EXTRACTING AND INCORPORATING CLINICAL INFORMATION FOR RADIOLOGY REPORT GENERATION

WENJUN HOU

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
Department of Computing

Extracting and Incorporating Clinical Information for Radiology Report Generation

Wenjun Hou

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
May 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Wenjun Hou

Abstract

Automated interpretation of medical images is essential in modern healthcare, particularly with the daily growing volume of medical imaging data. Among various imaging types, chest X-ray (CXR) is one of the most widely used modalities, and a key application of this interpretation is radiology report generation (RRG), which aims to produce free-text descriptions of relevant findings in CXR images. These findings may include anatomical structures, pathological conditions, or other significant observations. However, analyzing CXR images requires highly specialized domain knowledge to understand and interpret both the visual content and the clinical context of a medical case. Writing radiology reports can be time-consuming, often requiring considerable effort from radiologists, even for experienced professionals. Consequently, automating RRG has garnered significant interest from the research community due to its potential to alleviate radiologists' workload and expedite the diagnostic process. Existing RRG approaches typically process a CXR as input and employ an auto-regressive decoding strategy to generate reports sequentially from left to right. However, these methods often exhibit limited clinical accuracy, as they fail to adequately exploit and incorporate relevant clinical information, such as observations, disease progression, or relevant attributes. It is essential to properly extract and integrate diverse information sources, thereby enhancing the quality and utility of automated radiology reports.

In this thesis, we aim to extract and incorporate clinical information for radiology report generation, where different sources of information are effectively utilized to

improve the accuracy of generated reports. In particular, we identify three main research problems: (1) How to improve the disease/observation accuracy of generated reports given CXR images, especially when (large) language models can produce highly readable and coherent clinical texts? (2) How to properly model the attributes of diseases/observations that reflect both spatial characteristics and temporal progression, given sequential CXRs? (3) How to regulate a radiology report generation model to produce consistent reports at the attribute-level when semantically equivalent radiological studies are provided as input? Based on the categories of work carried out, this thesis is structured into three parts.

The first part of our work (Works 1 and 2) focuses on improving observation accuracy (problem 1). Observations represent high-level clinical information of CXRs, and enhancing this accuracy requires effective visual understanding and domain knowledge. To achieve this, we first construct an observation-specific graph from radiology reports, including three levels of nodes: observations, n-grams, and tokens. We then propose an observation-guided approach, ORGan, which first extracts observations from CXRs and then selects relevant information from the graph to enhance report generation. Building upon this, we further enhance clinical accuracy by leveraging large language models (LLMs), given their strong capabilities across various domains. However, LLMs still exhibit knowledge gaps when analyzing CXR studies, particularly complex cases. To address this, we introduce Radar, a method that first assesses and refines the knowledge already acquired by LLMs based on extracted observations, and then injects supplementary knowledge to complement the learned information. Extensive experiments demonstrate that our proposed methods significantly improve observation-level accuracy in radiology report generation.

The second part of our work (Work 3) addresses problem 2, which involves both incorporating prior study information and effectively integrating relevant attributes to generate spatiotemporally precise reports. To achieve this, we categorize attributes from sequential radiology reports into two types: spatial and temporal. Since these

attributes are closely linked to observations and disease progression, we construct a progression graph and propose a framework called RECAP. RECAP leverages prior CXR studies as additional input and reasons over the progression graph to accurately select relevant attributes, thereby enhancing radiology report generation. Extensive experiments demonstrate that our framework outperforms existing baselines in attribute modeling, highlighting its effectiveness in improving radiology report generation.

In the third part (Work 4), we address problme 3 by introducing two metrics to quantify inter-report consistency and developing a lesion-aware mixup for consistent radiology report generation. Building on extracted observation- and progression-aware attributes, we propose a framework called ICON, which models such consistency using regional information from CXRs. Given an X-ray, our approach extracts lesions and retrieves similar cases for mixup. The model is then trained to align shared representations of mixed lesions with relevant attributes, enabling ICON to effectively enhance inter-report consistency. Extensive experiments validate the effectiveness of our framework, demonstrating its ability to improve consistency in radiology report generation.

In summary, this thesis presents a comprehensive study of radiology report generation, advancing factual accuracy through the integration of clinical information. Our findings demonstrate the effectiveness of the proposed approaches, highlighting their significant potential to enhance medical image interpretation and support real-world diagnostic workflows.

Publications Arising from the Thesis

1. **Wenjun Hou**, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu, "ORGan: Observation-Guided Radiology Report Generation via Tree Reasoning", in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8108–8122, 2023.

2. **Wenjun Hou**, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu, "Recap: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning", in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2134–2147, 2023.

3. **Wenjun Hou**, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiang Liu, "ICon: Improving Inter-Report Consistency in Radiology Report Generation via Lesion-aware Mixup Augmentation", in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9043–9056, 2024.

4. **Wenjun Hou**, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. "Radar: Enhancing Radiology Report Generation with Supplementary Knowledge Injection", in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26366–26381, 2025.

Acknowledgments

As my PhD journey comes to an end, I reflect on what has been one of the most challenging yet rewarding experiences of my life.

First and foremost, I would like to express my deepest gratitude to my two supervisors, Prof. Jiang Liu from the Southern University of Science and Technology (SUSTech) and Prof. Maggie Wenjie Li from PolyU. I am truly grateful for the opportunity to learn from them, and for their invaluable guidance, patience, and unwavering support throughout my PhD studies.

I would also like to extend my heartfelt thanks to my lab members and collaborators: Kaishuai Xu, Yi Cheng, Jian Wang, Feiteng Mu, Shichao Sun, Yongqi Li, Jiashuo Wang, Wenge Liu, Dongding Lin, Chak Tou Leong, Heming Xia, Ruifeng Yuan, and many others from PolyU, as well as Yan Hu, Luoying Hao, Zongxi Qiu, Jiaqi Zhang, and colleagues from SUSTech. Their insightful feedback and suggestions have been instrumental in shaping my research.

Lastly, I am deeply appreciative of the unwavering support, encouragement, and love from my older sister and my parents. Their belief in me has been my greatest source of strength throughout this journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Medical images are visual representations of the internal structures of the human body, captured using various imaging modalities such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound. In particular, chest X-rays (CXRs) [117] offer detailed visualization of the lungs, airways, heart, diaphragm, and bones, by directing a controlled beam of X-rays toward the chest and capturing the radiation that passes through the body to create a two-dimensional image. As the most widely used medical imaging modality worldwide, chest radiography plays a crucial role in routine screenings, emergency diagnostics, and disease management. Its non-invasive nature and cost-effectiveness make it particularly valuable for early disease detection and widespread clinical applications. Despite its significance in modern healthcare, the increasing demand for chest X-ray interpretation has outpaced the available workforce, leading to a global shortage of radiologists [143, 77, 105]. This challenge is even more pronounced in underdeveloped areas, where the scarcity of trained specialists exacerbates delays in diagnosis and patient care [144]. The growing disparity between demand and availability has raised concerns about increased physician workload,

