CXR, the linked radiology reports can be integrated with the question-answer pairs, enabling the development of new solutions that combine ARR with

VQA.

For comparative purposes, the statistics of other ARR datasets are presented in Table 2.2, while the statistics for VQA datasets are outlined in Table 2.5.

In this thesis, we do not utilise uncurated data from non-open sources due to the lack of access to such datasets.

2.8.1 MIMIC-CXR

To the best of our knowledge, MIMIC-CXR [91, 92, 57] is the largest open-source English dataset for ARR, exclusively composed by CXRs. The dataset contains 227,835 imaging studies taken from 65,379 patients who visited the Beth Israel Deaconess Medical Center Emergency Department (United States) from 2011 to 2016. During each imaging study, one or more chest radiographs were acquired, including frontal and lateral views, for a total of 377,110 images. Each study is accompanied by a semi-structured free-text radiology report written by a practicing radiologist, detailing the radiological observations. All images and reports have undergone de-identification.

MIMIC-CXR also provides 14 medical findings labels automatically extracted from the radiology reports with CheXpert [83], a rule-based approach. These labels include: *Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices*. The CheXpert labeler assigns each label one of four values: *positive*, *negative*, *uncertain*, or *missing*, indicating whether the condition is mentioned as present, absent, unclear, or not discussed in the report, respectively. We present the frequency of the labels in Table 2.3.

Numerous other datasets are derived from MIMIC-CXR and include additional annotations, such as Chest ImaGenome [214] and Medical-Diff-VQA [72]. We have chosen MIMIC-CXR for our experiments on ARR generation due to its extensive collection of CXRs and corresponding radiology reports. Furthermore, MIMIC-CXR, along with its dedicated datasets,

Label	Positive	Negative	Uncertain	Missing
Atelectasis	45,808	1,531	10,327	170,161
Cardiomegaly	44,845	15,911	6,043	161,028
Consolidation	10,778	7,967	4,331	204,751
Edema	27,018	25,641	13,174	161,994
Enlarged Cardiomediastinum	7,179	5,283	9,375	205,990
Fracture	4,390	886	555	221,996
Lung Lesion	6,284	862	1,141	219,540
Lung Opacity	51,525	3,069	3,831	169,402
No Finding	75,455	-	-	-
Pleural Effusion	54,300	27,158	5,814	140,555
Pleural Other	2,011	126	765	224,925
Pneumonia	16,556	24,338	18,291	168,642
Pneumothorax	10,358	42,356	1,134	173,979
Support Devices	66,558	3,486	237	157,546

Table 2.3: Frequency of labels in MIMIC-CXR [92, 91, 57] on the 227,827 radiologic studies annotated using the CheXpert labeler [83]. Each label is categorised based on whether it has a *positive*, *negative*, or *uncertain* mention in the report, or is not discussed (*missing*).

contains a large and diverse set of annotations, including medical findings, bounding box annotations, and question-answer pairs. This rich set of labeled data enables the developing and evaluation of various innovative approaches.

While MIMIC-CXR and its derivative datasets are valuable resources, there are several limitations to consider. MIMIC-CXR is derived from patients in the Intensive Care Unit of a large academic hospital, meaning the patient population may not be fully representative of the general population. It is likely to over-represent patients with severe conditions while under-representing healthier individuals. Additionally, the data is primarily sourced from a specific geographic region, the United States, with the majority of patients coming from a hospital in Boston. This geographic concentration may result in a skewed representation of diseases and conditions, limiting the ability to generalise to other settings or populations. Furthermore, certain diseases, particularly rare conditions, may be underrepresented, which could cause models to overfit to more common diseases and fail to generalise effectively to less frequent ones.

2.8.2 Chest ImaGenome

Chest ImaGenome, derived from MIMIC-CXR, is designed to enhance the semantic understanding of chest radiographs by linking anatomical structures and medical findings to specific regions in the images. The dataset provides detailed annotations, including bounding boxes for 36 anatomical regions (e.g., heart, lungs, and abdomen) and 71 medical finding labels (e.g., pleural effusion, pneumothorax) tied to these regions. This enables tasks such as anatomy localisation, multi-label classification, and fine-grained pathology detection, with the anatomical regions and findings listed in Table 2.4.

The dataset employs NLP, a chest X-ray ontology, and image segmentation techniques to automate the annotation process. Using rule-based text analysis pipelines and atlas-based bounding box extraction, it processes 242,072 frontal chest X-rays and 217,013 radiology reports from MIMIC-CXR, linking radiological findings to anatomical regions. The annotations are structured into scene graphs, which visually and semantically represent relationships between anatomical locations and attributes such as findings, devices, and technical assessments.

An essential feature of Chest ImaGenome is its ability to capture comparative relationships (*e.g.*, *improved*, *worsened*, or *no change*) between sequential exams. This supports longitudinal modelling of disease progression, leveraging temporal information to enhance clinical understanding. To ensure annotation quality, bounding box data was refined using manual validation and corrections informed by a Faster R-CNN model [169].

Furthermore, the dataset maps its objects, attributes, and relationships to Concept Unique Identifiers (CUIs) in the Unified Medical Language System (UMLS), integrating radiology knowledge into a broader medical context. This makes it a valuable resource for clinical reasoning tasks and multi-modal analyses, offering robust annotations that facilitate research into both spatial and temporal aspects of CXRs.

Anatomical Regions		
abdomen	left hilar structures	right clavicle
aortic arch	left lower lung zone	right costophrenic angle
cardiac silhouette	left lung	right hemidiaphragm
carina	left mid lung zone	right hilar structures
cavoatrial junction	left upper abdomen	right lower lung zone
descending aorta	left upper lung zone	right lung
left apical zone	mediastinum	right mid lung zone
left cardiac silhouette	right apical zone	right upper lung zone
left cardiophrenic angle	right atrium	spine
left clavicle	right cardiac silhouette	svc
left costophrenic angle	right upper abdomen	trachea
left hemidiaphragm	right cardiophrenic angle	upper mediastinum
Findings		
airspace opacity	enteric tube	picc
alveolar hemorrhage	fluid overload/heart failure	pigtail catheter
aortic graft/repair	goiter	pleural effusion
artifact	granulomatous disease	pleural/parenchymal scarring
aspiration	hernia	pneumomediastinum
atelectasis	hydropneumothorax	pneumonia
bone lesion	hyperaeration	pneumothorax
breast/nipple shadows	ij line	prosthetic valve
bronchiectasis	increased reticular markings/ild pattern	pulmonary edema/hazy opacity
cabg grafts	infiltration	rotated
calcified nodule	interstitial lung disease	scoliosis
cardiac pacer and wires	intra-aortic balloon pump	shoulder osteoarthritis
chest port	linear/patchy atelectasis	spinal degenerative changes
chest tube	lobar/segmental collapse	spinal fracture
clavicle fracture	low lung volumes	sub-diaphragmatic air
consolidation	lung cancer	subclavian line
copd/emphysema	lung lesion	superior mediastinal mass/enlargement
costophrenic angle blunting	lung opacity	swan-ganz catheter
cyst/bullae	mass/nodule (not otherwise specified)	tortuous aorta
diaphragmatic eventration (benign)	mediastinal displacement	tracheostomy tube
elevated hemidiaphragm	mediastinal drain	vascular calcification
endotracheal tube	mediastinal widening	vascular congestion
enlarged cardiac silhouette	multiple masses/nodules	vascular redistribution
enlarged hilum	pericardial effusion	

Table 2.4: Complete set of 36 anatomical regions and 71 findings used to supervise the anatomy localisation and the finding detection tasks, as annotated in the Chest ImaGenome dataset (https://physionet.org/content/chest-imagenome/1.0.0/).

2.8.3 Medical-Diff-VQA

The Medical-Diff-VQA dataset was introduced to address, among others, the novel task of Difference Visual Question Answering (diff-VQA), which focuses on answering questions about changes between two sequential CXRs of the same patient. This dataset was constructed using a semi-automated pipeline based on the MIMIC-CXR dataset. The process involved

Dataset	# Images	# QA pairs	QA Type
VQA-Med-2018 [62]	2,866	6,413	open-ended
VQA-Med-2019 [21]	4,200	15,292	open-ended
VQA-Med-2020 [19]	5,000	5,000	open-ended
VQA-Med-2021 [22]	5,500	5,500	open-ended
VQA-RAD [105]	315	3,515	open-ended, multiple-choice
Path-VQA [66]	4,998	32,798	open-ended, multiple-choice
SLAKE [125]	642	14,028	open-ended, multiple-choice
PMC-VQA [230]	149,075	226,946	open-ended, multiple-choice
MIMIC-CXR-VQA [11]	156,090	377,391	multiple-choice
Medical-Diff-VQA [72, 70]	164,324 (pairs)	700,703	open-ended, multiple-choice

Table 2.5: Comparison of medical domain VQA datasets. This table contrasts Medical-Diff-VQA, the dataset used in this thesis, with other open-access datasets.

extracting keywords from radiology reports and manually and automatically verifying them for accuracy. These keywords were then used to generate question-answer pairs based on template questions designed to reflect clinicians' interests.

The dataset comprises a total of 700,703 QA pairs derived from 164,324 pairs of main (current) and reference (previous) images. The questions are categorised into seven types: abnormality (145,421), location (84,193), type (27,478), level (67,296), view (56,265), presence (155,726), and difference (164,324). Difference-based questions link main (current) and reference (past) images from the same patient, enabling longitudinal analysis of changes such as "What has changed in the right lower lobe?". While the other six types of questions only refer to the main CXR.

This dataset provides a rich resource for training and evaluating AI models in tasks requiring analysis of anatomical changes, leveraging the structured linkage between radiology findings, their attributes, and temporal comparisons. Medical-Diff-VQA was chosen for the final technical chapter of this thesis due to its large scale, clinical relevance, high-quality annotations, and direct connection with MIMIC-CXR images and reports.

2.9 Evaluation Metrics

In this section, we present different metrics considered throughout this thesis.

We begin by discussing commonly used metrics for classification-based approaches and object detection. In classification tasks, metrics such as accuracy, precision, recall, F1-score and Area Under the Receiver Operating Characteristic Curve (AUROC) are used for measuring a model's ability to correctly classify instances across different categories. Object detection, which involves not only classifying objects within an image but also determining their positions, requires more sophisticated evaluation metrics, like Intersection over Union (IoU) and mean Average Precision (mAP).

Next, we present some of the most commonly used metrics for evaluating NLP generative tasks such as VQA and ARR, and some metrics that are specifically tailored to the medical domain. Due to the nature of these tasks, it can be challenging to correctly evaluate the quality of the generated text. In the context of VQA, the choice of metric depends on the type of questions: open-ended or multiple-choice. For multiple-choice questions, accuracy is the most appropriate metric, as there is only one correct answer. However, for open-ended questions, multiple correct answers may exist, differing in phrasing or level of detail. For instance, the question "What animal is shown in the picture?" could correctly be answered as "cat" or "black cat". If accuracy is used as the evaluation metric, it would penalise any answer that does not exactly match the ground truth, regardless of its correctness. Similar challenges arise in ARR and image captioning tasks, where the same report can be phrased in various ways by including or omitting less relevant details, yet remaining semantically equivalent. [3]

Moreover, due to the large size of the datasets, it is often unfeasible to conduct a thorough human evaluation of the model's prediction, especially in the medical domain, which requires medical experts. Therefore, to compare and automatically evaluate how well these models perform, different automatic metrics are discussed, each showing some limitations. Selecting

[3]Despite its flaws, in this thesis, we use accuracy for the VQA task to ensure comparability with prior works, as it remains a standard baseline metric in the field. However, we limit its use to evaluating short answers, such as binary 'yes/no' responses or answers containing only a few words, where its shortcomings are less pronounced.

the appropriate evaluation metric is often difficult. As such, it is generally advisable to monitor multiple metrics simultaneously.

2.9.1 Classification Metrics

We present in detail the different metrics used throughout this thesis for classification tasks.

Accuracy Accuracy measures the overall correctness of the model's predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2.5}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Precision Precision measures the proportion of true positives among all positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.6}$$

Recall Recall, or sensitivity, evaluates the proportion of true positives among all actual positives:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.7}$$

F1-score The F1-score is the harmonic mean of precision and recall, providing a single metric to balance these two aspects:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.8}$$

Area Under the Receiver Operating Characteristic Curve AUROC measures the model's ability to discriminate between positive and negative classes. It plots the True Positive Rate,

or Recall, (TPR $= \frac{TP}{TP+FN}$) against the False Positive Rate (FPR $= \frac{FP}{FP+FN}$) across different threshold values (t):

$$\text{AUROC} = \int_0^1 \text{TPR}(t) d\text{FPR}(t) \tag{2.9}$$

2.9.2 Object Detection Metrics

We present in detail the different metrics used throughout this thesis for object detection tasks.

Intersection over Union IoU measures the overlap between the predicted bounding box and the ground truth bounding box:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{2.10}$$

where the area of overlap is the intersection of the predicted and ground truth boxes, and the area of union is their combined area.

Mean Average Precision mAP aggregates the precision-recall curve over different IoU thresholds and object categories:

$$\text{mAP} = \frac{1}{|C|} \sum_{c=1}^{|C|} \text{AP}_c \tag{2.11}$$

AP_c represents the average precision for class c and $|C|$ is the total number of classes. Average precision measures the area under the Precision-Recall curve and is defined as:

$$\text{AP} = \sum_{n=1}^{N} P(n) \times \Delta R(n) \tag{2.12}$$

where $P(n)$ is the precision and $\Delta R(n)$ is the change in recall at threshold n.

2.9.3 Natural Language Generation Metrics

Drawing from various tasks, several metrics are widely used in Natural Language Generation (NLG) to evaluate generated text. For VQA and ARR generation, popular metrics include the BiLingual Evaluation Understudy (BLEU) [150] and the Metric for Evaluation of Translation with Explicit ORdering (METEOR) [14], both of which are inspired by machine translation. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [122], which originates from text summarisation tasks, is another commonly used metric. Additionally, for image captioning, the Consensus-based Image Description Evaluation (CIDEr) [199] is frequently employed. These metrics are designed to assess different aspects of how well the generated text matches the ground-truth references, each offering unique methods for measuring overlap.

BLEU BLEU determines the co-occurrences of n-grams in a candidate text and a reference text and it is computed as:

$$\text{BLEU} = \text{BP}\exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{2.13}$$

where p_n is the n-gram precision of the entire corpus; w_n are positive weights, In the baseline proposed by [150], the authors set $N = 4$ and and used uniform weights $w_n = 1/N$. BP corresponds to a brevity penalty function computed as:

$$\text{BP} = \begin{cases} 1, & \text{if } c > r, \\ e^{(1-r)/c}, & \text{if } c \leq r; \end{cases} \tag{2.14}$$

with c and r the length of the candidate and reference answer, respectively.

METEOR METEOR tries to address some of the weaknesses of BLEU. Given the unigram recall (R) and the unigram precision (P) METEOR is based on the F_{mean} score—the harmonic mean of the P and the 9R—rather than the n-gram precision. Further, METEOR is sensitive to

the alignment of unigrams between the candidate and the reference text. This is computed as:

$$\text{METEOR} = F_{mean}(1-p); \tag{2.15}$$

where p is the penalty score:

$$p = 0.5\left(\frac{\text{\# chunks}}{\text{\# unigrams matched}}\right)^3, \tag{2.16}$$

considering a chunk as a set of unigrams that are adjacent in both the candidate and the reference text.

ROUGE ROUGE is a set of metrics, based on the recall of the overlap of n-grams between the reference (X) and candidate text (Y) as follows:

$$\text{ROUGE-n} = \frac{\text{\# overlapping n-grams}}{\text{\# all n-grams in a reference text}}, \tag{2.17}$$

or based on the length of the Longest Common Subsequent ($LCS(X,Y)$):

$$\text{ROUGE-L} = \frac{(1+\beta)\times R\times P}{R+P\times\beta^2}, \tag{2.18}$$

where $R = \frac{LCS(X,Y)}{m}$, $P = \frac{LCS(X,Y)}{n}$, m the length of X and n the length of Y, and β is a parameter that wights the importance of P and R.

CIDEr CIDEr is an evaluation metric for image captioning that measures the similarity between a predicted caption (c_i) and reference captions ($S_i = \{s_{i1},\ldots,s_{im}\}$) by comparing n-grams (1 to 4 words) and weighting them using Term Frequency - Inverse Document Frequency (TF-IDF) to emphasise informative and less common phrases. It accounts for variations in language by stemming words and reducing the influence of frequently occurring n-grams that are less descriptive of the image content. CIDEr$_n$ score for n-grams of length n is computed using the cosine similarity between the candidate caption and the reference captions

as follows:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m}\sum_{j=1}^{m}\frac{g^n(c_i)\cdot g^n(s_{ij})}{\|g^n(c_i)\|\|g^n(s_{ij})\|} \tag{2.19}$$

where $g^n(\cdot)$ represents the vector of all n-grams of length n in the caption (*i.e.*, c_i or s_{ij}), and $\|g^n(\cdot)\|$ its magnitude. CIDEr captures grammatical properties and richer semantics by using higher-order n-grams and combines the scores from different n-grams using the following equation:

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr}_n(c_i, S_i) \tag{2.20}$$

with $w_n = 1/N$ and $N = 4$, as it was empirically found to work best.

CIDEr can still be computed for ARR and VQA evaluation even when each image is associated with only one ground truth radiology report or answer, but there are limitations to its effectiveness in this context. While CIDEr measures n-gram similarity between the generated and reference text, its core strength lies in leveraging multiple references to compute TF-IDF weights, which highlight the importance of informative and distinctive n-grams. With only a single reference, this weighting loses its intended purpose, potentially reducing the metric's correlation with human judgment. In such cases, CIDEr provides a basic similarity measure but may not fully capture nuanced or contextually accurate text generation, highlighting the need to supplement it with other metrics or human evaluation. We only adopt CIDEr to evaluate the VQA task to ensure comparability with prior works.

2.9.4 Semantic Metrics

Semantic metrics evaluate the quality of generated medical text based on the presence or absence of medical terms. In this section, we focus on semantic metrics specifically designed for CXR radiology reports.

Clinical Efficiency One of the most widely used semantic metrics for CXR radiology is Clinical Efficiency (CE). These metrics assess the accuracy of generated reports by comparing them against a predefined set of common findings, such as *Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax*, and *Support Devices*.

To determine whether these findings are present in the reports, CheXpert [83] or CheXbert [184] multi-label classifiers are applied to both the reference (ground truth) and the generated reports. These classifiers assign one of four possible values to each finding: *positive*, *negative*, *uncertain*, or *missing* (not mentioned). The CE metrics then compute precision, recall, and F1-score by treating the ground truth labels as the reference for evaluating the generated content.

To compute the CE metric, the task is reformulated as a multi-label binary classification problem. This involves aggregating the four possible label values in different ways. Two common aggregation strategies are employed, differing in how they treat *uncertain* findings—either as part of the positive findings or grouped with the negative findings:

1. *positive* versus all others (*uncertain* + *negative* + *missing*).

2. *positive* + *uncertain* versus *negative* + *missing* findings.

In this thesis, both aggregation strategies are used, selecting the one used by the methods we are comparing against to ensure fair and consistent evaluation. The specific aggregation approach applied in each experiment is specified in the corresponding chapters.

Others Another semantic metric is RadGraph F1, which is based on the RadGraph entity and relation extraction schema [86]. RadGraph F1 evaluates the performance of generated radiology reports by examining the accuracy of extracting specific medical entities and the relationships between them, as defined in the RadGraph schema. This metric focuses on how well the generated text aligns with predefined entities and their interrelationships, offering a structured way to assess the semantic accuracy of radiology reports.

Additionally, more closely aligned metrics have been introduced recently, such as RadCliQ [226]. RadCliQ combines multiple evaluation approaches into a composite metric designed to enhance assessment accuracy. It integrates traditional metrics like ROUGE, BLEU, and CheXbert embedding similarities, and RadGraph to form a composite metric which aims to match expert-generated error counts. However, RadCliQ suffers from reduced interpretability, as it combines multiple metrics into a single score without providing clear insights into specific areas of performance.

2.9.5 Limitations

Automatically evaluating text generation systems is an open problem with several limitations. NLG metrics evaluate the fluency of generated text by comparing n-grams with the ground truth but fail to capture the semantic similarity between the generated reports and the ground truth. Moreover, in the medical domain, most of these metrics inadequately treat all words equally, failing to distinguish between stopwords and clinically relevant terms. This equal weighting can lead to misleading evaluations, as the presence of clinically significant information should be prioritised over common, less important words.

Semantic metrics have partially solved the limitations of evaluating factual correctness. However, these are limited to predetermined labels and show a low correlation with manual evaluations conducted by radiologists. For instance, CE metrics have the disadvantage of not measuring either the laterality or the severity of the findings, attributes that can influence both the diagnosis and the clinical treatment decision.

More recently, large language models have been adopted to identify and explain clinically significant errors in candidate reports, both quantitatively and qualitatively (GREEN [149]). However, they are computationally demanding and slow to compute. Moreover, the error quantification process lacks full control, resulting in a certain level of randomness in how errors are counted.

2.10 Conclusion

In summary, this chapter highlights the recent trends in ARR and medical VQA, by first introducing some of the relevant works in CV, NLP and multimodal learning and the differences characterising the general and medical domains. We categorise different works into subgroups showing common features, to present the most common research questions researchers are trying to address. Despite showing promising results and rapid advancement, these are still active research areas due to the many challenges faced in the medical domain, such as the need to detect subtle details in medical scans, the difficulty in predicting the dense and detailed descriptions in radiology reports, and the lack of large-scale and high-quality benchmark datasets compared to the general domain. Moreover, the limitations of metrics used for the automatic evaluation of generative methods in ARR and VQA pose some limitations on assessing the quality of these systems, which is especially crucial in the medical domain, where errors can have enormous consequences in the treatment of a patient.

In this thesis, we aim to address some of these challenges with a focus on chest radiographs, by focusing on some of the questions that have already been explored as well as investigating others that have just very recently drawn attention. In particular, we investigate several key strategies: leveraging self-supervision to improve the image encoding in multimodal networks when labelled data is limited; extracting structured representations to ground the ARR process; utilising anatomical representations to improve report quality while increasing interpretability and control over the generated content; enabling effective comparisons between longitudinal studies of the same patient; and, finally, addressing questions about individual CXR images and comparisons between successive images by integrating ARR solutions with VQA.

Chapter 3

Multimodal CXR Classification from Self-Supervised Image Encoders

The main findings outlined in this chapter have been published as "Improving Image Representations via MoCo Pre-training for Multimodal CXR Classification" [39] at the Annual Conference on Medical Image Understanding and Analysis, 2022 (MIUA 2022).

My contributions to this chapter include conceptualisation, methodological design, conducting experiments, evaluation and writing.

Chapter Summary

Multimodal learning, here defined as learning from multiple input data types, has exciting potential for healthcare. However, current techniques rely on large multimodal datasets being available, which is rarely the case in the medical domain. In this chapter, we focus on improving the extracted image features which are fed into multimodal image-text Transformer architectures, evaluating on a medical multimodal classification task with dual inputs of chest X-ray images (CXRs) and the indication text passages in the corresponding radiology reports. We demonstrate that self-supervised Momentum Contrast (MoCo) [63] pre-training of the image representation model on a large set of unlabelled CXR images improves multimodal performance compared to supervised ImageNet [175] pre-training. MoCo shows a 0.6% *absolute improvement in AUROC-macro, when considering the full MIMIC-CXR [92, 91, 57] training set, and* 5.1% *improvement when limiting to* 10% *of the training data.*
Contributions in this chapter are:

1. *Compare how different pre-training strategies of the image encoder perform in the multimodal setup: AutoEncoder (AE), MoCo, and standard (supervised) ImageNet pre-trained weights; finding MoCo to perform best.*

2. *Explore how these strategies degrade when the multimodal transformer is fine-tuned on a smaller subset of the training set, finding MoCo pre-trained weights to perform better in a limited data scenario.*

3. *Extend the work of [187] – which demonstrates the effectiveness of MoCo pre-training for image-only pleural-effusion classification – to a multimodal multi-label classification problem.*

4. *Apply Gradient-weighted Class Activation Mapping (Grad-CAM) [180] to evaluate the impact of the pre-training strategy on the image features that activate the model, and report quantitative results on a small subset of the ChestX-ray8 test set with annotated bounding boxes [208].*

3.1 Introduction

Multimodal learning has recently gained attention for healthcare applications [75], due to the rich patient representation enabled by the combination of different data sources (*e.g.*, images, reports, and clinical data). Recent works in multimodal learning have mainly focused on Transformer [198] architectures, with similar approaches adopted in the medical domain [85]. Whilst the role of the joint pre-training process has been widely explored [67], fewer works have focused on the single modality components of the models. In particular, the role of the image representation is frequently neglected. However, the task of multimodal representation learning is complex and one of the main challenges in the medical domain is the lack of large-scale, labelled datasets, compared to the millions of images available in computer vision tasks in the general domain. Therefore, we seek to mitigate the complexity of multimodal learning by providing robust image representation as input.

In the general multimodal domain, the "bottom-up top-down" [8] approach is a popular image representation paradigm for multimodal Transformer architectures such as VisualBERT [111] and ViLBERT [134]. These models use Region of Interest (RoI) feature maps extracted from Faster R-CNN [169], which is pre-trained on large object detection datasets (*e.g.* VisualGenome [103]). Other image representation strategies have been proposed. In Pixel-BERT [77], the image representation is defined as the feature map of the last convolutional layer of a Convolutional Neural Network (CNN). Similarly, the discrete latent space of a Variational AutoEncoder (VAE) has been adopted in DALL-E [164]. Alternatively, the Vision Transformer (ViT) [49] consists of directly feeding raw pixel patches as the input for Transformer architectures.

In this paper, we are interested in multimodal CXR multi-label classification of medical images supported by the medical history of the patient which is available in free-text radiology reports (indication field), as shown in Figure 3.1. We use MIMIC-CXR [92, 91, 57], which is the largest open-access multimodal medical dataset, to evaluate our proposed methodology, for the task of Chest X-Ray classification of 14 radiographic findings classes. The two most

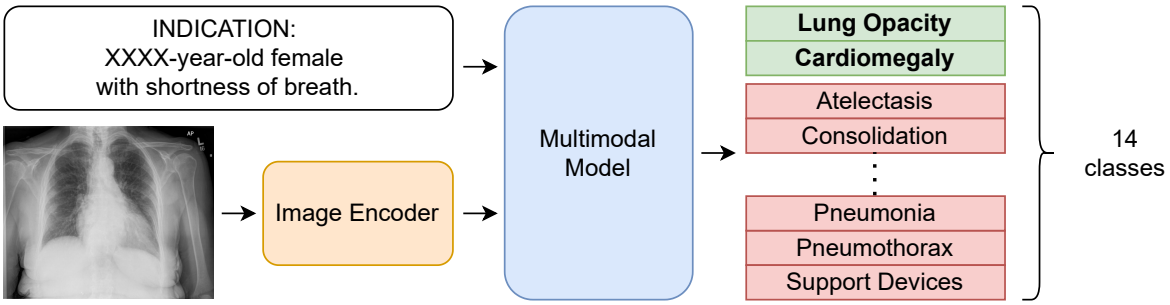


Figure 3.1: Illustration of the multimodal CXR multi-label classification pipeline. The indication field and CXR image are dual input modalities, and the output is a set of positive (green) or negative (red) predictions for 14 radiographic findings labels, as annotated in the MIMIC-CXR dataset [92, 91, 57]. In this data, taken from the IU-Xray dataset [44], ages and other patient-identifiable information are replaced by a placeholder, here indicated by XXXX. The **image encoder** is the component that we investigate in this paper, to discover a strategy for learning a good image representation.

relevant works performing multimodal classification of CXR using the text indication section as an additional inference-time input: ChestBERT [85] and what we denote as "Attentive" [186]. Following ChestBERT [85] we adopt the MultiModal BiTransformer model (MMBT) [99], which has a similar image representation to Pixel-BERT. Different to the previously described multimodal BERT models, MMBT does not include a joint pre-training step. More recently, Liao et al [120] have shown a method of joint modality pre-training to be effective by maximising the mutual information between the encoded representations of images and their corresponding reports. At inference time, the image only is used for classification. However, we consider the situation where we may have limited task-specific labelled multimodal (paired image and text) training data, but ample unlabelled unimodal (imaging) data available for pre-training and therefore we investigate image-only pre-training techniques.

For learning good visual representations, many self-supervised contrastive learning strategies have shown promising results in the medical domain, for instance, Momentum Contrast (MoCo) contrastive training [63] [187] and Multi-Instance Contrastive Learning [10] – an application of SimCLR [30] to medical imaging. In particular, MoCo pre-training has shown superior results in a similar chest X-Ray imaging classification task, outperforming other methods using standard supervised pre-training on ImageNet [187]. Similarly, MedAug [203]

Method	Ease of training	Medical image suitability
Supervised ImageNet *Supervised training on 1000 ImageNet classes.*	No training required. Pre-trained weights available for standard CNN architectures.	Weak – trained on natural images.
AutoEncoder *Encoder-decoder architecture trained on reconstruction loss.*	Easy – does not require large batches.	Flexible – can train on relevant medical image data (no labels required)
SimCLR *Contrastive learning approach*	Hard – requires high compute power to handle the large batches ($> 10^3$ images).	
MoCo *Contrastive learning approach*	Moderate – designed to work with a small batch size ($\sim 10^2$ images) & uses efficient updating of the large dynamic dictionary.	

Table 3.1: Summary of the considered image pre-training strategies suited to CXR image classification.

has extended the work of Sowrirajan et al [187], by considering different criteria to select positive pairs for MoCo. However, the best approach in [203] (which targets mixed-view classification) is to create pairs from lateral and frontal views of CXR, while we focus on frontal views only, making this method unsuitable for our task. MoCo works by minimising the embedding distance between positive pairs – generated by applying different data augmentations to an image – and maximising the distance to all other augmented images in the dataset [63]. MoCo maintains a large dynamic dictionary of negative samples as a queue with fixed length (set as a hyperparameter) which is updated every step by adding the newest batch of samples and removing the oldest. This allows the model to have a large number of negative samples without the need for very large batches, unlike other contrastive learning approaches (*e.g.* SimCLR [30]), making MoCo a sensible choice when training on fewer GPUs.[1] In this work, for the imaging component of MMBT we experiment with two strategies for training a

[1]Due to the limited computing power, we decided to neglect the Multi-Instance Contrastive Learning approach proposed by [10], trained on 16–64 Cloud TPU cores.

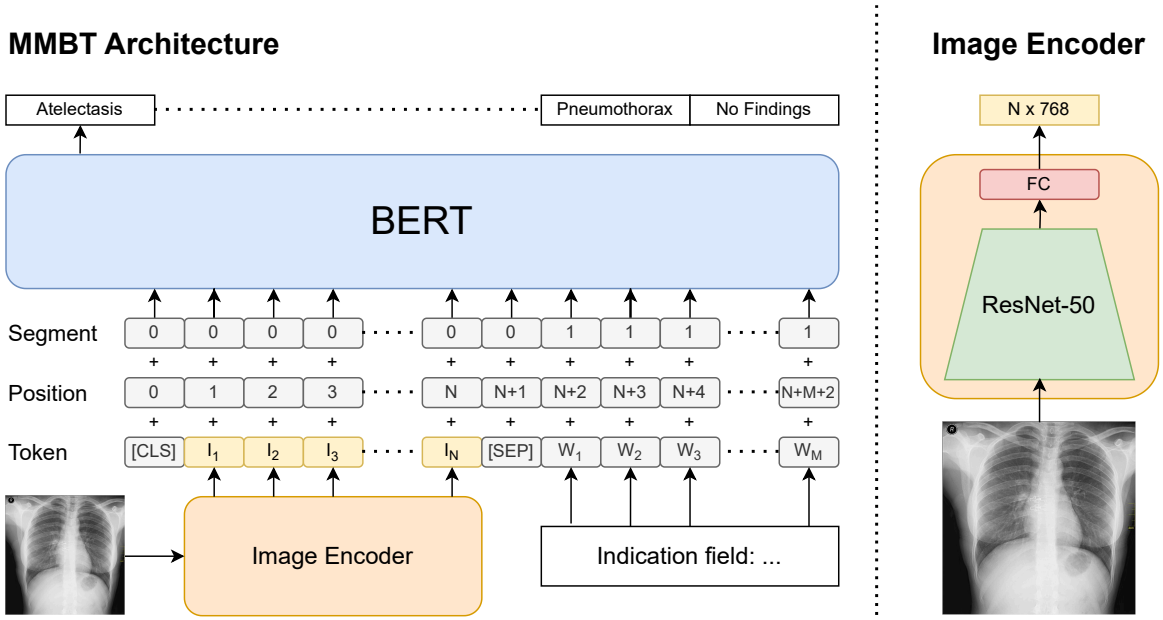


Figure 3.2: Illustration of the MMBT architecture [99] (left) and a closer look at the image encoder (right). The ResNet-50 backbone of the image encoder is the image feature extractor initialised using different pre-training strategies.

CNN image encoding: a) MoCo and b) a classic AutoEncoder strategy.

3.2 Method

We explore the effectiveness of different image representations for the model by considering different pre-training strategies.

3.2.1 Model

The overall architecture is based on the multimodal bitransformer model (MMBT) [99] as shown in Figure 3.2 (left). This builds on the BERT architecture [47] – the Transformer encoder backbone is initialized with pre-trained BERT weights – and adapts it for multimodal data by introducing an additional visual input.

Input Embedding The *text input embedding* is obtained by first tokenising the input text into M subword tokens, using the WordPiece tokenizer [215] with a 30,000 token vocabulary.

The text input tokens are embedded using the BERT embedding layer to obtain the vector representation of each token, indicated as $W = \{W_1 \dots W_M\} \in \mathbb{R}^{768}$, where 768 corresponds to the input embedding dimension of BERT.

The *visual input embedding*, indicated as $I = \{I_1 \dots I_N\} \in \mathbb{R}^{768}$, are obtained from the feature map outputted from the last convolutional layer of ResNet-50. This is flattened into $N = 49$ vectors and projected by a single fully connected layer into the input embedding space of BERT, Figure 3.2 (right).

The textual and visual input embedding are separated using the $[SEP]$ special token. Both the textual and the visual input embedding are summed with the related positional – to determine the position of each token of the input sequence – and the segment embedding – to discriminate between textual and visual inputs.

Classification A $[CLS]$ token is used at the beginning of the input sequence, and its final hidden vector $h_{[CLS]}$ is used as the multimodal sequence representation for classification. This is passed through a *fully-connected* classification layer and a *sigmoid* function to obtain the probability score of each class. The model is trained using binary cross-entropy loss for each output class. At inference, we select the classes with a probability greater or equal to 0.5 as positive.

3.2.2 Self-supervised Image Pre-training

We experiment with two self-supervised strategies: an AE and MoCo.

AutoEncoder The AE consists of a ResNet-50 encoder and decoder, Figure 3.3 (top). The model is trained by minimising the reconstruction loss, defined as the mean squared error between the input image x and the reconstructed image $\hat{x}$:

$$\mathcal{L}_{reconstruction} = \frac{||x - \hat{x}||_2^2}{W \cdot H} \tag{3.1}$$

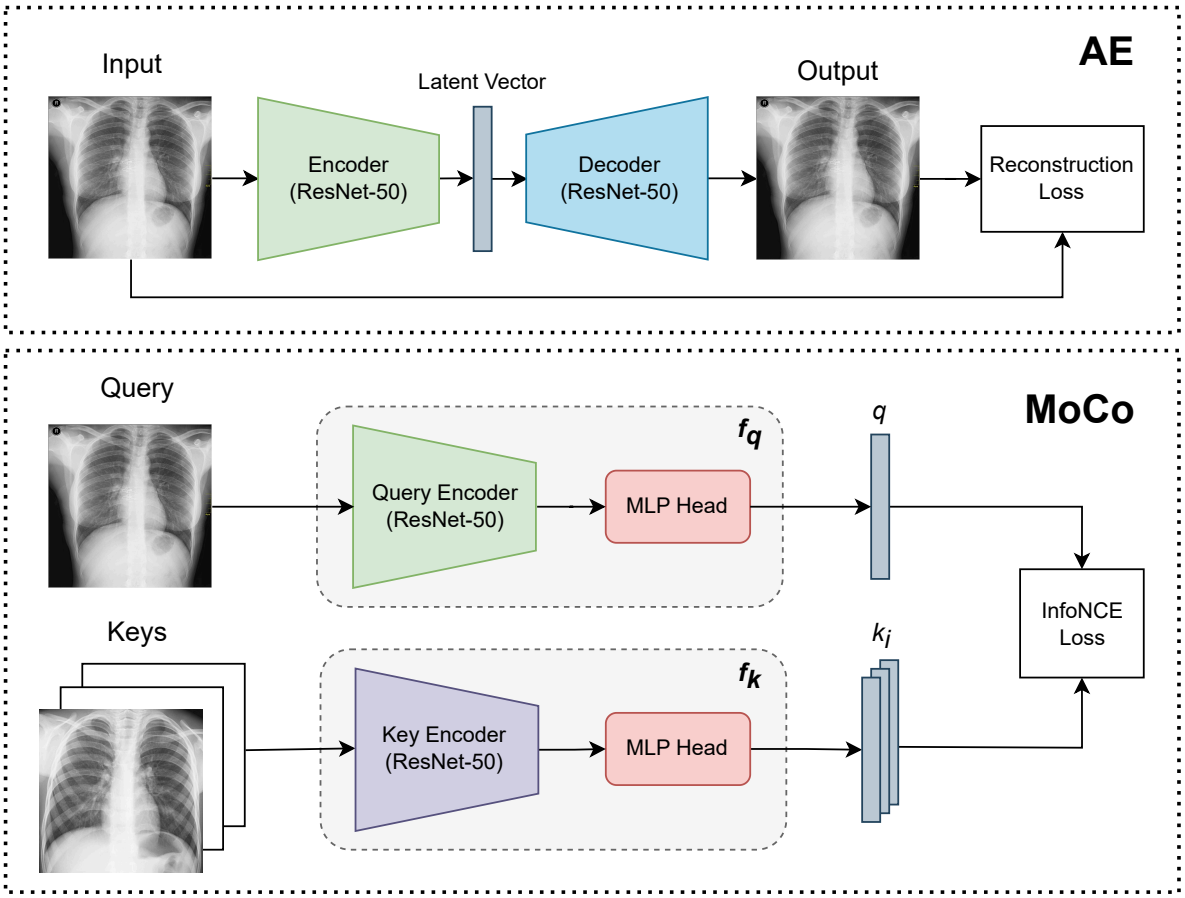


Figure 3.3: Self-supervised pre-training: AE vs. MoCo.

where W and H are the width and height of the image, respectively. Following pre-training, the decoder is discarded and the ResNet-50 encoder weights are used as initialisation for the MMBT image encoder.