diagnostic inefficiencies, and potential impacts on patient outcomes. To address this challenge, AI-assisted medical intelligence systems [150, 167] have gained significant attention from research and industrial communities for their ability to automate and enhance clinical workflows. Among various medical image analysis advancements, AI-driven medical report generation has emerged as a transformative solution, streamlining the interpretation process and improving communication between radiologists and referring physicians. These reports provide systematic and standardized assessments of imaging studies, facilitating accurate diagnoses, optimized treatment planning, and more efficient patient management. Over the years, producing medical reports, for radiology and other modalities, has evolved from traditional handwritten or dictated formats [115, 47] to sophisticated digital and AI-assisted systems [153, 219, 218, 73, 72], marking a significant leap toward more efficient and scalable medical imaging practices. In the past few years, numerous studies [119, 177] have focused on automating the report generation process using computer vision (CV) and natural language processing (NLP) techniques. As a result, the quality of generated reports, in terms of fluency, coherence, and clinical accuracy, has seen significant improvement. Despite this significant progress, the performance showcased by radiology report generation methods is still far behind the expertise of radiologists, posing a major obstacle to their practical deployment in real-world clinical settings. Specifically, [205] identified six common error categories in radiology report generation, including false predictions of findings and omissions of comparative descriptions that indicate changes from prior studies. These errors can be attributed to several underlying factors. On the one hand, unlike general image captions [95, 2], radiology reports are typically longer and consist of multiple narrative sentences that describe both normal and abnormal findings in CXRs. On the other hand, abnormal regions and relevant lesions in grayscale CXRs often exhibit subtle, highly patterned features and are sometimes small in size, making them difficult to detect using conventional image captioning techniques. Consequently, accurately identifying anatomical regions, extracting clinically relevant observations, and effectively incorporating this information into the generation process are essential

for producing factually correct radiology reports.

Radiology report generation is primarily categorized into two directions: retrieval-based methods and auto-regressive methods. Retrieval-based approaches [34, 200] leverage expert-written reports as a reference database, ensuring outputs adhere to clinical standards. In contrast, auto-regressive methods [73, 21] generally achieve better performance by generating reports dynamically. However, since diseases and abnormal findings occur less frequently than in normal cases, direct image-to-text learning signals are often sparse. This sparsity makes it challenging for AI systems to produce clinically accurate reports. Existing research has introduced specialized knowledge representations and modular architectures designed to capture complex patterns in medical samples, thereby enhancing radiology report generation. However, inherent challenges persist due to the complexity of medical imaging and the nuanced language required for precise clinical descriptions. Given radiology report generation involves two key steps, i.e., visual understanding and language generation, addressing these challenges requires more effective extraction of clinical information (e.g., observations, attributes, and diseases progressions) from medical images and the seamless incorporation of this information into the generated reports. Therefore, accurately modeling clinical information is essential for producing accurate and coherent radiology reports.

Overall, this thesis aims to extract and incorporate patient-specific clinical information and enhance the performance of radiology report generation. We believe that our explored research problems are of practical significance and hold academic value, contributing to the field of medical image interpretation. In the following sections, we will introduce the key research challenges addressed in this thesis, followed by a comprehensive overview of the work conducted.

1.2 Research Problems

The primary objective of this thesis is to extract and incorporate patient-specific clinical information for accurate, precise, and consistent radiology report generation. In pursuit of this goal, we aim to address the following key research challenges:

- Problem 1: How to improve the disease/observation accuracy of generated reports given CXR images, especially when (large) language models (LLMs) can produce highly readable and coherent clinical texts?
- Problem 2: How to properly model the attributes of diseases/observations that reflect both spatial characteristics and temporal progression, given sequential CXRs?
- Problem 3: How to regulate a radiology report generation model to produce consistent reports at the attribute-level when semantically equivalent radiological studies are provided as input?

The first problem primarily concerns observation-level accuracy, evaluating how well the model's generated reports capture the observations present in the reference reports. This issue is particularly critical, as the overall quality and clinical reliability of a radiology report heavily depend on its accuracy. Effectively addressing this challenge lays the groundwork for subsequent tasks, such as attribute modeling and ensuring report consistency. However, several challenges arise when tackling this problem. Firstly, conventional image captioning methods, as well as other auto-regressive language generation models designed for radiology report generation, often struggle to identify observations in CXRs accurately. A key reason is that these models typically generate reports primarily based on the CXRs, without incorporating additional clinical information. As a result, supervision signals for specific observations are often sparse. Secondly, due to the free-text nature of radiology reports, a single

observation can be expressed in various ways (e.g., "*The heart size is enlarged.*" or "*This is mild cardiomegaly.*"). It is challenging for models to recognize such variations as equivalent. Lastly, although radiology report generation models often demonstrate strong fluency and coherence, they frequently overlook abnormalities due to a lack of domain knowledge that could otherwise enhance their diagnostic accuracy.

The second problem mainly involves modeling the attributes of observations (e.g., status, location, size, severity, and progression), which requires the model to generate both coarse-grained observations and their corresponding fine-grained descriptions accurately. A closer examination of the free-text descriptions of observations reveals that each observation is associated with specific attributes. Improper attribute assignment and inadequate spatiotemporal modeling can lead to clinical errors and hallucinations, increasing the time required for proofreading and correction of the generated reports. A critical aspect in addressing this problem lies in effectively associating observations with their corresponding attributes and accurately modeling disease progression. Learning these relationships from image-report pairs is particularly difficult due to the sparse supervision signals available in medical datasets. There are two primary challenges associated with this problem. One major challenge is that sequential CXR studies of a patient are temporally grounded, as follow-up studies rely on prior clinical history, and radiology reports reflect changes over time. However, many previous approaches have ignored the history, resulting in hallucinations. Another challenge is that the attributes of observations are often missing or inaccurately captured, due to the complexity of observations and the evolving nature of longitudinal medical data.

The third problem pertains to inter-report consistency in radiology report generation methods, i.e., when two semantically equivalent studies are provided as input, a robust and reliable model should generate reports that are consistent in attributes of observations, reflecting the similarity of the underlying findings. There are mainly two challenges in this problem. On the one hand, inter-report consistency remains

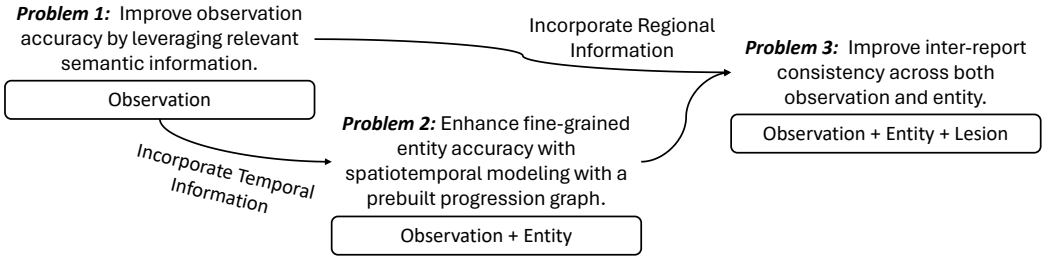


Figure 1.1: Relationships among three research problems.

unexplored, and it is crucial to properly quantify this aspect. On the other hand, since one study contains various information (e.g., prior/current CXRs and prior reports), handling this inherent complex structure is difficult. As a result, enhancing inter-report consistency requires effective extraction of semantically equivalent information between two studies for fine-grained attribute alignment.

These three research problems are closely related and should be addressed sequentially. Research problem 1 focuses on observation-level accuracy, laying the foundation for clinically accurate RRG models. Building on problem 1, problem 2 examines RRG at the entity level, taking a deeper perspective. It goes a step further by targeting the attributes of observations, thereby addressing a finer-grained aspect. Extending this, problem 3 focuses on inter-report consistency across both observations and entities, taking another step toward capturing fine-grained information. We displayed the relationships among these three research problems in Figure 1.1.