Momentum Contrast MoCo minimises the embedding distance between a query image x^q and a positive key image x^{k+} – generated by applying different data augmentations to the same original image – and maximises the distance to all other augmented images in the dataset. MoCo generates a large dynamic dictionary of samples $\{x_i^k\}_{i=1}^K$ as a queue with fixed length K (set as a hyperparameter) containing $K-1$ negative samples and a single positive. The dictionary is updated every step by adding the newest batch of samples and removing the oldest.

As shown in Figure 3.3 (bottom), we implement MoCo with two ResNet-50 models – a

query encoder and a key encoder. We define the two streams as two functions f_q and f_k. The model is then trained by optimising the Info Noise Contrastive Estimation (InfoNCE) loss function [146]:

$$\mathcal{L}_{InfoNCE} = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)} \tag{3.2}$$

where τ is the temperature hyper-parameter, $q = f_q(x^q)$ is the encoded query representation of the query image x^q, $k_i = f_k(x_i^k)$ the encoded key vector of a key image from the dictionary $\{x_i^k\}_{i=1}^K$ and k_+ is the encoded vector of the sole positive key from the dictionary whose similarity we aim to maximise. Given that the dictionary is often very large, training of the key encoder through backpropagation is computationally intractable; instead, the parameters of the key encoder are updated using momentum updates in tandem with the query encoder:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q; \tag{3.3}$$

where θ_k are the parameters of f_k, θ_q the parameters of f_q and $m \in [0,1)$ is a momentum coefficient.

Following pre-training, the weights of the query encoder (without the MLP head) are used to initialise the ResNet-50 image encoder, shown in Figure 3.2b.

3.3 Experimental Setup

3.3.1 Datasets

We evaluated our method on MIMIC-CXR [92, 91, 57], which contains 377,110 CXR images with the associated radiology reports of 227,827 radiologic studies from 65,379 patients. Using the CheXpert labeler [83], 14 different labels have been automatically extracted from the radiology report: *Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia,*

Pneumothorax, Support Devices. The CheXpert labeler assigns a value of whether the label has a *positive*, *negative*, or *uncertain* mention in the report, or is not discussed (*missing*).[2] For each label, we re-formulate the task as a multi-label binary classification task: *positive* vs. *others* (*negative*, *uncertain*, *missing*).

In this study, we select only images from a frontal view, either anteroposterior (AP) and posteroanterior (PA). Following the official MIMIC-CXR split, this yields 208,794 training pairs, 1,695 validation pairs and 2,920 test report/image pairs. As presented in [85], the text modality corresponds to the indication field (*i.e.* scan request text) extracted from the radiology reports. This is the part that would be available at imaging time and describes relevant medical history.

The self-supervised pre-training of the image encoder is performed on the CheXpert dataset [83] which consists of 224,316 CXR images from 65,240 patients; we ignore the available annotations and treat this dataset as a large unlabelled dataset. Input images are resized by matching the smaller edge to 224 pixels and maintaining the original aspect ratio.

3.3.2 Implementation Details

For the self-supervised pre-training, we adopt the AE and MoCo implementations available from the PyTorch Lightning library. During pre-training, the input images are resized by matching the smaller edge to 224 pixels and maintaining the original aspect ratio. Similar to Sowrirajan et al [187], we employ the following data augmentation techniques: random rotation ($-10^\circ \leq \theta \leq 10^\circ$), random horizontal flipping; and random crop of 224×224 pixels. The same data augmentations are also applied during the fine-tuning step.

At the fine-tuning stage, we adopt the MMBT implementation made available by the authors of ChestBERT [85][3], which uses the MultiModal Framework (MMF) [183]. We use the same training parameters as [85]: models are trained using a batch size of 128 and Adam

[2]The CheXpert labeler assigns the *No Finding* label to a study when none of the other labels has a *positive* mention in the report. For this reason, the *No Finding* label is not categorised as *negative*, *uncertain*, or *missing*.

[3]https://github.com/jacenkow/mmbt

optimiser with weight decay, with the learning rate set to 5×10^{-5}, and a linear warm-up schedule for the first 2000 steps, and micro F1 score computed on the validation set was used as the early stopping criterion and patience of 4000 steps, up to a maximum of 14 epochs. Each experiment was repeated 5 times using different random seeds to initialise the model weights and randomise batch shuffling.

3.3.3 Baselines

The chosen method is compared with two unimodal baselines, to verify the improvement brought by inputting both visual and textual modalities at once. Moreover, we compare MMBT with another multimodal approach which we denote "Attentive" [186], to justify the architecture design chosen for our multimodal experiments.

- **BERT** [47] - using a BERT model only (similar to the backbone of MMBT) an unimodal text classifier is trained, without the CXR image.
- **ResNet-50** [65] - using ResNet-50 only (similar to the network used for the image representation in MMBT) an unimodal image classifier is trained, without text information.
- **Attentive** [186] - this model follows a two-stream approach where a) the CXR image is processed by a ResNet-50 model and b) the indication field is encoded by BioWordVec embeddings [231] followed by two sequential bi-directional Gated Recurrent Units (GRUs) [34]. The visual and textual feature representations are then fused using two multimodal attention layers.

3.4 Results

3.4.1 Comparison of Self-Supervised Pre-training Strategies

Here we compare MMBT with the baselines, adopting different pre-training strategies for the image encoder, as described in Section 3.2.2. The AE and MoCo pre-trained ResNet-50 are

100% Training Set

Model	Image Pre-Training		F1		AUROC	
	Method	Dataset	Macro	Micro	Macro	Micro
BERT	-		$24.4_{\pm 1.0}$	$40.2_{\pm 0.5}$	$71.5_{\pm 0.3}$	$82.1_{\pm 0.4}$
ResNet-50	Supervised	ImageNet	$27.2_{\pm 0.6}$	$48.4_{\pm 0.9}$	$75.8_{\pm 1.2}$	$85.3_{\pm 0.8}$
ResNet-50	MoCo	CheXpert	$\mathbf{28.5}_{\pm 0.7}$	$\mathbf{49.5}_{\pm 0.6}$	$\mathbf{76.3}_{\pm 0.3}$	$\mathbf{85.5}_{\pm 0.1}$
Attentive	Supervised	ImageNet	$29.3_{\pm 0.5}$	$51.1_{\pm 0.6}$	$76.3_{\pm 0.5}$	$85.9_{\pm 0.5}$
Attentive	MoCo	CheXpert	$\mathbf{31.9}_{\pm 0.3}$	$\mathbf{53.2}_{\pm 0.5}$	$\mathbf{77.8}_{\pm 0.6}$	$\mathbf{86.4}_{\pm 0.4}$
MMBT	*Random Initialisation*		$32.0_{\pm 1.2}$	$49.7_{\pm 0.7}$	$76.1_{\pm 0.3}$	$85.2_{\pm 0.3}$
MMBT	Supervised	ImageNet	$34.3_{\pm 2.1}$	$54.7_{\pm 0.7}$	$79.8_{\pm 1.1}$	$87.4_{\pm 0.8}$
MMBT	AE	CheXpert	$34.5_{\pm 1.2}$	$52.4_{\pm 0.3}$	$77.9_{\pm 0.4}$	$86.3_{\pm 0.4}$
MMBT	MoCo	CheXpert	$\mathbf{36.7}_{\pm 1.4}$	$\mathbf{55.3}_{\pm 0.6}$	$\mathbf{80.4}_{\pm 0.3}$	$\mathbf{87.6}_{\pm 0.4}$

10% Training Set

Model	Image Pre-Training		F1		AUROC	
	Method	Dataset	Macro	Micro	Macro	Micro
BERT	-		$21.3_{\pm 2.7}$	$36.6_{\pm 1.7}$	$67.4_{\pm 0.4}$	$79.7_{\pm 1.3}$
ResNet-50	Supervised	ImageNet	$22.1_{\pm 0.9}$	$42.1_{\pm 0.7}$	$68.0_{\pm 1.9}$	$79.7_{\pm 3.4}$
ResNet-50	MoCo	CheXpert	$\mathbf{23.6}_{\pm 1.1}$	$\mathbf{43.8}_{\pm 1.8}$	$\mathbf{70.8}_{\pm 0.9}$	$\mathbf{81.3}_{\pm 0.9}$
Attentive	Supervised	ImageNet	$21.7_{\pm 0.9}$	$42.1_{\pm 1.4}$	$65.1_{\pm 1.1}$	$78.9_{\pm 0.6}$
Attentive	MoCo	CheXpert	$\mathbf{22.8}_{\pm 1.0}$	$\mathbf{44.3}_{\pm 1.9}$	$\mathbf{70.2}_{\pm 0.5}$	$\mathbf{82.7}_{\pm 0.4}$
MMBT	*Random Initialisation*		$25.1_{\pm 2.1}$	$40.7_{\pm 3.0}$	$69.6_{\pm 0.7}$	$81.6_{\pm 0.6}$
MMBT	Supervised	ImageNet	$26.4_{\pm 2.1}$	$44.3_{\pm 1.5}$	$69.0_{\pm 0.4}$	$79.3_{\pm 1.8}$
MMBT	AE	CheXpert	$27.6_{\pm 1.2}$	$44.2_{\pm 1.1}$	$70.5_{\pm 0.4}$	$82.1_{\pm 0.3}$
MMBT	MoCo	CheXpert	$\mathbf{28.5}_{\pm 2.4}$	$\mathbf{48.8}_{\pm 1.1}$	$\mathbf{74.1}_{\pm 0.7}$	$\mathbf{84.5}_{\pm 0.9}$

Table 3.2: Results on the MIMIC-CXR test set, comparing different ResNet-50 pre-training strategies. The models are fine-tuned on the full training set (top) and on 10% of the training set (bottom).

compared against (1) random initialisation – to verify the benefit of starting from pre-trained weights; (2) ImageNet initialisation – widely adopted in computer vision.

We report the F1 score and the Area Under the Receiver Operating Characteristic (AUROC), multiplying all metrics by 100 for ease of reading. To assess whether a pre-training strategy helps in a limited training data scenario, the same experiments are conducted using only a 10% random sample of the original training set.

As shown in Table 3.2 (top), both unimodal baselines (text-only BERT and image-only ResNet-50) obtain lower classification scores compared to the multimodal approaches (Attentive and MMBT); with MMBT achieving the best results, as previously reported in [85].

However, in the limited data scenario (Table 3.2 (bottom)), the gap between unimodal and multimodal approaches is reduced when considering the standard ImageNet initialisation. This suggests that the image modality is not processed effectively by the multimodal architectures, which motivates us to investigate how to improve the image representations to maintain the benefit of using both modalities with limited data.

Table 3.2 shows a consistent improvement from adopting MoCo initialisation of the image encoder (ResNet-50), which demonstrates that MMBT benefits from such domain-specific image pre-training strategy. The margin of improvement from ImageNet increases with a limited training set, aligned with the results in [187]. Compared to Sowrirajan et al [187] — who showed the benefit of MoCo pre-training only on pleural effusion classification, using an image-only CNN – we broaden the paradigm to multimodal classification of 14 different classes. Furthermore, we report the AUROC scores for each class in Table 3.4. This shows that MoCo pre-trained MMBT yields the highest scores for most classes when fine-tuned on the full MIMIC-CXR training set, and more obviously when fine-tuned on a 10% random subset of the training set.

On the contrary, AE seems to be a less effective pre-training strategy. This might be attributed to the reconstruction loss, which encourages the model to focus on the intensity variation of CXRs rather than other meaningful features (*e.g.* shapes and textures) to discriminate between different classes.

Table 3.2 shows a consistent improvement achieved by adopting MoCo pre-trained weights also for the image encoder of the Attentive model and the image-only ResNet-50. This confirms that both unimodal and multimodal models benefit from the MoCo pre-training of the image encoder.

3.4.2 Model Explainability

To investigate the impact of pre-training on the learned features, we visually assess the quality of the activation maps obtained by two of the pre-training strategies: supervised ImageNet

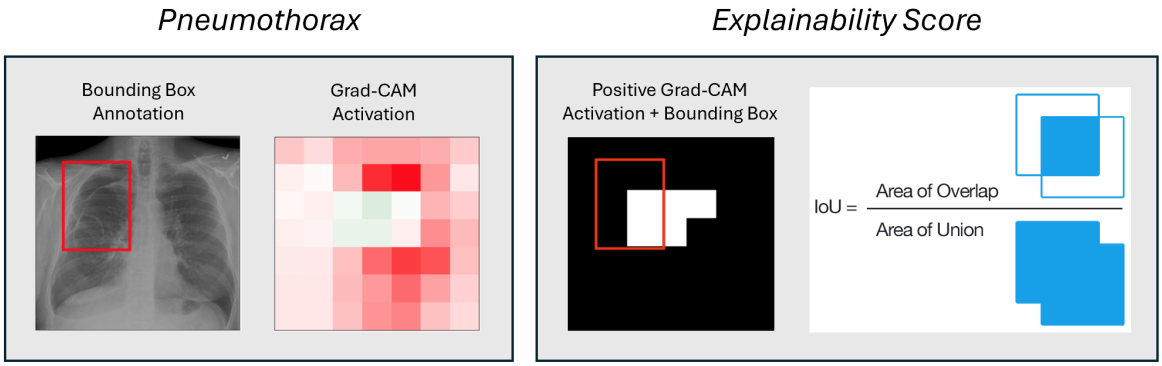


Figure 3.4: Model explainability pipeline. On the left, for the label Pneumothorax, we show the CXR image with the ground truth bounding box annotation (taken from ChestX-ray8 dataset [208]) and the Grad-CAM activation map. On the right, we isolate the positive regions of the activation map and the bounding box and compute the Intersection over Union (IoU) between the two.

Image Pre-Training		IoU				
Method	Dataset	Atelectasis	Cardiomegaly	Effusion	Pneumonia	Pneumothorax
Supervised	ImageNet	1.3	3.8	5.9	0.0	0.1
MoCo	CheXpert	**2.6**	**16.0**	**11.9**	**1.8**	**2.7**

Table 3.3: IoU results computed on the ChestX-ray8 test set, containing bounding box annotations. We evaluate only on the five classes that overlap with MIMIC-CXR.

pre-training and MoCo pre-training on CheXpert. First, we fine-tune the fully connected layer of the ResNet-50 architecture on the full training set of MIMIC-CXR, while freezing the remaining pre-trained weights. Second, we apply Grad-CAM [180] to the final 7×7 activation map, computed before the fully connected layer. Finally, we assess if the generated maps highlight the correct anatomical location of the pathology, by computing the Intersection over Union (IoU) between the bounding boxes – annotated in the ChestX-ray8 dataset [208] – and the regions in the activation map that contribute positively to the classification of a target label (Figure 3.4). In this final step, we only consider the subset of ChestX-ray8 labels overlapping with those in MIMIC-CXR: *Atelectasis, Cardiomegaly, Pleural Effusion, Pneumonia, Pneumothorax*.

The mean IoU scores for each class are reported in Table 3.3. Although the overlap between the positive areas of the activation maps and the bounding boxes is low for both

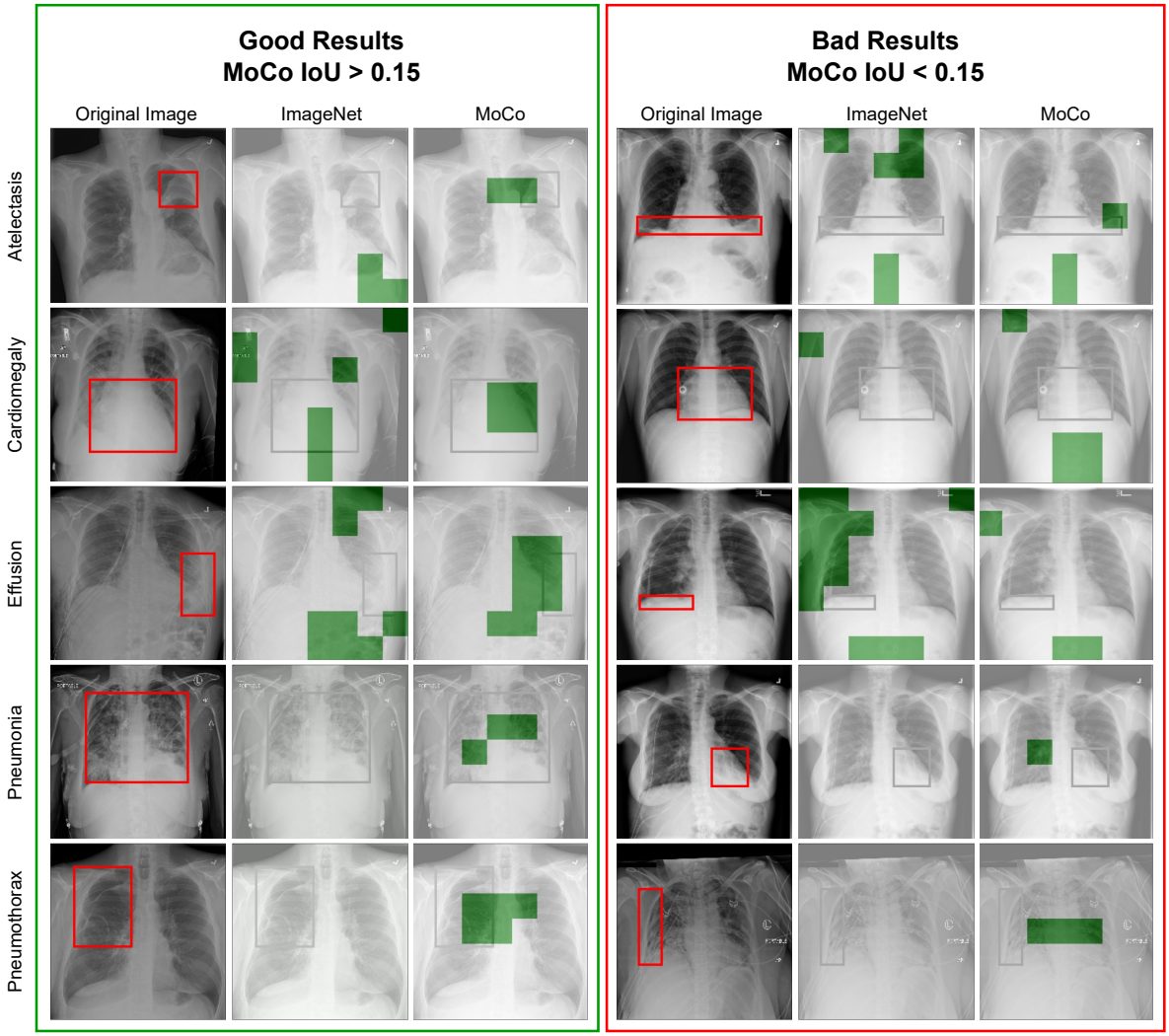


Figure 3.5: Examples of CXRs taken from ChestX-ray8 dataset with the corresponding bounding box annotations highlighted in red. Grad-CAM is computed on the last 7×7 activation map, before the fully connected layer of ResNet-50, for both ImageNet and MoCo pre-training. The green regions show the activations thresholded at 0 i.e. all positive activations (activations can also be negative). The left side images are selected having an IoU score greater than 0.15 between the bounding box and the positive regions, using MoCo pre-trained weights; the right side images are selected with an IoU score lower than 0.15.

pre-training strategies, it can be observed that MoCo pre-training outperforms ImageNet for each class. This suggests that, when adopting MoCo pre-training, the CNN learns more meaningful features of CXRs that can be effectively exploited by the model for the downstream classification task. This is shown visually in Figure 3.5, where MoCo pre-trained ResNet-50 focuses more accurately on the areas matching with the bounding boxes.

100% Training Set

Model	Attentive		MMBT			
Image Pre-Training	Supervised ImageNet	MoCo CheXpert	Random Init.	Supervised ImageNet	AE CheXpert	MoCo CheXpert
Atelectasis	$73.5_{\pm 1.0}$	$72.8_{\pm 0.9}$	$71.8_{\pm 0.6}$	$\mathbf{75.2}_{\pm 0.8}$	$74.2_{\pm 0.7}$	$74.9_{\pm 0.7}$
Cardiomegaly	$77.1_{\pm 0.5}$	$79.1_{\pm 0.5}$	$80.0_{\pm 0.4}$	$81.3_{\pm 0.9}$	$81.6_{\pm 0.4}$	$\mathbf{82.4}_{\pm 0.4}$
Consolidation	$72.3_{\pm 1.1}$	$75.0_{\pm 0.6}$	$71.3_{\pm 0.8}$	$\mathbf{77.2}_{\pm 1.8}$	$74.3_{\pm 0.6}$	$76.5_{\pm 0.6}$
Edema	$82.2_{\pm 0.7}$	$82.8_{\pm 0.5}$	$80.9_{\pm 0.8}$	$83.6_{\pm 1.1}$	$82.7_{\pm 0.4}$	$\mathbf{84.2}_{\pm 0.4}$
Enlarged Card.	$67.0_{\pm 1.6}$	$68.7_{\pm 0.9}$	$68.4_{\pm 0.6}$	$73.3_{\pm 2.1}$	$71.4_{\pm 1.8}$	$\mathbf{75.0}_{\pm 1.8}$
Fracture	$66.7_{\pm 3.0}$	$69.7_{\pm 1.6}$	$68.7_{\pm 1.8}$	$70.0_{\pm 1.2}$	$70.7_{\pm 0.8}$	$\mathbf{72.3}_{\pm 0.8}$
Lung Lesion	$68.9_{\pm 2.2}$	$71.0_{\pm 0.9}$	$69.6_{\pm 0.6}$	$74.5_{\pm 3.0}$	$70.2_{\pm 0.6}$	$\mathbf{76.7}_{\pm 0.6}$
Lung Opacity	$68.9_{\pm 0.4}$	$70.8_{\pm 0.9}$	$66.9_{\pm 0.6}$	$71.8_{\pm 0.5}$	$69.4_{\pm 0.5}$	$\mathbf{72.0}_{\pm 0.5}$
No Findings	$80.4_{\pm 0.4}$	$80.9_{\pm 0.9}$	$79.7_{\pm 0.8}$	$82.5_{\pm 1.2}$	$81.2_{\pm 0.6}$	$\mathbf{82.6}_{\pm 0.6}$
Pleural Effusion	$86.7_{\pm 0.9}$	$86.8_{\pm 0.6}$	$82.6_{\pm 0.3}$	$\mathbf{87.6}_{\pm 0.6}$	$85.0_{\pm 0.2}$	$\mathbf{87.6}_{\pm 0.2}$
Pleural Other	$78.8_{\pm 1.7}$	$80.2_{\pm 2.3}$	$89.4_{\pm 1.2}$	$\mathbf{86.1}_{\pm 4.4}$	$81.6_{\pm 1.9}$	$\mathbf{86.1}_{\pm 1.9}$
Pneumonia	$70.8_{\pm 0.9}$	$74.1_{\pm 1.5}$	$69.7_{\pm 0.9}$	$74.6_{\pm 0.6}$	$71.8_{\pm 0.6}$	$\mathbf{76.7}_{\pm 0.6}$
Pneumothorax	$84.8_{\pm 0.8}$	$86.9_{\pm 0.9}$	$87.4_{\pm 1.2}$	$\mathbf{87.9}_{\pm 0.4}$	$86.9_{\pm 1.0}$	$87.7_{\pm 1.0}$
Support Devices	$90.4_{\pm 0.2}$	$91.1_{\pm 0.2}$	$89.3_{\pm 0.2}$	$\mathbf{91.7}_{\pm 0.6}$	$90.1_{\pm 0.3}$	$\mathbf{91.7}_{\pm 0.3}$
Average	$76.3_{\pm 0.5}$	$77.8_{\pm 0.6}$	$76.1_{\pm 0.3}$	$79.8_{\pm 1.1}$	$77.9_{\pm 0.4}$	$\mathbf{80.4}_{\pm 0.3}$

10% Training Set

Model	Attentive		MMBT			
Image Pre-Training	Supervised ImageNet	MoCo CheXpert	Random Init.	Supervised ImageNet	AE CheXpert	MoCo CheXpert
Atelectasis	$66.9_{\pm 1.5}$	$69.3_{\pm 1.3}$	$64.6_{\pm 0.4}$	$65.5_{\pm 0.9}$	$67.2_{\pm 0.4}$	$\mathbf{71.4}_{\pm 1.3}$
Cardiomegaly	$67.3_{\pm 0.5}$	$72.4_{\pm 0.8}$	$71.8_{\pm 0.5}$	$70.7_{\pm 0.7}$	$74.0_{\pm 1.3}$	$\mathbf{77.0}_{\pm 0.9}$
Consolidation	$61.3_{\pm 0.3}$	$68.0_{\pm 0.8}$	$66.3_{\pm 1.1}$	$64.0_{\pm 1.3}$	$67.7_{\pm 0.8}$	$\mathbf{71.3}_{\pm 1.0}$
Edema	$76.1_{\pm 1.0}$	$78.4_{\pm 1.4}$	$74.5_{\pm 0.6}$	$76.5_{\pm 1.1}$	$77.1_{\pm 0.8}$	$\mathbf{80.7}_{\pm 1.3}$
Enlarged Card.	$58.5_{\pm 2.5}$	$63.3_{\pm 1.8}$	$62.6_{\pm 3.3}$	$62.8_{\pm 1.8}$	$61.5_{\pm 5.0}$	$\mathbf{67.8}_{\pm 3.0}$
Fracture	$52.2_{\pm 3.4}$	$51.8_{\pm 3.0}$	$61.6_{\pm 4.5}$	$58.7_{\pm 4.0}$	$60.1_{\pm 2.4}$	$\mathbf{62.4}_{\pm 2.6}$
Lung Lesion	$56.7_{\pm 1.0}$	$64.5_{\pm 2.9}$	$64.8_{\pm 2.2}$	$60.7_{\pm 2.0}$	$65.8_{\pm 2.0}$	$\mathbf{67.2}_{\pm 1.4}$
Lung Opacity	$60.4_{\pm 0.7}$	$65.7_{\pm 0.6}$	$61.7_{\pm 0.8}$	$62.2_{\pm 1.7}$	$62.2_{\pm 0.8}$	$\mathbf{67.2}_{\pm 1.1}$
No Findings	$72.1_{\pm 1.4}$	$75.2_{\pm 1.2}$	$74.5_{\pm 0.8}$	$74.1_{\pm 0.6}$	$75.8_{\pm 0.9}$	$\mathbf{78.7}_{\pm 0.6}$
Pleural Effusion	$80.0_{\pm 0.9}$	$82.6_{\pm 0.7}$	$73.0_{\pm 0.8}$	$79.2_{\pm 0.9}$	$76.8_{\pm 0.4}$	$\mathbf{84.7}_{\pm 0.3}$
Pleural Other	$60.0_{\pm 1.3}$	$62.2_{\pm 3.6}$	$61.2_{\pm 1.2}$	$65.0_{\pm 6.7}$	$62.1_{\pm 3.2}$	$\mathbf{67.6}_{\pm 3.3}$
Pneumonia	$57.6_{\pm 1.5}$	$64.1_{\pm 1.3}$	$66.4_{\pm 1.5}$	$60.4_{\pm 1.4}$	$66.0_{\pm 0.7}$	$\mathbf{68.6}_{\pm 2.0}$
Pneumothorax	$68.4_{\pm 3.7}$	$78.7_{\pm 2.8}$	$84.6_{\pm 2.0}$	$79.9_{\pm 2.0}$	$\mathbf{85.0}_{\pm 0.6}$	$84.7_{\pm 0.8}$
Support Devices	$78.0_{\pm 1.9}$	$86.5_{\pm 2.0}$	$87.1_{\pm 0.6}$	$86.0_{\pm 0.9}$	$86.9_{\pm 0.5}$	$\mathbf{88.8}_{\pm 1.2}$
Average	$65.1_{\pm 1.1}$	$70.2_{\pm 0.5}$	$69.6_{\pm 0.7}$	$69.0_{\pm 0.4}$	$70.5_{\pm 0.4}$	$\mathbf{74.1}_{\pm 0.7}$

Table 3.4: Per-class AUROC scores using different ResNet-50 initialisations. The models are fine-tuned on the full training set (top) and on 10% of the training set (bottom).

3.5 Limitations

In this chapter, we have only focused on CXRs and further research is needed if we want to transfer similar methods to other imaging modalities, especially if switching from 2D scans (*e.g.*, CXR) to 3D scans (*e.g.*, MRI, CT). The lack of availability of large-scale open-access datasets containing other imaging modalities dictates our focus on CXRs.

Moreover, the results of this study are limited by the errors of the CheXpert labeler, used to extract the categorical labels from the reports in MIMIC-CXR. This is an automatic label extractor prone to errors since the set of rules defined by its authors is imperfect. For instance, such a labeler fails when the report contains comparisons with prior scans. If we consider the following example *"There are no new findings."*, without taking into account the report associated with the prior scan, this is an ambiguous statement that could either imply that (1) there are no findings seen in both the current and prior scan, or (2) there are findings in both the current and prior scan but they did not change. Similar instances are widely present in MIMIC-CXR reports, but CheXpert does not address such cases. Another limitation of the CheXpert schema is that it only allows for 14 predefined labels. While these are among the most common types of findings, other findings or more granular findings should be defined to assess in more detail the quality of the methods.

Finally, we have considered Grad-CAM as a simple approach for visualising the activation map of the image encoder and assessing whether it activates at the correct location in the image. However, in computing the IoU score, we are comparing the low-resolution 7×7 activation maps with the bounding box coordinates associated with the high-resolution images. Furthermore, these bounding boxes approximate the locations of diseases in the image, but obtaining more detailed pixel-level annotations, like segmentation masks, is challenging as they are time-consuming and require expert annotators. For such limitations, we obtain low detection scores, as shown by the relatively low IoU scores in Table 3.3.

3.6 Conclusion

In this chapter, we have focused on showing how different initialisations of the image encoder can impact the results for multimodal multi-label CXR classification, demonstrating the benefit of domain-specific contrastive learning pre-training. We focus this study on comparing widely adopted initialisation techniques – random initialisation or using the ImageNet pre-trained weights – and self-supervised techniques – AE and MoCo. Due to the common limitation of annotating medical data, which requires expert annotators, we have only focused on self-supervised domain-specific pre-training, to simulate the real-world scenario where only a portion of the data is annotated.

Firstly, in this chapter, we re-demonstrate the value of the indication field in guiding the interpretation of CXRs. By incorporating information about the patient's clinical history and the reason for the scan, which is included in the indication field, we provide essential contextual information. This enables the model to better understand the clinical scenario and generate more accurate and clinically relevant predictions. However, while integrating the indication field offers advantages, it also introduces certain limitations that must be carefully considered. For example, there is a risk that the model may develop biases based on patterns associated with specific indications. It might, for instance, overpredict certain findings that are commonly linked to particular clinical scenarios, leading to systematic errors. Such biases could affect the model's generalisability and performance, particularly in cases that deviate from typical patterns. This issue underscores the need for future research to better understand and mitigate these biases, as well as to explore methods to enhance the robustness of the model in diverse clinical contexts. Addressing these challenges remains a key area for future investigation, as it is crucial to ensure that the use of the indication field truly enhances CXR interpretation.

Our results show that the choice of the initialisation of the image encoder component of the multimodal network plays a substantial role, especially with limited annotated data. We present how pre-training the image encoders on domain-specific data can improve the

performances on the downstream task, with MoCo pre-training showing the overall best performances. This is further validated when applying Grad-CAM to visualise which part of the image is activated to classify some specific findings, showing a better localisation on the affected area of the CXR when using MoCo compared to ImageNet initialisation.

Different conclusions are presented when pre-training ResNet-50 using an AE, which appears to be a less efficient pre-training approach. We hypothesise that the reason can be attributed to the reconstruction loss, which prioritises the intensity fluctuation in CXRs over other significant features used to discriminate between various classes.

Overall, the results from this chapter show that self-supervised pre-training techniques are effective strategies to initialise the image encoder of a multimodal Transformer model when having an unlabelled pre-training dataset much larger than the fine-tuning dataset. However, large unlabelled datasets are often not open access, due to privacy concerns. This limits the use of such pre-training techniques when the given dataset is both the largest and highly curated with dense annotations, such as MIMIC-CXR.

Chapter 4

CXR Automated Reporting using Intermediate Triples Representations

The main findings outlined in this chapter have been published as "Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations" [38] in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2022).

My contributions to this chapter include conceptualisation, methodological design, technical implementation, data analysis, conducting experiments, evaluation, and writing. Thanks to Dr William Clackett, Dr Chaoyang Wang, and Dr Giovana Klefti for collaborating on designing the annotation schema and the human evaluation.

Chapter Summary

*Automated radiology reporting has the potential to assist radiologists with the time-consuming procedure of generating text radiology reports. Most existing approaches generate the report directly from the radiology image, however, we observe that the resulting reports exhibit a realistic style but lack clinical accuracy. Therefore, we propose a two-step pipeline that subdivides the problem into factual triple extraction followed by free-text report generation. The first step comprises the supervised extraction of clinically relevant structured information from the image, expressed as triples of the form (*entity1, relation, entity2*). In the second step, these triples are input to condition the generation of the radiology report. In particular, we focus on Chest X-ray report generation. The proposed framework shows state-of-the-art results on the MIMIC-CXR dataset according to most of the standard text generation metrics that we employ (BLEU, METEOR, ROUGE) and to clinical efficiency metrics (recall, precision and F1 assessed using CheXpert [83]), also giving a 23% reduction in the total number of errors and a 29% reduction in critical clinical errors as assessed by expert human evaluation. Contributions in this chapter are:*

1. *Propose, using a clinically informed schema, to express the information in CXR radiology reports in a structured form, using triples (entity1, relation, entity2).*

2. *Propose a two-step pipeline – called TE-RG – for CXR radiology report generation: Triples Extractor (TE) followed by Report Generator (RG).*

3. *Conduct extensive experiments on the MIMIC-CXR dataset [91, 92, 57], showing state-of-the-art results for NLG and clinical efficiency metrics.*

4. *Conduct a human evaluation to assess the quality of reports, by counting the number of errors, divided into 6 different categories (hallucinations, omissions, attribute errors, impression errors, grammatical errors, and critical errors)*

4.1 Introduction

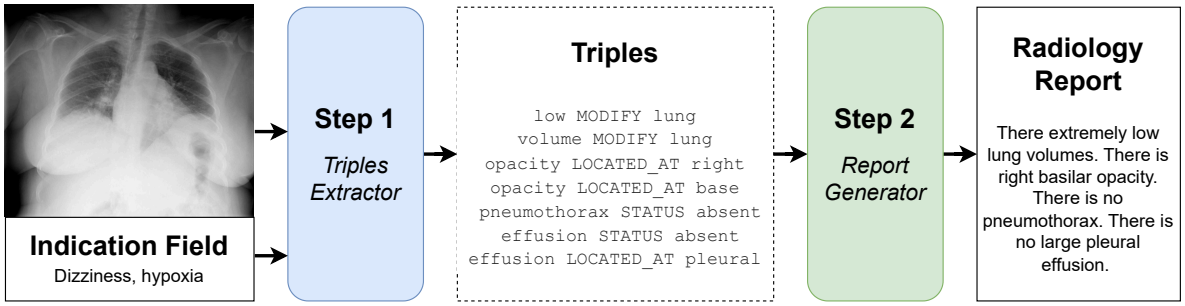


Figure 4.1: Illustration of the proposed TE-RG two-step pipeline. **Step 1** – a Triples Extractor (TE) is implemented to extract a set of triples associated with each CXR scan. **Step 2** – a Report Generator (RG) is implemented to generate a radiology report, based on the extracted triples. The CXR image and report shown in this example are both taken from the IU-Xray dataset [44], while the triples are extracted as described in Section 4.2.

This work focuses on the task of Automated Radiology Reporting (ARR) for Chest X-Ray (CXR) images. Recent studies on CXR report generation have explored various ways to enhance model architecture. For instance, relational memory modules [32] have been introduced to enable models to retain information from previous generations, while cross-modal memory modules [31, 155] promote alignment between visual and textual information. Another line of research has investigated incorporating external knowledge into models [128, 221], either through pre-constructed knowledge graphs or by retrieving similar reports within the dataset.

However, these approaches typically generate radiology reports directly from images using standard cross-entropy loss for supervision, which primarily rewards verbatim replication of target text (style) without prioritising the accurate reporting of clinically significant findings (content). This concern was partially treated by including a classification module of a pre-defined set of findings and pathologies that are present in the image [6], as an auxiliary task. However, in this approach, there is no direct link between the classification and reporting outputs, and the transfer of information relies on multi-tasking functioning effectively. Further, this approach does not consider the relations between different classes. Overall, there is a limited effect on the generation process.

We focus on improving the clinical utility of the generated reports, by introducing an intermediate step to the generation process. It consists of extracting, from a CXR image, factual information in a structured format, expressed in the form of triples (entity1, relation, entity2). We further categorise the entities and relations according to a clinical schema, in order to remove heterogeneity of expression.

This is particularly relevant in the field of radiology, where radiologists can express similar clinical concepts using different phrases i.e. the following phrases all relate to the same clinical concept of edema: "pulmonary oedema", "cardiac decompensation", "fluid overload" and "evidence of acute heart failure". We adopt RadGraph [86] to extract four predefined clinically relevant relations (*Suggestive of*, *Located at*, *Modify* and *Status*), and we map medical entities to medical concepts (*e.g.,* "fluid overload" to *«edema»*) according to a scheme devised by a junior physician. Our two-step pipeline is shown in Figure 4.1, where the first step consists of the Triples Extraction (TE) process which aims at extracting factual information from a CXR image, and the second step corresponds to Report Generation (RG) which uses the image as input alongside (i.e. conditioned by) the extracted triples. In this chapter and throughout the rest of this thesis, we refer to this approach as TE-RG.