1.3 Research Overview and Contributions

In this thesis, we aim to address the challenges of enhancing the factual accuracy of radiology report generation by extracting and incorporating relevant clinical information. Based on three research problems, the studies in this thesis are organized into

Research Problem	Research Work	Publication Venue
Problem 1	Work 1: ORGAN: Observation-Guided Radiology Report Generation via Tree-Reasoning	ACL 2023 [57]
	Work 2: RADAR: Enhancing Radiology Report Generation with Supplementary Knowledge Injection	ACL 2025 [55]
Problem 2	Work 3: RECAP: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning	Findings of EMNLP 2023 [56]
Problem 3	Work 4: ICON: Improving Inter-Report Consistency in Radiology Report Generation via Lesion-aware Mixup Augmentation	Findings of EMNLP 2024 [54]

Table 1.1: Overview of Research Work in this Thesis.

three parts, and the overview of these works is displayed in Table 1.1. Specifically, the first part concentrates on enhancing the clinical accuracy of generated reports by introducing observation-relevant information. The second part addresses the modeling of observation attributes and disease progression in sequential CXR studies. The third part explores methods for improving the inter-report consistency using lesion-attribute information, ensuring that semantically equivalent CXRs yield consistent report outputs. For problem 1, we first develop an observation-guided report generation model to lay the foundation (work 1), and then propose an observation-aware knowledge-enhanced model to effectively address this problem (work 2). For problem 2, we incorporate historical records and model spatiotemporal attributes to achieve accurate and precise report generation (work 3). For problem 3, we first quantify inter-report consistency and then propose a lesion-aware mixup method to enhance such consistency in the generated outputs (work 4).

Work 1 & 2: Observation-aware Radiology Report Generation

To enhance clinical accuracy of radiology report generation, we propose leveraging clinical information extracted from CXRs, specifically observation-level information. This type of information provides a high-level abstraction of CXRs and encodes rich semantic content. Our primary research objective is to extract clinically relevant observations from CXRs and effectively utilize their semantic representations to generate more accurate and informative radiology reports. In work 1, we propose an observation-guided radiology report generation framework (ORGan), which fully exploits observation-relevant information. The framework first extracts observations from input CXRs and then generates corresponding reports using a pre-constructed observation graph. In particular, this graph comprises multi-level nodes, i.e., observations, n-grams, and tokens, to represent hierarchical semantic information. By learning to select and integrate relevant nodes from the graph, ORGan can effectively translate observations into coherent and clinically meaningful reports.

Building upon ORGan, we further enhance clinical accuracy by leveraging LLMs, given their strong capabilities across various domains. However, LLMs still exhibit knowledge gaps when analyzing CXR studies, especially in complex cases. Additionally, prior knowledge-enhanced methods often overlook the knowledge already embedded in LLMs, resulting in redundant information. To address these issues, in work 2 we introduce Radar, a method that injects supplementary knowledge based on observation information. Radar first assesses and refines the internal knowledge of LLMs using observations extracted by an expert model, then augments this with external information to generate clinically accurate radiology reports.

Contributions. In work 1, we introduce observation information as a guiding signal to enhance radiology report generation and propose ORGan. To fully exploit this high-level semantic information and improve surface realization for observations, we construct a three-level observation graph from the training corpus, consisting of

observations, n-grams, and tokens. We then perform hierarchical reasoning over this graph to dynamically select observation-relevant content. Extensive experiments on two publicly available benchmarks demonstrate the effectiveness of our model.

In work 2, we leverage LLMs for report generation and explore effective knowledge integration strategies to supplement the existing knowledge of a model, aiming to further address research problem 1. We propose a novel supplementary knowledge injection approach, RADAR, which first extracts the knowledge already learned by an LLM and then retrieves additional observation-aware knowledge to complement it. The internal knowledge and retrieved external information are collectively leveraged to enhance radiology report generation. We conduct extensive experiments on three publicly available datasets, and the results demonstrate the effectiveness of our approach.

Work 3: Spatiotemporally Precise Radiology Report Generation

Our previous studies focus on observation extraction, incorporation, and its surface realization in radiology report generation, which significantly improve the clinical accuracy of the generated outputs. Building on this, we investigate a finer-grained aspect of radiology reports, namely, the attributes of observations (e.g., *small pleural effusion*). Previous radiology report generation models often overlook the spatial and temporal attributes of observations, leading to clinically incorrect reports at the entity level. Our research aims to introduce fine-grained clinical information for spatiotemporally precise report generation. Recall that radiology reports serve as narrative descriptions of CXRs, and these descriptions are typically modified by positional or status-related attributes. Moreover, when a CXR study is a follow-up record, the report should reflect changes in observations across longitudinal CXRs. With this in mind, we propose a radiology report generation model with dynamic disease progression reasoning (RECAP). RECAP first extracts clinical information (i.e.,

observations and progressions) from the given sequential CXRs, and then incorporates this information for report generation. To effectively convert this information in the generated report, we first construct a disease progression graph based on the training corpus, where nodes represent observations and their attributes, and edges represent progression relationships. RECAP then performs reasoning over this graph to dynamically select the most relevant attributes that best describe the observations and their temporal changes.

Contributions. In work 3, we propose RECAP, a model that captures both spatial and temporal information to generate precise and accurate free-text reports. To achieve fine-grained attribute modeling, we construct a disease progression graph that incorporates both observations and positional or severity-quantifying attributes, and then we devise a reasoning mechanism that selects relevant attributes for precise report generation. Extensive experiments on two datasets validate the effectiveness of our model in terms of attributes and temporal modeling.

Work 4: Consistent Radiology Report Generation

Our previous studies focused on enhancing clinical accuracy by extracting and incorporating relevant clinical information (e.g., observations and progressions) into the report generation process. Based on this foundation, we now address another critical aspect: *inter-report consistency*. This property requires a report generation model to produce accurate and consistent outputs when provided with semantically equivalent CXRs as input. While clinical accuracy ensures the quality of individual generated outputs, inter-report consistency goes a step further by evaluating quality across different similar studies, contributing to the development of a more reliable and trustworthy system. To achieve this goal, we address this largely unexplored aspect by introducing two quantitative metrics for assessing inter-report consistency. We then propose a report generation model with regional feature integration, called ICON,

which extracts lesions from CXRs and learns to align lesions with their attributes. To further improve consistency, we augment the extracted lesions with retrieved lesions from similar cases, enabling the model to align similar features with shared attributes more effectively.

Contributions. In work 4, we introduce inter-report consistency as a critical aspect of radiology report generation and propose two novel metrics to measure it. We develop ICon, a model that captures abnormalities at the region level while requiring only coarse-grained image labels for training. With lesion-aware mixup augmentation, ICon effectively aligns similar lesions with shared attributes, thereby enhancing inter-report consistency. Extensive experiments confirm the effectiveness of our model in improving both consistency and accuracy in generated reports.

1.4 Structure of Thesis

This thesis is organized as follows to provide a comprehensive overview:

- Chapter 1 begins by introducing the background of research in radiology report generation, highlighting existing challenges and the motivations behind the problems explored in this work. It then outlines the three key research problems, offers an overview of the study, and presents the main contributions of this thesis.

- Chapter 2 presents a comprehensive literature review, covering medical report generation approaches, evaluation methods and datasets, as well as various technologies introduced for medical image analysis.

- Chapter 3 focuses on extracting and incorporating observations for radiology report generation, and proposes a two-stage observation-guided framework, ORGan. The first stage generates observation plans, while the second stage

converts these observations using pre-constructed knowledge.