To the best of our knowledge, only [113] have proposed a similar approach for the automatic generation of ophthalmic reports. In their work, they show an improvement by extracting, from an ophthalmic image, entities and relations (they consider the extracted triples to represent a latent clinical graph), and injecting them into the text generation process. This varies from our work in three aspects: the definition and generation of triples, the model architecture, and the medical domain application (Ophthalmology *vs.* CXR). In terms of triples annotation, their approach is granular, using the original linguistic terms and relations, without further categorisation and processing: the entities are represented by single words as written in the source text, and they consider the verbs extracted with a dependency parser as the relations. Thus, our annotation pipeline generates a much lower number of entities and relations, standardising and simplifying the triples. Moreover, in terms of model architecture,

whilst they train the model end-to-end using a triples restoration loss, we keep the two steps independent from one other and frame each step as a sequence-to-sequence task.

4.2 Triples Representation

We hereby describe how the ground truth triples are extracted from the finding section of each original radiology report. These triples represent the intermediate structured representation of the original report and are used to supervise the first step of the proposed TE-RG two-step pipeline. The triples are represented as (e_1, r, e_2), where e_1 and e_2 are two entities linked by the relationship r.

The order of the extracted triples is determined by the order in which their corresponding text spans appear in the original radiology report, to reflect the sequence of the text in the report. For triples that share an entity and thus reference the same text span, their order is further determined by the position of the text span associated with the other entity.

4.2.1 Extracting Ground Truth Triples

The overall annotation pipeline is shown in Figure 4.2. We use two publicly available tools to annotate the ground truth triples – RadGraph [86] and ScispaCy [139] – which are then refined with the help of a junior physician with 2 years of clinical experience.

We consider only sentences that can be extracted from a single CXR image, therefore we filter out mentions of comparisons with previous scans since they are not always available in the MIMIC-CXR dataset.

RadGraph Entity & Relation Extraction We first apply RadGraph [86], which extracts entities and relations from a radiology report. RadGraph classifies the extracted entities as *Anatomy* corresponding to anatomical concepts (*e.g., heart* or *lung*), or *Observation* referring to words associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications. The *Observation* entities are further categorised as *Definitely Present*,

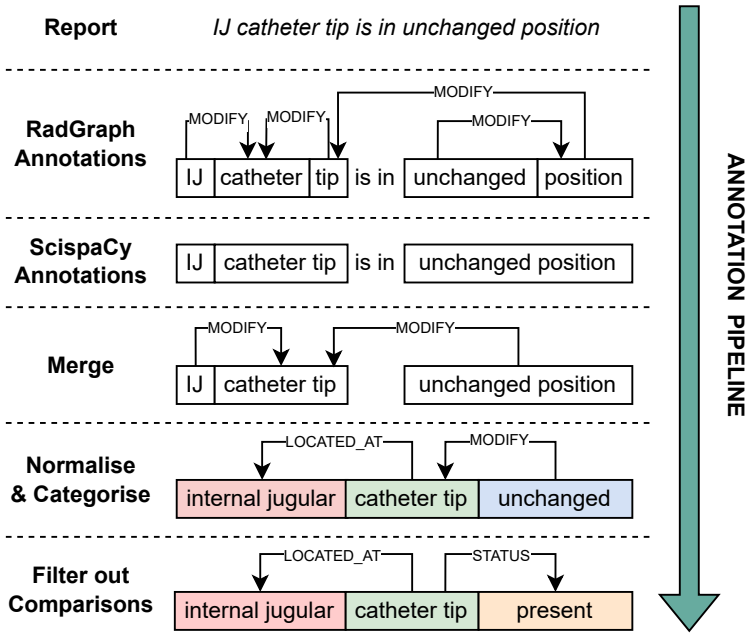


Figure 4.2: Example of the annotation pipeline to extract the ground truth triples from the radiology report. In the last two steps, we adopt the same color scheme as indicated in Figure 4.3, to categorise the entities.

Uncertain, and *Definitely Absent*. The schema proposed by RadGraph includes three different relations: *Suggestive Of* – which links two *Observation* entities, where the second entity is implied based on the first entity (*e.g.,* «*opacity* → *SUGGESTIVE_OF* → *pneumonia*»); *Located At* – which indicates where an *Observation* entity is located (*e.g.,* «*fracture* → *LOCATED_AT* → *rib*»); and *Modify* – indicating that the first entity modifies the scope of, or quantifies the degree of, the second entity (*e.g.,* «*dense* → *MODIFY* → *consolidation*»). We use the pre-trained model[1] to extract the entities and relations from the Finding section of MIMIC-CXR radiology reports. Given that we aim to represent each report as a set of triples, we introduce another relation named *Status*, to include the three categorisations that RadGraph associates to each *Observation* entity: *Definitely Present* becomes *STATUS present*, *Uncertain* becomes *STATUS uncertain*, and *Definitely Absent* becomes *STATUS absent*.

ScispaCy Entity Extraction The RadGraph schema was designed to prefer granular entities (mostly represented by single words), linked to one other with many relations, in order to

[1] https://physionet.org/content/radgraph/1.0.0/

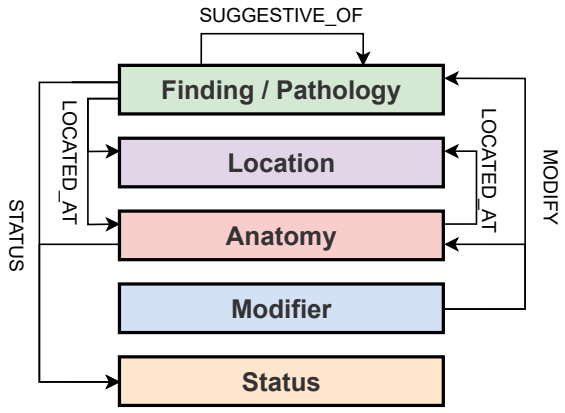


Figure 4.3: Triples schema. The relations correspond to the edges of the graph, and the type of relation is indicated in capital letters. The entity labels are represented by the nodes of the graph. These represent the triples to which our annotation pipeline is constrained.

have dense annotations associated with each report. However, to simplify the task and reduce the number of triples, we want to merge triples which could be sensibly represented as a single entity (*e.g., «enteric* $\rightarrow$ *MODIFY* $\rightarrow$ *tube»* can be merged into a single medical entity called *«enteric tube»*). Therefore, we additionally use a named-entity recognition model which extracts less granular medical entities, namely ScispaCy's [139] `en_core_sci_scibert` model[2].

Merge RadGraph & ScispaCy Entities The third step consists of merging together the two sets of entities associated with the same report while keeping the relations extracted with RadGraph. This is performed by prioritising entities extracted using ScispaCy (E_{sc}) over those extracted using RadGraph (E_{rg}). Formally, if there exists $e_{sc} \in E_{sc}$ and $e_{rg} \in E_{rg}$ such that $e_{rg} \subset e_{sc}$ (*i.e.* e_{rg} is a substring of e_{sc}), then we substitute e_{rg} with e_{sc} and assign to it all the relations originally associated with e_{rg}. Moreover, if $e_{rg,1}$ and $e_{rg,2}$ are linked together with a relation – $(e_{rg,1}, r, e_{rg,2})$ – and $e_{rg,1}, e_{rg,2} \subset e_{sc}$, then we remove the relation r and only keep e_{sc} as a single entity. Otherwise, if $e_{rg} \not\subset e_{sc}\ \forall e_{sc} \in E_{sc}$, then we keep e_{rg} and its associated relations.

[2] https://github.com/allenai/scispacy

Normalise Entities & Categorise Relations The final step of our annotation process comprises the refinement of the merged entities. With the help of a junior physician, we defined five entity categories: *Anatomy* (*e.g., «heart»*), *Finding/Pathology* (*e.g., «pneumothorax»*, *«effusion»*), *Location* (*e.g., «left»*, *«top»*), *Modifiers* (*e.g., «large»*, *«left»*) and *Status* (*e.g., «present»*, *«normal»*). For each entity term, we defined a set of synonyms. We then associate the term when one of the synonyms is detected in an entity span. Further, we constrain the triples to a fixed schema, based on the entity labels, as shown in Figure 4.3, and filter out the triples whose entity types and relations do not appear in that schema. If more than one of the manually selected terms is found inside an entity name, we split the entity and assign the relation based on the same schema. This occurs when ScispaCy detects entities that can be expressed as the combination of two or more separate entities (*e.g., «pulmonary vascular engorgement»* can be expressed as «*engorgement* $\rightarrow$ *LOCATED_AT* $\rightarrow$ *pulmonary vascular*»).

Remove Comparisons with Prior Reports Finally, we substitute the triples that express a change from previous studies of the same patient, since we are aiming to generate the report from a single CXR image, without having access to previous images. We identify the triples expressed as «$e_1 \rightarrow$ *MODIFY* $\rightarrow e_2$», where e_1 corresponds to *«unchanged»*, *«new»*, *«increase»* or *«decrease»*; we then substitute the triple with «$e_2 \rightarrow$ *STATUS* $\rightarrow$ *present*», based on the assumption that if the radiologist mentions a change of a pathology or a finding, this is still present and visible in the image.

4.2.2 Statistics

We show some statistics of the extracted ground-truth triples. In Table 4.1, we present the number of unique entities, relations, and triples obtained from our annotation pipeline on the MIMIC-CXR reports. The relatively low number of entities (302) and relations (4) still results in a high number of unique triples (8,672), as entities and relations can be assorted in multiple ways while still adhering to the fixed schema of allowed triples, presented in Figure 4.3.

# Entities	# Relations	# Triples
302	4	8,672

Table 4.1: Number of unique entities, relations and triples using our annotation pipeline on the MIMIC-CXR reports.

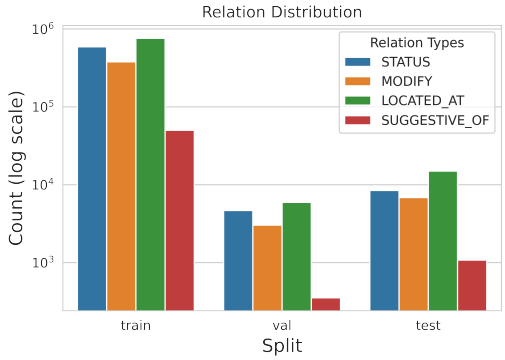


Figure 4.4: Distribution of the four different relation types of the triples in the train/validation/test split of MIMIC-CXR.

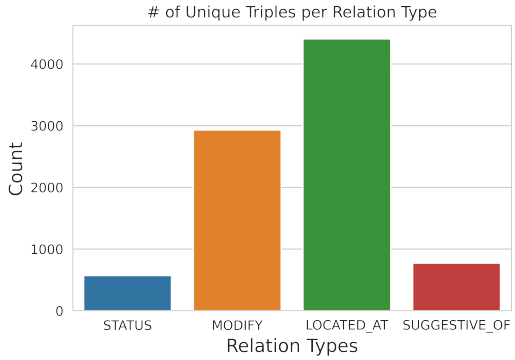


Figure 4.5: Number of unique triples per relation type.

In Figure 4.4, we illustrate the distribution of the different relation types of the triples within the train/validation/test split. In each split, *LOCATED_AT* has the most number of occurrences due to the laterality attributes associated with most medical findings or the position of the support devices, described in the radiology reports. Followed by *STATUS* and *MODIFY*, while *SUGGESTIVE_OF* emerges as the least common relation type.

In Figure 4.5, we visualise the number of unique triples per relation type. This shows that *LOCATED_AT* has the highest variance, in terms of triples, compared to the other relations, due to the many anatomical regions visualised in a CXR and the number of entities (*e.g.*, findings or support devices) which position is relevant to assess. On the other end, *STATUS* has the least variability since, in our schema, each *Observation* entity can only be linked by the *STATUS* relation to *present*, *uncertain*, or *absent*.

Figure 4.6 shows the distribution of the 302 entities extracted from MIMIC-CXR reports. We see that it follows a long-tail distribution, with some entities - such as *absent*, *present*, *pleural*, *edema*, *pneumonia* etc. – dominating the dataset. This is expected since some medical terms are very frequent in radiology reports, and some diseases are more commonly assessed

than others when looking at CXR scans.

4.3 Model

We propose a novel framework to perform ARR in two steps: *Triples Extraction* and *Report Generation*. Similarly to [32], we design and train Transformer models with custom architectures from scratch. Figure 4.7 shows a detailed diagram of the TE-RG two-step pipeline.

4.3.1 Triples Extractor

The first step consists of extracting the triples associated with each CXR image, whose semi-automated annotation process is described in Section 4.2. We treat this problem as a sequence-to-sequence task, using a multimodal encoder-decoder Transformer as the backbone, with both the CXR image and the indication field (*i.e.*, scan request text) as inputs. The benefit of using the indication field as context for CXR classification in an encoder Transformer model was previously shown by [85].

The multimodal input sequence is the concatenation of the CXR image embedding and the indication field text embedding. The image embedding, denoted $I = \{I_1 \dots I_N\}$, corresponds to the feature map extracted from the last convolutional layer of ResNet-101 and flattened into a 49×2048 image embedding. The text input is tokenised into a $M \times 2048$ token embedding, indicated as $W = \{W_1 \dots W_M\}$. Further, we sum to the input sequence a segment embedding – to allow the model to discriminate between visual and textual inputs – and position embedding – needed by the Transformer to access the order of the input embedding. A `[SEP]` token is used to separate the two input modalities. The target sequence $Trp = \{Trp_1 \dots Trp_K\}$ corresponds to the concatenation of the ground truth triples, each separated by a `[SEP]` token.

We compare two different setups of the triples extractor model *TE-Transformer* to generate the triples:

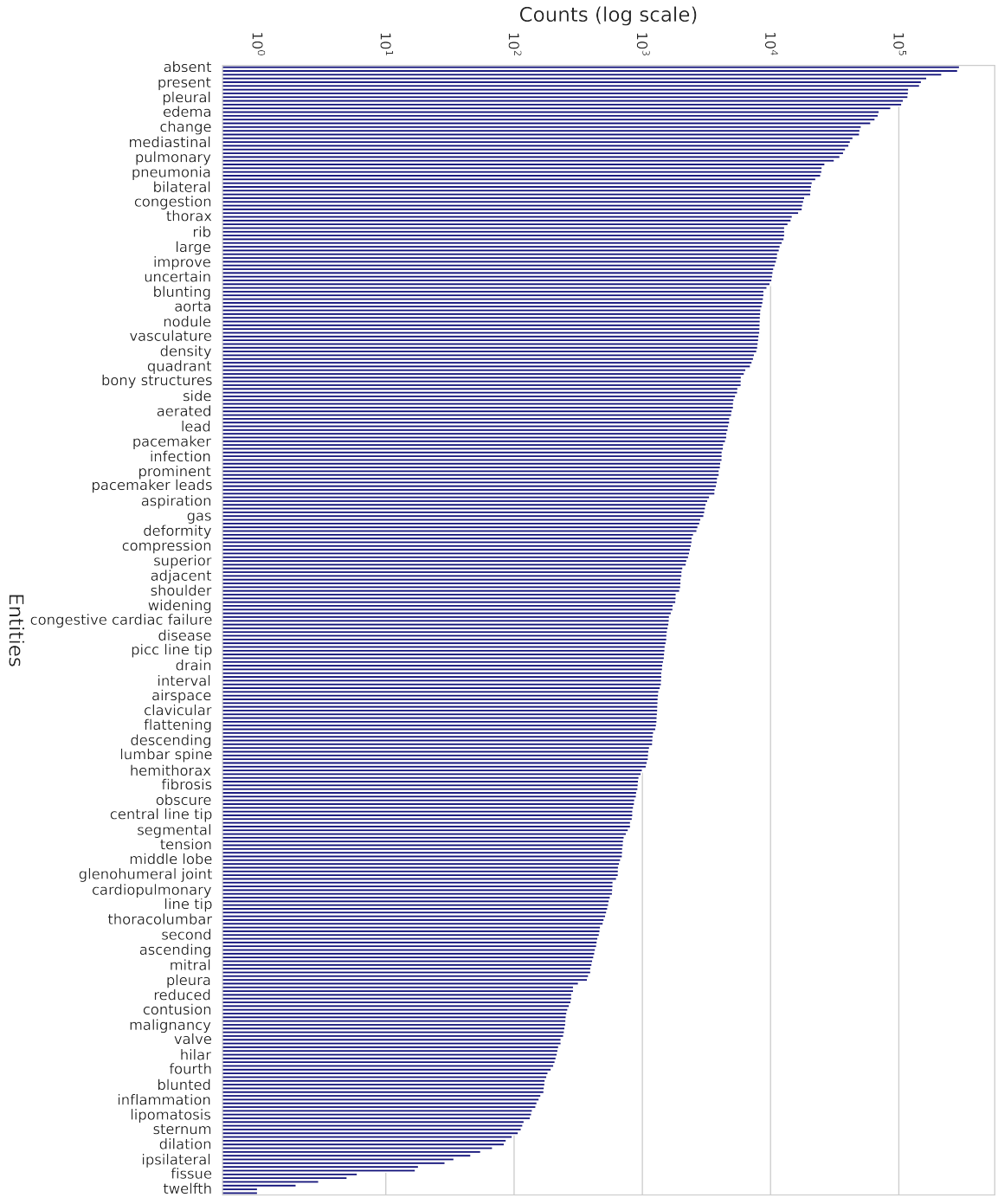


Figure 4.6: Distribution of the 302 entities extracted using our annotation pipeline on the Finding section of the MIMIC-CXR reports.

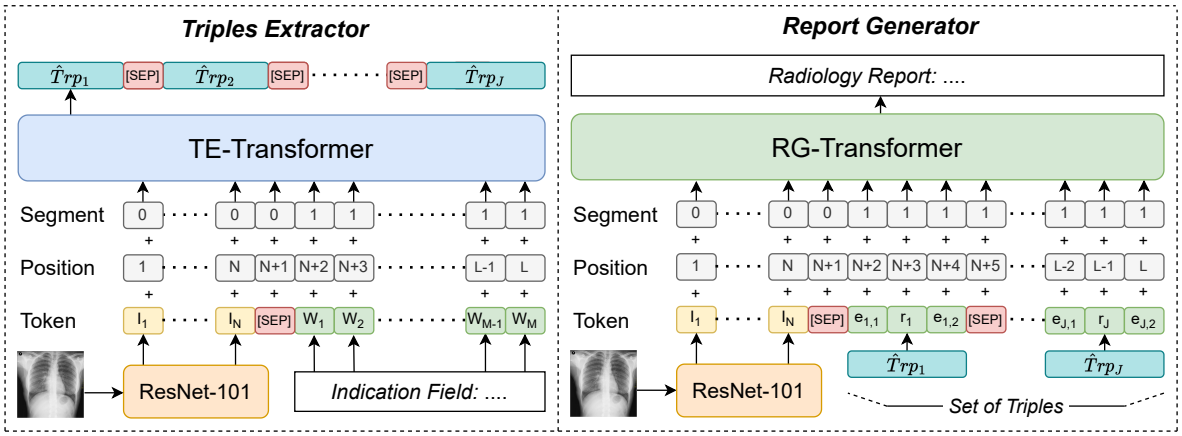


Figure 4.7: Architecture design of the two models of our TE-RG pipeline: *Triples Extractor* and *Report Generator*.

- **CXR TE-Transformer (CXR → Trp)**: a visual Transformer, which only takes a single CXR image as input.
- **MM TE-Transformer (CXR + Ind → Trp)**: a multimodal Transformer which takes as input the *Indication Field* (Ind), along with the CXR image, to provide additional context to the model.

4.3.2 Report Generator

The second step of the pipeline corresponds to the generation of the radiology report. The problem is again framed as a sequence-to-sequence task, using a multimodal encoder-decoder Transformer as the model backbone. The multimodal input sequence comprises the CXR image embedding $I = \{I_1 \dots I_N\}$, computed as in step 1; and the text embedding $\hat{T}rp = \{\hat{T}rp_1 \dots \hat{T}rp_J\}$ represents the extracted triples from step 1, which correspond to a single string of text, where the triples are separated by a `[SEP]` token.

During the training phase, we use the concatenation of the ground truth triples $Trp = \{Trp_1 \dots Trp_K\}$, to train our model. To prevent the model from focusing only on the triples – which already contain a comprehensive set of information, sufficient to generate a clinically accurate report – and ignoring the CXR image, we also consider randomly masking out 40% of the triples (this percentage was selected empirically based on the performance on the

validation set). This way, we expect the model to also learn representative features from the image to compensate for the missing information. We adopt such a training strategy because step 1 is not expected to be performed perfectly, thus we force the model to still consult the image when generating the final report.

During this step, we compare three different setups of the report generator model *RG-Transformer*, to generate the radiology report (RR):

- **Trp** $\rightarrow$ **RR**: a Transformer which takes only triples as input to generate radiology report.
- **Trp + CXR** $\rightarrow$ **RR**: a multimodal Transformer taking both triples and CXR as inputs.
- **Trp + CXR** $\rightarrow$ **RR (w/ Mask)**: a multimodal Transformer, similar to the above, trained on a random subset of the input triples.

4.4 Experimental Setup

4.4.1 Dataset

We conducted our experiments on the MIMIC-CXR dataset, which comprises 377,110 CXR images from 65,379 patients and the associated radiology reports. We adopted the same training/validation/test split as used by [32][3] and [31][4], for a fair comparison with their methods. This results in 270,790 training images, 2,130 validation images and 3,858 test images, alongside the associated radiology reports. All the images are resized by matching the smaller edge to 256 pixels and maintaining the original aspect ratio.

Following previous methods, we consider only the *Finding Section* of each report as the target text output of our pipeline; this is the section in the report which contains a free-text description of the radiographic findings and/or pathologies which are visualised within the image. Further, we extract the *Indication Field* (sometimes termed *Clinical History*) from the radiology reports, when this is present, as it contains relevant medical history. We use this as

[3] https://github.com/cuhksz-nlp/R2Gen
[4] https://github.com/cuhksz-nlp/R2GenCMN

additional context for the Triples Extraction step since this is the part of the report that would be available at imaging time.

4.4.2 Baselines

We compare our TE-RG two-step pipeline with:

- **Lower Bound (CXR $\rightarrow$ RR)**: we train an encoder-decoder Transformer architecture which generates the reports from the CXR in one step, without extracting the triples first. This defines the Lower Bound, and we expect our TE-RG pipeline to outperform this.
- **Upper Bound (GT-Trp $\rightarrow$ RR)**: we train an encoder-decoder Transformer to generate the radiology report from the ground truth triplets (GT-Trp). This sets an Upper Bound to our problem, as it mimics the scenario where all the triples are perfectly extracted in step 1. This allows us to understand the feasibility of generating a report from the set of triplets.

4.4.3 Implementation Details

We consider the same model architecture for both steps of the proposed pipeline. A vanilla encoder-decoder Transformer is used as the backbone of our models. Both its encoder and decoder are composed of three Attention Layers, as described by [198], each composed of 8 heads and 512 hidden units, and we initialise them randomly. For both steps, the vocabulary of the tokeniser is defined independently, where each token corresponds to a single word appearing either in the input or output text of the training set; with an additional `[SEP]` token used in the input to separate the image vs text (first step), or image vs triples (second step).

We adopt ResNet-101 as the visual extractor, initialised using ImageNet pre-trained weights [46], with the scope of encoding a single CXR image and feeding the embedding to the Transformer as the visual input. During training, we adopt standard data augmentation of

Model	**F1**
CXR TE-Transformer: CXR $\rightarrow$ Trp	0.275
MM TE-Transformer: CXR + Ind $\rightarrow$ Trp	**0.307**

Table 4.2: F1 scores for triples (Trp) extracted in step 1 on the test set of MIMIC-CXR. We compare two different versions of the Triples Extractor, as defined in Section 4.3.

the image: random 224×224 crop; random horizontal flip; and random rotation within the range $(-10^{\circ}, +10^{\circ})$. During inference, we take a 224×224 central crop of the image.

At each step, the whole model is trained end-to-end using a cross-entropy loss with Adam optimiser [100]. The learning rate for the visual extractor is set to 5×10^{-5} and 1×10^{-4} for the remaining parameters, and we decay them by a factor of 0.8 every three epochs.

4.4.4 Metrics

To evaluate the goodness of step 1, we compute the F1 score between the set of extracted triples $\hat{T}rp$ and the set of ground truth triples Trp.

Step 2 is evaluated using common Natural Language Generation (NLG) metrics: BLEU score [150], ROUGE score [122] and METEOR [14]. Given that these often fail to capture the semantic meaning of the text, we also consider Clinical Efficiency (CE) metrics. These are computed by applying the CheXpert labeler [83] to the generated reports, which extracts 14 labels: *Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax*, and *Support Devices*. Generated labels are then compared with the ground truth labels, provided in the MIMIC-CXR dataset, by computing precision, recall and F1 scores. The CE metrics are computed by condensing the four classes extracted from CheXpert (*positive*, *negative*, *uncertain*, or *missing*) into binary classes of *positive* versus others (*uncertain*, *negative*, *missing*). We note that the CheXpert labeler provides only a partial assessment of clinical accuracy since attributes are ignored, as well as entities outside of the 14 defined labels. Therefore we also perform a qualitative human evaluation of a subset of the generated reports.

Model		NLG Metrics						CE Metrics		
Step 1	**Step 2**	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**P**	**R**	**F1**
Lower Bound: CXR → RR		0.341	0.212	0.145	0.106	0.136	0.280	0.373	0.33	0.334
CXR + Ind → Trp	Trp → RR	0.322	0.219	0.159	0.122	0.150	0.311	**0.454**	0.431	0.442
CXR + Ind → Trp	CXR + Trp → RR	0.336	0.226	0.164	0.125	0.149	0.307	0.439	0.398	0.417
CXR + Ind → Trp	CXR + Trp → RR (w/ Mask)	**0.363**	**0.245**	**0.178**	**0.136**	**0.161**	**0.313**	0.428	**0.459**	**0.443**
Upper Bound: GT-Trp → RR		0.523	0.408	0.332	0.276	0.251	0.466	0.523	0.581	0.551

Table 4.3: NLG and CE results on the MIMIC-CXR test set, where BL=BLEU, MTR=METEOR, RG=ROUGE, P=Precision and R=Recall. We adopt the TE-RG pipeline, considering a multimodal TE-Transformer to extract the triples in the 1st step, and comparing different implementations of the 2nd step, defined in Section 4.3. These results are also compared with the Lower Bound and the Upper Bound models, described in Section 4.4.2.

Model	NLG Metrics						CE Metrics		
	BL-1	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**P**	**R**	**F1**
ST [201]	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.204
Att2In [170]	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
AdaAtt [135]	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.181
TopDown [8]	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
R2Gen [32]	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
CA [129]	0.350	0.219	0.152	0.109	0.151	0.283	-	-	-
CMCL [127]	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-
PPKED [128]	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
R2Gen CMN [31]	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
R2Gen CMM+RL [155]	**0.381**	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
Ours	0.363	**0.245**	**0.178**	**0.136**	**0.161**	**0.313**	**0.428**	**0.459**	**0.443**

Table 4.4: NLG and CE results on the MIMIC-CXR test set. All the results of the comparison methods are taken from [155].

4.5 Results

Here we evaluate our proposed method on the MIMIC-CXR dataset at each step: *Triples Extraction* and *Report Generation*. Every experiment is repeated 3 times using different random seeds to initialise the model weights and randomise batch shuffling; we report the average scores between the 3 different runs. We also conduct some human evaluation on the generated reports, to further assess their clinical accuracy.

4.5.1 Results on Triples Extraction

In Table 4.2, we compare the two models – CXR TE-Transformer and MM TE-Transformer – by computing the F1 score on both the MIMIC-CXR validation and test set. This shows that

introducing the *Indication Field* as additional context to the model helps to restore the triples more accurately. This result confirms what has previously been found by [85], and extends their results on a more difficult task.

4.5.2 Results on Report Generation

In Table 4.3, we show a comparison of three variants of the Report Generator, which are described in Section 4.3. We also compare the results with a Lower Bound and an Upper Bound model, defined in Section 4.4.2. During inference, for all three models, we input the triples extracted by the MM TE-Transformer, as it yields the highest F1 scores.

It can be seen that the models trained without masking do not consistently outperform the Lower Bound metrics. The reason could be attributed to the fact that, during training, we input to the model the ground truth triples, which contain the necessary information to generate a good quality report. Therefore, the model tends to focus solely on the triples and always expects to see a set of triples perfectly matching the final report. However, this is not true, as seen from the results in Table 4.2. We overcome this by masking out some of the ground truth triples during training, which encourages the model to leverage also the CXR image when generating the radiology report. Moreover, it can be noticed that all three models show significantly lower performance compared to the upper bound. This suggests that there is still a considerable margin of improvement.

In Table 4.4, we benchmark our pipeline against existing state-of-the-art ARR methods. Our TE-RG approach outperforms other methods for most of the NLG metrics and all the CE metrics, suggesting a good compromise between clinical accuracy and text fluency of the generated radiology reports.

4.5.3 Human Evaluation

Both NLG and CE metrics allow us to automatically evaluate the quality of the generated reports. However, they both present some limitations. While NLG metrics measure the fluency

Original Report	Generated Report
The heart is normal in size. The cardiomediastinal contours are stable. There are stable bilateral pleural effusions with partial right-sided loculation. Biapical scarring and pleural thickening appears stable. There is again right-sided superior hilar retraction and mild rightward XXXX deviation. No acute infiltrate is appreciated.	As compared to the previous radiograph there is no relevant change. The extent of the **right** pleural effusion is constant. Constant size of the cardiac silhouette. No newly appeared parenchymal opacities. Omission errors = **Biapical scarring, hilar retraction, pleural thickening, XXXX deviation**
Large left lower lobe opacity is present. There does not appear to be significant mediastinal shift. There is no pneumothorax. The cardiac silhouette is not definitively identified and not fully evaluated. The mediastinal contours are unremarkable.	PA and lateral views of the chest were reviewed and compared to the prior studies. A right **pleural effusion** has increased in size since the prior study. **The left lung is clear**. There is no pneumothorax. Omission errors = **Left lower lobe opacity, mediastinal shift, mediastinal contours**

Figure 4.8: Example of human evaluation undertaken on generated reports. Errors: **Hallucination**, **Omission**, **Attribute error**, **Impression error**. In this data, taken from the IU-Xray dataset [44], ages (and other patient-identifiable information) are replaced by a placeholder, here indicated by XXXX.

of the generated reports, based on n-gram matching with the ground truth, they fail to capture the semantic meaning and they weight all the words equally, whether these are stopwords or clinically relevant findings or attributes. For instance, let us consider the two sentences (1) "the patient presents signs of pulmonary edema" and (2) "the patient presents *no* signs of pulmonary edema". These sentences share most of the words, resulting in high NLG scores, despite having opposite meanings. On the contrary, CE metrics, computed using CheXpert, capture the semantic meaning of the generated reports but are limited to the fixed 14 labels and measure neither the laterality nor the severity of the findings, which can influence both the diagnosis and the clinical treatment decision.

Due to these limitations of the adopted metrics, we additionally evaluated the quality of reports using two human evaluators, who compared the reports generated by the Lower Bound baseline model and our TE-RG approach to the original report. The evaluators were junior physicians with 2 and 3 years of clinical experience respectively, including experience of reading CXR reports. Evaluators were blinded to the model type used to generate reports during the exercise. For each example, evaluators were shown the radiologist's report and treated this as the gold standard (they were not shown the underlying CXR image). In line with human evaluation methods used to assess voice recognition software [166, 157, 172], evaluators counted types of errors which occurred in generated reports. The types of errors

Error Type	**Baseline**	**TE-RG**	**RC**
Hallucinations	101	66	-0.35
Omissions	103	86	-0.17
Attribute Errors	29	25	-0.14
Impression Errors	4	6	+0.50
Grammatical Errors	3	1	-0.67
Total Errors	240	184	-0.23
Critical Errors	31	22	-0.29

Table 4.5: Number of errors found by the clinical evaluators in 60 reports generated with the baseline and the TE-RG approach. We indicate with RC the relative change between the two models' errors.

evaluated are:

- **Hallucination** – when the predicted report includes findings not discussed in the ground truth report.
- **Omission** – when the predicted report omits findings present in the ground truth report.
- **Attribute** – when the predicted report describes the presence/absence of a finding correctly but the associated attribute is wrong or omitted (*e.g.*, wrong laterality).
- **Impression** – when the predicted report gives the wrong assessment (*i.e.*, wrong diagnosis).
- **Grammatical** – non-clinical errors such as repetitions.

Examples of the use of these errors is shown in Figure 4.8. There was also the option for evaluators to assign a *critical error* to the first four errors if this was felt to significantly alter the clinical course of action. For example, if a generated report erroneously described a region as being suggestive of pneumonia, this might result in a patient unnecessarily receiving antibiotics. Alternatively, if a report fails to describe a mass, this might result in possible cancer being missed.

The evaluators discussed and agreed on the evaluation protocol prior to the exercise. Evaluators received a combined total of 60 ground truth reports alongside the reports generated

with the baseline and the TE-RG approach, including 10 reports shown to both evaluators to compute the inter-annotator agreement. We found a moderate agreement between the two annotators with a Gwet's AC1 score [61] equal to 0.53.

The number of detected errors is displayed in Table 4.5. Most of the errors are reduced when using our TE-RG approach, which is consistent with the results in Section 4.5.2. This shows that the TE-RG approach generates more clinically accurate radiology reports compared to the single-step baseline. However, the number of clinical errors is still significant, which makes this method still unsuitable for real-life diagnostic applications.

4.6 Limitations

This chapter only focuses on CXR ARR due to the availability of large-scale open-access multimodal datasets (containing both images and reports), such as MIMIC-CXR, that are currently not accessible for other image modalities. However, a similar approach of extracting triples from the image could be extended to other image modalities, as this is not specific to CXRs. Because different medical images serve different purposes and are used to assess different pathologies, depending on each specific image modality and the region of the body that is assessed, the triples would contain entities and relations accordingly.

Another limitation of this work corresponds to the semi-automatic nature of the annotation pipeline. Due to the large number of reports, we still had to rely on automatic tools (*i.e.*, RadGraph and SciSpacy) to annotate triples. These are prone to errors, which can result in noisy triples. To alleviate this, we involved a junior physician to normalise entities and categories relations, as discussed in Section 4.2.

Despite showing results better or comparable to state-of-the-art methods, our proposed TE-RG approach is prone to error propagation between step 1 and step 2, since each step is treated independently. When the TE-Transformer incorrectly predicts triples, this results in incorrect reports predicted by the RG-Transformer. To mitigate this, we introduced a simple method of adding noise during training by masking a percentage of the triples. This simulates

the scenario where some triples are omitted in the prediction. However, other sources of errors could arise in step 1, such as incorrect triples (*i.e.* describing wrong findings) or incorrect entities (e.g., wrong laterality or incorrect severity). Therefore, further exploration of noise injection methods is needed, which we leave as future work.

Finally, similar to the other ARR methods in the literature we compared against, our method does not address the longitudinal comparison between multiple scans. Radiologists often mention in the reports how some specific finding has evolved between two subsequent findings, to assess whether it improved, worsened or remained stable. Since the ground-truth reports used to train our method contain such comparisons, the predicted reports often contain hallucinations about comparisons with prior CXRs. We will address such limitation in a later chapter by introducing a latent representation of prior scans into the network.

4.7 Conclusion

In this chapter, we present a two-step framework for CXR ARR, which splits the task into *Triples Extraction* and *Report Generation*, as opposed to directly generating the radiology report from the image. The intermediate triples extraction step serves for the model to focus on clinically relevant concepts, rather than directly focusing on predicting the final reports from the radiology images, which often exhibit a realistic style but lack clinical accuracy. To this end, we propose a semi-automated annotation schema, which extracts structured information from a radiology report in the form of triples and serves to supervise the first step of our approach.

Our method shows state-of-the-art performances on the MIMIC-CXR dataset for most of the NLG metrics and all the CE metrics. This is further validated by the human evaluation conducted to assess different error types in the generated text, comparing our proposed TE-RG approach with an image-to-report baseline, showing that our approach generates 23% fewer errors and 29% fewer critical errors compared to the baseline. In our results, we show how the intermediate step plays a key role in improving the reports' clinical accuracy. Therefore,

further research should be focused on how to extract less noisy triples or to adopt more sophisticated structure representations of the reports such as knowledge graphs. As suggested by Table 5.1, where the upper-bound significantly outperforms all other solutions, we show that focusing more on *Triples Extraction* is an important step for generating clinically accurate reports.

Nevertheless, end-to-end supervised report generation from images requires further research on improving clinical accuracy to have utility as a diagnostic tool.

In future, this solution can easily integrate more advanced model architectures – to both improve the triple extraction and the report generation – and can be applied to other complex image captioning tasks, such as those found in the medical domain.

Chapter 5

CXR Automated Reporting using Finding-Aware Anatomical Tokens

The main findings outlined in this chapter have been published as "Finding-Aware Anatomical Tokens for Chest X-Ray Automated Reporting" [41] at the Machine Learning in Medical Imaging Workshop (MLMI 2023) held in conjunction with the 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2023).

My contributions to this chapter include conceptualisation, methodological design, technical implementation, data analysis, conducting experiments, evaluation, and writing.