- Chapter 4 further enhances the clinical accuracy by leveraging LLMs and introduces a report generation framework with supplementary knowledge injection (RADAR). Building upon the framework proposed in the previous chapter, this work integrates both the model's internal information and externally retrieved knowledge for report generation.

- Chapter 5 explores finer-grained modeling at the entity level and presents a spatiotemporally precise report generation method (RECAP). This approach extracts observations and disease progression from CXRs and incorporates attribute-level information during the generation process, resulting in more precise and clinically accurate radiology reports. In contrast to the previous two chapters, this work emphasizes entity-level accuracy and progression modeling.

- Chapter 6 addresses another crucial aspect beyond clinical accuracy, i.e., *inter-report consistency*, which ensures that a radiology report generation model produces consistent reports for semantically equivalent cases. To evaluate this property, we introduce two quantitative consistency metrics and propose ICON, a framework that enhances inter-report consistency through lesion-aware mixup at the entity level.

- Chapter 7 summarizes the proposed approaches, key findings, and overall contributions of this thesis. It also discusses potential future research directions.

Chapter 2

Literature Review

This chapter provides a review of studies relevant to the research presented in this thesis. Our main objective is to offer a comprehensive overview of related work, with a particular focus on medical report generation and various technologies of medical image analysis. By summarizing these studies, we aim to derive valuable insights and build a solid foundation for our research.

2.1 Medical Report Generation Approaches

Automating the process of medical report generation can greatly reduce the heavy strain on radiologists and has high practical value in real-world environments. Similar to image captioning [170, 4], which involves generating descriptive textual summaries for images, medical report generation aims to produce free-text descriptions of findings based on given medical images. Many studies have employed various medical image analysis techniques to improve the quality of generated reports. Due to the variety of imaging modalities, previous research has primarily focused on CXR [28, 74], Computed Tomography (CT) scans [44], Whole-Slide images (WSI) [18, 42], Electroencephalograph (EEG) [11], Fundus Fluorescein Angiography (FFA) images [87],

and other types of fundus imaging [59], such as Optical Coherence Tomography (OCT) and Color Fundus Photography (CFP). In this section, we will first introduce different types of methods for radiology report generation, and then provide a brief summarizes of report generation approaches for other modalities.

2.1.1 RNN-Based Radiology Report Generation

Recurrent Neural Networks (RNNs) [147] and their variants, e.g., Long Short-Term Memory (LSTM) networks [53], have demonstrated strong language modeling capabilities. Consequently, RNN/LSTM-based architectures were widely adopted for image captioning and radiology report generation. However, due to the distinct characteristics of CXR studies, such as higher structure complexity and longer report lengths, directly extending conventional image captioning models to radiology report generation may result in suboptimal performance [170, 141, 21]. Furthermore, unlike image captioning, the focus of radiology report generation shifts from lexical similarity to clinical accuracy. Simply migrating existing models to report generation does not adequately address this challenge. To overcome this limitation, researchers have proposed various techniques and incorporated domain-relevant information into RNN-based models to enhance their clinical performance.

[153] presented a CNN-LSTM cascade model that first detects diseases using a CNN, and subsequently generates image annotations by producing a joint image/text context vector, achieving significant improvements in image annotation performance. Given the lengthy nature of radiology reports, many studies have adopted hierarchical LSTMs to capture both sentence-level and word-level semantic information. For instance, [73] introduced a multi-task learning framework that simultaneously predicts Medical Text Indexer (MTI) tags and generates reports using a hierarchical LSTM, thereby enhancing the quality of the generated outputs. Additionally, [208] proposed improved LSTM-based encoder-decoder models for radiology report generation by incorporating

multi-view images and medical concepts. To further improve accuracy, [202] introduced a hierarchical RNN with a topic matching mechanism, which facilitates better visual-textual alignment in the semantic space. To address the inherent data bias present in medical samples, [48] proposed a dual-word LSTM that generates normal and abnormal sentences separately, enabling the model to place greater focus on abnormalities in CXRs. [182] proposed a self-boosting framework that adopts image-text matching to enhance radiology report generation. In addition, incorporating domain knowledge has become a common strategy to bridge the knowledge gap in radiology report generation. [216] proposed leveraging knowledge graphs (KGs) to address this gap and introduced a novel evaluation metric based on KGs. To mitigate the visual-semantic gap, [201] proposed a framework that jointly encodes visual features and complementary semantic embeddings of medical tags, enabling the generation of more accurate and fine-grained reports. Another important research direction is applying Reinforcement Learning (RL) to optimize the generated outputs, as it provides flexible objective towards either lexical similarity or clinical accuracy. Inspired by the workflow of radiologists, [92] proposed leveraging manually curated templates and developed a hybrid retrieval-generation approach via RL to improve the selection of relevant information. Similarly, [72] proposed to exploit the structured information within report sections, and devised a cooperative multi-agent system that learns to capture key information through RL. While earlier studies primarily employed RL to enhance language fluency, [103] introduced a clinically coherent reward that improves clinical accuracy by maximizing the correlation of disease distributions between generated reports and ground truth. Similarly, [114] proposed leveraging CheXpert [67] as a source of clinical information to generate clinically coherent radiology reports.

2.1.2 Transformer-Based Radiology Report Generation

Since the Transformer architecture [168] has demonstrated remarkable performance across a wide range of tasks and domains, particularly in modeling long-range de-

pendencies, it has been widely adopted for medical report generation. Research in this area generally falls into three main categories: cross-modal alignment models, knowledge enhanced models, and clinical information guided models. Cross-modal alignment models focus on designing various alignment strategies and modules to strengthen the connection between visual and textual modalities, knowledge-enhanced models latter aim to incorporate domain-specific knowledge to bridge the existing knowledge gap in the medical domain, and clinical information guided models extract and incorporate clinical information from CXRs to enhance clinical accuracy.

Cross-modal Alignment Models

[21] proposed a memory-driven Transformer model, incorporating a memory module designed to retain key information throughout the generation process. Building on this work, [20] introduced a cross-modal memory mechanism that explicitly leverages cross-modal mappings to enhance radiology report generation. Further extending these approaches, [133] utilized RL to optimize the model toward specific evaluation metrics, aiming for improved performance. To fully utilize study-level information, [223] proposed a cross-supervision method that learns joint representations of images and reports. In addition to these methods, [102] proposed the contrastive attention model, comparing the given image with normal images to distill information. [91] introduced a multi-level cross-modal alignment framework that leverages various features to enhance report generation. While the aforementioned studies facilitate cross-modal alignment, they still face the challenge of severe data bias in the training set. To address this issue, [100] proposed a competence-based multimodal curriculum learning approach, which ranks training samples based on visual and textual difficulty, leading to improved performance. Additionally, [178] introduced a purely Transformer-based approach optimized with multi-criteria objectives, including multi-label disease classification, image-text matching, and weighted report generation, to mitigate this issue and further enhance cross-modal alignment. To generate informative content, [193] developed a

weakly-supervised contrastive learning method by identifying hard negative samples during training. [195] proposed a vision-language pretraining framework called Clinical-BERT for disease diagnosis and report generation, with masked language modeling objective [30].

Knowledge Enhanced Models

Given the existing knowledge gap in the medical domain, enhancing radiology report generation with domain-specific knowledge represents a promising research direction. Additionally, the attention mechanism, which dynamically captures and encodes contextual information, allows for the flexible integration of diverse knowledge sources, thereby significantly improving the performance of radiology report generation models. For example, [81] proposed a knowledge-driven approach that decomposes the report generation process into two stages: abnormality learning and language generation. This decomposition significantly enhances overall performance. The authors constructed an abnormality graph that links visual features with the final radiology reports, substantially improving the detection and description of abnormal findings. Similarly, [199] proposed incorporating both general and specific knowledge extracted from RadGraph [69], which is a knowledge graph constructed from radiology reports, to improve the factual accuracy of generated reports. To emulate the diagnostic patterns of radiologists, who draw upon prior medical knowledge and experience, [101] proposed a method called PPKED to extract and integrate both posterior and prior knowledge from similar studies for improved report generation. While PPKED employs only limited knowledge, [181] introduced a broader set of medical concepts for semantic concept prediction and proposed a memory-augmented sparse attention mechanism to capture fine-grained visual features, thereby improving performance. Similarly, [112] introduced disease tags and medical concepts as knowledge sources, and devised a compatible decoder to fuse multi-view knowledge. To further bridge the domain knowledge gap, [198] proposed a knowledge-enhanced model with multi-

modal alignment, in which a knowledge base is learned during training to supply relevant contextual information. Unlike earlier approaches that rely on a single visual representation, [65] introduced U-Transformer, which effectively leverages multi-level visual features to enrich the report generation process. Additionally, they incorporated a symptom graph to inject domain knowledge during generation. As static KGs may be inadequate during the dynamic training process, [89] proposed a dynamic graph-enhanced contrastive learning framework to improve learned representations, thereby enhancing the quality and accuracy of generated reports. Inspired by multi-specialist consultation, [179] devised a diagnostic captioning framework using multiple learnable expert tokens. [196] introduced an attributed abnormality graph and adopted it contextual information for clinical accurate report generation.

Clinical Information Guided Models

Despite the effectiveness of external knowledge in providing contextual information to bridge the domain gap, it often requires additional effort to process, such as learning to align the knowledge with visual representations. Consequently, many researchers have proposed extracting clinical information directly from CXRs. For example, [118] proposed to use the entity matching score as a reward to encourage the model to generate factually complete and consistent radiology reports. Besides entities, observations play a crucial role in radiology reports, as they summarize high-level clinical content. In this regard, [125] proposed a planning-based approach that first generates a content plan and then employs RL to realize the plan. Subsequently, [57] proposed an observation-guided radiology report generation framework that produces observation plan and constructs observation graph to improve surface realization, enhancing the observation accuracy of generated reports. To enhance clinical understanding and disease identification, [71] proposed diagnosis-driven prompts extracted from CXRs for medical report generation. Some studies have focused on extracting anatomical regions or lesions from CXRs to enhance report generation. Specifically,

[162] proposed a region-guided framework that first identifies anatomical regions and then generates descriptions for each region to compose the final report. Similarly, [54] introduced a novel framework that extracts coarse-grained lesions and aligns them with corresponding attributes, aiming to improve inter-report consistency. However, several prior studies have overlooked the importance of historical records when generating CXR reports, which can result in inappropriate or inconsistent descriptions in follow-up examinations. As pointed out by [138], ignoring prior records can lead to hallucinations in generated reports. To address this issue, [10] leveraged temporal information for vision-language pretraining, demonstrating that incorporating prior records effectively enhances performance in radiology report generation. Given the importance of temporal information, [225] proposed a hierarchical longitudinal memory mechanism that encodes previous images and reports to pre-fill the reports of current CXRs, demonstrating promising results. Despite the effectiveness of previous research, most studies have modeled temporal information at a coarse-grained level while overlooking spatial information. To incorporate both spatial and temporal information at a finer-grained level, [56] proposed a spatiotemporal-aware model that captures these two aspects at the attribute level, resulting in more precise report generation.

2.1.3 VLM-based Radiology Report Generation

Foundation Models

Pretraining to learn good visual and textual representations is key to foundation models (FMs). Given the cost and labor-intensive nature of labeling medical images and texts, FMs hold significant value in the medical domain by enabling the exploitation of large-scale unlabeled data. Various research works have designed different learning objectives for training FMs and they are mainly categorized as CLIP-style contrastive learning [134], GPT-style next-token prediction [135], and their combinations.

Contrastive learning [61, 184, 204] has demonstrated strong zero-shot transfer ca-

pabilities across a variety of medical tasks. [12] developed a contrastive learning framework with improved text modeling, achieving better performance on various downstream tasks. Building on this, [10] incorporated prior images and reports of given inputs to learn temporal changes across sequential studies. While earlier methods typically ignored fine-grained details during contrastive learning, [173] first pretrained a phenotype-based CLIP model and then adopted its vision model for report generation, showing improved performance. Similarly, [224] proposed breaking down reports into concise descriptions during pretraining, enabling the model to capture fine-grained details more effectively during report generation. [97] introduced a multi-grained report generation framework with sentence-level contrastive learning, which boosts performance with dual decoders. [151] proposed a novel framework that combines both CLIP and diffusion model [14] representations for report generation. [63] leveraged contrastive learning to enhance image-report alignment for fine-grained representations.

Optimized through next-token prediction, GPT-style FMs are capable of performing diverse tasks in a unified language generation framework [128]. A medical FM typically undergoes two stages of training: pretraining and instruction tuning. Designed to perform diverse biomedical tasks, [212] developed a lightweight BiomedGPT pretrained on large-scale corpus, and this model demonstrates significant improvements over strong baselines. Inspired by LLaVA [104], [82] developed a similar model for biomedicine (LLaVA-Med), which undergoes two-stage training, i.e., medical concept alignment and medical instruction tuning. LLaVA-Med requires only one day of training, significantly reducing the cost of building a clinical assistant. Based on pretrained LLMs, [22] developed a general-purpose vision-language foundation model, i.e., CheXagent, for CXR interpretation, which can perform various types of analysis. [222] developed a generalist model, i.e., MedVerse, which supports multimodal inputs and is suited for clinical practice. [226] proposed Uni-Med that can perform six medical tasks using a mixture-of-experts module. [166] proposed Med-PaLM M and [197] proposed

Med-Gemini, both showing strong performance on various medical tasks. To bridge the medical knowledge gap in vision-language models, [121] proposed a framework called VILA-M3, which incorporates domain knowledge from expert models.

Fine-tuned LLMs

LLMs demonstrate strong capabilities across various tasks and domains. By accepting multimodal inputs and producing unified language sequences, they significantly reduce the need for additional model architecture designs. Despite their effectiveness, LLMs still face several challenges in generating clinically relevant outputs. A major reason is that these models are pretrained on diverse general-domain corpora and tend to perform poorly on medical-specific tasks. Consequently, many studies have focused on fine-tuning LLMs for the medical domain, particularly for radiology report generation. Furthermore, accepting medical images as inputs necessitates alignment between visual and textual modalities, which typically requires substantial computational resources and large-scale multi-modal corpora, particularly due to the complexity inherent in CXR studies. Many studies have made efforts to bridge this gap and benchmarked the performance of LLMs. As one of the earliest LLM-based models for report generation, [78] proposed a unified CXR-LLM capable of performing both CXR-to-report and report-to-CXR tasks through instruction tuning, thereby extending the capabilities of LLMs to a broader range of clinical scenarios. To incorporate pragmatic intents, [122] introduced indications as additional inputs for report generation. To unlock the potential of LLMs in radiology report generation, [180] developed a model based on a frozen Llama 2 [165] and fine-tuned it using Low-Rank Adaptation (LoRA) [58], resulting in a significant performance improvement. Similarly, [66] introduced MAIRA-1, which consists of a CXR-specific image encoder and a Vicuna-7B [220] language model, demonstrating impressive lexical and clinical performance. Building on this foundation, [9] presented MAIRA-2, which extends the capabilities of its predecessor by generating grounded findings, where textual descriptions are linked to

corresponding spatial annotations.