Chapter Summary

The task of Automated Radiology Reporting (ARR) comprises interpreting the medical findings in radiographic images, including descriptions of their location and appearance. Automated approaches to radiology reporting requires the image to be encoded into a suitable token representation for input to the vision-language model. Previous methods commonly use convolutional neural networks or vision Transformers to encode an image into a series of image-level *feature map representations. In pursuit of enhanced report quality, we explore the use of localised tokens that correspond to specific anatomical structures. Our approach is inspired by recent advancements in image captioning, where each visual token is linked to a detected object. We propose a novel adaptation of Faster R-CNN in which* finding detection *is performed for the candidate bounding boxes extracted during anatomical structure localisation. We use the resulting bounding box feature representations as our set of* finding-aware anatomical tokens*. This encourages the extracted anatomical tokens to be semantically informative about the findings they contain (required for the final task of radiology reporting). Our evaluation on the MIMIC-CXR dataset [91, 92, 57] of chest X-Ray images demonstrates that finding-aware anatomical tokens improve the performance of ARR systems compared to state-of-the-art, leading to clinically more accurate reports.*

Our contributions are:

1. *The development of a novel multi-task Faster R-CNN that extracts* finding-aware anatomical tokens *by performing finding detection on the candidate bounding boxes identified during anatomical region localisation. The model is trained on an extensive set of anatomy regions and associated findings from the Chest ImaGenome dataset [214, 57] to ensure the tokens convey rich information.*

2. *The integration of the extracted finding-aware anatomical tokens as the visual input in the Triples Extraction - Report Generation (TE-RG) two-stage pipeline for ARR, introduced in the previous chapter.*

3. *The demonstration of the effectiveness of these tokens in generating chest X-ray reports, supported by extensive experiments on the MIMIC-CXR dataset [91, 92, 57].*

5.1 Introduction

A radiology report is a detailed text description and interpretation of the findings present in a medical scan, including a description of their anatomical location and appearance. For example, a Chest X-Ray (CXR) report may describe an opacity (a type of finding) in the left upper lung (the relevant anatomical location) which is diagnosed as a lung nodule (interpretation). The combination of a finding and its anatomical location influences both the diagnosis and the clinical treatment decision, since the same finding may have a different list of possible clinical diagnoses depending on the location. For this reason, we focus on improving the Automated Radiology Reporting (ARR) task for CXR images by incorporating fine-grained, region-specific image representations that better capture the anatomical regions within the scan.

Recent ARR methods have adopted Convolutional Neural Networks (CNNs) [65, 73] or Vision Transformers [49] to encode the medical scan (*e.g.*, CXR) into global image-level features, which are then used as input to a Transformer language model [198] to generate the radiology report. Drawing inspiration from image captioning approaches in the general domain [8, 118, 229], where each visual token corresponds to an object detected in the input image, we investigate whether replacing the image-level tokens with local tokens – corresponding to anatomical structures – can improve the clinical accuracy of the generated reports. Moreover, we aim to extract local representations that are also semantically informative about the findings within each anatomical token, which is essential for the final task of ARR.

5.2 Related Works

5.2.1 Automated Radiology Reporting

Previous works on ARR have mostly focused on CXR scans, due to the large availability of open datasets compared to other medical imaging scans. These have examined the model architecture [32, 31], the use of additional loss functions [137], retrieval-based report generation

[190, 52], and grounding report generation with structured knowledge [38, 221]. However, no specific focus has been given to the image encoding. Inspired by these works in image captioning in the general domain [8, 118, 229], where each visual token corresponds to an object detected in an image, we propose to replace the image-level representations with local representations corresponding to anatomical structures detected in a CXR. To the best of our knowledge, only [210, 193] have considered anatomical feature representations for CXR automated reporting. In [210], they extract anatomical features from an object detection model trained solely on the anatomy localisation task. In [193], they train the object detector through multiple steps – anatomy localisation, binary abnormality classification and region selection – and feed each anatomical region individually to the language model to generate one sentence at a time. This approach simplistically assumes that each anatomical region is described in exactly one sentence in the report. However, most sentences in the report often refer to multiple regions (*e.g.* "Low lung volumes." refer to both lungs).

5.2.2 Finding Detection

Prior works have tackled the problem of finding detection in CXR images via weakly supervised approaches [227, 238]. However, the architectural design of these approaches does not allow the extraction of anatomy-specific vector representations, making them unsuited for our purpose. Agu et al [2] proposed AnaXnet, comprising two modules trained independently: a standard Faster R-CNN trained to localise anatomical regions, and a Graph Convolutional Network trained to classify the pathologies in each anatomical region bounding box. This approach assumes that the finding information is present in the anatomical representations after the first stage of training. However, as we will show later in this chapter, our results suggest that this assumption does not hold.

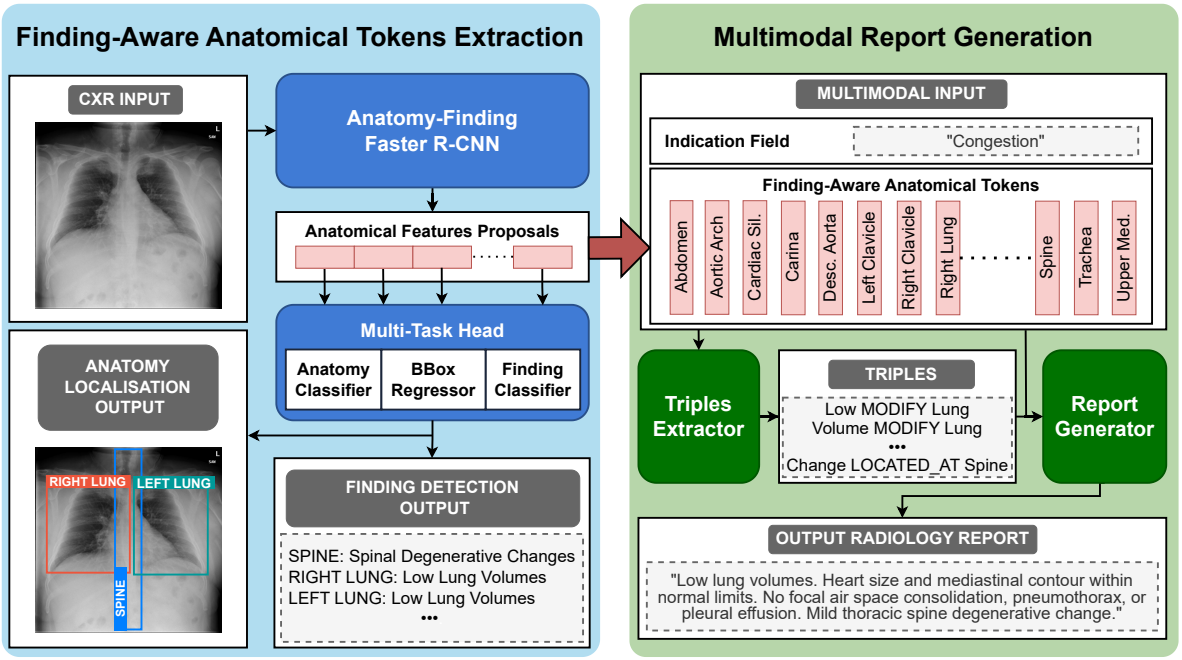


Figure 5.1: Finding-aware anatomical tokens integrated into the TE-RG CXR automated reporting pipeline. The CXR image and report are taken from the IU-Xray dataset [44].

5.3 Methods

We describe our method in two parts: (1) Finding-aware anatomical token extraction (Figure 5.1, left) – a custom Faster R-CNN which is trained to jointly perform *anatomy localisation* and *finding detection*; and (2) Multimodal report generation (Figure 5.1, right) – the TE-RG two-step pipeline (introduced in Chapter 4) which is adapted to perform *triples extraction* and *report generation*, using the anatomical tokens extracted from the Faster R-CNN as the visual inputs for the multimodal Transformer backbone [198].

5.3.1 Finding-Aware Anatomical Token Extraction

Let us consider $A = \{a_n\}_{n=1}^{N}$ as the set of anatomical regions in a CXR and $F = \{f_m\}_{m=1}^{M}$ the set of findings we aim to detect. We define $f_{n,m} \in \{0,1\}$ indicating the absence or presence of the finding f_m in the anatomical region a_n, and $f_n = \{f_{n,m}\}_{m=1}^{M}$ as the set of findings in a_n. We define *anatomy localisation* as the task of predicting the top-left and bottom-right bounding box coordinates $c = (c_{x1}, c_{y1}, c_{x2}, c_{y2})$ of the anatomical regions A; and *finding detection* as

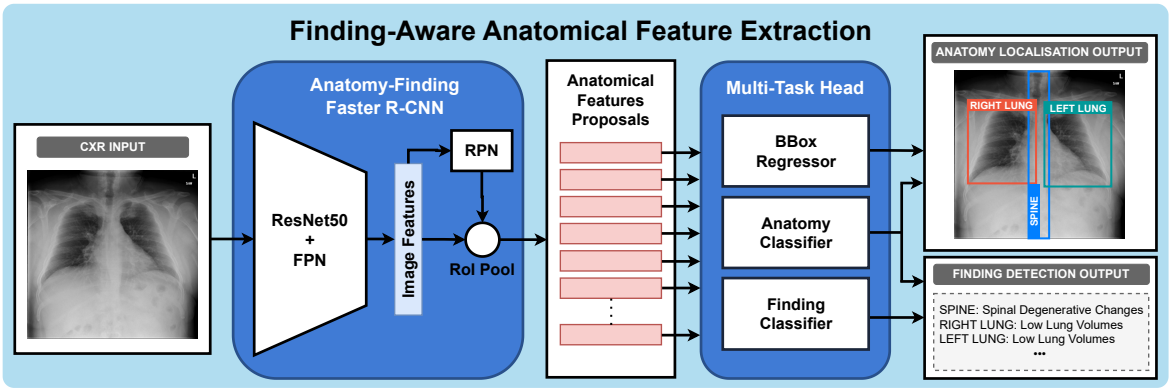


Figure 5.2: The proposed anatomy-finding Faster R-CNN trained jointly on anatomy localisation and finding detection.

the task of predicting the findings f_n at each location a_n.

We frame anatomy localisation as a general object detection task, employing the Faster R-CNN framework to compute the coordinates of the bounding boxes and the anatomical labels assigned to each of them. First, the image features are extracted from the CNN backbone, composed of a ResNet-50 [65] and a Feature Pyramid Network (FPN) [123]. Second, the multi-scale image features extracted from the FPN are passed to the Region Proposal Network (RPN) to generate the bounding box coordinates $c_k = (c_{k,x1}, c_{k,y1}, c_{k,x2}, c_{k,y2})$ for each proposal k and to the Region of Interest (RoI) pooling layer, designed to extract the respective fixed-length vector representation $l_k \in \mathbb{R}^{1024}$. Each proposal's local features l_k are then passed to a classification layer (*Anatomy Classifier*) to assign the anatomical label (a_k) and to a bounding box regressor layer to refine the coordinates. In parallel, we insert a multi-label classification head (*Findings Classifier*) – consisting of a single fully-connected layer with sigmoid activation functions – that classifies a set of findings for each proposal's local features (Figure 5.2).

During training, we use a multi-task loss comprising three terms: *anatomy classification loss*, *box regression loss*, and (multi-label) *finding classification loss*. Formally, for each predicted bounding box, this is computed as

$$\mathcal{L} = \mathcal{L}_{anatomy} + \mathcal{L}_{box} + \lambda \mathcal{L}_{finding}, \tag{5.1}$$

where $\mathcal{L}_{anatomy}$ and $\mathcal{L}_{box}$ correspond to the anatomy classification loss and the bounding box regression loss described in [55][1] and $L_{finding}$ is the finding classification loss that we introduce; λ is a balancing hyper-parameter set to $\lambda = 10^2$. We define

$$\mathcal{L}_{finding} = -\sum_{k=1}^{K}\sum_{m=1}^{M} w_m f_{k,m} \log(p_{k,m}) \tag{5.2}$$

a binary cross-entropy loss between the predicted probability $p_k = \{p_{k,m}\}_{m=1}^{M}$ of the k-th proposal and its associated ground truth $f_k = \{f_{k,m}\}_{m=1}^{M}$. We class weight using $w_m = (1/\nu_m)^{\alpha}$, where ν_m is the frequency of the finding f_m in the training dataset and we empirically set α to 0.25.

At inference time, for each CXR image, we extract the finding-aware anatomical tokens $A_{tok} = \{l_n\}_{n=1}^{N}$. For each anatomical region, we consider a proposal as a valid positive detection if its anatomical classification score is greater than 0.5. Additionally, for each region, we retain only the proposal with the highest anatomical classification score and take the associated latent vector representation l_n. Regions with no valid detection or non-detected regions are assigned a zero vector of dimension 1024. The set of anatomical tokens A_{tok} is then provided as input to the report generation model.

5.3.2 Multimodal Report Generation

We adopt the multimodal knowledge-grounded approach for ARR on CXR images, as proposed in Chapter 4. Firstly, *triples extraction* is performed to extract structured information from a CXR image in the form of triples, given the indication field (*Ind*) as context. Secondly, *report generation* is performed to generate the radiology report from the triples with the CXR image and indication field again provided as context.

Each step is treated as a sequence-to-sequence task; for this purpose, the triples are concatenated into a single text sequence (in the order they appear in the ground truth report) separated by the special `[SEP]` token to form Trp, and the visual tokens are concatenated

[1]The sum of $\mathcal{L}_{anatomy}$ and $\mathcal{L}_{box}$ correspond to the detector loss defined in Equation 2.4.

in a fixed order of anatomical regions. Two multimodal encoder-decoder Transformers are employed as the Triples Extractor (TE) and Report Generator (RG). The overall approach is:

$$\begin{aligned} &\text{STEP 1} && Trp = TE(seg_1 = A_{tok}, seg_2 = Ind) \\ &\text{STEP 2} && R = RG(seg_1 = A_{tok}, seg_2 = Ind \ \texttt{[SEP]} \ Trp) \end{aligned} \tag{5.3}$$

where seg_1 and seg_2 are the two input segments which are themselves concatenated at the input. In step 2, the indication field and the triples are merged into a single sequence of text by concatenating them, separated by the special `[SEP]` token. Similarly to [47], the input to a Transformer corresponds to the sum of the textual and visual *token embeddings*, the *positional embeddings*—to inform about the order of the tokens—and the *segment embeddings*—to discriminate between the two modalities.

5.4 Experimental Setup

5.4.1 Datasets

Our experiments are based on MIMIC-CXR [91, 92, 57], which consists of CXR image-report pairs and is used for the report generation task. Additionally, we use Chest ImaGenome [214, 57], which extends MIMIC-CXR, by including additional automatically extracted annotations for 242,072 anteroposterior and posteroanterior CXR images, which we use to train the finding-aware anatomical token extractor. We follow the same train/validation/test split as proposed in the Chest ImaGenome dataset. We extract the *Findings* section of each report as the target text[2]. For the textual input, we extract the *Indication field* from each report.[3] We annotate the ground truth triples for each image-report pair following a semi-automated pipeline using RadGraph [86] and sciSpaCy [139], as described in Chapter 4.

[2] https://github.com/MIT-LCP/mimic-cxr/blob/master/txt/create_section_files.py
[3] https://github.com/jacenkow/mmbt/blob/main/tools/mimic_cxr_preprocess.py

5.4.2 Metrics

To assess the quality of the generated reports, we compute Natural Language Generation (NLG) metrics: BLEU [150], ROUGE [122] and METEOR [14]. We further compute Clinical Efficiency (CE) metrics by applying the CheXbert labeller [184] which extracts 14 findings to the ground truth and the generated reports, and evaluate F1, precision and recall scores. The CE metrics are computed by aggregating the four classes extracted from CheXbert (*positive*, *negative*, *uncertain*, or *missing*) into binary classes of *positive+uncertain* versus *negative+textit*. We repeat each experiment 3 times using different random seeds, reporting the average in our results.

5.4.3 Implementation

Finding-aware anatomical token extractor

We adapt the Faster R-CNN implementation from [119][4], by including the finding classifier. We initialise the network with weights pre-trained on the COCO dataset [124], then fine-tune it to localise 36 anatomical regions and to detect 71 findings within each region, as annotated in the Chest ImaGenome dataset (Table 2.4). The CXR images are resized by matching the shorter dimension to 512 pixels (maintaining the original aspect ratio) and cropping to a resolution of 512×512 (random crop during training and centre crop during inference). We train the model for 25 epochs with a learning rate of 10^{-3}, decayed every 5 epochs by a factor of 0.8. We select the model with the highest finding detection performances for the validation set, measured by computing the AUROC score for each finding at each anatomical region.

Report generator

We implement a vanilla Transformer encoder-decoder at each step of the ARR pipeline. Both the encoder and the decoder consist of 3 attention layers, each composed of 8 heads and 512

[4] https://pytorch.org/vision/main/models/generated/torchvision.models.detection.fasterrcnn_resnet50_fpn_v2.html

hidden units. All the parameters are randomly initialised. We train step 1 for 40 epochs, with the learning rate set to 10^{-4} and we decay it by a factor of 0.8 every 3 epochs; and step 2 for 20 epochs, with the same learning rate as step 1. During training, we follow [38] in masking out a proportion of the ground-truth triples (50%, determined empirically), while during inference we use the triples extracted at step 1. We select the model with the highest CE-F1 score on the validation set.

Baselines

We benchmark against other CXR automated reporting methods: R2Gen [32], R2GenCMN [31], $\mathcal{M}^2$ Tr.+fact$_{\text{ENTNLI}}$ [137], CNN+Two-Step [38] and RGRG [193]. All these methods (except RGRG) adopt a CNN-Transformer and have shown state-of-the-art performances in report generation on the MIMIC-CXR dataset. All reported values are re-computed using the original code based on the same data split and image resolution as our method, except for [193] who already used this data split and image resolution, therefore we cite their reported results. We keep the remaining hyperparameters as the originally reported values.

CNN Pre-Training on Chest ImaGenome Findings

We train ResNet-101 on the set of 71 findings labels listed in Table 2.4 and use it to initialise the image encoders of the TE-RG model. This is done to validate that the improvements do not arise from the finding classification supervision alone but from the joint anatomy localisation and finding detection supervision. Differently from Faster R-CNN, ResNet-101 is trained only to classify whether findings are present or absent in a CXR without any further anatomical localisation of each finding. We train ResNet-101 for 50 epochs, with an initial learning rate set to 10^{-3} and decrease every 2 epochs by a factor of 0.5. The loss term corresponds to a binary cross-entropy:

$$\mathcal{L}_{class} = -\sum_{m=1}^{M} w_m f_m \log(p_m)$$

Method	NLG						CE		
	BL-1	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**F1**	**P**	**R**
R2Gen [32]	0.381	0.248	0.174	0.130	0.152	0.314	0.431	0.511	0.395
R2GenCMN [31]	0.365	0.239	0.169	0.126	0.145	0.309	0.371	0.462	0.311
$\mathcal{M}^2$ Tr. + $\text{fact}_{\text{ENTNLI}}$ [137]	0.402	0.261	0.183	0.136	0.158	0.300	0.458	0.540	0.404
ResNet-101 + $TE-RG$ [38]	0.468	0.343	0.271	0.223	0.200	0.390	0.477	0.556	0.418
RGRG [193]	0.400	0.266	0.187	0.135	0.168	-	0.461	0.475	0.447
A_{tok} + RG (*ours*)	0.422	0.324	0.265	0.225	0.201	**0.426**	0.515	0.579	0.464
A_{tok} + $TE-RG$ (*ours*)	**0.490**	**0.363**	**0.288**	**0.237**	**0.213**	0.406	**0.537**	**0.585**	**0.496**

Table 5.1: Comparison of our proposed solution with previous approaches. TE = Triples Extractor, RG = Report Generator.

between the predicted probability of each class p_m and the ground truth f_m. We class weight with $w_m = (1/\nu_m)^{0.25}$, where ν_m corresponds to the frequency of the finding f_m in the training set.

5.5 Results

5.5.1 Report Generation Results

In Table 5.1, we benchmark against other state-of-the-art CXR automated reporting methods and compare with the A_{tok} integrated into the full *TE-RG* pipeline versus a simpler approach of the report generator model only, *RG*, which generates the report directly from image and indication field (omitting triples extraction). The proposed solution, which integrates the finding-aware anatomical tokens with the *TE-RG* pipeline, generates reports with state-of-the-art fluency (NLG metrics) and clinical accuracy (CE metrics). Moreover, the superior results of our A_{tok} + RG approach compared to RGRG [193] suggest that providing the full set of anatomical tokens together, instead of separately, gives better results. The broader visual context is indeed necessary when describing findings that span multiple regions (*e.g.*, assessing the position of a tube).

Visual Input	Supervision	NLG						CE		
		BL-1	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**F1**	**P**	**R**
ResNet-101	ImageNet	0.468	0.343	0.271	0.223	0.200	0.390	0.477	0.556	0.418
ResNet-101	Findings	0.472	0.346	0.273	0.225	0.202	0.396	0.495	0.565	0.440
Naive A_{tok}	Anatomy	0.436	0.320	0.253	0.208	0.187	0.387	0.392	0.487	0.329
A_{tok}	Anatomy+Findings	**0.490**	**0.363**	**0.288**	**0.237**	**0.213**	**0.406**	**0.537**	**0.585**	**0.496**

Table 5.2: Comparison of different visual input representations (ResNet-101 *vs.* A_{tok}) using different pre-training supervision (ImageNet, Findings, Anatomy and Anatomy+Findings), integrated with the TE-RG two-step pipeline.

Supervision		Anatomy Localisation	Finding Detection
Anatomy	**Finding**	**mAP@0.5**	**AUROC**
✓		0.938	-
✓	✓	0.918	0.863

Table 5.3: Anatomy localisation and finding detection results of different configurations of the proposed Faster R-CNN: anatomy localisation only (Anatomy); and including the finding classification head (Finding) (*our proposed solution*).

5.5.2 Anatomy Localisation & Finding Detection Results

We evaluate the anatomy localisation performance of our proposed Faster R-CNN by computing the mean Average Precision (mAP@0.5), with positive detections when the Intersection over Union score between the predicted bounding boxes and the ground truth is above 0.5. Finding detection performance is measured by computing the Area Under the Receiver Operating Characteristic (AUROC) for each finding at each anatomical region.

Table 5.3 shows that our proposed anatomy-finding Faster R-CNN performs worse than a standard Faster R-CNN trained solely on anatomy localisation in terms of mAP@0.5 while achieving a good finding detection score. This is expected, as we select the best anatomy-only Faster R-CNN model based on the highest mAP@0.5 on the validation set and the anatomy-finding Faster R-CNN model based on the finding detection AUROC score. We observe that the trade-off between anatomy localisation and finding detection performance can be tuned by adjusting the weighting hyperparameter λ in the multi-task loss.

5.5.3 Ablation Study

Table 5.2 shows the results of adopting different visual representations. Firstly, we use a CNN (*ResNet-101*) trained end-to-end with TE-RG and initialised two ways: pre-trained on *ImageNet* versus pretrained on the *Findings* labels of Chest ImaGenome. Secondly, we extract anatomical tokens (A_{tok}) with different supervision of Faster R-CNN: anatomy localisation only (*Anatomy*) or anatomy localisation + finding detection (*Anatomy+Findings*). The results show the positive effect of including supervision with finding detection either when pre-training ResNet-101 or as an additional task for Faster R-CNN. However, these results show that the *Anatomy+Findings* achieve the highest scores across all metrics, showing the benefit of the proposed finding-aware anatomical tokens. Example reports are shown in Figure 5.3.

5.5.4 Anatomical Embedding Distributions

In Figure 5.4, we visualise the impact of the finding detection task on the extracted anatomical tokens. To generate these plots, for 3000 randomly selected test set scans, we first perform principle component analysis [93] for dimensionality reduction of the token embeddings (from $\mathbb{R}^{1024}$ to $\mathbb{R}^{50}$), then use t-distributed stochastic neighbour embedding (t-SNE) [196], colour coding the extracted embeddings by their anatomical region and additionally categorising as *normal* or *abnormal* (a token is considered abnormal if at least one of the 71 findings is positively labeled). For most anatomical regions, the normal and abnormal groups are better separated by the finding-aware tokens, suggesting these tokens successfully transmit information about findings. We also compute the mean distance between normal and abnormal clusters using Fréchet Distance (mFD) [50], measuring mFD=8.80 (naive anatomical tokens) and mFD=78.67 (finding-aware anatomical tokens).

GT Report	ResNet-101 + TE + RG	Naive A_{tok} + TE + RG	A_{tok} + TE + RG
Cardiac size is top normal. Aside from atelectasis in the left base the lungs are clear.[1] There is no pneumothorax or pleural effusion. There is hardware material in the cervical spine.	Cardiac mediastinal and hilar contours are normal. Pulmonary vasculature is normal. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. There is no evidence of pulmonary edema. There is minimal atelectasis in the right mid lung. There is no evidence of pneumoperitoneum.	The cardiomediastinal and hilar contours are within normal limits. The lungs are well expanded and clear. There is no focal consolidation pleural effusion or pneumothorax. There is minimal left basilar atelectasis.[1]	There is mild left basilar atelectasis.[1] No focal consolidation is seen. There is no pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are stable and unremarkable.
There is linear opacification of the left lung base[1] and right mid lung likely representing atelectasis.[2] The lungs are otherwise clear. The cardiac is top normal in size. There is no pleural effusion or pneumothorax.	Pa and lateral views of the chest provided.	Pa and lateral views of the chest provided. There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.	A linear opacity at the left lung base is most consistent with subsegmental atelectasis.[1,2] The lungs are otherwise clear. No focal consolidation pulmonary edema pleural effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.
Left-sided pacemaker device is noted with leads terminating in the right atrium and right ventricle.[1] Mild to moderate enlargement of cardiac silhouette is re-demonstrated.[2] The mediastinal and hilar contours are unremarkable. Pulmonary vasculature is not engorged. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities detected.	Left-sided pacemaker device is noted with leads terminating in the right atrium and right ventricle.[1] Cardiac silhouette size is normal. The mediastinal and hilar contours are unremarkable. The pulmonary vasculature is normal. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.	Heart size is normal. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.	Left-sided dual-chamber pacemaker is noted with leads terminating in the right atrium and right ventricle.[1] Heart size is mildly enlarged.[2] Mediastinal and hilar contours are normal. Pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is present. No acute osseous abnormalities detected.
As compared to the previous radiograph the patient has received a nasogastric tube.[1] The course of the tube is unremarkable the tip of the tube projects over the middle parts of the stomach.[2] There is no evidence of complications notably no pneumothorax. Left plate-like atelectasis.[3] No pleural effusions. No pneumonia no pulmonary edema.	As compared to the previous radiograph the patient has received a nasogastric tube.[1] The course of the tube is unremarkable the tip of the tube projects over the middle parts of the stomach.[2] There is no evidence of complications notably no pneumothorax.	As compared to the previous radiograph there is no relevant change. Low lung volumes. Borderline size of the cardiac silhouette with tortuosity of the thoracic aorta. No pulmonary edema. No pneumonia no pleural effusions. No pneumothorax.	A nasogastric tube terminates within the stomach.[1,2] The heart is normal in size. The mediastinal and hilar contours are normal. Plate-like atelectasis is present in the left mid lung.[3] There is no pleural effusion or pneumothorax.

Figure 5.3: Examples of predicted reports with different visual representations. From left to right: the ground truth (GT) report, the predicted reports using a CNN, the naive anatomical tokens and the finding-aware anatomical tokens as the visual representations. In generated reports, correctly detected positive findings are highlighted in green, and errors are highlighted in red. The equivalent text spans in the ground truth report are also highlighted; we number corresponding descriptions.

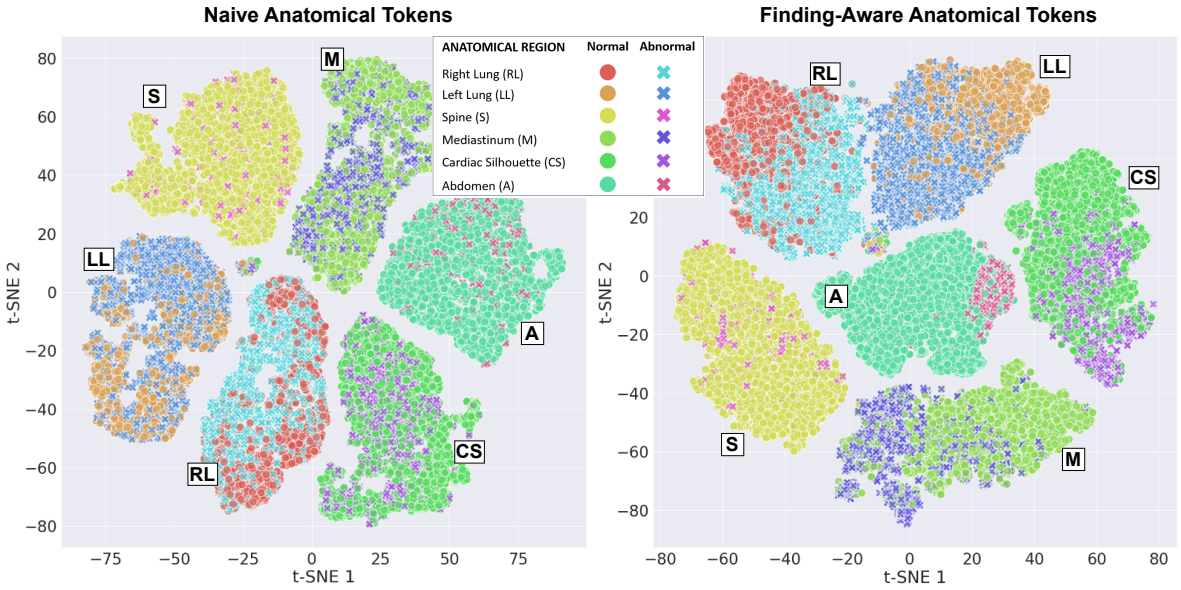


Figure 5.4: T-SNE visualisation of normal and abnormal embeddings for a subset of visual tokens. Left: *naive anatomical token* embeddings extracted from Faster R-CNN trained solely on anatomy localisation. Right: *finding-aware anatomical token* embeddings extracted from Faster R-CNN trained also on the finding detection task.

5.6 Limitations

This approach is specific to CXRs, in the sense that both the anatomical regions and finding categories are specific to what is visualised in CXRs. However, by re-defining these categories, a similar approach could be transferred to other image modalities or body regions; this might require ad-hoc solutions for 3D scans such as CT and MRI. Nonetheless, we hypothesise that extracting such local anatomical tokens is especially effective for CXR scans due to the complex, multi-organ nature of this modality, which is used to monitor a wide variety of conditions. More localised and specific imaging types might benefit less from such local visual representations.

Another limitation is that our model requires strong supervision, provided by the bounding box annotations and finding categories for each anatomical region. Chest ImaGenome is currently the only dataset that provides such annotations. However, in Chest ImaGenome these were automatically extracted, mitigating the intensive effort that would be required for manually annotating a large-scale dataset such as MIMIC-CXR. Similar solutions could be

applied to other multimodal datasets that provide imaging scans and humanly written radiology reports.

Even though our method uses visual tokens assigned to specific anatomical regions as input, it requires the full set of extracted regions to be present when generating the final radiology report. This means that the end user cannot control which region of the CXR the model should focus on when generating the report. Allowing this control could potentially unlock new capabilities of the ARR system, especially when a radiologist is only interested in describing specific abnormal regions.

Likewise in Chapter 4, we do not take into account the longitudinal evolution of different findings, such as the removal of support devices or worsening of pathologies, often reported by radiologists. As a result, our method still generates multiple hallucinations in the predicted report. We will address this limitation, as well as propose a more controllable solution in the next chapter.

5.7 Conclusion

This chapter explores how to extract and integrate anatomical visual representations with language models, targeting the task of ARR. We propose a novel multi-task Faster R-CNN adaptation that performs finding detection jointly with anatomy localisation, to extract *finding-aware anatomical tokens*. We then integrate these tokens as the local visual representation for a multimodal image+text report generation pipeline, as opposed to using global representations of the CXRs. We show that finding-aware anatomical tokens improve both the fluency (NLG metrics) and clinical accuracy (CE metrics) of the generated reports, giving state-of-the-art results on MIMIC-CXR.

Moreover, we argue that finding detection supervision is essential when training Faster R-CNN to extract informative anatomical tokens. This is shown in the ablation study, where we demonstrate that when adopting the naive anatomical tokens (extracted from Faster R-CNN trained solely on anatomy localisation) for report generation degrades performances across

all metrics. We further visualise through the t-SNE plot, how finding detection improves the token embeddings, which better discriminate between normal and abnormal regions.

Adopting local anatomy-specific representation for ARR can allow for more control over what is passed to the language model, hence having more control over the desired output. This as well as how to effectively compare longitudinal scans will be presented in the next chapter.

Chapter 6

Longitudinal and Controllable CXR Automated Reporting

The main findings outlined in this chapter have been published as "Controllable Chest X-Ray Report Generation from Longitudinal Representations" [40] in Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2023 (EMNLP 2023).

My contributions to this chapter include conceptualisation, methodological design, technical implementation, data analysis, conducting experiments, evaluation, and writing.

Chapter Summary

Radiology reports describe the presence/absence and location of relevant clinical findings, commonly including comparison with prior exams of the same patient to assess how they evolved. Therefore, to effectively adopt Automated Radiology Reporting (ARR) systems, they must be both accurate and easily interpretable, capable of assessing and tracking the evolution of a patient's condition over time. Previous approaches to ARR generally do not provide the prior study as input, precluding comparison which is required for clinical accuracy in some types of scans. Moreover, these methods do not allow for targeted analysis of specific regions and often rely on heatmaps for interpretation, which can be imprecise, making it difficult to identify the exact areas influencing the predictions. In this chapter, we propose a solution to model the temporal evolution of the findings of two subsequent scans, leveraging the finding-aware anatomical tokens, described in Chapter 5. To this end, we exploit the anatomical tokens of both the current and prior scan, proposing a method to align, concatenate and fuse the current and prior visual information into a joint longitudinal representation, provided as the visual input to the multimodal report generator. Moreover, we propose a novel training strategy to achieve controllability in the reporting generation process, where the report generator is trained to predict only the correct portion of the report specific to the subset of anatomical regions given as input. We show through in-depth experiments on the MIMIC-CXR dataset [91, 92, 57] how the proposed approach achieves state-of-the-art results while enabling anatomy-wise controllable report generation. Contributions in this chapter are:

1. *The development of a novel solution to model the evolution of two subsequent scans of the same patient, by aligning and concatenating representations for equivalent anatomical regions in prior and current CXRs and projecting them into a joint representation.*

2. *The development of a novel training strategy,* sentence-anatomy dropout*, in which the model is trained to predict a* partial report *based on a sampled subset of anatomical regions, thus training the model to associate input anatomical regions with the corre-*

sponding output sentences, giving controllability over which anatomical regions are reported on.

3. *Empirically demonstrate state-of-the-art performance of the proposed method on both full and partial report generation via extensive experiments on the MIMIC-CXR dataset.*

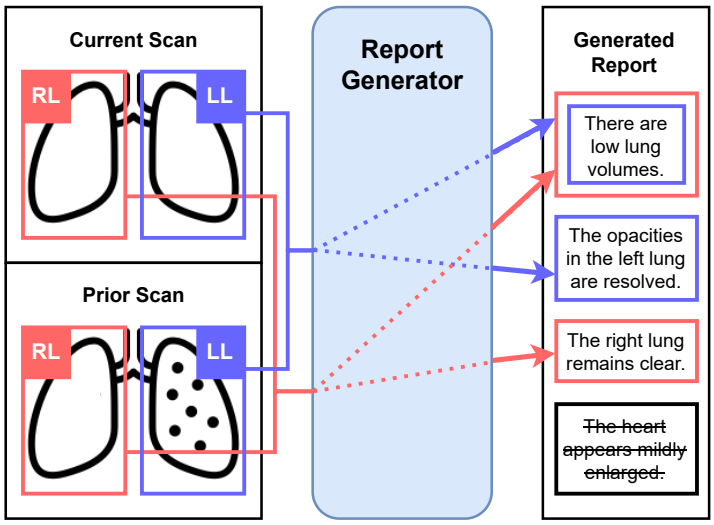


Figure 6.1: Illustration of our controllable ARR system using longitudinal representations. The report generator is trained to generate only sentences corresponding to the selected input anatomical regions. LL indicates the left lung and RL the right lung and we colour-match each region with the corresponding sentence. ~~Strikethrough text~~ represents the section of the report that we do not want the generated report to include when only the LL and RL are selected as inputs.

6.1 Introduction

In clinical practice, when prior studies are available, radiologists commonly compare the current clinical findings of a patient with the prior clinical findings to assess their evolution over time (*e.g.* "the heart remains enlarged", "the catheter has been removed"); this is especially critical in follow-up exams performed for monitoring.

For an Automated Radiology Reporting (ARR) system to be employed in real-world clinical scenarios, an automated system must be accurate, controllable and explainable. Meeting these criteria is challenging, particularly when the task involves complex clinical reasoning across various input image features and targeting a complex text output.

While radiology reporting is typically standardised, there are scenarios where a more flexible, controllable approach can be valuable. A controllable ARR system allows end users to select specific regions within an image, offering a deeper level of interaction and interpretability. This enables radiologists to generate detailed descriptions of specific anatomical areas, such as tracking the progression of findings in targeted regions. In these cases, the system can create

tailored reports for the selected regions, providing more precise insights. Additionally, this control enhances the understanding of model predictions by enabling comparisons between outputs generated from all anatomical regions and those based on a subset. This helps identify whether the model's predictions are influenced by irrelevant or incorrect areas, improving both interpretability and trustworthiness.

Previous works on CXR automated reporting have mostly focused on solutions to improve clinical accuracy, e.g. [137]. However, they generally use a single radiology study as input to generate the full report, precluding comparison with prior scans. They also do not allow the end user control over what parts of the image are reported on, leading to limited transparency on which image features prompted a specific clinical finding description. Interpretability is commonly achieved by generating heatmaps that highlight the areas of the input that influence the model's output [32, 31]. However, these heatmaps can be diffused or highlight large areas, making it unclear which specific features or regions are most important. They can also contain noise or be sensitive to small input changes, questioning their robustness.

In this chapter, we focus on two novel aspects: (1) **longitudinal representations** – the most recent previous CXR from the same patient is passed as an additional anatomically aligned input to the model to allow effective comparison of current and prior scans; (2) **controllable reporting** – to encourage the language model to describe only the subset of anatomical regions presented as input: this might be single anatomical regions (*e.g.* {`cardiac silhouette`} $\rightarrow$ "the cardiac silhouette is enlarged"), multiple anatomical regions (*e.g.* {`left lung`, `right lung`} $\rightarrow$ "low lung volumes") or the full set of anatomical regions (in which case the target regresses to the full report, as in previous methods). A high-level representation is shown in Figure 6.1.