Despite recent progress, these LLMs still exhibit a domain gap, as they acquire relevant knowledge solely from training data. Therefore, incorporating external knowledge and context is a promising research direction to address this limitation. [194] proposed a two-step approach, which first extracts content defined in KGs from images, and then verbalizes these elements into coherent reports using GPT-3.5. Since visual–textual alignment typically requires large-scale image–text pairs, [99] proposed a bootstrapping method that generates synthetic data using an LLM and then feeds it back into the model for further training, resulting in significant performance improvements. To fully exploit visual–textual relationships, [19] proposed an adaptive patch–word matching model integrated with LLMs, enabling explainable cyclic image-to-report and report-to-image generation. Leveraging the context comprehension capabilities of LLMs, [106] introduced an in-context learning framework utilizing a multimodal contextual vector. Addressing the issue of hallucinations commonly produced by LLMs, [50] proposed a fact-checking mechanism for report generation models based on a query–code–update paradigm.

In addition, [174] conducted a comprehensive evaluation of various models on the CheXpert Plus dataset and proposed an LLM enhanced with multiple training strategies. To incorporate longitudinal patient information, [107] developed a historically constrained LLM that utilizes prior studies for temporally accurate report generation. Similarly, [215] developed Libra, aiming for sequential radiographs analysis. Recognizing the diversity in input formats, ranging from single-view CXR to multi-view CXRs with associated reports, [183] introduced a flexible framework capable of handling various input combinations. To improve alignment between visual and textual representations, [64] proposed a cross-modal alignment adapter to enhance the performance of Multimodal LLMs (MLLMs). Furthermore, [175] presented a memory-enhanced report generation model designed to better capture disease associations. Finally, [51] developed a preference learning approach to align generated reports with

radiologists' expectations.

2.1.4 Medical Report Generation for Other Modalities

As previously noted, several commonly used imaging modalities have also attracted research attention. Although different modalities and their corresponding reports exhibit distinct characteristics, many of the proposed approaches share common characteristics with those developed for CXRs. For CT scans, [45] proposed CT2Rep, which employs a novel autoregressive causal transformer alongside a cutting-edge 3D vision encoder. Inspired by [86], [221] devised an abnormality-aligned pretraining framework to capture abnormal information for accurate report generation. [186] introduced RadFM, a foundation model for interpreting both 2D and 3D medical images, capable of processing multiple imaging modalities as input. To further advance the application of LLMs in medical image analysis, [7] collected a large-scale 3D dataset and trained a model for various medical tasks, including report generation. [17] leveraged the advanced capabilities of LLMs for chest CT report generation. In addition, [23] incorporated region-level information into CT report generation by utilizing masks produced by a universal segmentation module. Beyond CT imaging, prior works have also explored fundus images. [87] introduced an explainable and reliable FFA report generation benchmark, where each sample consists of multiple images, and evaluated several baseline models on this benchmark. Given the complexity of FFA report samples, [88] proposed a knowledge-driven cross-modal Transformer to bridge the knowledge gap in the ophthalmic domain. Recognizing the effectiveness of pretraining in leveraging unlabeled data, [189] collected a multimodal dataset and developed a novel knowledge-enhanced pretraining model. [59] released a dataset containing both FFA and CFP images, and presented a deep neural network (DNN)-based module that predicts diseases and generates clinical descriptions. WSI is another important imaging modality. [219] developed a smantically and visuallly interpretable diagnosis network based on a pathology bladder cancer dataset. [18] collected and

released a pathology report generation dataset, and proposed a multiple instance generative model to produce reports for gigapixel WSIs. Besides, [161] trained a multimodal LLM to generate captions for WSIs and created a dataset containing over one million samples.

Although these modalities and their associated datasets share similarities with CXRs, there are several distinct differences. Firstly, CXRs are quick, low-cost imaging tools primarily used for chest-related examinations, with each study typically consisting of one or two images. In contrast, CT scans are high-resolution 3D radiographs composed of multiple images; WSIs capture entire pathology slides at microscopic resolution; and FFAs involve the injection of fluorescein dye, producing a series of images over time. These modalities have different structures, and the approaches developed for each modality exhibit distinct features. Thus, each requires specific designs based on its characteristics, resulting in different types of approaches compared to radiology report generation.

2.2 Medical Report Generation Datasets and Evaluation

2.2.1 Medical Report Generation Datasets

In the early stages of medical AI development, datasets for medical report generation were relatively scarce for several reasons. In specific, patient data is highly sensitive, and protecting privacy is a legal and ethical necessity. As a result, hospitals and related institutions are often cautious about sharing such data. The release of medical datasets typically requires a rigorous ethical review process, during which all data must be de-identified, which is a time-consuming and complex task. Furthermore, collecting medical samples is significantly more challenging than in other domains.

Unlike general datasets, where a single source can contribute numerous examples, in the medical field, one patient often corresponds to just one case or study. Consequently, assembling large-scale datasets can take an long period. In this section, we provide a review of medical report generation datasets, and offer the statistics of these dataset in Table 2.1.

Dataset	Modality	#Image	#Report
IU X-RAY [28]	Chest X-ray	7,470	3,955
MIMIC-CXR [124]	Chest X-ray	377,110	276,778
MIMIC-ABN [124]	Chest X-ray	38,551	38,551
CHEXPERT PLUS [15]	Chest X-ray	223,462	187,711
PadChest [13]	Chest X-ray	160,868	22,710
CT-RATE [46]	CT Scans	50,188	25,692
M3D-Cap [7]	CT Scans	120,092	42,496
FFA-IR [87]	FFA	1,048,584	10,790
DEN [59]	CFP+FFA	15,709	15,709
WsiCaption [18]	WSI	1,041	1,041
PathGen [161]	WSI	1.6M	1.6M

Table 2.1: Overview of medical report generation datasets. The first four datasets are mainly adopted for experiments in this thesis.

Radiology Report Generation Datasets

The IU X-RAY dataset [28] is a publicly available resource that comprises 3,955 studies and 7,470 CXRs. Each study includes two CXRs captured from different perspectives, typically frontal and lateral views. Collected by Indiana University, this dataset provides corresponding radiology reports for each study, which are structured into multiple sections, including Indication, Findings, and Impression. In addition to these

sections, the authors also encoded the findings using MeSH terms [37] and annotated them with the MTI, thereby providing rich, structured metadata that facilitates more comprehensive analysis and downstream processing. Subsequently, the MIMIC-CXR dataset was released by [74], and it has become one of the most widely used datasets in the field. It contains 377,110 CXRs and 227,827 corresponding free-text radiology reports. All data has been thoroughly de-identified to ensure patient privacy. Unlike the IU X-Ray dataset, where each study consists of exactly two CXRs (frontal and lateral), a single study in MIMIC-CXR may include one or more CXRs, depending on the patient's clinical status. Moreover, the dataset includes both first-visit and follow-up cases, adding to its structural complexity compared to the IU X-Ray dataset. In addition to CXRs and reports, MIMIC-CXR dataset provides rich meta data as well as report annotations from CheXpert [67]. As a result, this dataset can support wide range of research in the medical field. Since normal observations dominate the content of the MIMIC-CXR dataset, [124] proposed a method to extract and cluster the abnormal findings, resulting in the creation of the MIMIC-ABN dataset. This curated dataset is significantly smaller than MIMIC-CXR, comprising 38,551 samples that focus specifically on abnormal cases. Recently, [15] released the CheXpert Plus dataset, which is currently the largest publicly available dataset of its kind. It serves as an enhanced version of the original CheXpert dataset and includes not only CXRs with corresponding radiology reports but also associated patient information. In addition, the authors have provided a rich set of supplementary resources, including observation annotations from CheXbert, entity and relation annotations from RadGraph, and a collection of pretrained models developed using this dataset. While the previously mentioned datasets contain radiology reports written in English, PadChest [13] is a large-scale dataset with reports written in Spanish. It comprises 109,931 studies and 168,861 CXRs, offering valuable linguistic and geographic diversity for multilingual and cross-lingual medical imaging research.

Other Medical Report Generation Datasets

Datasets for other imaging modalities have been collected in a manner similar to those used for radiology report generation. Broadly, there are three main types of medical report generation datasets beyond chest X-rays: those based on CT scans, CFP/FFA, and WSI. Among these, CT scans employ imaging technology similar to CXRs, utilizing X-ray radiation to produce cross-sectional images of the body. Specifically, [7] introduced a large-scale 3D multi-modal medical dataset that supports a variety of tasks, including visual question answering, report generation (M3D-Cap), and segmentation. The M3D-Cap subset contains 120,092 images paired with 42,496 radiology reports, making it a valuable resource for multi-modal learning in 3D medical imaging. In addition, the CT-RATE dataset [46] comprises chest CT volumes and corresponding reports. It includes 50,188 CT volumes derived from 25,692 imaging studies conducted on 21,304 patients, and is designed to support the development of generalist foundation models in medical imaging.

Fundus images (CFP and FFA) have also garnered significant attention in the medical imaging research community. [59] introduced the DEN dataset to support medical report generation for retinal imaging. This dataset includes both CFP and FFA images, comprising a total of 15,709 samples. Furthermore, [87] released the large-scale FFA-IR dataset, which contains over 1 million FFA images and 10,790 corresponding radiology reports. Notably, the FFA-IR dataset also provides lesion-level annotations, enhancing its utility for explainable and interpretable machine learning models. WSI report generation datasets are less acquirable compared to radiology report generation datasets. WisCaption [18] is one of the most widely used datasets for medical report generation based on WSI. It contains 1,041 high-resolution WSI samples paired with corresponding textual descriptions. In contrast, PathGen is a large-scale dataset that includes approximately 1.6 million automatically annotated image–caption pairs.

2.2.2 Medical Report Generation Evaluation

Evaluation plays a critical role in advancing the field of medical report generation. In its early stages, evaluation metrics were largely borrowed from the image captioning domain due to the shared characteristics between the two tasks. However, directly applying these metrics has proven to be insufficient for capturing the clinical relevance and complexity inherent in medical reports. As a result, evaluation strategies have evolved to better align with the specific requirements of this task. Currently, automatic evaluation metrics can be broadly categorized into three types: natural language generation (NLG) metrics, clinical metrics, and LLM-based metrics. In the following sections, we describe how each of these metric types assesses the quality of generated reports, taking into account the unique challenges and objectives of medical report generation.

NLG Metrics

Commonly used NLG metrics include BLEU [130], ROUGE [93], METEOR [8], CIDEr [169], and BERTScore [214], which evaluate the similarity between generated and reference reports by measuring lexical overlap and coverage. Specifically, BLEU was originally introduced for machine translation (MT) and remains one of the most widely used metrics. It primarily evaluates the n-gram precision between the generated text and reference text, assessing how many n-grams in the candidate output are also present in the reference. BLEU scores typically range from BLEU-1 (unigram precision) to BLEU-4 (four-gram precision), with BLEU-4 being the most commonly reported variant due to its balance between short- and long-range textual overlap. In contrast, ROUGE emphasizes n-gram recall rather than precision, and was originally developed for text summarization tasks. Among its various variants (i.e., ROUGE-1, ROUGE-2, and ROUGE-L), ROUGE-L is the most commonly reported metric in the medical report generation literature. It measures the longest common sequence between

the generated and reference texts, effectively capturing sequence-level similarity and reflecting the model's ability to preserve the structure and order of relevant content. METEOR, also designed for MT evaluation, aims to balance precision and recall. Unlike BLEU, METEOR incorporates linguistic features such as synonym matching, stemming, and paraphrase recognition, allowing for more flexible and semantically meaningful alignments between generated and reference texts. CIDEr, introduced for image captioning, evaluates TF-IDF-weighted n-gram similarity [136], capturing both term frequency and informativeness. By incorporating both term frequency and inverse document frequency, CIDEr emphasizes the informativeness of n-grams, rewarding content that is both relevant and specific to the context. Finally, BERTScore emphasizes semantic similarity by leveraging contextual embeddings from pretrained language models such as BERT [30]. BERTScore aligns tokens based on their semantic representations, enabling more nuanced evaluations of meaning preservation, even when surface forms differ.

Clinical Metrics

Beyond lexical similarity, factual accuracy is of greater importance in medical report generation, as it evaluates the clinical correctness of the information conveyed in the generated reports. In specific, medical abnormality terminology detection accuracy was first proposed in [92], aiming to evaluate the model's ability to identify clinically significant terms. It computes the average precision and average false positive for the 10 most frequent medical abnormality terminologies selected from medical reports. While this metric can reflect the quality of generated reports to some extent, it only covers limited clinical content. In contrast, Clinical Efficacy (CE) metrics [21] evaluate model performance based on its ability to accurately identify thoracic diseases and support devices. This evaluation is performed across 14 predefined clinical categories (e.g., *Cardiomegaly* and *Pneumothorax*), providing a more task-specific assessment of the model's performance. To support this evaluation, tools such as CheXpert [67] and

CheXbert [157] are widely used to extract clinically relevant observations from both reference and generated reports, where the former one is an automated rule-based labeler and the latter one is a BERT-based observation classifier. While CE metrics assess observation-level accuracy, they primarily capture coarse-grained aspects of model performance. To enable a more nuanced and clinically meaningful evaluation, finer-grained metrics are essential. [26] proposed leveraging KGs to assess generated radiology reports, and introduced RadGraph [69, 27] to assess entity-level accuracy. It captures both medical entities and their relations using a named entity recognizer [171], enabling structured comparisons between generated and reference reports. Although these automatic metrics provide valuable insights into clinical accuracy, they may fail to capture radiologists' preferences and deviate from real-world clinical practices. To bridge this gap, [205] proposed RadCliQ, a composite metric that integrates multiple automatic evaluation strategies and demonstrates a strong correlation with expert radiologist assessments, offering a more reliable proxy for human judgment in radiology report evaluation.

LLM-based Metrics

Recently, many studies proposed LLM-as-a-judge approaches [83], which leverage LLMs as evaluators. Given their strong capabilities in context comprehension, LLMs enable more flexible and nuanced evaluations across various dimensions, such as factual accuracy, coherence, and preference alignment in generated outputs. Several works have been developed to adopt LLMs as evaluators for medical report generation. For instance, [129] introduced GREEN, an LLM fine-tuned with preference data and validated by board-certified radiologists. The evaluations produced by GREEN are well-aligned with expert preferences, and the model is capable of identifying and explaining clinically significant errors in generated reports. This enables more interpretable and informative assessments compared to traditional NLG and clinical evaluation metrics. Similarly, [109] proposed MRScore, which demonstrates high

alignment with human judgment and achieves superior performance in model selection. Recognizing the criticality and assistive nature of clinical texts, [190] proposed a fine-grained evaluation framework, DocLen, which assesses three key dimensions: completeness, conciseness, and attribution. DocLen is compatible with both closed- and open-source LLMs, and exhibits strong alignment with expert preferences, offering a reliable approach for evaluating clinical language generation.