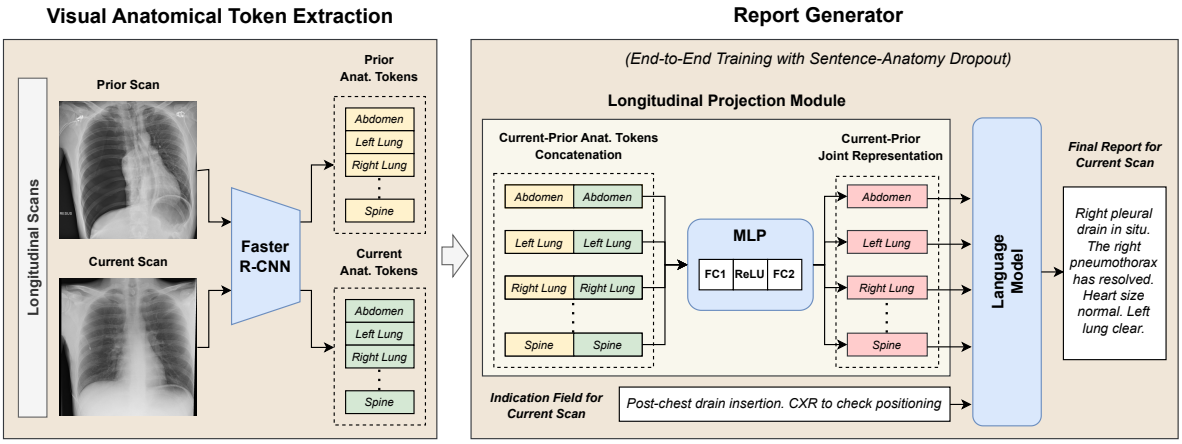


Figure 6.2: Architecture overview. The anatomical region representations of the current and prior CXRs are extracted from Faster R-CNN (*visual anatomical token extraction*). These are aligned, concatenated and projected into a joint representation (*longitudinal projection module*), then passed alongside the tokenised indication field as input to the language model to generate the report for the current scan. The Report Generator is trained end-to-end using *sentence-anatomy dropout*.

6.2 Related Works

6.2.1 Longitudinal CXR Representation

The problem of tracking how a patient's clinical findings evolve over time in CXRs has received limited attention either generally or for the application of CXR reporting, despite this being a critical component of a CXR report. [165] avoid the problem by proposing a method to remove comparison references to priors from the ground truth radiology reports to alleviate hallucinations about unobserved priors when training a language model for the downstream report generation task. [16] introduce a self-supervised multimodal approach that models longitudinal CXRs from image-level features as a joint temporal representation to better align text and image. [97] have proposed an anatomy-aware approach to classifying if a finding has improved or worsened by modelling longitudinal representations between CXRs with graph attention networks [200]. Similar to [97], we project longitudinal studies into a joint representation based on anatomical representations rather than image-level features. However, we extract the anatomical representations from Faster R-CNN [169], as in [41].

6.2.2 Controllable Automated Radiology Reporting

We define a *controllable* ARR system as one which allows the end users to select what regions in the image they want to report on, giving a level of interpretability. This has partially been tackled using hierarchical approaches [130], by introducing a multi-head transformer with each head assigned to a specific anatomical region and generating sentences exclusively for that region [204], and contemporaneously to our work by [193] who (similarly to us) generate sentences based on region-level features extracted through Faster R-CNN [169].

[193] make the assumption that each sentence in the report describes at most one anatomical region. Conversely, we acknowledge that there may be multiple anatomical regions which are relevant to the target text output, *e.g.*, a sentence "No evidence of emphysema." requires information from both left and right lungs; this requires us to identify valid subsets of anatomical regions in each CXR report for our dropout training strategy (Section 6.3.4).

6.3 Method

We first extract anatomical visual feature representations $\vec{v} \in \mathbb{R}^d$ with d dimensions for N predefined anatomical regions $A = \{a_n\}_{n=1}^{N}$ appearing in a CXR. To model longitudinal relations, we perform feature extraction for both the scan under consideration and for the most recent prior scan, and then combine region-wise using our proposed longitudinal projection module. We then input the features to the language model (LM) alongside the text indication field, which is trained using our anatomy-sentence dropout strategy. We show the proposed architecture in Figure 6.2 and describe these steps in more detail below.

6.3.1 Visual Anatomical Token Extraction

For the anatomical representations, we extract the bounding box feature vectors from the Region of Interest (RoI) pooling layer of a trained Faster R-CNN model. Faster R-CNN is trained on the tasks of *anatomical region localisation* in which the bounding box coordinates

Original Report	Anatomical region	Target output	Mapping type
"The mediastinum is mildly enlarged. Blunting of right costophrenic angle noted. No suspicious nodules seen. No pneumothorax or infective consolidation. Bilateral atelectasis, likely post -operative. Degenerative changes seen in both shoulders. NG tube tip positioned correctly in stomach. No free air under diaphragm"	`mediastinum`	"The mediastinum is mildly enlarged."	one-to-one
	`abdomen`	"NG tube tip positioned correctly in stomach. No free air under diaphragm."	one-to-many
	`left clavicle`, `right clavicle`	"Degenerative changes seen in both shoulders."	many-to-one
	`right lung`, `left lung`	"Blunting of right costophrenic angle noted. No suspicious nodules seen. No pneumothorax or infective consolidation. Bilateral atelectasis, likely post-operative."	many-to-many

Table 6.1: Example correspondences between anatomical regions and target output, for a synthesised CXR "Original Report" (Findings section). Note the different types of correspondence mapping.

of the N=36 anatomical regions (*e.g.* `abdomen`, `aortic arch`, `cardiac silhouette`) are detected in each CXR image, and *finding detection* in which presence/absence is predicted in each proposed bounding box region for a set of 71 predefined findings (*e.g.* `pleural effusion`, `lung cancer`, `scoliosis`).[1] Specifically, we augment the standard Faster R-CNN architecture head — comprising an anatomy classification head and a bounding box regression head — with a multi-label classification head, following [41], to extract *finding-aware anatomical tokens* $V = \{\vec{v}_n\}_{n=1}^{N}$ with $\vec{v}_n \in \mathbb{R}^d$ where d=1024. Then, for each anatomical region we select the bounding box representation proposal with the highest confidence score. When the anatomical region is not detected, we assign a d-dimensional vector of zeros. For more details about the model architecture, the loss term and other implementation details, we refer to Chapter 5.

[1] Full lists of anatomical regions and clinical findings are provided in Table 2.4.

6.3.2 Longitudinal Projection Module

Taking the *current scan* (the most recent scan at a specific time point) and the CXR from the most recent study (*prior scan*)[2], we extract from Faster R-CNN the anatomical tokens of both CXRs. There is one token for each of the N anatomical regions: $V_{current} = \{\vec{v}_{c,n}\}_{n=1}^{N}$ and $V_{prior} = \{\vec{v}_{p,n}\}_{n=1}^{N}$. When the current scan is part of an initial exam: $\vec{v}_{p,n} := \vec{0} \in \mathbb{R}^d$ $\forall n = 1, \ldots, N$. We select indices for the subset of regions that we want to report on $A_{target} \subseteq A$, and we obtain the longitudinal representation by concatenating the anatomical tokens for each anatomical region a_n of the two CXRs, and passing them through the longitudinal projection module f:

$$\vec{v}_{joint,n} = \begin{cases} f([\vec{v}_{c,n}, \vec{v}_{p,n}]) & \text{if } a_n \in A_{target} \\ f([\vec{0}, \vec{0}]), & \text{otherwise.} \end{cases}$$

The projection module f is a Multi-Layer Perceptron (MLP) consisting of a stack of a Fully-Connected layer (FC1), a ReLU function, and another Fully-Connected layer (FC2). We refer to the resulting output as the current-prior joint representation $V_{joint} = \{\vec{v}_{joint,n}\}_{n=1}^{N}$.

6.3.3 Report Generator

The Report Generator (RG) consists of a multimodal Transformer encoder-decoder, which takes the current-prior joint representation V_{joint} and the *indication field*[3] I as the visual and textual input respectively, to generate the output report Y:

$$Y = \text{RG}(V_{joint}, I)$$

where Y corresponds to the partial report if $A_{target} \subset A$ or the full report if $A_{target} = A$.

Similarly to [47], the input to the RG corresponds to the sum of the textual and visual

[2] A prior scan is only available if the current scan is not part of an initial exam.

[3] The indication field contains relevant medical history in the form of free text and it is available at imaging time.

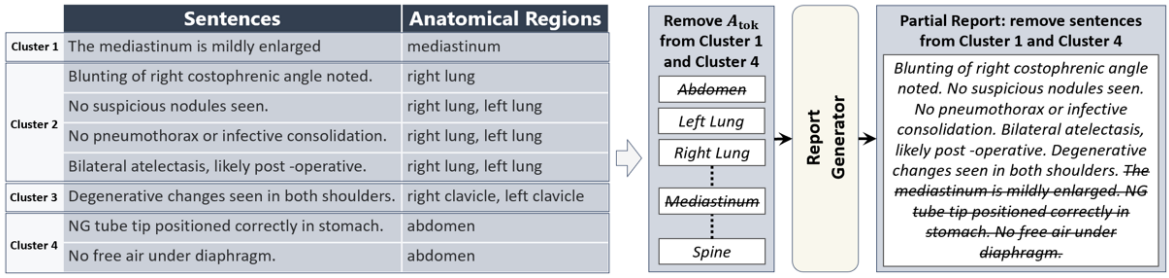


Figure 6.3: Sentence-anatomy dropout. We first cluster sentences in a report describing overlapping sets of anatomical regions. During training, one or more clusters are randomly selected and the corresponding anatomical tokens and sentences are removed.

token embeddings, the *positional embeddings* (for the position of tokens) and the *segment embeddings* (for modality type: vision or text).

6.3.4 Training with Sentence-Anatomy Dropout

During training, for each instance in each batch, we randomly drop a subset of anatomical tokens from the input and omit the corresponding sentences from the target radiology report; we term this training strategy **sentence-anatomy dropout**. In practice, for each training sample not all combinations of anatomical regions will be fit for dropout, since they must satisfy the following conditions:

1. Given a subset of anatomical tokens as input, the target output must be the full subset of sentences in the report that describe the corresponding anatomical regions;

2. Given a subset of sentences as the target output, anatomical tokens must be input for the full subset of described anatomical regions.

The above conditions are necessary since we reject the assumption that each sentence in the report describes only one anatomical region, as made by [193]. We illustrate with examples of the different mappings in Table 6.1.

Let us consider a radiology report as a set of L sentences $S = \{s_l\}_{l=1}^{L}$, each one describing the findings appearing in a different subset of anatomical regions $A_l \subseteq A$; and $P = \{\langle s_l, A_l \rangle\}_{l=1}^{L}$ the set of sentence-anatomy pairs of a report. To satisfy the two conditions above, we seek

Algorithm 1 Find set of valid sentence-anatomy subsets.

Input: set of $\langle$`sentence`, `regions`$\rangle$ pairs from a single CXR report, P
Output: set of valid sentence-anatomy subsets, $\mathcal{F}$

function FINDVALIDSUBSETS(P)
  $\mathcal{F} \leftarrow$ empty set
  $P_i \leftarrow$ set populated with the first $\langle$`sentence`, `regions`$\rangle$ pair in P
  $P_{remaining} \leftarrow$ set of $\langle$`sentence`, `regions`$\rangle$ pairs in P not assigned to P_i
  $R(P_i) \leftarrow$ regions in P_i
  $R(P_{remaining}) \leftarrow$ regions in $P_{remaining}$
  while $P_{remaining} \neq \{\}$ **do**
    while $R(P_i) \cap R(P_{remaining}) \neq \{\}$ **do**
      for $\langle$`sentence`, `regions`$\rangle \in P_{remaining}$ **do**
        if `regions` $\cap R(P_i) \neq \{\}$ **then**
          $P_i \leftarrow$ include $\langle$`sentence`, `regions`$\rangle$
        end if
      end for
      $P_{remaining} \leftarrow$ set of $\langle$`sentence`, `regions`$\rangle$ in P not assigned to P_i nor any $P_k \in \mathcal{F}$
      $R(P_i) \leftarrow$ regions in P_i
      $R(P_{remaining}) \leftarrow$ regions in $P_{remaining}$
    end while
    $\mathcal{F} \leftarrow$ include P_i
    $P_i \leftarrow$ set populated with the first $\langle$`sentence`, `regions`$\rangle$ pair in $P_{remaining}$
  end while
  return $\mathcal{F}$
end function

to discover the connected components in a graph where sentences are the nodes and an edge between two nodes represents an overlap of described anatomical regions between the two sentences. We describe in Algorithm 1 the algorithm to identify the connected components for each CXR report and how we group the corresponding sentence-anatomy pairs to each connected component into $P_k \subseteq P$. We then define as $\mathcal{F} = \{P_k\}_{k=1}^{K}$ the *set of valid sentence-anatomy subsets* (see Figure 6.4 for an example). During training, we randomly select one or more elements of $\mathcal{F}$ and then use the anatomical tokens as input and concatenate the corresponding sentences to create the target output. We show an example of the sentence-anatomy dropout pipeline in Figure 6.3

GT Report

'The lungs are hyperinflated with flattening of the diaphragms suggestive of underlying COPD. The heart is mildly enlarged. The aorta is tortuous and diffusely calcified. Mediastinal and hilar contours otherwise are unremarkable. Pulmonary vascularity is not engorged. No focal consolidation, pleural effusion or pneumothorax is identified. There are minimal streaky bibasilar atelectatic changes. No acute osseous abnormalities are present. Mild multilevel degenerative changes are seen in the thoracic spine.'

Chest ImaGenome sentence-anatomies pairs:

$$P = \{\langle s_l, A_l \rangle\}_{l=1}^{L}$$

Sentence	Anatomical Regions
'The lungs are hyperinflated with flattening of the diaphragms suggestive of underlying COPD.'	['right lung', 'left lung', 'right hemidiaphragm', 'left hemidiaphragm']
'Pulmonary vascularity is not engorged.'	['right lung', 'right hilar structures', 'left lung', 'left hilar structures']
'No focal consolidation, pleural effusion or pneumothorax is identified.'	['right lung', 'right costophrenic angle', 'left lung', 'left costophrenic angle']
'There are minimal streaky bibasilar atelectatic changes.'	['right lung', 'right lower lung zone', 'left lung', 'left lower lung zone']
'Mediastinal and hilar contours otherwise are unremarkable.'	['right hilar structures', 'left hilar structures', 'mediastinum', 'upper mediastinum']
'The aorta is tortuous and diffusely calcified.'	['mediastinum', 'aortic arch']
'The heart is mildly enlarged.'	['cardiac silhouette']
'No acute osseous abnormalities are present.'	['right clavicle', 'left clavicle', 'spine'],
'Mild multilevel degenerative changes are seen in the thoracic spine.'	['spine']

Set of valid sentence-anatomy subsets:

$$\mathcal{F} = \{P_k : P_k \subseteq P\}_{k=1}^{K}$$

Sentences	Anatomical Regions
'The lungs are hyperinflated with flattening of the diaphragms suggestive of underlying COPD.', 'Pulmonary vascularity is not engorged.', 'No focal consolidation, pleural effusion or pneumothorax is identified.', 'There are minimal streaky bibasilar atelectatic changes.', 'Mediastinal and hilar contours otherwise are unremarkable.', 'The aorta is tortuous and diffusely calcified.'	['right lung', 'left lung', 'right hemidiaphragm', 'left hemidiaphragm'] ['right lung', 'right hilar structures', 'left lung', 'left hilar structures'] ['right lung', 'right costophrenic angle', 'left lung', 'left costophrenic angle'] ['right lung', 'right lower lung zone', 'left lung', 'left lower lung zone'] ['right hilar structures', 'left hilar structures', 'mediastinum', 'upper mediastinum'] ['mediastinum', 'aortic arch']
'No acute osseous abnormalities are present.', 'Mild multilevel degenerative changes are seen in the thoracic spine.'	['right clavicle', 'left clavicle', 'spine'] ['spine']
'The heart is mildly enlarged.'	['cardiac silhouette']

Figure 6.4: Example of sentence-anatomy annotations of a report and its set of valid sentence-anatomy subsets.

6.4 Experimental Setup

6.4.1 Datasets

We consider two open-source CXR imaging datasets: MIMIC-CXR [91, 92, 57] and Chest ImaGenome [214, 57]. The MIMIC-CXR dataset comprises CXR image-report pairs. The Chest ImaGenome dataset includes additional annotations based on the MIMIC-CXR images and reports. In this chapter, we train Faster R-CNN with the automatically extracted anatomical bounding box annotations from Chest ImaGenome, provided for 242,072 AnteroPosterior (AP)

and PosteroAnterior (PA) CXR images. Chest ImaGenome also contains sentence-anatomy pairs annotations that we use to perform sentence-anatomy dropout. The longitudinal scans of each patient are obtained by ordering different studies based on the annotated timestamp and for each study, taking the most recent previous study as the prior. For this purpose, we only select AP or PA scans as priors (*i.e.*, ignore lateral views). If multiple scans are present in a study, we consider the one with the highest number of non-zero anatomical tokens. In case of a tie, we select it randomly. In all experiments, we follow the train/validation/test split proposed in the Chest ImaGenome dataset.

6.4.2 Data pre-processing

We extract the *Findings* section of each report as the target text[4]. For the text input, we extract the *Indication field* from each report[5].

When training Faster R-CNN, CXRs are resized by matching the shorter dimension to 512 pixels (maintaining the original aspect ratio) and then cropped to a resolution of 512×512.

6.4.3 Metrics

We assess the quality of our model's predicted reports by computing three Natural Language Generation (NLG) metrics: BLEU [150], ROUGE [122] and METEOR [14]. To better measure the clinical correctness of the generated reports, we also compute Clinical Efficiency (CE) metrics [184], derived by applying the CheXbert labeller to the ground truth and generated reports to extract 14 findings — and hence computing F1, precision and recall scores. In line with previous studies [137], this is computed by condensing the four classes extracted from CheXbert (*positive*, *negative*, *uncertain*, or *missing*) into binary classes of positive (*positive*, *uncertain*) versus negative (*negative*, *missing*).

[4] https://github.com/MIT-LCP/mimic-cxr/blob/master/txt/create_section_files.py
[5] https://github.com/jacenkow/mmbt/blob/main/tools/mimic_cxr_preprocess.py

6.4.4 Implementation

We adopt the `torchvision` Faster R-CNN implementation, as proposed in [119]. This consists of a ResNet-50 [65] and a Feature Pyramid Network [123] as the image encoder. We modify it and select the hyperparameters following [41].

The two FC layers in the MLP projection layer have input and output feature dimensions equal to 2048 and include the bias term.

Both the encoder and the decoder of the Report Generator consist of 3 attention layers, each composed of 8 heads and 512 hidden units.

The MLP and the Report Generator are trained end-to-end for 100 epochs using a cross-entropy loss with Adam optimiser [100] and the sentence-anatomy dropout training strategy. We set the initial learning rate to 5×10^{-4} and reduce it every 10 epochs by a factor of 10. The best model is selected based on the highest F1-CE score.

We repeat each experiment 3 times using different random seeds, reporting the average in our results.

6.4.5 Baselines

We compare our method with previous state-of-the-art works in CXR automated reporting: R2Gen [32], R2GenCMN [31], $\mathcal{M}^2$ Transformer+fact$_{\text{ENTNLI}}$ [137], A^{tok}+TE+RG [41] and RGRG [193]. For all baselines, we keep the hyperparameters as the originally reported values. For a fair comparison, we use the same text and image pre-processing used for our method and we re-train the baselines based on the Chest ImaGenome dataset splits.[6]

[6]We did not re-train [41] and [193], as they already use that same dataset split.

Method	NLG Metrics						CE Metrics		
	BL-1	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**F1**	**P**	**R**
R2Gen [32]	0.381	0.248	0.174	0.130	0.152	0.314	0.431	0.511	0.395
R2GenCMN [31]	0.365	0.239	0.169	0.126	0.145	0.309	0.371	0.462	0.311
$\mathcal{M}^2$ Tr. + $\text{fact}_{\text{ENTNLI}}$ [137]	0.402	0.261	0.183	0.136	0.158	0.300	0.458	0.540	0.404
A^{tok} + TE + RG [41]	**0.490**	0.363	0.288	0.237	0.213	0.406	0.537	0.585	0.496
RGRG [193]	0.400	0.266	0.187	0.135	0.168	-	0.461	0.475	0.447
Ours	0.486	**0.366**	**0.295**	**0.246**	**0.216**	**0.423**	**0.553**	**0.597**	**0.516**

Table 6.2: Comparison of our proposed approach with previous methods. We show the **best results in bold**.

Configuration		NLG Metrics						CE Metrics		
Priors	**SA drop**	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**F1**	**P**	**R**
-	-	0.430	0.327	0.266	0.224	0.202	0.420	0.534	0.593	0.485
✓	-	0.456	0.347	0.283	0.239	0.210	**0.428**	0.548	0.577	**0.522**
-	✓	0.473	0.358	0.289	0.243	0.213	0.426	0.550	**0.597**	0.510
✓	✓	**0.486**	**0.366**	**0.295**	**0.246**	**0.216**	0.423	**0.553**	**0.597**	0.516

Table 6.3: Ablation study on incorporating prior CXR scans as input and adopting sentence-anatomy dropout during training. We report the NLG and CE results on the MIMIC-CXR test set.

6.5 Results

6.5.1 Automated Radiology Reporting

Table 6.2 shows the effectiveness of the proposed method over the baselines on the ARR task, showing superior performance for most NLG and CE metrics. Compared to [41], our method can achieve similar BLEU metrics and superior scores on the remaining metrics, whilst providing also better controllability (and interpretability). Whilst both our method and that proposed by [193] tackle the controllability aspect, we show superior results in all metrics.

6.5.2 Ablation Study

We investigate the effect of incorporating prior CXR scans as input (*Priors*) and adopting sentence-anatomy dropout during training (*SA drop*).

Full Reports

We evaluate the different configurations of our method using the full set of anatomical regions as input ($A_{target} = A$) and the full report as the target text. Table 6.3 shows the results; it can be seen that both adding priors and using sentence-anatomy dropout during training boost most metrics, with the best overall performance obtained when combining the two mechanisms. This is illustrated qualitatively in Figure 6.6.

Initial *vs* Follow-up Scans

We study the effect of the different components by dividing the test set into initial scans versus follow-up scans, resulting in 11,951 and 20,735 CXR report pairs respectively. The results are shown in Table 6.4. We note the improvement of our method over the baseline on both subsets, with the best results obtained when adding the priors alone or in combination with the sentence-anatomy dropout. It is worth noting that the benefit of including priors is also present for initial studies with no prior scans. We hypothesise that since the model can infer which are initial scans (using the fact that the prior anatomical tokens are all zero-vectors), it will correctly generate a more comprehensive report rather than focusing on the progression or change of known findings.

Partial Reports

To measure the controllability of our method, we evaluate the ability of the different configurations to generate partial reports given a subset of anatomical regions. For this purpose, we divide each report in the test set into its set of valid sentence-anatomy subsets $\mathcal{F}$ (Algorithm 1). We take the anatomical regions contained in each subset $\chi \subset \mathcal{F}$ as input and the corresponding sentences as the target output. We then obtain a total of 71,698 partial reports. Figure 6.5 illustrates the distribution of partial reports per full report. Table 6.5 presents the NLG and CE metrics for partial reporting, averaged across all partial reports. These results demonstrate that using sentence-anatomy dropout provides a controllable approach, accurately reporting on the

Configuration		NLG Metrics			CE Metrics		
Priors	**SA drop**	**BL-4**	**MTR**	**RG-L**	**F1**	**P**	**R**
Initial Scans							
-	-	0.283	0.234	0.479	0.532	0.570	0.499
✓	-	0.303	0.244	**0.490**	**0.543**	0.563	**0.524**
-	✓	0.303	0.244	0.485	0.541	0.582	0.507
✓	✓	**0.306**	**0.245**	0.479	0.542	**0.589**	0.502
Follow-up Scans							
-	-	0.194	0.187	0.377	0.533	0.600	0.479
✓	-	0.206	0.194	**0.383**	0.550	0.583	**0.521**
-	✓	0.210	0.197	0.378	0.552	**0.602**	0.510
✓	✓	**0.216**	**0.202**	0.382	**0.557**	0.599	0.520

Table 6.4: NLG and CE results on the **Initial** and the **Follow-up** subsets of the MIMIC-CXR test set.

Configuration		NLG Metrics			CE Metrics		
Priors	**SA drop**	**BL-4**	**MTR**	**RG-L**	**F1**	**P**	**R**
-	-	0.113	0.187	0.291	0.549	0.519	0.583
✓	-	0.127	0.182	0.301	0.587	0.604	0.571
-	✓	**0.225**	**0.226**	**0.467**	0.667	0.651	**0.683**
✓	✓	0.223	0.225	0.462	**0.672**	**0.663**	0.680

Table 6.5: NLG and CE results for partial reporting, averaged across all **partial reports** of the MIMIC-CXR test set, by dividing each report into its set of valid sentence-anatomy subsets.

anatomical regions provided as input while avoiding hallucination of missing regions. This effect is further visualized in Figure 6.7.

6.5.3 Report Length

To further assess the quality of our method and each of its components, we measure the length distribution of the predicted reports, similar to [32], showing that our method more closely matches the ground truth distribution than baseline methods.

The length of a report corresponds to the number of words contained. We compute this for the full reports generated from the full set of anatomical regions and for the partial reports derived from the set of valid sentence-anatomy subsets. In Figure 6.8 (left) we see that the distribution of the proposed method is closer to the distribution of the GT reports. In Figure 6.8 (right), we note that adopting the sentence-region dropout strategy allows the method to generate partial reports with a length distribution closer to the GT partial reports. These

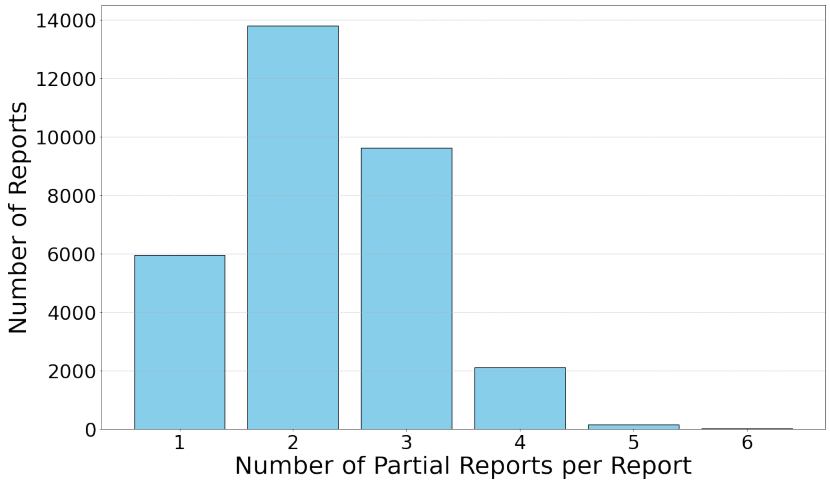


Figure 6.5: Frequency distribution of partial reports per report. The x-axis represents the number of partial reports in each report, while the y-axis indicates the number of reports with that many partial reports.

Ground Truth	**Baseline**	**Ours**	***Comments***
There are persisting small to moderate bilateral pleural effusions with subjacent atelectasis as well as pulmonary vascular engorgement and mild interstitial septal thickening. The right infrahilar opacity is unchanged. There is enlargement of the cardiac silhouette unchanged. Calcification of the aortic arch is present.	Moderate to severe cardiomegaly is stable. Moderate bilateral pleural effusions are unchanged. There is mild pulmonary edema. There is no focal consolidation or pneumothorax.	Moderate to large right and small left pleural effusions are unchanged. There is mild pulmonary vascular congestion and interstitial edema. There is a moderate right pleural effusion with associated right lower lobe atelectasis. Moderate cardiomegaly is stable. The aortic knob is calcified. No pneumothorax. No acute osseous abnormalities.	*The report predicted by our method shows a more accurate description of the findings and higher recall compared to the baseline.*
In comparison to the prior chest radiograph all lines and tubes have been removed. The bilateral lung aeration has improved dramatically. There is a small left pleural effusion. There is a subtle right basilar opacity. Heart size is normal. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. No pneumothorax is seen.	Compared to the prior study there is no significant interval change.	Interval removal of the endotracheal tube. The feeding tube has been removed. The right internal jugular catheter has been removed. The cardiomediastinal and hilar contours are within normal limits. The aorta is tortuous. There is a small left pleural effusion with adjacent atelectasis. There is no pneumothorax.	*The report predicted by our method correctly describes the removal of the different tubes.*
Stable chronically elevated right hemidiaphragm with moderate right lower lobe atelectasis. Small stable breast calcification projects adjacent to the left ventricle as is seen on ct. No pneumothorax or pleural effusion. Left lung clear. Heart size mediastinal contour and hila appear normal. No bony abnormality.	No focal consolidation pleural effusion pneumothorax or pulmonary edema is seen. Heart size is normal. Mediastinal contours are normal.	Elevation of the right hemidiaphragm is unchanged from prior. The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. Heart and mediastinal contours are within normal limits with tortuous aortic contour. There is elevation of the right hemidiaphragm.	*Our proposed method shows an improvement compared to the baseline, but there is still missing information such as a comparison with a previous CT scan.*

Figure 6.6: Qualitative results of full reports generation. We compare the reports generated by the baseline (without adding prior scans and sentence-anatomy training) and the proposed solution with the ground truth. We highlight using different colours the segments of the reports that are commented on in the right column.

results provide further evidence of the improvement of the proposed method over the baseline (without adding prior scans and sentence-anatomy training).

Anat. Regions	Ground Truth	Baseline	Ours
spine	Moderate anterior osteophytes along the lower thoracic spine appear unchanged.	**The heart is at the upper limits of normal size. The aortic arch is calcified. The mediastinal and hilar contours appear unchanged. There is no pleural effusion or pneumothorax. The lungs appear clear.**	Mild degenerative changes are similar along the thoracic spine.
cardiac silhouette	Moderate to severe enlargement of the cardiac silhouette is unchanged.	Moderate enlargement of the cardiac silhouette is re-demonstrated. **Mediastinal and hilar contours are unchanged. Pulmonary vasculature is not engorged. No focal consolidation pleural effusion or pneumothorax is present.There are no acute osseous abnormalities**.	Moderate to severe cardiomegaly is re-demonstrated.
abdomen, cardiac silhouette, mediastinum, upper mediastinum	The nasogastric tube has been removed. The heart and mediastinum are within normal limits.	The heart size is normal. The **hilar** and mediastinal contours are normal. **There has been interval resolution of the previously seen small left-sided pneumothorax. There has been slight interval improvement of the previously seen small right pleural effusion. There is no evidence of focal consolidations concerning for pneumonia.**	The nasogastric tube has been removed. The cardiomediastinal silhouette is unremarkable.
left lung, left lower lung zone left costophrenic angle, left hilar structures, right lung, right lower lung zone, right costophrenic angle, right hilar structures	There are areas of streaky atelectasis at the bilateral lung bases. There are persistent prominent interstitial markings which suggest chronic interstitial abnormality versus mild interstitial edema. The lungs remain hyperinflated. There is no pleural effusion or pneumothorax. No focal consolidation is seen.	Lung volumes are low. Linear opacities in the bilateral lower lungs are most consistent with subsegmental atelectasis. There is no focal consolidation pleural effusion or pneumothorax. **The cardiomediastinal silhouette is unchanged.**	There is pulmonary vascular congestion and mild interstitial pulmonary edema. Linear bibasilar opacities are most consistent with atelectasis. There is no pleural effusion or pneumothorax.

Figure 6.7: Qualitative results of partial reports generation. From left to right: the subset of anatomical regions A_{target} we want to report, the ground truth partial reports, the reports generated by the baseline (without adding prior scans and sentence-anatomy training) based on A_{target} and those generated by our proposed method. We indicate in **red** the hallucination on the missing anatomical regions.

6.6 Limitations

Similar to previous chapters, our method focuses only on CXR and adapting it to other types of medical scans might be challenging. First, due to the 2D nature of CXRs compared to other types of 3D scans (*e.g.*, CT, MRI). Second, we strongly rely on the Chest ImaGenome dataset and its annotations. These are automatically extracted and similar sentence-anatomy annotations could be extracted for radiology reports from other types of scans. However, as the same authors pointed out, there are some known limitations of their NLP and the region extraction pipelines; for instance, clinical findings may not be properly extracted from a follow-up report which may be as simple as "No change is seen". Hence, some refinement of the pipelines with or without additional manual input might be required.

A limitation of the sentence-anatomy dropout strategy is the potential mismatch between the combinations of anatomical regions seen during training and those that could be selected interactively by the end users. The training process relies on the natural co-occurrence patterns of anatomical regions found in radiology reports, whereas users might request rare or

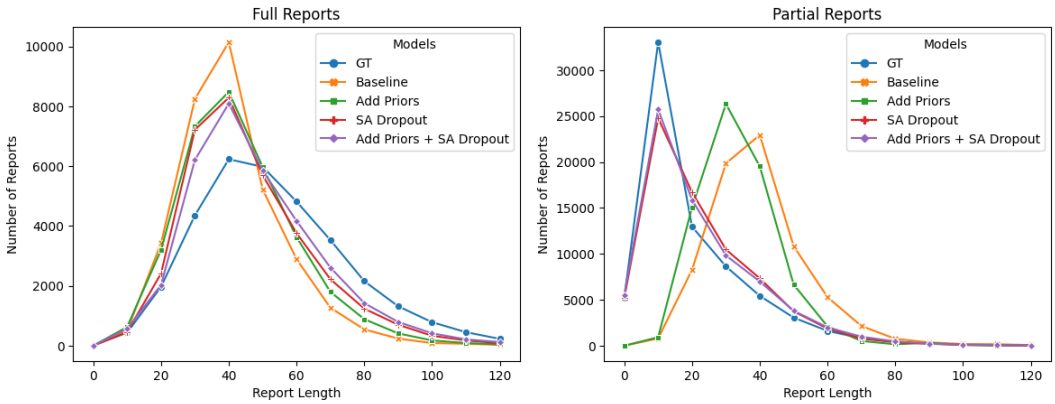


Figure 6.8: Length distribution of the predicted reports compared to the GT reports. The length of a report corresponds to the number of words.

atypical combinations. This could lead to a distributional shift, affecting the model's ability to generalise effectively to unseen combinations. Future work could address this by augmenting the training data with synthetic examples, incorporating combinations of anatomical regions that are absent in the original training set.

6.7 Conclusion

This chapter focuses on two key aspects of CXR automated reporting: *controllability* and *longitudinal CXR representation*. We proposed a simple yet effective solution to align, concatenate and fuse the anatomical representations of two subsequent CXR scans into a joint representation used as the visual input to a language model for ARR. We then proposed a novel training strategy termed *sentence-anatomy dropout*, to supervise the model to link each anatomical region to the corresponding output sentences. This gives the user more control and easier interpretability of the model predictions. We showed the effectiveness of the proposed solution on the MIMIC-CXR dataset where it gives state-of-the-art results. Moreover, we evaluated through extensive ablations how the different components help to generate better reports in different setups: *full report generation*, *partial report generation*, and *Initial vs. Follow-up report generation*.

In future, this method could be integrated with more advanced language models such as

[195] or [148]. These models are not considered in this study due to their higher computational requirements. Further, when considering the patient history, the prior CXR scan might usefully be augmented with other types of imaging and associated radiology reports, clinical notes, clinical letters, and lab results. In future work, our method could be extended to a broader multimodal approach considering more data inputs.

In this chapter, we have used as the report generator a single Transformer encoder-decoder, without adopting the TE-RG framework as presented in Chapter 4 and later adopted in Chapter 5. In such a way, we aim to highlight the specific advantages of the two solutions proposed in this chapter on the model's report generation output. In the following chapter, we combine all the different solutions proposed in this thesis for the ARR task. We perform a final human evaluation of the combined methods to assess the quality of the generated reports.

Chapter 7

Assessing Integrated Automated Reporting Solutions: A Human Evaluation

My contributions to this chapter include conceptualisation, methodological design, data analysis, conducting experiments, evaluation, and writing. Thanks to Dr Chaoyang Wang and Dr Noon Altijani for collaborating on designing the evaluation protocol and conducting the human evaluation.

Chapter Summary

In this chapter, we aim to provide a more in-depth and reliable evaluation of the Automated Radiology Reporting (ARR) solutions presented in previous chapters. We first combine all the solutions developed in previous chapters, relevant to automated reporting: the Triples Extractor-Report Generator framework from Chapter 4, the finding aware anatomical tokens from Chapter 5, and the longitudinal projection module and the sentence-anatomy dropout training strategy from Chapter 6 into a single model, that we call Integrated Model. *We present ablation results,* i.e. *how each solution impacts accuracy, as measured by our NLP and CE metrics. The results demonstrate the cumulative improvement of each solution, leading to the Integrated Model achieving the highest performance on most metrics. We then evaluate the quality of the generated reports from the Integrated Model with the help of two expert human evaluators. The proposed evaluation protocol is designed to measure whether the predicted reports accurately describe the findings (*e.g.*, opacities, cardiomegaly, etc.), the correctness of relevant attributes (e.g., the laterality and temporal evolution), and whether any errors are likely to result in patient harm (*i.e.*, missing infection or giving incorrect clinical advice). Our evaluation shows that the method achieves an F1 score of 0.656, and effectively associates the corresponding attributes. However, the number of errors that could impact patient treatment and lead to harmful outcomes remains significant, with approximately 70% of the reports containing one or more critical errors. In summary, the contributions of this chapter are:*

1. *To integrate the ARR solutions proposed in previous chapters into a unified method, named Integrated Model.*

2. *To propose a detailed protocol for expert human evaluation of generated reports, and present results of human evaluation using this protocol on a subset of reports generated by the Integrated Model.*

7.1 Introduction

The objective of this chapter is to provide a comprehensive and in-depth human evaluation of the Automated Radiology Reporting (ARR) solutions presented in previous chapters.

First, we developed the *Integrated Model* by combining key solutions for ARR. This model extracts finding-aware anatomical tokens (A_{tok}) from current and prior scans (Chapter 5), processes them through the Triples Extraction-Report Generation (TE-RG) pipeline (Chapter 4), incorporates prior scans (Chapter 6), and refines report generation using the sentence-anatomy dropout strategy (Chapter 6).

We proceed to evaluate the quality of the generated reports from the Integrated Model. This evaluation process involves two expert human evaluators. This is to better understand the quality of the generated reports compared to the automatic NLG and CE metrics reported in previous chapters, which present some limitations. NLG metrics evaluate fluency through n-gram matching but overlook semantic meaning and treat all words the same. For instance, sentences with opposite meanings can receive similar scores. On the other hand, CE metrics measure whether the predicted reports correctly mention the 14 labels defined in CheXpert [83]. However, CE metrics are limited since the CheXbert labeller reportedly has an accuracy of only 0.798 F1 [184] and thus we do not expect to measure perfect scores. Moreover, CE metrics do not account for laterality, severity and progression, which are essential attributes for accurate diagnosis and treatment decisions. Alternatively, RadGraph [86] captures errors related to laterality and severity but does not account for progression, a critical factor in tracking the evolution of findings over time. However, RadGraph's correlation with manual evaluations by radiologists remains low, which underscores the need for manual assessment of the predictions to ensure a more accurate and comprehensive evaluation.

To overcome the limitations introduced when using automatic metrics, we propose a human evaluation protocol which aims to measure several key aspects of the generated reports. Firstly, it examines whether the reports accurately describe the medical findings, such as opacities, cardiomegaly, and other significant clinical observations. This aspect is crucial for ensuring

the reports are clinically relevant and useful. Secondly, the protocol assesses the correctness of various attributes related to the findings. This includes checking the accuracy of details such as the laterality (*i.e.*, whether the finding is on the left or right side of the body) and the temporal evolution (*i.e.*, how the finding has changed over time). Accurate description of these attributes is essential for proper clinical decision-making. Lastly, the evaluation protocol is designed to identify any potential errors in the reports that could lead to patient harm. This includes errors that might result in missing critical conditions like infections or providing incorrect clinical advice. The goal is to ensure that the predicted reports are accurate and safe for use in a clinical setting.

Compared to the human evaluation described in Chapter 4 (Section 4.5.3), which involved only comparing the predicted and original reports, the evaluators also had access to the original image. This allowed them to better verify whether something stated in the prediction was a hallucination of the model or omitted from the original report. Moreover, the evaluation protocol for this round of evaluation was more extensive, evaluating not only the errors but also the correct mentions in the predictions. In this way, we can gain a better understanding of the model's behaviour by analysing not only the number of errors it makes but also what it correctly predicts. Finally, we also evaluated additional fine-grained attributes, such as the laterality and the progression of the findings, as we provide specific solutions for these aspects throughout this thesis, such as anatomical tokens (Chapter 5) and the longitudinal projection module (Chapter 6).

Overall, this chapter provides a detailed evaluation of the Integrated Model. Through this comprehensive evaluation, we aim to assess the effectiveness and safety of our Integrated Model in a real-world clinical environment.

7.2 Related Works

Many works in ARR rely solely on automatic metrics to assess the quality of their methods [32, 31, 128, 193]. However, as discussed in Section 2.9.5, these metrics have intrinsic limitations.

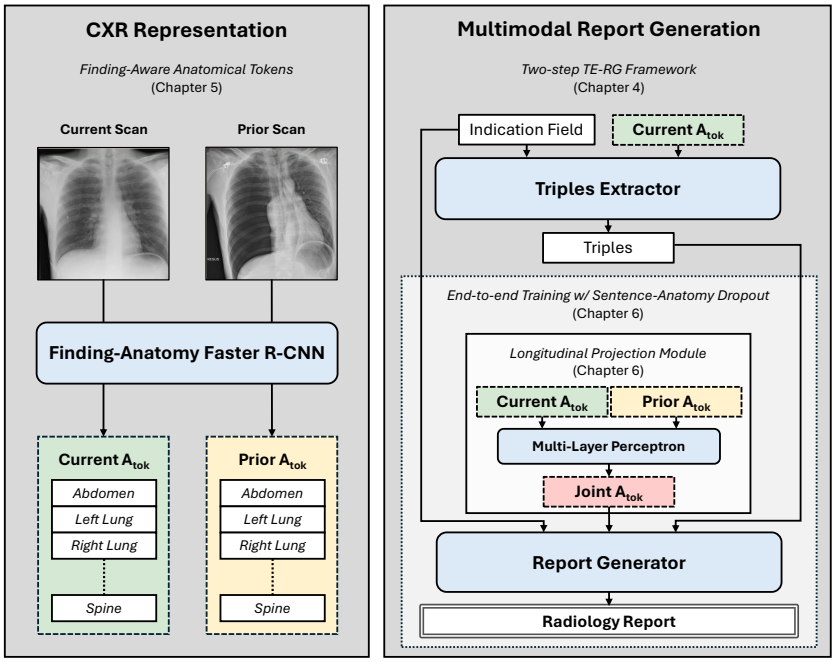


Figure 7.1: High-level diagram of the Integrated Model, in which we combine the different solutions presented in previous chapters. We indicate the chapters where each solution is discussed in more detail.

Some studies have attempted to address these issues by incorporating human evaluations. Miura et al. [137] provide a human evaluation to compare their proposed method with a baseline, by proposing a preference-based human evaluation, where two board-certified radiologists decide which predicted report is factually more similar to the original reference report. This evaluation is intended to provide additional evidence of the improved performance of their solution. However, it does not include comparisons with reports written by radiologists, thereby lacking any assessment of the quality of the predicted reports in real clinical practice.

In [87] the authors conduct a human evaluation on 60 impression sections generated from their proposed method, a baseline method [53] and human-generated. Each report was divided into lines and four radiologists scored the quality of each line based on five possible categories: "no error", "not actionable", "actionable non-urgent error", "urgent error", or "emergent error". For each report, they compute the maximum error severity – the worst error in the report – and the average error severity – the sum of error severity across lines standardised by the number of lines. Their evaluation provides a more detailed score compared to report-level scores but

does not specify the source of errors (e.g., hallucinations, omissions, impressions, etc.).

In [33], the authors propose a high-level human evaluation to compare their method (CheXAgent) against text written by a physician for both the ARR and finding summarization tasks. For both tasks, five radiologists review the same set of 20 randomly selected samples, to compare the generated output text with the original text written by a physician. The radiologists then evaluate these outputs using a five-point scale based on "completeness", "correctness", and "conciseness". Although this evaluation provides an overall score for the quality of the generated text, they do not assess more fine-grained error categories.

Similarly, in [189, 240] the authors assessed the use of GPT-4 [1]in generating impression sections of the report based on the corresponding finding sections from the report [189] and/or CXRs [240]. The assessment involved radiologists comparing human- and AI-generated text on criteria such as "coherence", "factual consistency", "comprehensiveness", and "medical harmfulness", using a 5-point Likert scale. However, impression sections are generally less detailed compared to finding sections, as they provide a concise summary of the radiologist's overall conclusions based on the findings. Given that our method generates the finding section, we define a more detailed human evaluation protocol.

7.3 Integrated Model

First, we developed a comprehensive model, named the *Integrated Model*, that integrates all the key solutions proposed in this thesis related to ARR. The Integrated Model is built by systematically joining various components, each designed to enhance different aspects of the ARR process:

1. **Finding-Aware Anatomical Tokens Extraction:** We begin by extracting specialised tokens, referred to as finding-aware anatomical tokens (A_{tok}), from both the current and prior CXRs. This is achieved using the finding-anatomy Faster R-CNN model described in Chapter 5. These tokens capture the essential anatomical and pathological

information from the CXR and are used as the visual representation for subsequent stages of the model.

2. **Triples Extraction-Report Generation Pipeline:** We then adopt the TE-RG pipeline detailed in Chapter 4. This pipeline leverages the structured information expressed as triples (entity1, relation, entity2) extracted during the TE step, which is then converted by the RG step into coherent and clinically relevant radiology reports.

3. **Longitudinal Data Integration:** To enhance the model's performance, we incorporate prior CXR scans into the report generation process. This is done through the introduction of the *longitudinal projection module* (Chapter 6), which enables the model to assess changes over time, providing a richer context for interpreting the current scan and improving the quality of the reports.

4. **Enhanced Controllability:** Finally, the report generator is refined by implementing the *sentence-anatomy dropout* training strategy (Chapter 6). This strategy improves the model's ability to effectively link anatomical regions with corresponding output sentences, enhancing both user control and the interpretability of the model's predictions.

A high-level diagram of the Integrated Model is presented in Figure 7.1

7.4 Evaluation Protocol

We assess the quality of reports by having two junior physicians compare the reports generated by our Integrated Model with the original (ground truth) reports. The evaluators had respectively 4.5 years (evaluator A) and 2 years (evaluator B) of clinical experience, including experience in reading CXR reports. We randomly selected 100 reports from the MIMIC-CXR test set, with each evaluator assessing a total of 60 reports. Of these, 40 reports were unique to each evaluator, and 20 reports were assessed by both evaluators to calculate the inter-annotator

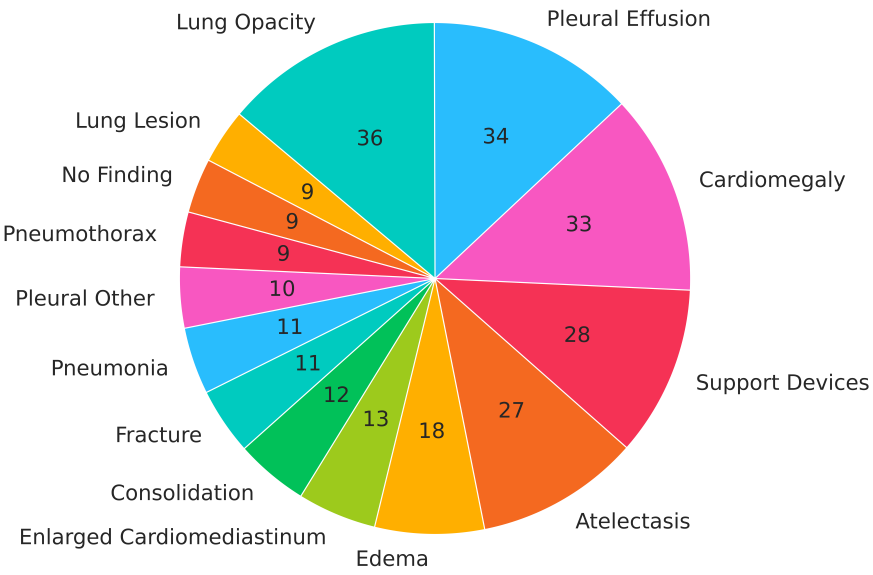


Figure 7.2: Distribution of 14 CheXpert labels in the selected reports for the human evaluation.

agreement. The reports were randomly selected, with the only constraint that they need to cover all 14 CheXpert labels, resulting in the distribution shown in Figure 7.2. [1]

During the evaluation, evaluators compared the predicted and the ground truth reports and assessed the semantical differences between the two. To better assess the correctness of the predicted reports, evaluators had also access to the original CXR, which could be consulted when the model predicted plausible findings. This allowed evaluators to determine if something mentioned in the predicted report was correct but omitted in the ground truth report. Finally, we provided the evaluators with the ground truth report of the prior scan, to better assess the evolution of findings. This is especially crucial in cases where the ground truth reports only state "No changes". Without context, this is ambiguous as it could refer to either persisting findings or an absence of findings. We show an example of the evaluation interface used by the two evaluators in Figure 7.3.

The evaluation protocol was designed to include both errors and correct mentions in the predictions, as well as more fine-grained attributes. The final evaluation schema is organised

[1] Some labels are more frequently represented given that they are more likely to co-occur with other labels.

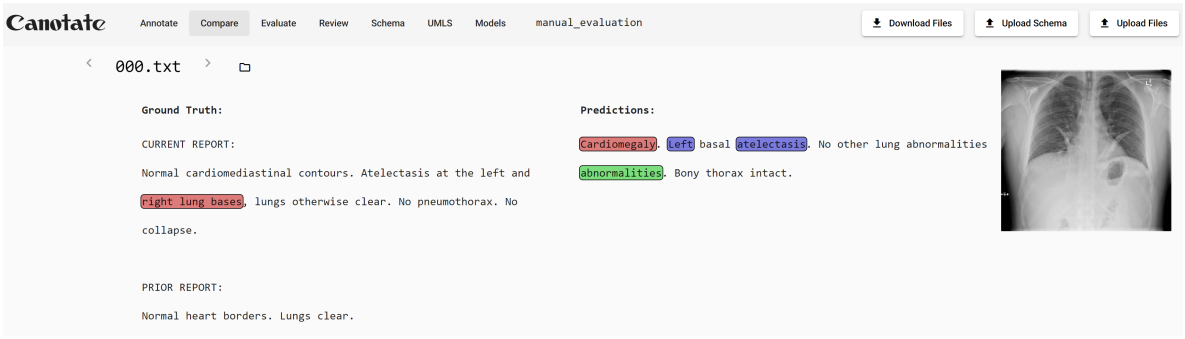


Figure 7.3: Example of the evaluation interface used during the human evaluation, displaying the ground truth current and prior radiology reports, the report predicted by our model, and the frontal CXR of the current study. The interface highlights different types of annotations: red spans indicating factual errors made by the integrated model (e.g., omissions, hallucinations, attribute errors, etc.), blue spans marking correct predictions (e.g., accurate entities, correct attributes, etc.), and green spans identifying grammatical or repetition errors. The evaluation tool depicted in this figure is an existing tool and was not developed specifically for this thesis.

into two levels. Level 1 comprises a high-level set of generic error types and correctness categories. We present the different Level 1 categories, their definitions and examples in Table 7.1. Level 2 covers more specific categorisations and is further divided into a) attributes of anatomical location and disease progression b) whether the error is clinically critical, and c) omissions in the ground truth that is correctly described in the predicted report. We present the different Level 1 categories, their definitions and examples in Table 7.2. We summarise the annotation schema in Table 7.3, where we specify on the same row how Level 1 annotations can be further classified into more specific Level 2 annotations.

7.5 Results

We first present the results of the ablation study using automatic NLG metrics, followed by the detailed human evaluation of generated reports from the Integrated Model only.

7.5.1 Ablation Study

We provide an ablation study to evaluate how the different components introduced in the previous chapters contribute to performing ARR. These components include the *TE-RG*

Category	Definition	Example
hallucination	Model predicts an incorrect finding not included in the ground truth.	GT: *"The lungs are clear."* PRED: *"**Multiple opacities present in the lungs.**"*
omission	Model does not predict a finding included in the ground truth.	GT: *"There are bilateral opacities."* PRED: "***The lungs are clear.***"
impression error	Model predicts the wrong assessment of a finding.	GT: *"Opacity is consistent with atelectasis."* PRED: *"Opacity may represent **infection**."*
attribute error	Model predicts an incorrect attribute associated with a finding.	GT: *"Large pleural effusion."* PRED: *"**Small** pleural effusion."*
repetition	Model repeats a sentence or a word.	PRED: *"Right pleural effusion is moderate in size. **Moderate right pleural effusion.**"*
grammatical error	Model generates a sentence with incorrect structure.	PRED: *"**Signs of remains enlarged.**"*
correct entity	Model predicts a finding included in the ground truth.	GT: *"Low lung volumes."* PRED: "***Lung volumes are low.***"
correct impression	Model predicts the correct assessment of a finding.	GT: *"Consolidation is concerning for pneumonia."* PRED: *"Consolidation present. **Difficult to exclude pneumonia.**"*
correct attribute	Model predicts a correct attribute associated with a finding.	GT: *"The cardiac silhouette is moderately enlarged."* PRED: *"The heart is **moderately** enlarged."*

Table 7.1: Definitions and examples of Level 1 categories in the annotation schema. We provide both ground truth (GT) and the model's prediction (PRED) for categories that require a comparison between the two. Otherwise, only the prediction is included. In the examples, we highlight the text spans corresponding to errors in **red** and the text spans corresponding to correct mentions in **green**.

Category	**Definition**	**Example**
anatomy	Correct/Incorrect mention of anatomical locations.	GT: *"Right atelectasis present."* PRED1: *"Atelectasis present in the **left** lung."* PRED2: *"Atelectasis present in the **right** lung."*
progression	Correct/Incorrect mention of temporal evolution.	GT: *"Improving appearances of the pleural effusion"* PRED1: *"Pleural effusion has **worsened**."* PRED2: "Pleural effusion has ***reduced in size***."
critical error	Errors that could potentially lead to patient harm.	GT: *"Pulmonary edema present."* PRED: *"**No signs of pulmonary edema.**"*
ground truth omission	Model's prediction is correct but not included in the ground truth.	GT: *"Right lobar consolidation"* PRED: *"Right lobar consolidation **most likely represents infection**."*

Table 7.2: Definitions and examples of Level 2 (a,b,c) categories in the annotation schema. We provide ground truth (GT) and the model's prediction (PRED). In the examples, we highlight the text spans corresponding to errors in **red** and the text spans corresponding to correct mentions in **green**.

Level 1	Level 2a	Level 2b	Level 2c
hallucination	-	critical error	-
omission	-	critical error	-
impression error	-	critical error	-
attribute error	anatomy, progression	critical error	-
repetition	-	-	-
grammatical error	-	-	-
correct entity	-	-	ground truth omission
correct impression	-	-	ground truth omission
correct attribute	anatomy, progression	-	ground truth omission

Table 7.3: Annotation schema of the human evaluation divided into 1) generic error types and correctness categories, 2a) finding-specific attributes of laterality and progression, 2b) critical error, and 2c) ground truth omissions. We colour-coded the 1st level based on the annotation types: correct (**green**) and incorrect (**red**). We specify on the same row how Level 1 annotations can be categorised into more specific Level 2 annotations.

pipeline (Chapter 4), the *finding-aware anatomical tokens* (A_{tok}) (Chapter 5), incorporating *prior CXR scans* as input (Priors) (Chapter 6), and *sentence-anatomy dropout* (SA drop) (Chapter 6). The results of this study are presented in Table 7.4[2]. These results show that adopting anatomical tokens (A_{tok}), sentence-anatomy dropout (SA drop), and longitudinal studies (Prior) consistently improve the overall performance. Moreover, When adopting TE-RG framework, there is a substantial improvement compared to the baseline. When this is incorporated into the Integrated Model (Figure 7.1), adopting the TE-RG pipeline results in an increment in the CE-Recall (CE-R) and a decrement in CE-Precision (CE-P). This suggests that the Triples Extraction step encourages the model to include more findings in the predicted report, but this leads to an increase in the number of hallucinations. Based on these results, we select the Integrated Model for the human evaluation as it yields the highest CE-F1 score and ranks highest or second-highest in most NLG metrics.

Method	TE-RG	A_{tok}	SA drop	Prior	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	CE-F1	CE-P	CE-R
Baseline	-	-	-	-	0.401	0.307	0.252	0.214	0.192	0.417	0.451	0.561	0.377
Chapter 4	✓	-	-	-	0.468	0.343	0.271	0.223	0.200	0.390	0.477	0.556	0.418
Chapter 5	✓	✓	-	-	0.490	0.363	0.288	0.237	0.213	0.406	0.537	0.585	0.496
Chapter 6	-	✓	✓	✓	0.486	0.366	**0.295**	**0.246**	0.216	**0.423**	0.553	**0.597**	0.516
Integrated Model	✓	✓	✓	✓	**0.496**	**0.367**	0.290	0.239	**0.221**	0.412	**0.558**	0.570	**0.546**

Table 7.4: Ablation results of all different components proposed in Chapters 4, 5, and 6. These components include the *TE-RG* pipeline, *finding-aware anatomical tokens* (A_{tok}), *prior CXR scans* as input (Prior), and the *sentence-anatomy dropout* (SA drop) strategy.

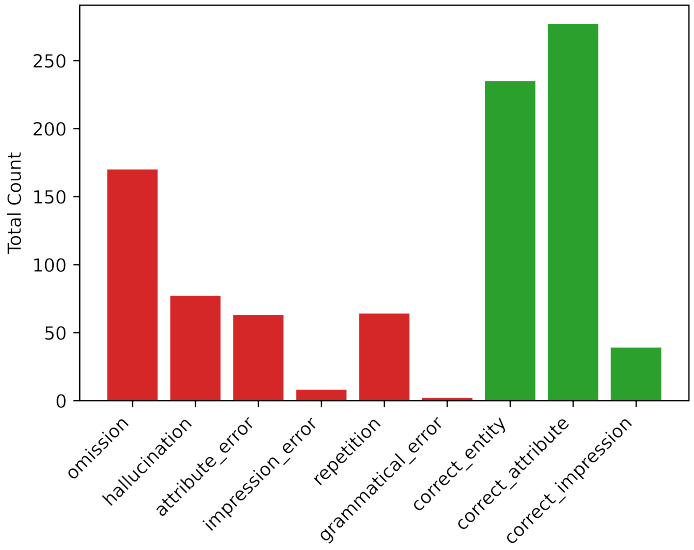


Figure 7.4: Distribution of the Level 1 annotations. We colour-coded the bars based on the annotation types: incorrect (**red**) and correct (**green**).

Precision	Recall	F1
0.753	0.580	0.656

Table 7.5: Precision, Recall and F1 scores between *correct entities* (true positives), *hallucinations* (false positives) and *omission* (false negatives).

7.5.2 Human Evaluation

To gain a deeper insight into our model's behaviour, we present the results of the human evaluation conducted on the 100 reports in Figures 7.4 and 7.5. For the 20 reports evaluated by both evaluators, we report only the annotations from evaluator A, who is the most experienced.

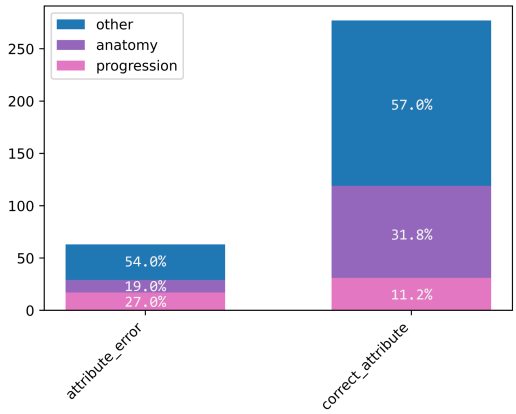
[2]Some of the results shown in this table are different to those presented in the previous chapters, as they were computed using different dataset splits; CE metrics were computed using CheXpert [83] instead of CheXbert [184]; and different image resolutions. In this table, we keep the training/validation/test split consistent with the one proposed in Chest ImaGenome; we adopt CheXbert to compute the CE metrics; and an image resolution set to 512×512 pixels.

Figure 7.4 highlights the total count of Level 1 annotations across all 100 reports, revealing many omissions and hallucinations. To quantify this further, we compute Precision, Recall and F1 scores between *correct entities* (true positives), *hallucinations* (false positives) and *omissions* (false negatives), presented in table 7.5. Although the predicted reports capture most of the findings, the number of mistakes is notable. Moreover, as shown in Figure 7.5b, the majority of errors are not critical. However, there is still a significant number of critical errors, averaging about 1.1 critical errors per report and around 70% of the reports contain one or more critical errors. This frequency of critical errors makes our method unsuitable for clinical practice.

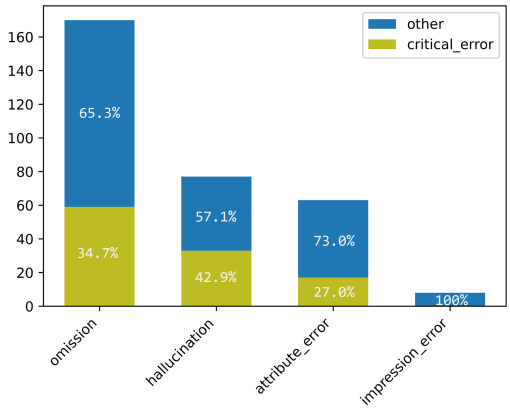
A positive outcome shown by this evaluation is that the attributes described in the reports, particularly the progression and laterality, are mostly correct (Figure 7.5a). This indicates that our method effectively assesses the visual appearance and evolution of findings. This further validates the effectiveness of the anatomical tokens and the longitudinal representation module introduced in Chapter 5 and 6, respectively.

Moreover, the results illustrated in Figure 7.5c demonstrate that our method can occasionally identify additional findings not described in the original reports, as indicated by the extra mentions in the predictions (ground truth omission). These discrepancies might arise from inconsistencies in radiologists' reporting practices, where certain findings are not consistently documented. We suspect these inconsistencies may also contribute to some omissions, as the model has learned to not always include certain findings.

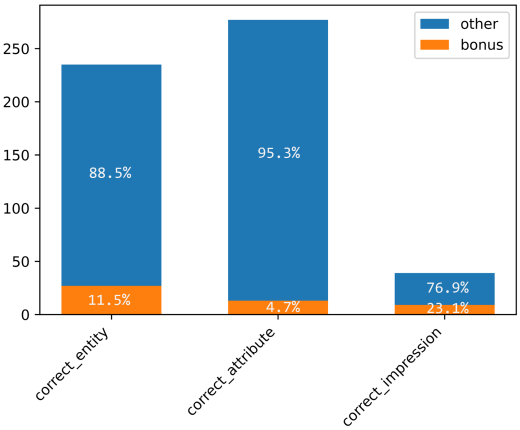
Finally, in Figure 7.6 we show the distribution of the Level 1 annotations made by evaluators A and B on the 20 reports assessed by both, showing consistent results. Across the reports, 68.8% of the text spans annotated by both evaluators overlapped, indicating moderate agreement on the relevant spans of the text. Among these overlapping spans, 83.9% received identical annotations, reflecting a good level of consensus on how to label these sections. Overall, these findings indicate a moderate level of agreement between the two evaluators.



(a) Level 2a: number of *anatomy* and *progression* attributes compared to *other* types of attributes (*e.g.* severity) for attribute error and correct attribute categories.



(b) Level 2b: *critical errors* compared to *other* errors (*i.e.,* non-critical) for omission, hallucination, attribute error and impression error categories.



(c) Level 2c: *ground truth omissions* compared to *other* correct mentions for correct entity, correct impression and correct attribute categories.

Figure 7.5: Distribution of Level 2 annotations. We present the total count and the frequency (%) for each annotation category.

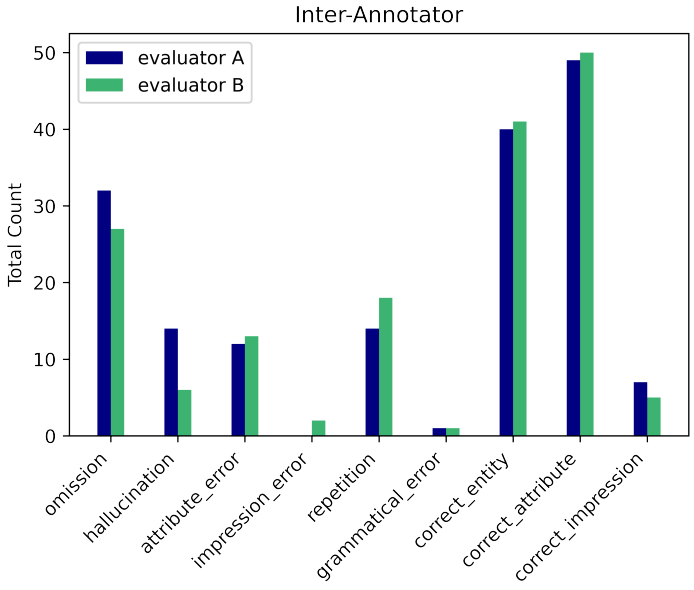


Figure 7.6: Distribution of the annotation of evaluator A and evaluator B on the 20 overlapping reports.

7.6 Limitations

This human evaluation is limited in scope, as only a small percentage of the test set was considered. Despite efforts to select a diverse set of reports based on the 14 CheXpert labels, some rarer findings not covered by these labels are likely not included in the reports selected for this evaluation.

Moreover, some of the categories defined in our evaluation schema could be further divided into more fine-grained subcategories. For example, in this evaluation, we do not disambiguate between different finding categories (e.g., *Cardiomegaly*, *Pleural Effusion*, etc.) when counting errors and correct mentions. This is partly because we already compute precision, recall and F1 score between these categories using CheXbert, and also to reduce the annotation effort for the two evaluators. This evaluation is specifically designed to assess some of the key aspects addressed in this thesis, including laterality and temporal evolution, as well as the clinical utility of our method by measuring the critical errors.

7.7 Conclusion

This evaluation aims to analyse the behaviour of the model integrated with all the solutions presented in Chapters 4, 5, and 6. We designed the protocol to include generic error types and correct mentions, to measure the overall performance of our method. Some of these high-level categories were further subdivided into more specific attributes and labelled as critical errors when a mistake could result in harmful outcomes for the patient.

With the help of two junior physicians, we have then evaluated a total of 100 unique reports predicted by our model. We observe that our method shows promising results, accurately identifying the medical findings (showing an F1-score equal to 0.656) and the corresponding attributes. However, the frequency of critical errors remains concerning (approximately 1.1 per report and 70% of the reports containing at least one), which makes our method unsuitable for clinical practice. In addition, we note in some cases that our method correctly identifies some findings not included in the original reports. We suspect that both these additional findings and some omissions might be due to inconsistencies in radiologists' reporting practices, which lead our model to omit some of the findings from the predicted report. Having multiple ground truth reports for the same images would allow for higher consistency in the training data and better quantification of human error. Unfortunately, this limitation is primarily due to a lack of resources, such as the time and cost required from clinical experts to generate multiple reports.

These results indicate that future efforts must focus on reducing the number of critical errors to enable the safe use of ARR in real-life clinical applications.

Chapter 8

Grounding CXR Visual Question Answering with Radiology Reports

My contributions to this chapter include conceptualisation, methodological design, technical implementation, data analysis, conducting experiments, evaluation, and writing.

Chapter Summary

In this chapter, we focus on the CXR Visual Question Answering (VQA) task, addressing both questions referring to individual CXRs ("What abnormalities are seen in image X?"*) and those concerning differences between two CXRs acquired at different times (*"What are the differences between image X and Y?"*). We further explore how the integration of radiology reports can enhance the performance of VQA models. While previous approaches have demonstrated the utility of radiology reports during the pre-training phase, we extend this idea by showing that the reports can also be leveraged as additional input to improve the VQA model's predicted answers. First, we propose a unified method that handles both types of questions and auto-regressively generates the answers. For single-image questions, the model is provided with the single CXR that the question refers to. For image-difference questions, the model is provided with two CXRs from the same patient, captured at different time points, enabling the model to detect and describe temporal changes. For this purpose, we leverage two solutions we previously proposed for the automated radiology reporting task: the finding-aware anatomical tokens (Chapter 5) and the longitudinal projection module (Chapter 6). Furthermore, in line with the current trend in QA and VQA, where the generation of answers is supported by relevant text explanations, known as Chain-of-Thought reasoning, we demonstrate how a similar approach can be applied to improve the CXR VQA task. This involves grounding the answer generator module with the radiology report predicted from the same CXR. This is achieved by dividing the VQA model into two steps: Report Generation (RG) and Answer Generation (AG). Our contributions in this chapter are to:*

1. *Propose a unified approach for single-image and image-difference CXR VQA.*
2. *Demonstrate the effectiveness of automatically generating a radiology report for the CXR and using this as additional input to improve the answer generation.*
3. *Show state-of-the-art performance of the proposed method on the Medical-Diff-VQA dataset [71, 72].*

8.1 Introduction

This chapter focuses on Visual Question Answering (VQA) applied to Chest X-Ray (CXR) radiology. This task is challenging because of the subtle differences in the appearance of normal and abnormal findings, which can be difficult to detect even for experienced radiologists. Additionally, the thoracic cavity contains various overlapping structures (e.g., ribs, heart, and lungs), making it challenging to isolate and identify specific abnormalities. We address questions based on a single image (*e.g.*, "Is there any sign of pneumonia in the given scan?") to assist radiologists in making precise and rapid diagnoses. Additionally, we explore questions that compare scans of the same patient taken at different time points (*e.g.*, "What has changed compared to the prior scan?"), which are essential for monitoring disease progression or treatment response. We refer to the second as *image difference question answering*. This task has received limited attention in both the general [156, 222] and medical [72, 35] domains, and is particularly relevant in the medical domain, where radiologists often compare scans from different timepoints to assess the progression of findings.

Unlike ARR and medical finding classification models, which are designed to perform a comprehensive analysis of the full input scan—identifying all present findings or generating a detailed radiology report describing the scan's full visual appearance—the VQA task is highly dependent on the specific question posed, directing the model to focus on particular aspects of the scan. Moreover, VQA can be open-ended, requiring the generation of free-form answers, offering the most flexibility; closed-ended, requiring the generation of short specific responses; or multiple-choice, where the model selects the correct answer from a set of predefined options. Additionally, questions may refer to a single image or comparisons between two images to identify differences.

To effectively handle diverse VQA scenarios, we propose a flexible vision-language model that processes dual visual and textual information inputs, tailoring its input configuration based on the specific VQA task we want to perform. We do so by integrating and adapting solutions proposed in previous chapters, such as *anatomy-finding anatomical tokens* (Chapter 5) and

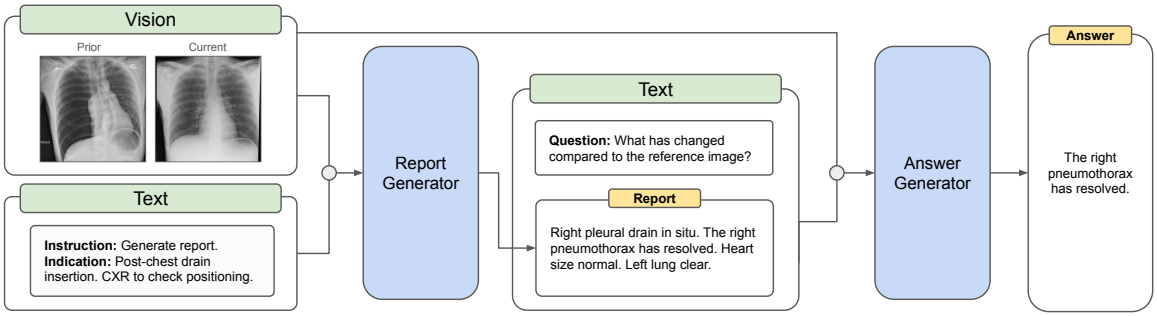


Figure 8.1: Overview of Report Generator-Answer Generator (RG-AG) pipeline. The answer generation is grounded using the predicted radiology report from the same CXR.

the *longitudinal projection module* (Chapter 6). Moreover, we study the effect of grounding the answer generation of the VQA model using the radiology report predicted from the same image. This is similar to Chain-Of-Thought (CoT) reasoning [213, 102], where the answer inference is improved by having a language model first generate the reasoning process to arrive at a specific answer. While this has been widely studied in language-only QA [174, 54, 223, 234], only a few works have studied its effect on VLMs [136, 232]. We argue that a similar solution could be applied to medical VQA, by having the model generate a full description of the appearance of a CXR, in the form of radiology reports. To the best of our knowledge, this is the first study to comprehensively examine the impact of grounding radiology reports for medical VQA and to provide evidence of its effectiveness. Additionally, we will explore how this affects the differential VQA task.

8.2 Related Works

8.2.1 Medical Visual Question Answering

Until recently, the small size of medical VQA datasets [125, 105] often led researchers to approach medical VQA as a classification task [140, 98, 48], due to the limited number of possible answers in these datasets. However, treating VQA as a classification task limits the solution to a predetermined set of answers, whilst treating VQA as a text generation

task naturally yields more detailed and wide-ranging responses expressed in the form of one or multiple sentences. Generative approaches [168, 181, 197] have been enabled by the availability of open-access datasets [230, 70] and the rise of generative large language models [195, 148]. For instance, some authors [168, 181] have proposed medical VQA methods that perform both classification and answer generation by designing their model with a specific head depending on the type of answers, whether closed-ended or open-ended. More recently, van Sonsbeek et al. [197] proposed a generative method, based on GPT-2 [159], and fine-tuned it on three medical VQA datasets [125, 66, 76].

8.2.2 Medical Image Difference Question Answering

Image difference question answering, which we refer to as *diff-VQA*, is the task where questions refer to the differences between two or more images. This has received minimal attention in the medical domain due to the lack of a suitable dataset until Hu et al. [72] collected a CXR VQA dataset including these types of questions. To the best of our knowledge, only two works [72, 35] have tackled diff-VQA. In [72], they propose a method that utilises anatomical features and a multi-relationship image-difference graph feature representation learning method to extract image-difference features. Subsequently, Cho et al. [35] have focused on adopting a pre-trained vision-language model [206] and propose an effective pretraining pipeline (PLURAL) to perform the medical image difference question answering task. Their method is pre-trained on the ARR task, but they do not utilise the predicted radiology reports in the VQA task.

8.2.3 Grounding CXR-VQA with Radiology Reports

To our knowledge, the use of predicted radiology reports to enhance VQA performance has not been explored in the literature. The most relevant work is by [207], who proposed a method that integrates the output of Computer-Aided Diagnosis (CAD) networks with a Large Language Model (LLM) to leverage the LLMs' medical domain knowledge and logical reasoning. However, their approach focuses on leveraging LLMs to improve the interactivity

of a CAD network. We instead focus on demonstrating how CAD networks can enhance the VQA performance of a VLM. Our work addresses this gap by applying a similar approach to CoT reasoning, as introduced by [232] in the general domain.

8.3 Method

In this chapter, we present a flexible generative VQA method that avoids the requirement for specific heads (classification and generation heads) for different question types (closed and open-ended questions). We present the RG-AG solution (Figure 8.1), named after its two main components: the *Report Generator* (RG) and the *Answer Generator* (AG) model. Our approach incorporates successful strategies from previous works on the diff-VQA task, such as anatomical feature representations [72] and a pre-training strategy based on radiology report generation [35]. We enhance our solution by using predicted radiology reports from each CXR to ground the generation of the answers, similar to CoT reasoning methods in the general domain. In Figure 8.1, we present an overview of our approach.

8.3.1 CXR Anatomical Tokens

We begin by extracting the CXR features in the form of finding-aware anatomical tokens, as introduced previously (see Chapters 5 and 6). These comprise vector representations corresponding to each anatomical region in the CXR, trained to contain information about the findings within each region, which serve as the input visual representations for both the RG and VQA modules.

To extract tokens, we train a Faster R-CNN model [169] to perform two tasks: (1) *anatomical region localisation*—detecting the bounding box of $N = 36$ anatomical regions; and (2) *finding detection*—determining the presence or absence of 71 findings within each region. For each CXR, we then define *finding-aware anatomical tokens* as the feature vectors extracted from the Region of Interest pooling layer of the Faster R-CNN, selecting the bounding

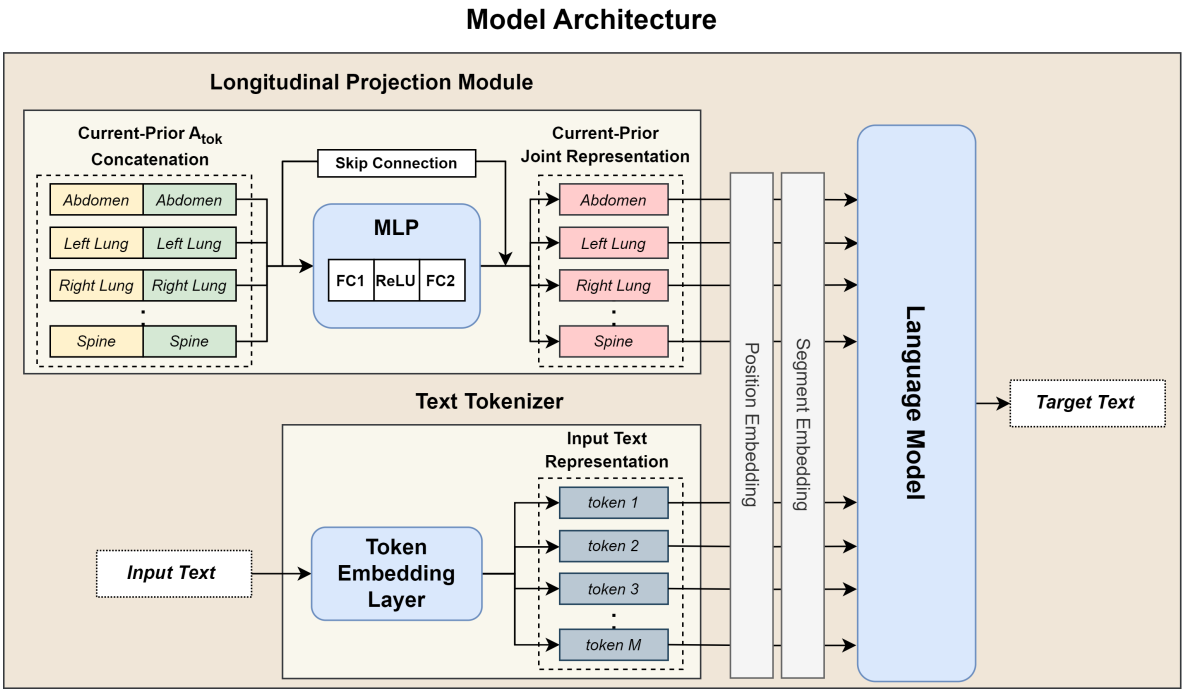


Figure 8.2: VLM architecture of the RG and AG model.

box representation with the highest confidence score for each anatomical location. This results in N vectors $V = \{\vec{v}_n\}_{n=1}^{N}$ with $\vec{v}_n \in \mathbb{R}^d$ and d=1024. If an anatomical region i is not detected in a CXR, the corresponding token $\vec{v}_i$ is a zero vector. For more details about the training and the model architecture, we refer back to Chapter 5.

8.3.2 Model Architecture

The RG and AG use the same VLM architecture, composed of the *Longitudinal Projection Module* (LPM) and a Transformer-based language model. This architecture has a total of 68M trainable parameters. The VLM is defined as a function f which takes visual features (V) and text (T) as inputs, and generates output text (Y):

$$Y = f(V, T), \tag{8.1}$$

which is achieved in an autoregressive manner. The input and output texts vary based on the specific task (RG or AG). The VLM architecture of the RG and AG models are shown in

Figure 8.2.

Longitudinal Projection Module

The LPM is responsible for projecting the current and prior CXR scans into a joint representation. The input visual components of the LPM correspond to the finding-aware anatomical tokens from both the current and – depending on the task – the prior scan. These are denoted as $V_{current} = \{\vec{v}_{c,n}\}_{n=1}^{N}$ and $V_{prior} = \{\vec{v}_{p,n}\}_{n=1}^{N}$, respectively. Whenever we do not intend to use the prior scan as input, we set $V_{prior} = \{\vec{0}\}_{n=1}^{N}$.

We first concatenate the anatomical tokens of the current and prior scans assigned to the same anatomical region ($[\vec{v}_{c,n}, \vec{v}_{p,n}]$). We then pass them through the Multi-Layer Perceptron (MLP), which consists of a stack of a Fully-Connected layer (FC1), a ReLU function and another Fully-Connected layer (FC2). Differently from our previous implementation for ARR (see Chapter 6), we also include a residual connection. This gives us

$$\vec{v}_{joint,n} = MLP([\vec{v}_{c,n}, \vec{v}_{p,n}]) + [\vec{v}_{c,n}, \vec{v}_{p,n}] \tag{8.2}$$

and we refer to the output of the MLP as the current-prior joint representation $V_{joint} = \{\vec{v}_{joint,n}\}_{n=1}^{N}$.

Language Model

The Language Model (LM) consists of a vanilla encoder-decoder Transformer [198]. Both the encoder and the decoder are composed of 3 attention layers with 8 heads and 512 hidden units. The LM takes as input the current-prior joint representation V_{joint} and the tokenised input text embeddings T and generates the target text Y:

$$Y = LM(V_{joint}, T). \tag{8.3}$$

The input visual V_{joint} and text embeddings T are concatenated, and summed to the *position*

embeddings – which establish the position of each token within the input sequence. The *segment embedding* is used to enable the model to distinguish between the modality of each set of input tokens: vision *vs.* text.

While the RG and AG models share the same architecture design, the input and output components are different, as described in the following sections.

8.3.3 Report Generator

The Report Generator (RG) consists of the LPM and the LM. The visual input comprises the current ($V_{current}$) and prior anatomical tokens (V_{prior}), when available. The target text Y is *either* the finding (F) *or* the impression (I) section of a radiology report, and the text input T comprises not only the indication field ($\{ind\}$) but also an instruction specifying the section to be generated: (1) $Inst_f =$ *"generate the finding section"* or (2) $Inst_i =$ *"generate the impression section"*. Following Eq. 8.1, the RG can be defined as a function f_{RG}:

$$\begin{aligned} F &= f_{RG}(V = [V_{current}, V_{prior}], T = \text{“}\texttt{[ARR]}\ \{ind\}\ \texttt{[Q]}\ Inst_f\text{”}), \\ I &= f_{RG}(V = [V_{current}, V_{prior}], T = \text{“}\texttt{[ARR]}\ \{ind\}\ \texttt{[Q]}\ Inst_i\text{”}). \end{aligned} \tag{8.4}$$

where `[ARR]` corresponds to a special token indicating the ARR task, and `[Q]` is the special token used in front of the instruction.

8.3.4 Answer Generator

The Answer Generator (AG) is responsible for performing VQA. Similar to the RG, it consists of the LPM and the LM, and is initialised using the pre-trained RG weights. The AG model can be defined as a function $f_{AG}(V,T)$, whose input visual features (V) and input text (T) vary depending on the question type.

When the question asks about the difference in appearance between two scans, the AG takes as input two sets of anatomical tokens (*current* and *prior*). Otherwise, the AG takes only

one set of anatomical tokens (considered to be *current*) and sets $V_{prior} = \varnothing = \{\vec{0}\}_{n=1}^{N}$. This results in the visual component being:

$$V = \begin{cases} (V_{current}, V_{prior}), & \text{if diff-VQA;} \\ (V_{current}, \varnothing), & \text{otherwise.} \end{cases} \tag{8.5}$$

Additionally, the input text T varies depending on whether the question $\{q\}$ is open-ended or closed (multiple choice):

$$T = \begin{cases} \text{“[OE\_VQA] } \{rr\} \text{ [Q] } \{q\}\text{”}, & \text{if } \textit{\{q\}} \text{ is open-ended} \\ \text{“[MC\_VQA] } \{rr\} \text{ [Q] } \{q\} \text{ [MC] } \{a_1\} \ \ldots \ \text{[MC] } \{a_M\}\text{”}, & \text{if } \textit{\{q\}} \text{ is multiple-choice.} \end{cases} \tag{8.6}$$

where $\{rr\}$ refers to the predicted radiology report for the given scan. We define `[OE_VQA]` and `[MC_VQA]` as special tokens, used to specify the task as open-ended VQA or multiple-choice VQA to the model. `[MC]` is a special token placed before each possible answer $\{a_j\}$, from which the model has to pick.

8.4 Experimental Setup

8.4.1 Datasets

We conduct our experiment on the publicly available Medical-Diff-VQA dataset [71, 72], which is derived from MIMIC-CXR [91, 92, 57]. The dataset contains a total of 700,703 question-answer pairs (QA) related to 109,923 pairs of current and prior CXRs. We use the official split, the number of QA pairs and CXR pairs for each data split are presented in Table 8.1. The dataset is divided into training, validation, and testing sets in an 8:1:1 ratio at the

Split	QA pairs	CXR pairs
Training	560,563	88,098
Validation	70,070	10,864
Test	70,070	10,963

Table 8.1: Number of QA pairs and CXR pairs for each data split (training/validation/test) in the Medical-Diff-VQA dataset [71, 72].

study level, ensuring that studies from the same patient appear in only one split to prevent data contamination. To ensure the availability of a second image for differential comparison, only patients from the MIMIC-CXR dataset with more than one prior radiology visit were included.

The questions in Medical-Diff-VQA are categorised into seven types:

1. **abnormality** (*"What abnormalities are seen in the image?"*)
2. **location** (*"Where in the image is the <abnormality> located?"*)
3. **type** (*"What type is the <abnormality>?"*)
4. **level** (*"What level is the <abnormality>?"*)
5. **view** (*"Which view is this image taken?"*)
6. **presence** (*"Is there any evidence of <abnormality>?"*)
7. **difference** (*"What has changed compared to the reference image?"*)

The *difference* questions, which we refer to as diff-VQA, ask about differences in appearance between the current and a prior scan. In accordance with previous works using this dataset [72, 35], we classify closed questions as those with answers limited to "yes" or "no", treating them as multiple choice as detailed in Section 8.3.4. The remaining questions are considered open-ended, with free-form answers.

We use the MIMIC-CXR dataset to train the RG model and the Chest ImaGenome annotation [214] to extract anatomical tokens, following the data split indicated in Medical-Diff-VQA at all stages.

8.4.2 Implementation Details

The Report Generator is initialised with random weights and is trained end-to-end for 100 epochs using a cross-entropy loss and Adam optimiser [100]. We set the initial learning rate to 1×10^{-4} and reduce it every 10 epochs by a factor of 0.8. The RG is trained to predict both the finding and impression sections as detailed in Section 8.3.3. The best-performing model is selected based on the highest BLEU-4 score computed across both the finding and impression sections of the validation set.

The Answer Generator is initialised using the RG weights and is fine-tuned for 100 epochs using the same loss, optimiser and learning rate as the RG. We select the best model based on the highest BLEU-4 score computed across all questions of the validation set.

Each experiment is repeated three times using different random seeds and we report the average in our results.

8.4.3 Metrics

We adopt different metrics based on the type of question, in line with previous studies [72, 35]. For "difference" type questions, we report natural language generation metrics including BLEU [150], METEOR [14], ROUGE [122], and CIDEr [199]. We calculate exact-match accuracy for other types of questions, differentiating between open-ended and closed (yes/no) questions.

8.4.4 Baselines

For the diff-VQA task, we compare our method against two existing approaches that focused on the general image difference captioning task: MCCFormers [156] and IDCPCL [222]. Furthermore, we compare our method against the only two other approaches that have specifically targeted the diff-VQA task: EKAID [72] and PLURAL [35].

For all other question types, we compare our method against state-of-the-art medical VQA methods which were previously evaluated on the Medical-Diff-VQA dataset, such as MMQ

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MCCFormers [156]	0.214	0.190	0.170	0.153	0.319	0.340	0
IDCPCL [222]	0.614	0.541	0.474	0.414	0.303	0.582	0.703
EKAID [72]	0.628	0.553	0.491	0.434	0.339	0.577	1.027
PLURAL [35]	0.704	0.633	0.575	0.520	0.381	0.653	1.832
AG (w/o report)	0.678	0.619	0.569	0.525	0.372	0.659	2.102
RG-AG (w/ report)	**0.711**	**0.650**	**0.600**	**0.551**	**0.384**	**0.668**	**2.198**

Table 8.2: Comparison results between our proposed approach, both with the report generation step (RG-AG) and without it (AG), and previous methods on the *difference* questions of the MIMIC-diff-VQA dataset [72]. We show the **best results in bold**. All the results of the comparison methods are taken from [35].

Model	Open Question	Closed Question	All Questions
MMQ [48]	0.115	0.108	0.115
EKAID [72]	0.264	0.799	0.525
PLURAL [35]	0.512	**0.873**	0.688
AG (w/o report)	0.509	0.865	0.683
RG-AG (w/ report)	**0.523**	0.871	**0.693**

Table 8.3: Comparison results between our proposed approach, both with the report generation step (RG-AG) and without it (AG), and previous methods on all but the *difference* questions of MIMIC-diff-VQA dataset [72]. We compute the accuracy (exact match) on the open-ended and the closed-ended questions (yes/no). We show the **best results in bold**. All the results of the comparison methods are taken from [35].

[48], EKAID [72] and PLURAL [35].

8.5 Results

8.5.1 VQA Reults: Difference & Non-Difference

We present the diff-VQA results in Table 8.2 and the non-diff-VQA results in Table 8.3, comparing our method with and without the report generation step (RG-AG and AG, respectively) as well as with other state-of-the-art approaches. Our RG-AG method achieves state-of-the-art performance, demonstrating superior results on all NLG metrics for *difference* type questions and improved overall accuracy on the remaining questions. These results suggest that using the pre-trained model on ARR not only to initialise the VQA model, as done in PLURAL [35] and AG, but also to predict the reports and use them to ground the answer generation

Visual	**Text**	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	Acc
C	-	0.686	0.520	0.373	0.634	1.854	0.668
C + P	-	0.678	0.525	0.372	0.659	2.102	0.683
C + P	I	0.679	0.523	0.369	0.654	2.111	0.690
C + P	F	0.690	0.533	0.374	0.655	2.117	0.691
-	F + I	0.633	0.479	0.344	0.595	1.734	0.630
C + P	F + I	**0.711**	**0.551**	**0.385**	**0.668**	**2.198**	**0.693**
C + P	*F + I (Ground Truth)*	*0.723*	*0.570*	*0.398*	*0.685*	*2.484*	*0.751*

Table 8.4: Ablation results. We test various visual inputs to the Answer Generation (AG) model: the current scan only (C), both current and prior scans (C + P), and no scan (-). Additionally, we test different textual inputs provided alongside the question: the findings section (F), the impression section (I), both sections combined (F + I), and no additional input text (-). In the final row, we present results when the AG model is given the ground truth findings and impression sections. This serves as an upper bound for performance, excluded from direct comparison. We show the **best results in bold**.

step helps generate more precise answers, especially for *difference* type questions. For other question types, the RG-AG model consistently outperforms AG across all metrics. Furthermore, compared to PLURAL, our RG-AG method achieves higher accuracy for open-ended questions but lower accuracy for closed-ended questions (yes/no). This suggests that the intermediate report generation step has a more pronounced impact on open-ended questions, while its influence on closed-ended questions is comparatively limited.

8.5.2 Ablation Study

In this ablation study, we explore the impact of various input components provided to the AG model. The quantitative results, which are detailed in Table 8.4, illustrate the positive effect of grounding the AG model with relevant sections of radiology reports. By breaking down the components individually, we observe that the Finding section (F) has the most substantial impact on the model's performance when compared to the Impression section (I). This outcome is as expected, as the Finding section typically contains more granular and detailed information about the CXR, offering richer data for the model to process. However, the study also reveals that when both sections are provided in conjunction (F + I), the VQA results show improvement beyond what is achieved with either section alone. This synergy

suggests that the combined information from both sections provides a more comprehensive context to the AG model, enhancing its predictions.

Additionally, we examined the scenario where the AG model is provided exclusively with textual information from the report, omitting any visual input from the CXR, to determine whether the generated report alone suffices for accurate answer prediction. The results, however, indicate a noticeable decrement in performance across all metrics, underscoring that the textual reports alone are insufficient. This can be attributed to several factors: predicted reports may lack certain critical details, contain inaccuracies, or even present contradictory information. These issues highlight that the visual data from the CXR is still necessary for the AG model to generate accurate and reliable answers.

Furthermore, we extended our investigation to assess whether enhancing the quality of the textual input could further boost the model's performance. To this end, we incorporated the original radiology reports from the MIMIC-CXR dataset, which were manually written by expert radiologists, as the auxiliary textual input instead of the predicted radiology reports. As demonstrated in Table 8.4, providing the AG model with high-quality curated reports led to improvements across all metrics. This finding underscores the pivotal role of report quality on VQA.

In Figure 8.3, we present the accuracy across different question types, excluding *difference* questions. Our proposed RG-AG model is compared with the baseline model, which does not incorporate the predicted CXR radiology report as input. The results indicate that not all question types are affected equally when we provide the model with the predicted reports as additional context. Notably, questions related to *location*, *type*, and *level* show the greatest improvement from this additional input.

Finally, we present qualitative results in 8.4 to provide a more comprehensive understanding of the model's behaviour. This figure compares the outputs of a baseline model—trained without the predicted CXR radiology report—with those of our proposed *RG-AG* model. In these results, we highlight the segments in the predicted report that correctly contain the

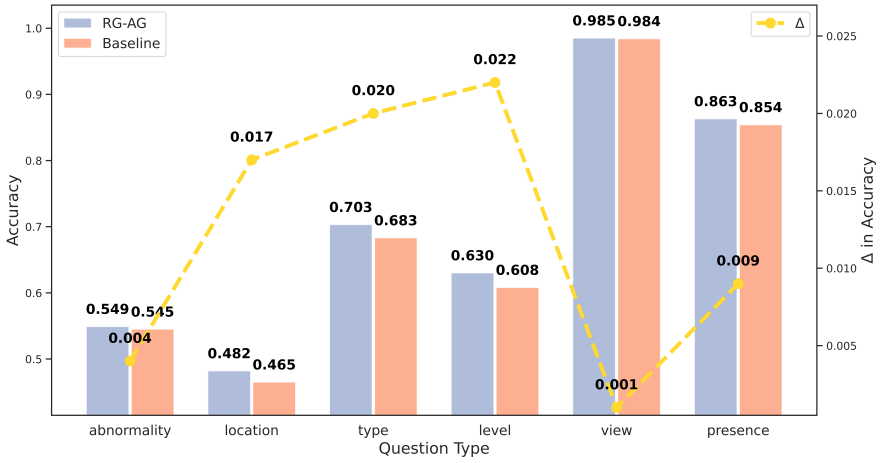


Figure 8.3: We compare the accuracy of our proposed RG-AG model with the baseline model (which does not include the predicted CXR radiology report as input) for each question type, except for the *difference* questions. We highlight the difference in accuracy (Δ) for each question type.

information needed to predict the correct answer. These results underscore the significance of radiology reports in providing additional evidence for VQA, further validating the importance of this input in enhancing model performance.

8.6 Limitations

This work applies a two-stage approach, following the work on multimodal CoT reasoning [232]. Despite showing improved performance on the Medical-Diff-VQA dataset, our RG-AG approach is prone to error propagation between the first and second stage of our pipeline, *i.e.* the predicted reports provided to the AG model might contain errors which can lead to the generation of wrong answers.

Furthermore, we generated the *findings* and *impression* sections of the report separately, following the methodology proposed in [35]. However, this approach can result in inconsistencies between the two sections, potentially leading to incorrect answer predictions (*e.g.*, discrepancies in the laterality or severity of a finding described in each section). A more effective strategy might involve generating the *impression* section based on the predicted *findings*, as is done in the task of summarising radiology reports [43].

Question & Ground Truth Answer	Predicted Answer w/o Rad Report	Predicted Answer w/ Rad Report
Q: "What has changed compared to the reference image?" **A:** "The main image has additional findings of pneumonia, edema, and lung opacity than the reference image. The main image is missing the finding of cardiomegaly than the reference image."	**A:** "The main image has additional findings of lung opacity, and pneumonia than the reference image. The main image is missing the finding of atelectasis than the reference image."	**R:** "*FINDING: Left subclavian central venous catheter tip terminates in the mid svc. **The heart size is normal. mediastinal contours are unremarkable. There is mild pulmonary edema**, new from the prior study. **Patchy bibasilar airspace opacities could reflect atelectasis but infection or aspiration is not excluded**. No large pleural effusion or pneumothorax is seen. IMPRESSION: **Bibasilar airspace opacities could reflect aspiration or infection**. Small bilateral pleural effusions."* **A:** "The main image has additional findings of pneumonia, edema, and lung opacity than the reference image. The main image is missing the finding of cardiomegaly than the reference image."
Q: "Where in the image is the lung opacity located?" **A:** "Bibasilar area"	**A:** "Bilateral area"	**R:** "*FINDING: Left subclavian central venous catheter tip terminates in the mid svc. The heart size is normal. Mediastinal contours are unremarkable. There is mild pulmonary edema, new from the prior study. **Patchy bibasilar airspace opacities** could reflect atelectasis but infection or aspiration is not excluded. No large pleural effusion or pneumothorax is seen. IMPRESSION: **Bibasilar airspace opacities** could reflect aspiration or infection. Small bilateral pleural effusions.*" **A:** "Bibasilar area"
Q: "What abnormalities are seen in this image?" **A:** "Pleural effusion, atelectasis, cardiomegaly"	**A:** "Atelectasis, lung opacity, pleural effusion"	**R:** "*FINDING: As compared to the previous radiograph, there is no relevant change. The monitoring and support devices are in constant position. **Unchanged moderate cardiomegaly with retrocardiac atelectasis and small bilateral pleural effusions**. No newly appeared parenchymal opacities. No pneumothorax. IMPRESSION: No significant change."* **A:** "Atelectasis, cardiomegaly, pleural effusion"
Q: "What abnormalities are seen in this image?" **A:** "Lung opacity, pneumonia"	**A:** "Lung opacity"	**R:** "*FINDING: The heart is moderately enlarged. **There are diffuse bilateral airspace opacities**. There is no pneumothorax or pleural effusion. IMPRESSION: **Multifocal pneumonia**."* **A:** "Lung opacity, pneumonia"
Q: "What level is the pleural effusion?" **A:** "Small"	**A:** "Moderate"	**R:** "*FINDING: Left internal jugular central venous catheter terminates in the low svc. Endotracheal tube terminates 4. 5 cm above the carina. Enteric tube courses below the diaphragm and out of view. Lung volumes remain low. Moderate cardiomegaly is unchanged. Mild pulmonary edema is unchanged. bibasilar opacities likely reflect atelectasis. **Small bilateral pleural effusions are unchanged**. There is no pneumothorax.* *IMPRESSION: Interval improvement in pulmonary edema and bilateral pleural effusions.*" **A:** "Small"
Q: "What level is the cardiomegaly?" **A:** "Moderate"	**A:** "Severe"	**R:** "*FINDING: As compared to the previous radiograph, there is no relevant change. **Moderate cardiomegaly** without pulmonary edema. No pleural effusions. No pneumonia. No pneumothorax. IMPRESSION: No acute cardiopulmonary radiographic abnormality.*" **A:** "Moderate"

Figure 8.4: We compare the quality of our predicted answers without (*Baseline*) and with the predicted CXR radiology report (our *RG-AG* model). For each question (Q), we highlight the correct parts of the answer (A) in green and the errors in red. Similarly, we colour-coded the text in the predicted radiology reports (R) as green for the segments that correctly contributed to the answer prediction.

8.7 Conclusion

We have explored the use of radiology reports as additional context to ground the answer generation task of a CXR VQA system. Our RG-AG method shows state-of-the-art results on the Medical-Diff-VQA dataset, with the most notable improvements noted on the *difference*

types of questions compared to prior methods.

We have further investigated the role of the radiology report for VQA in the ablation study, which shows how our RG-AG method achieves the highest overall performance when combining the Finding and Impression sections predicted from a CXR. Furthermore, our results highlight how the quality of the radiology report plays an important role—shown by using the original reports written by expert radiologists—and how providing the AG with visual clues from the CXR is still necessary.

In this chapter, we have only addressed whether using the predicted radiology reports can enhance the VQA performance. However, other types of clinical information related to a patient could be provided as evidence to the VQA model to improve its answer generation capability. This strongly depends on the type of questions we want our VQA model to be able to respond to.

Moreover, we have investigated the problem of grounding the AG with radiology reports adopting a two-step approach, following the approach proposed in [232] for multimodal CoT reasoning. We hypothesise that using a two-stage approach, with each stage implemented using a different model, may only be necessary for smaller models like ours. Larger and more capable models may be able to perform all tasks using a single model, as shown in text-only question-answering [209, 234].

We leave these two directions as open questions for future works in this space.

Chapter 9

Conclusion

9.1 Summary

In this thesis, we have explored the potential of visio-linguistic AI methods for CXR radiology, and their potential to be employed as diagnostic tools to support radiologists in many time-consuming tasks. We have discussed how they can be applied in different tasks including medical finding classification, Automated Radiology Reporting (ARR), and medical Visual Question Answering (VQA).

In this section, we summarise the contributions and conclusions of each chapter.

9.1.1 Chapter 3 – Multimodal CXR Classification from Self-Supervised Image Encoders

Multimodal learning has recently gained traction in healthcare applications due to its ability to process diverse data types. This is especially relevant in the medical field, where rich multimodal clinical data is routinely collected from patients, including images, reports, and clinical records. In this thesis, we began by exploring the task of medical finding classification within a multimodal framework. While many studies approach this as a vision-only task, detecting medical findings solely from images, we enhanced the model by incorporating the indication field from radiology reports as auxiliary textual input in a vision-language model. This approach, previously shown to boost classification performance [85], enriches the model by providing it with crucial context about the patient's medical history and the rationale behind the scan, leading to more accurate and informed predictions.

The study further investigated how different initialisations of the image encoder impact multimodal, multi-label CXR classification, emphasising the benefits of domain-specific contrastive learning pre-training. The comparison focused on commonly adopted initialisation techniques, including random initialisation, ImageNet pre-trained weights, and self-supervised methods. Due to the challenge of obtaining annotated medical data, the focus was on self-supervised domain-specific pre-training to mimic real-world scenarios with limited annotations.

The findings indicated that different initialisation techniques influence the performance of the fine-tuned model, particularly when labelled data is scarce. Overall, the results suggested that self-supervised pre-training is a viable strategy for initialising image encoders in multimodal models, particularly when large, unlabelled pre-training datasets are available. However, in the medical domain, access to such datasets is often limited due to privacy concerns, restricting the practical application of these techniques.

These findings suggest the importance of both the choice of input sources and the methods used to encode them, as they are critical factors influencing the performance of vision-language models. This motivated us to explore different input representations (*e.g.*, *triples* in Chapter 4 and *anatomical tokens* in Chapter 5) to enhance the effectiveness of vision-language models across various tasks.

9.1.2 Chapter 4 – CXR Automated Reporting using Intermediate Triples Representations

We proceeded by examining the ARR task, which is inherently a multimodal task, as it involves generating a textual description of the medical findings from a medical image. Traditionally, most approaches have adhered to this image-to-text paradigm, focusing on refining the model architecture or optimising the loss function. However, we adopted a different strategy by demonstrating that incorporating textual information from other available sources, such as the indication field (which has proven effective for the finding classification task) and other predicted textual information, can enhance performance while maintaining a simple model architecture and loss function.

Our method reformulated the ARR task as a two-stage process, dividing it into Triples Extraction (TE) and Report Generation (RG), rather than directly generating the radiology report from the image. In the first TE stage, we designed a vision-language model to predict a set of structured information, including all the clinically relevant information in the report. In the second RG stage, another vision-language model was employed to generate the radiology

report, using the structured information predicted in the first stage as an auxiliary input. To support this approach, we proposed a semi-automated annotation scheme that extracts structured information from radiology reports in the form of triples, providing supervision for the TE step of our method.

Our solution showed improved performance on the ARR task. This suggested that dividing the task into two subtasks is more effective than directly predicting final reports from the CXR image, which requires attention to both sentence syntax and clinically relevant information. In this way, the TE model can only focus on predicting clinically relevant concepts, which are then used to support the RG step in generating the free-text radiology report.

9.1.3 Chapter 5 – CXR Automated Reporting using Finding-Aware Anatomical Tokens

In the literature, most vision-language models for CXR applications have typically encoded images into global feature maps representing the full images through the use of convolutional neural networks, similar to what we have adopted in Chapters 3 and 4. While effective in capturing overall image characteristics, these global feature maps may overlook the nuanced details specific to various anatomical regions that are crucial for accurate medical interpretation. Therefore, we explored how to effectively extract local feature representations from each CXR, each corresponding to a specific anatomical region of the chest.

This was achieved by extracting finding-aware anatomical tokens from an object detection network, which was adapted to perform finding detection jointly with anatomy localisation in a multi-task setting. We demonstrated that finding detection supervision is crucial when training the object detector, as it enables the encoding of subtle information about abnormalities in each anatomical structure, rather than focusing solely on high-level visual features of the structures themselves. This is particularly relevant when the finding-aware anatomical tokens are adopted as the visual input representation for the final ARR task, where accurately describing the medical findings within each region is essential.

Our solution showed how using such anatomical representations helps improve the ARR performances, with the best performances obtained when integrated into the TE-RG approach described in Chapter 4, showing cumulative improvements.

9.1.4 Chapter 6 – Longitudinal and Controllable CXR Automated Reporting

By using anatomy-specific vector representations of the CXRs, as detailed in Chapter 5, we linked which anatomical region each input visual token corresponds to. This approach offers several significant advantages. First, it enables the modelling of the temporal evolution of the findings, by simply comparing the anatomical regions of two subsequent scans. Second, it provides greater control over which regions are used as visual input for the vision-language model, allowing us to specify the exact regions we want the model to report on. Both of these aspects are largely overlooked in existing literature.

To effectively model the evolution between two subsequent scans of the same patient, we developed a simple yet effective solution. This involves aligning and concatenating the finding-aware anatomical tokens from equivalent anatomical regions in both the prior and current CXRs, which are then projected into a joint representation. This method allows the model to track and describe changes over time within specific regions, leading to more accurate and relevant reporting.

In addition, to offer more control over the ARR task, we developed a training strategy in which the model is trained to predict a partial report based on a sampled subset of anatomical regions. This approach trains the model to associate specific input anatomical regions with the corresponding output sentences, enabling it to generate more targeted and relevant reports based on selected areas of interest.

Our solutions demonstrated improved performance in the ARR task, particularly when these techniques are combined. By incorporating longitudinal scan data and enabling controllable reporting, our approach not only enhanced the accuracy and relevance of the generated

reports but also introduced a level of precision and flexibility that is often missing in current vision-language models. This allows for more clinically useful radiology reporting, tailored to specific anatomical regions and changes over time.

9.1.5 Chapter 7 – Assessing Integrated Automated Reporting Solutions: A Human Evaluation

In the previous chapters, we primarily evaluated the performance of ARR solutions using automatic metrics. While these metrics provide a quantitative assessment, they have significant limitations, the most notable being their inability to fully capture the semantic meaning and clinical relevance of the generated reports. To address this shortcoming, we conducted a human evaluation to gain deeper insights into the model's behaviour.

To begin, we combined all the solutions discussed in Chapters 4, 5 and 6, which yielded the highest performance across most automatic metrics. We then designed an evaluation protocol that included generic error types and correct mentions, aiming to provide a comprehensive measure of our method's overall effectiveness. Two junior physicians performed this evaluation, reviewing a total of 100 unique reports generated by our model.

The human evaluation revealed that our method demonstrates promising results in accurately identifying medical findings and their associated attributes. However, we also identified a concerning frequency of errors that could potentially lead to patient harm, highlighting a significant gap between current model performance and the stringent requirements for clinical practice. These findings underscore the need for future work to focus on reducing the occurrence of critical errors. This is essential for enabling the safe and reliable use of ARR systems in real-world clinical settings.

9.1.6 Chapter 8 – Grounding CXR Visual Question Answering with Radiology Reports

In this chapter, we shifted our focus from ARR to VQA for CXR radiology and demonstrated how the two tasks can be connected into a unified solution to enhance VQA. Traditionally, most research has approached VQA as a standalone task, where the model takes images and corresponding questions as inputs and in response generates answers. However, we argued that we can leverage the radiology reports predicted by an ARR system and use them to provide evidence to generate the final answer.

To this end, we proposed a novel two-stage approach for the CXR VQA task, which sequentially performs Report Generation (RG) and Answer Generation (AG). In the first step, the model generates an overall description of the visual appearance of the CXRs presented in the form of free-text radiology reports that emphasise any abnormalities detected. In the second step, the AG model takes the report alongside the question and the CXR images to predict the answers. By incorporating the radiology report, we provided the AG model with additional evidence to support the prediction of an answer.

Our proposed method showed improved performance on the VQA task, for questions that either involve comparing two images and identifying their differences or inquire about specific visual features within a single image. The results indicate that integrating radiology reports into the VQA process effectively enhances the model's ability to generate accurate answers.

9.2 Validation of Thesis Statement

This thesis argues that effectively integrating VLMs into the radiological workflow can enhance the effectiveness of decision-support tools in CXR radiology. We hereby validate the specific claims of our thesis statement presented in Section 1.6, based on the results presented throughout Chapters 3-8.

- **Claim 1:** ***By processing and generating semantically rich textual information, VLMs***

can better align with radiologists' workflows. This claim is supported throughout the results in the technical chapters of this thesis. Our experiments show that incorporating the text *clinical indication field* as additional contextual input to a VLM improves the overall performance in the finding classification task (Chapter 3). Similar results have been presented on ARR, highlighting how this task benefits when the *indication field* is provided as input to the model (see in Chapter 4). These findings have led us to consistently integrate this additional information source alongside CXR data in all chapters when performing ARR. In addition, we have shown how VLMs can generate CXR radiology reports that closely resemble the format and content of radiologists' written reports, reducing the gap between AI tools and radiologists' workflow (see Chapters 4-7).

- **Claim 2:** ***Employing multiple VLMs—each dedicated to a specific subtask—can improve overall task performance, including Automated Radiology Reporting, and Visual Question Answering.*** We have presented a solution that breaks down the ARR task into two separate steps: Triples Extraction (TE) and Report Generation (RG) (Chapter 4). Our results indicate that the TE-RG pipeline yields improved performance compared to a single-step approach. We observe similar performance gains when applying the same pipeline in the following chapters (Chapters 5 and 7). Moreover, we demonstrate how a similar approach can be applied to VQA (Chapter 8), dividing the task into Report Generation (RG) followed by Answer Generation (AG). Our findings demonstrate that using the predicted radiology reports as an auxiliary input for answer generation helps to ground the answer inference, leading to improved responses.

- **Claim 3:** ***The strategic extraction of meaningful image representations enhances VLM performance in CXR interpretation.*** We have shown how adopting domain-specific self-supervised approaches to pre-train the image encoder of a VLM provides better initialisations, resulting in improved performance in the finding classification task (Chapter 3). This is particularly relevant in the scenario of a large unlabelled image

dataset used at pre-training and a small labelled one for fine-tuning. Following, we have investigated how to improve the image representations for CXR automated reporting. Specifically, we focus on extracting localised representations of CXRs as opposed to relying on image-level representations (Chapter 5). Our findings indicate that using anatomy-specific representations as the visual input to a VLM improves the quality of the generated reports. Moreover, we have shown that, by comparing the anatomical regions of two subsequent scans, a VLM is capable of effectively modelling the temporal evolution of the medical findings (Chapter 6).

- **Claim 4:** ***The flexibility of VLMs in handling diverse inputs and outputs allows for greater control over generated outputs and deeper interaction with imaging data.*** We have leveraged the link between anatomical regions and their corresponding sentences within a radiology report to develop a training strategy that teaches the VLM to only report on selected regions (see Chapter 5). This provides better control over which areas of the CXR we want the VLM to report on. Furthermore, we have explored the potential of VLMs for VQA by incorporating predicted reports as auxiliary inputs to ground the answer-generation process (Chapter 8). We have presented a flexible solution capable of responding to questions about the visual characteristics of individual CXRs, as well as comparing the visual differences between successive scans of the same patient. These results demonstrate the flexibility of VLMs in interacting with CXR data through text queries, offering a more controlled and dynamic approach to medical imaging analysis.

9.3 Future Work

We discuss potential avenues for advancing the current work, focusing on addressing existing limitations and expanding the applicability of our methods.

9.3.1 Adopting Larger Models

Throughout this thesis, our research has primarily focused on using relatively small models, containing only a few attention layers. We believe that small models offer several distinct advantages, particularly in medical applications, where they can be beneficial due to their reduced computational overhead, faster training times, and lower energy consumption, which are crucial for deployment in healthcare settings with limited resources. Small models also present the advantage of being easier to fine-tune, which makes them a suitable choice for tailored applications for precise medicine, where specific adjustments are often required to meet the stringent demands of clinical practice.

However, the field of natural language processing (NLP) and multimodal applications has recently witnessed remarkable advancements through the development of Large Language Models (LLMs) [195, 148, 51]. These models, with their vast parameter sizes and extensive pretraining on diverse datasets, have demonstrated unprecedented capabilities across a wide range of NLP tasks, including text generation, comprehension, and reasoning. More recently, LLMs have shown great promise in multimodal tasks that integrate both images and text in the general domain [5, 148], as well as in the medical domain [79, 15, 176, 220].

Due to limitations in our computational resources and limited dataset size, we were unable to effectively adopt LLMs in our research. Nevertheless, we hypothesise that the methodologies proposed in this thesis for finding classification, ARR, and VQA could be enhanced when integrated with LLMs. Leveraging the sophisticated language understanding and generative capabilities of LLMs could potentially lead to more accurate results compared to smaller models. Future research should explore the integration of LLMs into the proposed frameworks, evaluating their performance and scalability in more complex and resource-intensive environments.

9.3.2 Expanding to Other Imaging Modalities

This thesis has primarily focused on CXR radiology, presenting a variety of methodologies tailored specifically for this imaging modality. However, the concepts and solutions developed here have broad applicability and can be extended to other imaging techniques, including Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound, and Histology, as well as to different anatomical regions beyond the chest, such as the head, legs, and abdomen. Such extensions are feasible provided that similar datasets to those used in this research are available. However, transitioning from 2D imaging modalities like CXR to 3D modalities, such as CT and MRI, will require substantial modifications. In particular, 3D imaging necessitates more sophisticated model architectures capable of processing volumetric data, as well as enhanced computational resources to handle the larger and more complex datasets inherent in 3D scans.

A major challenge in generating reports for 3D medical images stems from the increased complexity of anatomical structures when compared to 2D X-rays. While bounding boxes can effectively isolate regions of interest in 2D images, this method is much less effective for 3D data, where anatomical structures are often irregular, elongated, or distributed across multiple slices. For instance, structures like blood vessels can have intricate shapes that a simple bounding box fails to encompass properly or drawing bounding boxes around anatomical regions like the skull may enclose excessive surrounding tissue, leading to inaccuracies in automated analysis.

Furthermore, 3D imaging requires models that not only analyse individual slices but also account for the spatial relationships between slices, which can differ from one patient to another. For example, the shape and extent of a tumor or organ may vary across slices, making it insufficient to apply a bounding box around a single region. This complexity calls for advanced techniques such as 3D segmentation, which demands deep learning models capable of accurately delineating complex shapes and structures in 3D space.

In the context of report generation for 3D medical images, the objective is to extract

meaningful information while maintaining the spatial context provided by the full 3D structure. This involves describing the shape, size, and location of abnormalities, such as tumors or lesions, which may span multiple slices or possess irregular boundaries. Achieving accurate and consistent report generation in this context requires models that not only identify these anomalies but also understand their spatial relationships across all three dimensions.

Advanced techniques like 3D CNNs can be employed to model volumetric data. These methods enable the network to learn the spatial relationships of anatomical structures, thus providing more precise delineations of abnormalities. However, such approaches introduce their own challenges, including the need for large volumes of annotated data, high computational demands, and the development of effective training strategies.

However, despite the challenges associated with 3D scans, some of the solutions presented throughout this thesis may be adapted or require minimal modifications to be effectively integrated into 3D imaging applications, or for use in other anatomical regions.

For instance, the self-supervised pre-training strategy discussed in Chapter 3 could be adapted to enhance the imaging encoding of VLMs for findings classification tasks across various imaging modalities, especially in scenarios where labelled data is limited. This approach would facilitate the development of more robust models, thereby improving diagnostic accuracy in diverse medical contexts.

The triples extraction framework presented in Chapter 4, originally tailored for CXR reports, could be re-designed to identify and extract relevant triples from other imaging modalities. This would involve adapting the framework to recognise and prioritise modality-specific entities and relationships.

Furthermore, the anatomical region representations developed in Chapter 5 for the chest area could be generalised to other body regions, provided these regions are compatible with bounding box representations or by adapting our method using segmentation masks. This would involve redefining the regions of interest based on the specific scan being evaluated. In this way, similar techniques could be employed to enhance the quality of radiology re-

ports, facilitate the modelling of longitudinal representations from sequential scans, and offer improved control over report content, as discussed in Chapters 5 and 4.

In the context of VQA, as explored in Chapter 8, the approach of grounding answer generation on predicted radiology reports is not inherently limited to CXR. This strategy could be extended to other types of medical imaging, allowing the VQA model to generate more accurate and contextually relevant answers across various diagnostic scenarios.

In summary, although this thesis has focused on VLM solutions for CXR radiology, and despite the inherent differences between 2D and 3D scans and the associated challenges, there is the potential to extend these methods to other imaging techniques and anatomical areas, which we leave as directions for future research.

9.3.3 Refining our Solutions

The solutions proposed throughout this thesis have demonstrated promising improvements for the specific tasks addressed. However, as discussed based on the human evaluation results of ARR (Chapter 7), further refinement is necessary before these methods can be effectively applied in real clinical practice. Below, we outline areas where our solutions could be enhanced.

For instance, the triple extraction pipeline, presented in Chapter 4, was designed using existing data and annotation tools which were not specifically tailored to our task. We adopted these tools with minimal additional clinical input to suit our specific needs. As a result, the extracted triples may not be optimal for effectively capturing structured information. To address this, we could create a new triple annotation schema in collaboration with clinicians and then train a model to accurately annotate triples from the reports. Improved triples could result in more accurate predictions of radiology reports, which is our ultimate goal.

We proposed a relatively simple longitudinal projection module (Chapter 6) to model the temporal evolution of findings between two subsequent scans. While our results indicated a consistent improvement in predicting reports for follow-up scans, further research is needed to

devise solutions that can capture more subtle changes between scans.

Finally, we demonstrated how VQA can benefit from incorporating radiology reports as additional context (Chapter 8). We discussed the impact of the quality of predicted reports on the answer-generation process. Thus, by enhancing the initial step of the pipeline responsible for generating the radiology report, we can improve the VQA outcomes. Furthermore, our study has focused solely on the impact of radiology reports on the VQA task. However, depending on the specific questions our system is designed to answer, additional clinical data such as clinical indication, image orientation, patient's age, gender, etc. could be incorporated to facilitate the answer-generation process.

These are some of the areas that should be considered to improve overall performance for the tasks of interest and to bring us closer to real-world applications.

9.3.4 Relevance of VQA for Key Applications

We have addressed VQA by considering questions that focus on critical aspects of medical image interpretation, relevant to various use cases. These questions concentrate on identifying abnormalities, their locations, and severity, as well as tracking changes over time, which are fundamental to medical diagnostics and decision-making. However, their utility can be enhanced through targeted adaptations for specific applications.

In clinical decision support, the questions we have considered align closely with the needs of healthcare professionals by providing actionable insights into diagnostic findings and disease progression. For example, questions about abnormalities or differences between images can assist radiologists in identifying pathologies or assessing treatment efficacy. To enhance their applicability, the scope of the questions could be expanded to include treatment-related queries, such as "What are the best treatment options based on this abnormality?" or "What is the risk level associated with this condition?". Such extensions could directly support clinical workflows and decision-making.

Moreover, in patient assistance, the proposed VQA system could help provide comprehen-

sible explanations to patients about their medical images, although the language and framing would need simplification to suit non-expert audiences. Questions about the presence or severity of abnormalities, for example, can inform patients about their conditions when translated into layman's terms. Similarly, questions exploring changes over time could clarify the effectiveness of treatments or highlight areas requiring further attention. Including examples or analogies in the answers could further improve understanding for patients with no medical background.

In the context of medical training, the proposed VQA system could serve as an effective tool for improving clinical diagnostic skills. By posing questions about specific abnormalities, their locations, and the changes observed between medical images, learners can simulate real-world diagnostic tasks. To maximise its usefulness in medical education, providing step-by-step explanations of why certain answers are correct could help learners grasp the diagnostic reasoning process in greater detail. This would enable students to understand the thought processes behind each decision, enhancing their ability to analyse and interpret medical images accurately in real-world clinical scenarios.

Therefore, a VQA dataset could benefit from further tailoring to these applications. Simplifying and rephrasing questions for patient-facing systems, expanding the question scope to include treatment-related insights for clinical decision support, and incorporating explanatory feedback for educational use are promising avenues for future work. These enhancements would strengthen the dataset's applicability across clinical, educational, and patient-centred domains, making it a more versatile tool for advancing healthcare technology.

9.4 Final Remarks

Given the inherently multimodal nature of clinical data—encompassing imaging and textual information, among others—using VLMs for healthcare applications has enormous potential. In this thesis, we have investigated the application of VLMs across a range of tasks in CXR radiology, including medical finding classification, Automated Radiology Reporting, and

Visual Question Answering. Our work has demonstrated the flexibility of these models to various tasks, each presenting its own set of challenges, and we have introduced novel solutions to address these.

While the potential of VLMs to enhance the accuracy and efficiency of radiological practice is evident, several performance gaps remain that must be addressed before these models can be reliably integrated into clinical workflows. In addition to the limitations discussed in this thesis, such as the high frequency of critical errors in the generated reports, several other challenges must be overcome to make VLMs viable for clinical use. One such challenge is the difficulty of evaluating the clinical quality of AI-generated predictions, which is essential for accurately assessing their utility in real-world applications.

Another hurdle lies in generalisation to unseen data. Current models often struggle to perform reliably across diverse populations and clinical contexts, such as differences between intensive care units and in/outpatient care settings. Addressing this requires robust fine-tuning and testing on a broader range of clinical scenarios to ensure reliability and minimise disparities in performance. Furthermore, research is needed to determine what levels of error might be acceptable in practice and how these can be managed to reduce radiology report turnaround times without compromising quality or patient safety.

An important area for future investigation is the role of VLMs as assistive tools in clinical workflows. For instance, studies could explore workflows where radiologists review and correct automatically generated reports to assess whether this setup reduces reporting time, enhances report accuracy, or inadvertently introduces risks of overreliance on AI predictions. Such studies would help clarify the balance required for successful clinician-AI collaboration.

Beyond improving reporting, the scope of VLMs could also be expanded to include use cases such as flagging potential errors, identifying missing findings, or serving as a second check to highlight inconsistencies in reports. These capabilities would provide additional layers of support to radiologists, further demonstrating the models' value as complementary tools. Addressing these challenges and exploring these avenues will be crucial for transitioning

VLMs from experimental models to trusted, practical solutions in clinical radiology workflows.

Through this research, we have not only highlighted the potential of VLMs to enhance the accuracy and efficiency of radiological practice but also underscored the importance of further development and refinement for their successful integration into clinical workflows. It is our hope that this thesis will inspire future research efforts aimed at refining these models and expanding their applicability, ultimately contributing to the evolution of radiological practice.

Bibliography

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[2] Nkechinyere N Agu, Joy T Wu, Hanqing Chao, Ismini Lourentzou, Arjun Sharma, Mehdi Moradi, Pingkun Yan, and James Hendler. "AnaXNet: Anatomy Aware Multi-label Finding Classification in Chest X-Ray". In: *MICCAI*. Springer. 2021.

[3] Abhimanyu S Ahuja. "The impact of artificial intelligence in medicine on the future role of the physician". In: *PeerJ* 7 (2019), e7702.

[4] Aisha Al-Sadi, Hana' Al-Theiabat, and Mahmoud Al-Ayyoub. "The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG". In: CEUR Workshop Proceedings 2696 (2020). Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. "Flamingo: a visual language model for few-shot learning". In: *Advances in neural information processing systems* 35 (2022), pp. 23716–23736.

[6] Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. "Automated radiology report generation using conditioned transformers". In: *Informatics in Medicine Unlocked* 24 (2021), p. 100557.

[7] Ariana Anderson, Pamela K Douglas, Wesley T Kerr, Virginia S Haynes, Alan L Yuille, Jianwen Xie, Ying Nian Wu, Jesse A Brown, and Mark S Cohen. "Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD". In: *NeuroImage* 102 (2014), pp. 207–219.

[8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6077–6086.

[9] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. "VQA: Visual Question Answering". In: (2015).

[10] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. "Big self-supervised models advance medical image classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3478–3488.

[11] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. "EHRXQA: A multi-modal question answering dataset for electronic health records with chest x-ray images". In: *Advances in Neural Information Processing Systems* 36 (2024).

[12] Long Bai, Mobarakol Islam, and Hongliang Ren. "CAT-ViL: Co-attention Gated Vision-Language Embedding for Visual Question Localized-Answering in Robotic Surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 397–407.

[13] Ivo Baltruschat, Leonhard Steinmeister, Hannes Nickisch, Axel Saalbach, Michael Grass, Gerhard Adam, Tobias Knopp, and Harald Ittrich. "Smart chest X-ray work-

list prioritization using artificial intelligence: a clinical workflow simulation". In: *European radiology* 31 (2021), pp. 3837–3845.

[14] Satanjeev Banerjee and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005, pp. 65–72.

[15] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. "MAIRA-2: Grounded Radiology Report Generation". In: *arXiv preprint arXiv:2406.04449* (2024).

[16] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. "Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing". In: *arXiv preprint arXiv:2301.04558* (2023).

[17] Andrew L Beam and Isaac S Kohane. "Translating artificial intelligence into clinical care". In: *Jama* 316.22 (2016), pp. 2368–2369.

[18] Iz Beltagy, Arman Cohan, and Kyle Lo. "SciBERT: Pretrained Contextualized Embeddings for Scientific Text". In: *CoRR* abs/1903.10676 (2019).

[19] Asma Ben Abacha, Vivek V. Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. "Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain". In: CEUR Workshop Proceedings (2020).

[20] Asma Ben Abacha, Soumya Gayen, Jason J. Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. "NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain". In: CEUR Workshop Proceedings 2125 (2018). Ed. by Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier.

[21] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019". In: CEUR Workshop Proceedings (2019).

[22] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. "Overview of the VQA-Med Task at ImageCLEF 2021: Visual Question Answering and Generation in the Medical Domain". In: *CLEF 2021 Working Notes*. CEUR Workshop Proceedings. Bucharest, Romania: CEUR-WS.org, Sept. 2021.

[23] Xia-an Bi, Xi Hu, Hao Wu, and Yang Wang. "Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest". In: *IEEE Journal of Biomedical and Health Informatics* 24.10 (2020), pp. 2973–2983.

[24] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. "An introduction to vision-language modeling". In: *arXiv preprint arXiv:2405.17247* (2024).

[25] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020.

[26] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. "Padchest: A large chest x-ray image dataset with multi-label annotated reports". In: *Medical image analysis* 66 (2020), p. 101797.

[27] Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. "Paniniqa: Enhancing patient education through interactive question answering". In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 1518–1536.

[28] Daniel J Cao, Casey Hurrell, and Michael N Patlas. "Current status of burnout in Canadian radiology". In: *Canadian Association of Radiologists Journal* 74.1 (2023), pp. 37–43.

[29] Care Quality Commission. “A national review of radiology reporting within the NHS in England”. In: (2018), pp. 1–26.

[30] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[31] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. “Cross-modal Memory Networks for Radiology Report Generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 5904–5914.

[32] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. “Generating Radiology Reports via Memory-driven Transformer”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 1439–1449.

[33] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. “Chexagent: Towards a foundation model for chest x-ray interpretation”. In: *arXiv preprint arXiv:2401.12208* (2024).

[34] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

[35] Yeongjae Cho, Taehee Kim, Heejun Shin, Sungzoon Cho, and Dongmyung Shin. “Pretraining Vision-Language Model for Difference Visual Question Answering in Longitudinal Chest X-rays”. In: *Medical Imaging with Deep Learning*. 2024.

[36] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. "Structural Scaffolds for Citation Intent Classification in Scientific Publications". In: (June 2019), pp. 3586–3596.

[37] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-memory transformer for image captioning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10578–10587.

[38] Francesco Dalla Serra, William Clackett, Hamish MacKinnon, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q O'Neil. "Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*. 2022, pp. 615–624.

[39] Francesco Dalla Serra, Grzegorz Jacenków, Fani Deligianni, Jeff Dalton, and Alison Q O'Neil. "Improving Image Representations via MoCo Pre-training for Multimodal CXR Classification". In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2022, pp. 623–635.

[40] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. "Controllable Chest X-Ray Report Generation from Longitudinal Representations". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4891–4904.

[41] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeffrey Dalton, and Alison Q. O'Neil. "Finding-Aware Anatomical Tokens for Chest X-Ray Automated Reporting". In: *Machine Learning in Medical Imaging*. Ed. by Xiaohuan Cao, Xuanang Xu, Islem Rekik, Zhiming Cui, and Xi Ouyang. Cham: Springer Nature Switzerland, 2024, pp. 413–423.

[42] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. "Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022, pp. 4348–4360.

[43] Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. "Overview of the RadSum23 Shared Task on Multi-modal and Multi-anatomical Radiology Report Summarization". In: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Ed. by Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 478–482.

[44] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. "Preparing a collection of radiology examinations for distribution and retrieval". In: *Journal of the American Medical Informatics Association* 23.2 (2016), pp. 304–310.

[45] Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. "Consumer health information and question answering: helping consumers find answers to their health-related information needs". In: *Journal of the American Medical Informatics Association* 27.2 (Oct. 2019), pp. 194–201.

[46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[48] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. "Multiple meta-model quantifying for medical visual question answering".

In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 64–74.

[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021.

[50] DC Dowson and BV666017 Landau. "The Fréchet distance between multivariate normal distributions". In: *JMA* 12.3 (1982), pp. 450–455.

[51] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. "The llama 3 herd of models". In: *arXiv preprint arXiv:2407.21783* (2024).

[52] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. "Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model". In: *MLH*. PMLR. 2021, pp. 209–219.

[53] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. "Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model". In: *Proceedings of Machine Learning for Health*. Ed. by Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer. Vol. 158. Proceedings of Machine Learning Research. PMLR, Apr. 2021, pp. 209–219.

[54] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. "Complexity-based prompting for multi-step reasoning". In: *The Eleventh International Conference on Learning Representations*. 2022.

[55] Ross Girshick. “Fast R-CNN”. In: *ICCV*. 2015, pp. 1440–1448.

[56] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[57] 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”. In: *circulation* 101.23 (2000), e215–e220.

[58] Travis R Goodwin and Sanda M Harabagiu. “Medical question answering for clinical decision support”. In: *Proceedings of the 25th ACM international on conference on information and knowledge management*. 2016, pp. 297–306.

[59] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs”. In: *JAMA* 316.22 (Dec. 2016), pp. 2402–2410.

[60] Yubraj Gupta, Ramesh Kumar Lama, Goo-Rak Kwon, and Alzheimer’s Disease Neuroimaging Initiative. “Prediction and classification of Alzheimer’s disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers”. In: *Frontiers in computational neuroscience* 13 (2019), p. 72.

[61] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[62] Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, and Henning Müller. "Overview of the ImageCLEF 2018 Medical Domain Visual Question Answering Task". In: CEUR Workshop Proceedings (2018).

[63] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[64] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. "Mask R-CNN". In: *CoRR* abs/1703.06870 (2017).

[65] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[66] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. *PathVQA: 30000+ Questions for Medical Visual Question Answering*. 2020.

[67] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. "Decoupling the role of data, attention, and losses in multimodal transformers". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 570–585.

[68] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[69] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations*. 2021.

[70] Xinyue Hu et al. "Medical-Diff-VQA: A Large-Scale Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images". In: ().

[71] Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. *Medical-Diff-VQA: A Large-Scale Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images*. 2023.

[72] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. "Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 4156–4165.

[73] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[74] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". In: *CoRR* abs/1904.05342 (2019).

[75] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines". In: *NPJ digital medicine* 3.1 (2020), p. 136.

[76] Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. "OVQA: A clinically generated visual question answering dataset". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 2924–2938.

[77] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. *Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers*. 2020.

[78] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. "KiUT: Knowledge-Injected U-Transformer for Radiology Report Generation". In: *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR). 2023, pp. 19809–19818.

[79] Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. "MAIRA-1: A specialised large multimodal model for radiology report generation". In: *arXiv preprint arXiv:2311.13668* (2023).

[80] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. "Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation". In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) (2018).

[81] Bogdan Ionescu et al. "ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature". In: Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019) (2019).

[82] Bogdan Ionescu et al. "Overview of the ImageCLEF 2020: Multimedia Retrieval in Medical, Lifelogging, Nature, and Internet Applications". In: Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020) 12260 (2020).

[83] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[84] Allan Jabri, Armand Joulin, and Laurens van der Maaten. "Revisiting Visual Question Answering Baselines". In: *CoRR* abs/1606.08390 (2016).

[85] Grzegorz Jacenków, Alison Q O'Neil, and Sotirios A Tsaftaris. "Indication as Prior Knowledge for Multimodal Disease Classification in Chest Radiographs with Transformers". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–5.

[86] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports". In: *arXiv preprint arXiv:2106.14463* (2021).

[87] Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. "Multimodal Image-Text Matching Improves Retrieval-based Chest X-Ray Report Generation". In: *Medical Imaging with Deep Learning*. Ed. by Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant. Vol. 227. Proceedings of Machine Learning Research. PMLR, Oct. 2024, pp. 978–990.

[88] Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. "Probing Biomedical Embeddings from Language Models". In: *CoRR* abs/1904.02181 (2019).

[89] Baoyu Jing, Pengtao Xie, and Eric Xing. "On the Automatic Generation of Medical Imaging Reports". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2577–2586.

[90] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3.1 (May 2016).

[91] Alistair EW 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". In: *Scientific Data* 6.1 (2019).

[92] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs". In: *arXiv preprint arXiv:1901.07042* (2019).

[93] Ian Jolliffe. *Principal component analysis*. New York: Springer Verlag, 2002.

[94] Tobias Jorg, Moritz C Halfmann, Fabian Stoehr, Gordon Arnhold, Annabell Theobald, Peter Mildenberger, and Lukas Müller. "A novel reporting workflow for automated integration of artificial intelligence results into structured radiology reports". In: *Insights into Imaging* 15.1 (2024), p. 80.

[95] Bumjun Jung, Lin Gu, and T. Harada. "bumjun_jung at VQA-Med 2020: VQA Model Based on Feature Extraction and Multi-modal Feature Fusion". In: (2020).

[96] Kaveri Kale, Pushpak Bhattacharyya, and Kshitij Jadhav. "Replace and Report: NLP Assisted Radiology Report Generation". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 10731–10742.

[97] Gaurang Karwande, Amarachi B Mbakwe, Joy T Wu, Leo A Celi, Mehdi Moradi, and Ismini Lourentzou. "CheXRelNet: An Anatomy-Aware Model for Tracking Longitudinal Relationships Between Chest X-Rays". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*. Springer. 2022, pp. 581–591.

[98] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. *MMBERT: Multimodal BERT Pretraining for Improved Medical VQA*. 2021.

[99] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. "Supervised Multimodal Bitransformers for Classifying Images and Text". In: *arXiv preprint arXiv:1909.02950* (2019).

[100] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[101] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. "Multimodal machine learning in precision health: A scoping review". In: *npj Digital Medicine* 5.1 (2022), p. 171.

[102] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. "Large language models are zero-shot reasoners". In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.

[103] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". In: *CoRR* abs/1602.07332 (2016).

[104] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: 25 (2012). Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger.

[105] Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. "A dataset of clinically generated visual questions and answers about radiology images". In: *Scientific Data* 5.1 (Nov. 2018).

[106] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *CoRR* abs/1901.08746 (2019).

[107] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. "Knowledge-driven encode, retrieve, paraphrase for medical image report generation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 6666–6673.

[108] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. "Llava-med: Training a large language-and-vision assistant for biomedicine in one day". In: *Advances in Neural Information Processing Systems* 36 (2024).

[109] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *arXiv preprint arXiv:2301.12597* (2023).

[110] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International conference on machine learning*. PMLR. 2022, pp. 12888–12900.

[111] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "VisualBERT: A Simple and Performant Baseline for Vision and Language". In: *Arxiv*. 2019.

[112] Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, et al. "Ffa-ir: Towards an explainable and reliable medical report generation benchmark". In: *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*. 2021.

[113] Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. "Cross-modal Clinical Graph Transformer for Ophthalmic Report Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20656–20665.

[114] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. "Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report

Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3334–3343.

[115] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. "Dynamic graph enhanced contrastive learning for chest x-ray report generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3334–3343.

[116] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. "Auxiliary signal-guided knowledge encoder-decoder for medical report generation". In: *World Wide Web* 26.1 (2023), pp. 253–270.

[117] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. "Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 374–383.

[118] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer. 2020, pp. 121–137.

[119] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. "Benchmarking detection transfer learning with vision transformers". In: *arXiv preprint arXiv:2111.11429* (2021).

[120] Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. "Multimodal representation learning via maximization of local mutual information". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Stras-*

bourg, France, September 27–October 1, 2021, Proceedings, Part II 24. Springer. 2021, pp. 273–283.

[121] Zhibin Liao, Qi Wu, Chunhua Shen, A. V. Hengel, and J. Verjans. "AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-based Sentence Mapping Approach for Medical Domain Visual Question Answering". In: (2020).

[122] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81.

[123] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[124] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft COCO: common objects in context". In: *ECCV*. Springer. 2014.

[125] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. *SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering*. 2021.

[126] Chang Liu, Yuanhe Tian, and Yan Song. "A Systematic Review of Deep Learning-based Research on Radiology Report Generation". In: *arXiv preprint arXiv:2311.14199* (2023).

[127] Fenglin Liu, Shen Ge, and Xian Wu. "Competence-based multimodal curriculum learning for medical report generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 3001–3012.

[128] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. "Exploring and distilling posterior and prior knowledge for radiology report generation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 13753–13762.

[129] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. "Contrastive attention for automatic chest x-ray report generation". In: *arXiv preprint arXiv:2106.06965* (2021).

[130] 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: *Machine Learning for Healthcare Conference*. PMLR. 2019, pp. 249–269.

[131] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. *Visual Instruction Tuning*. 2023.

[132] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual instruction tuning". In: *Advances in neural information processing systems* 36 (2024).

[133] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. "Ssd: Single shot multibox detector". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.

[134] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *CoRR* abs/1908.02265 (2019).

[135] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 375–383.

[136] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. "Learn to explain: Multimodal reasoning via thought chains for science question answering". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 2507–2521.

[137] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. "Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, pp. 5288–5304.

[138] Maram Mahmoud A. Monshi, Josiah Poon, and Vera Chung. "Deep learning in generating radiology reports: A survey". In: *Artificial Intelligence in Medicine* 106 (2020), p. 101878.

[139] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing". In: *BioNLP Workshop and Shared Task*. 2019, pp. 319–327.

[140] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. "Overcoming data limitation in medical visual question answering". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 522–530.

[141] NHS England and NHS improvement. "Diagnostic Imaging Dataset Statistical Release". In: (2022), pp. 1–17.

[142] Aaron Nicolson, Jason Dowling, and Bevan Koopman. "Improving chest X-ray report generation by leveraging warm starting". In: *Artificial intelligence in medicine* 144 (2023), p. 102633.

[143] Toru Nishino, Ryota Ozaki, Yohei Momoki, Tomoki Taniguchi, Ryuji Kano, Norihisa Nakano, Yuki Tagawa, Motoki Taniguchi, Tomoko Ohkuma, and Keigo Nakamura. "Reinforcement learning with imbalanced dataset for data-to-text medical report generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 2223–2236.

[144] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. "Progressive Transformer-Based Generation of Radiology Reports". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2824–2832.

[145] Piotr Obara, Merlijn Sevenster, Adam Travis, Yuechen Qian, Charles Westin, and Paul J Chang. "Evaluating the referring physician's clinical history and indication as a means for communicating chronic conditions that are pertinent at the point of radiologic interpretation". In: *Journal of digital imaging* 28 (2015), pp. 272–282.

[146] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[147] OpenAI. *ChatGPT-3.5: Language Model*. https://chat.openai.com. https://chat.openai.com. 2023.

[148] OpenAI. "GPT-4 Technical Report". In: *ArXiv* abs/2303.08774 (2023).

[149] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. "GREEN: Generative Radiology Report Evaluation and Error Notation". In: *arXiv preprint arXiv:2405.03595* (2024).

[150] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[151] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. “Radiology Objects in COntext (ROCO): A Multimodal Image Dataset”. In: (2018). Ed. by Danail Stoyanov, Zeike Taylor, Simone Balocco, Raphael Sznitman, Anne Martel, Lena Maier-Hein, Luc Duong, Guillaume Zahnd, Stefanie Demirci, Shadi Albarqouni, Su-Lin Lee, Stefano Moriconi, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, Eric Granger, and Pierre Jannin, pp. 180–189.

[152] Chantal Pellegrini, Matthias Keicher, Ege Özsoy, and Nassir Navab. “Rad-restruct: A novel vqa benchmark and method for structured radiology reporting”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 409–419.

[153] Yalei Peng and F. Liu. “UMass at ImageCLEF Medical Visual Question Answering(Med-VQA) 2018 Task”. In: (2018).

[154] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. “ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data”. In: *CoRR* abs/2001.07966 (2020).

[155] Han Qin and Yan Song. “Reinforced Cross-modal Alignment for Radiology Report Generation”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 448–458.

[156] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. “Describing and Localizing Multiple Changes With Transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 1971–1980.

[157] Leslie E Quint, Douglas J Quint, and James D Myles. "Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology". In: *Journal of the American College of Radiology* 5.12 (2008), pp. 1196–1199.

[158] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[159] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[160] European Society of Radiology (ESR) communications@ myesr. org. "ESR paper on structured reporting in radiology". In: *Insights into imaging* 9 (2018), pp. 1–7.

[161] Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. "Transfusion: Understanding Transfer Learning with Applications to Medical Imaging". In: *CoRR* abs/1902.07208 (2019).

[162] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning". In: *arXiv preprint arXiv:1711.05225* (2017).

[163] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* 1.2 (2022), p. 3.

[164] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[165] Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. "Improving Radiology Report Generation Systems by Removing Hallucinated References to Non-existent Priors". In: *Machine Learning for Health*. PMLR. 2022, pp. 456–473.

[166] DS Rana, G Hurst, L Shepstone, J Pilling, J Cockburn, and M Crawford. "Voice recognition for radiology reporting: is it good enough?" In: *Clinical radiology* 60.11 (2005), pp. 1205–1212.

[167] J Redmon. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[168] F. Ren and Y. Zhou. "CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering". In: *IEEE Access* 8 (2020), pp. 50626–50636.

[169] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[170] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. "Self-critical sequence training for image captioning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7008–7024.

[171] Abi Rimmer. "Radiologist shortage leaves patient care at risk, warns royal college". In: *BMJ: British Medical Journal (Online)* 359 (2017).

[172] Michael D Ringler, Brian C Goss, and Brian J Bartholmai. "Syntactic and semantic errors in radiology reports associated with speech recognition software". In: *Health informatics journal* 23.1 (2017), pp. 3–13.

[173] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *CoRR* abs/1505.04597 (2015).

[174] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. "Learning To Retrieve Prompts for In-Context Learning". In: *Proceedings of the 2022 Conference of the North Amer-*

ican Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022, pp. 2655–2671.

[175] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.

[176] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. "Capabilities of gemini models in medicine". In: *arXiv preprint arXiv:2404.18416* (2024).

[177] Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. "DualNet: Domain-Invariant Network for Visual Question Answering". In: *CoRR* abs/1606.06108 (2016).

[178] Mehmet Umut Salur and İlhan Aydın. "A soft voting ensemble learning-based approach for multimodal sentiment analysis". In: *Neural Computing and Applications* 34.21 (2022), pp. 18391–18406.

[179] Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. "SurgicalGPT: End-to-end language-vision GPT for visual question answering in surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 281–290.

[180] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[181] Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain". In: *Scientific Reports* 11.1 (2021), p. 19826.

[182] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: (2015). Ed. by Yoshua Bengio and Yann LeCun.

[183] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. *MMF: A multimodal framework for vision and language research*. https://github.com/facebookresearch/mmf. 2020.

[184] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. "Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT". In: *EMNLP*. 2020, pp. 1500–1519.

[185] Rebecca Smith-Bindman, Diana L Miglioretti, and Eric B Larson. "Rising use of diagnostic medical imaging in a large integrated health system". In: *Health affairs* 27.6 (2008), pp. 1491–1502.

[186] Tom van Sonsbeek and Marcel Worring. "Towards automated diagnosis with attentive multi-modal learning using electronic health records and chest x-rays". In: *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures: 10th International Workshop, ML-CDS 2020, and 9th International Workshop, CLIP 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 9*. Springer. 2020, pp. 106–114.

[187] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. "MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models". In: *Medical Imaging with Deep Learning*. PMLR. 2021, pp. 728–744.

[188] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. "VL-BERT: Pre-training of Generic Visual-Linguistic Representations". In: *CoRR* abs/1908.08530 (2019).

[189] Zhaoyi Sun, Hanley Ong, Patrick Kennedy, Liyan Tang, Shirley Chen, Jonathan Elias, Eugene Lucas, George Shih, and Yifan Peng. "Evaluating GPT-4 on impressions generation in radiology reports". In: *Radiology* 307.5 (2023), e231259.

[190] Tanveer Syeda-Mahmood, Ken CL Wong, Yaniv Gur, Joy T Wu, Ashutosh Jadhav, Satyananda Kashyap, Alexandros Karargyris, Anup Pillai, Arjun Sharma, Ali Bin Syed, et al. "Chest X-ray Report Generation through Fine-Grained Label Learning". In: *MICCAI*. Springer. 2020, pp. 561–571.

[191] Hao Tan and Mohit Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *CoRR* abs/1908.07490 (2019).

[192] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[193] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. "Interactive and Explainable Region-guided Radiology Report Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 7433–7442.

[194] Eric J Topol. "High-performance medicine: the convergence of human and artificial intelligence". In: *Nature medicine* 25.1 (2019), pp. 44–56.

[195] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv preprint arXiv:2302.13971* (2023).

[196] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *JMLR* 9.11 (2008).

[197] Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. "Open-ended medical visual question answering through prefix tuning of language models". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 726–736.

[198] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[199] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

[200] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. "Graph Attention Networks". In: *Proceedings of the International Conference on Learning Representations*. 2018.

[201] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164.

[202] Minh H. Vu, R. Sznitman, T. Nyholm, and T. Löfstedt. "Ensemble of Streamlined Bilinear Visual Question Answering Models for the ImageCLEF 2019 Challenge in the Medical Domain". In: (2019).

[203] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. "Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation". In: *Machine Learning for Healthcare Conference*. PMLR. 2021, pp. 755–769.

[204] Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. "An Inclusive Task-Aware Framework for Radiology Report Generation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th Inter-*

national Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII. Springer. 2022, pp. 568–577.

[205] Linda Wang, Zhong Qiu Lin, and Alexander Wong. "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images". In: *Scientific Reports* 10.1 (Nov. 2020).

[206] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 23318–23340.

[207] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. "Chatcad: Interactive computer-aided diagnosis on medical image using large language models". In: *arXiv preprint arXiv:2302.07257* (2023).

[208] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

[209] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *The Eleventh International Conference on Learning Representations*. 2023.

[210] Yuhao Wang, Kai Wang, Xiaohong Liu, Tianrun Gao, Jingyue Zhang, and Guangyu Wang. "Self adaptive global-local feature enhancement for radiology report generation". In: *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2023, pp. 2275–2279.

[211] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. "METransformer: Radiology Report Generation by Transformer With Multiple Learnable Expert Tokens". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 11558–11567.

[212] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. "A self-boosting framework for automated radiographic report generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2433–2442.

[213] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[214] Joy T Wu, Nkechinyere Nneka Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, William G Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. "Chest ImaGenome Dataset for Clinical Reasoning". In: *NeurIPS: Datasets and Benchmarks Track (Round 2)*. 2021.

[215] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016).

[216] Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Attila Kiraly, Sahar Kazemzadeh, Zakkai Melamed, et al. "ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders". In: *arXiv preprint arXiv:2308.01317* (2023).

[217] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. "Multimodal recurrent model with attention for automated ra-

diology report generation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer. 2018, pp. 457–466.

[218] An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. "Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4009–4015.

[219] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. "Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domain". In: CEUR Workshop Proceedings 2380 (2019). Ed. by Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller.

[220] 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". In: *arXiv preprint arXiv:2405.03162* (2024).

[221] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. "Knowledge Matters: Chest Radiology Report Generation with General and Specific Knowledge". In: *Medical Image Analysis* (2022), p. 102510.

[222] Linli Yao, Weiying Wang, and Qin Jin. "Image difference captioning with pre-training and contrastive learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 3. 2022, pp. 3108–3116.

[223] Shunyu Yao, Jeffrey Zhao, Dian Yu, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. "ReAct: Synergizing Reasoning and Acting in Language Models". In: *NeurIPS 2022 Foundation Models for Decision Making Workshop*. 2022.

[224] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. "Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer. 2021, pp. 72–82.

[225] Jingyi You, Dongyuan Li, Manabu Okumura, and Kenji Suzuki. "JPG - Jointly Learn to Align: Automated Disease Prediction and Radiology Report Generation". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 5989–6001.

[226] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. "Evaluating progress in automatic chest x-ray radiology report generation". In: *Patterns* 4.9 (2023).

[227] Ke Yu, Shantanu Ghosh, Zhexiong Liu, Christopher Deible, and Kayhan Batmanghelich. "Anatomy-Guided Weakly-Supervised Abnormality Localization in Chest X-rays". In: *MICCAI*. Springer. 2022.

[228] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer. 2019, pp. 721–729.

[229] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. "Vinvl: Revisiting visual representations in vision-language models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5579–5588.

[230] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. "Pmc-vqa: Visual instruction tuning for medical visual question answering". In: *arXiv preprint arXiv:2305.10415* (2023).

[231] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. "BioWordVec, improving biomedical word embeddings with subword information and MeSH". In: *Scientific data* 6.1 (2019), p. 52.

[232] Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. "Multimodal Chain-of-Thought Reasoning in Language Models". In: *Transactions on Machine Learning Research* (2024).

[233] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. "Simple Baseline for Visual Question Answering". In: *CoRR* abs/1512.02167 (2015).

[234] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models". In: *The Eleventh International Conference on Learning Representations*. 2023.

[235] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. "Unified Vision-Language Pre-Training for Image Captioning and VQA". In: *CoRR* abs/1909.11059 (2019).

[236] Yangyang Zhou, X. Kang, and F. Ren. "Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering". In: (2018).

[237] Yangyang Zhou, X. Kang, and F. Ren. "TUA1 at ImageCLEF 2019 VQA-Med: a Classification and Generation Model based on Transfer Learning". In: (2019).

[238] Xiongfeng Zhu, Shumao Pang, Xiaoxuan Zhang, Junzhang Huang, Lei Zhao, Kai Tang, and Qianjin Feng. “PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization”. In: *CMIG* 102 (2022), p. 102137.

[239] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. “Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books”. In: *CoRR* abs/1506.06724 (2015).

[240] Sebastian Ziegelmayer, Alexander W Marka, Nicolas Lenhart, Nadja Nehls, Stefan Reischl, Felix Harder, Andreas Sauter, Marcus Makowski, Markus Graf, and Joshua Gawlitza. “Evaluation of GPT-4’s Chest X-Ray Impression Generation: A Reader Study on Performance and Perception”. In: *J Med Internet Res* 25 (Dec. 2023), e50865.