2.3 Medical Image Analysis

Medical image analysis aims to extract meaningful clinical information from medical images. Since medical reports are narrative descriptions based on these images, the extracted information is particularly relevant to the factuality of the reports. In this section, we briefly introduce three key technologies that are widely used and effective in medical report generation: image classification, object detection and image segmentation, and study retrieval.

2.3.1 Medical Image Classification

Medical image classification, which transforms medical images into predefined labels (e.g., diseases or clinical observations), has been developed to support clinical decision-making and streamline the clinical workflow [96]. Consequently, numerous studies have undertaken annotation efforts to facilitate such analyses, especially for CXRs [227, 127]. For example, [176] presented a chest X-ray database labeled with eight common thoracic diseases extracted via text mining, and evaluated several pretrained backbone models on this dataset. Building on this work, [137] introduced CheXNet, a deep learning model for pneumonia detection, which was shown to outperform practicing radiologists. Similarly, [152] collected and released the RSNA Pneumonia Detection Challenge Dataset, which comprises 30,000 frontal-view chest radiographs

annotated for pneumonia-related findings. Subsequently, [67] released a large-scale chest radiograph dataset annotated with 14 clinical observations, explicitly modeling the uncertainty inherent in radiographic interpretation. They also conducted extensive experiments to explore strategies for managing this uncertainty. These datasets have become the mainstream benchmarks for model development and evaluation [217, 62, 184, 204]. Earlier datasets mainly relied on automatic annotation methods and did not provide the precise locations of abnormalities. To address this limitation, [123] released a manually annotated dataset, curated by experienced radiologists, with detailed localization of abnormalities. In addition, various studies [80, 188, 52] proposed and devised convolutional neural networks (CNNs) for abnormality detection in CXR. Given the simplicity and effectiveness of CXR image classification, its structured outputs provide high-level semantic information that can serve as valuable guidance for generating accurate clinical reports.

The aforementioned studies primarily focus on spatial disease classification, which involves identifying abnormalities within medical images. Another important area of research is progression classification and prediction, which concentrates on analyzing how a disease evolves over time. [187] introduced the Chest ImaGenome dataset, which utilizes a scene graph data structure to represent anatomical-level disease progression, categorized as better, stable, or worsen. The dataset includes automatically generated annotations using an atlas-based bounding box detection pipeline and rule-based NLP tools. Building on this dataset, [206] developed a twin neural network to classify disease progression at the anatomical level. They employed a two-step weakly supervised learning strategy consisting of pretraining followed by fine-tuning. [36] developed a representation learning framework for localized disease progression classification using anatomical region features. Their findings show that even a relatively simple model can achieve competitive performance on this task. In the context of disease progression prediction, the task involves developing models that can forecast future disease states based on current medical images. [164] proposed a deep learning framework

for predicting disability progression using multi-modal MRI data. [38] proposed a framework for predicting the progression of diabetic retinopathy by leveraging temporal regional features extracted from retinal images.

2.3.2 Medical Object Detection and Image Segmentation

While medical image classification provides only image-level information, medical image analysis has advanced toward finer-grained tasks. These include object detection, object localization, and image segmentation, each offering more detailed and clinically meaningful insights. Commonly used models for medical object detection include Faster R-CNN [140] and RetinaNet [94]. For instance, [191] proposed an enhanced version of Faster R-CNN, while [146] utilized RetinaNet for lung nodule detection. As previously mentioned, the Chest ImaGenome dataset provides bounding box annotations for various anatomical structures, primarily intended to support clinical reasoning. Given its rich and detailed annotations, several studies have leveraged this dataset to train object detectors for extracting relevant clinical information [162, 40, 213].

Since obtaining object-level annotations is labor-intensive and requires domain expertise, weakly supervised localization, which relies only on image-level annotations, has emerged as a promising direction. A representative study is Grad-CAM [148]. Specifically, Grad-CAM utilizes the gradients of the target concept flowing into the final convolutional layer to produce a coarse localization map, highlighting important regions in the input image. This class activation map (CAM) can then be overlaid on the image to visualize the areas of interest or used to assist in drawing bounding boxes for the specific class. Due to its effectiveness and practicality, CAM has become a widely adopted technique for extracting relevant clinical information to enhance medical report generation [90, 172].

Another important and effective technique for extracting clinical information is medical image segmentation [6], which focuses on identifying objects and delineating their

boundaries to provide more detailed and precise information into medical images. In particular, one of the most commonly used models in the medical field is U-Net [142], which captures fine-grained details of medical images through its contracting and expansive paths. Although U-Net was originally designed for light microscopy images, many studies have adapted it and its variants to other imaging modalities, including CXRs. For instance, Gaal et al. [39] employed Attention U-Net, a variant of U-Net, for lung segmentation and introduced an adversarial scheme to further enhance segmentation performance. Similarly, [108] developed an improved U-Net architecture incorporating pretrained vision models, demonstrating considerable improvements and good robustness in the lung segmentation task. Given the effectiveness of medical image segmentation, several studies have adopted it to extract fine-grained features for medical report generation [41, 149].

2.3.3 Medical Study Retrieval

Medical study retrieval refers to the task of retrieving similar studies using a given study as a query. Since a medical study may comprise both medical images and an associated report, study retrieval encompasses image-to-image retrieval and cross-modal retrieval (e.g., image-to-text retrieval). Given the importance of medical knowledge in bridging the knowledge gap between visual data and clinical interpretation, retrieving similar studies as additional context can be beneficial for medical report generation. A promising direction is to leverage deep learning models as feature extractors for image-to-image retrieval. For instance, [154] first trained a CNN for disease classification with disease-related saliency maps and subsequently repurposed it to extract disease-relevant features for retrieval, resulting in more disease-aligned search outcomes. Furthermore, [43] investigated various feature representations extracted from CNNs and ViTs to assess their effectiveness in medical image retrieval tasks.

In terms of cross-modal retrieval, the advancements introduced by CLIP [134] have

significantly facilitated the development of knowledge retrieval by enabling effective alignment between visual and textual modalities. In this regard, various CLIP variants specified for the medical field. [61] proposed GLoRIA, a framework that integrates local and global representation learning, demonstrating promising performance in image-to-text retrieval tasks. Similarly, [217] introduced ConVIRT, a contrastive learning framework designed to learn visual representations from paired medical image-text data. Despite their effectiveness, existing contrastive learning approaches treat medical reports with similar semantics as negatives during training, which can lead to suboptimal performance. To address this issue, [184] proposed MedCLIP, introducing a learning objective that leverages unpaired data. They employed a medical knowledge extractor to mine positive pairs from other samples and incorporated knowledge-driven semantic similarity as additional supervision to guide the learning process. Similarly, [98] constructed inter-report similarity by leveraging representations obtained from BERT [30], aiming to explore the latent semantic structures within radiology reports. In addition, [204] introduced additional contrastive loss functions to independently learn image and text representations, aiming to address the problem of data scarcity. Their approach significantly improved performance in retrieval tasks.

Chapter 3

Observation-aware Radiology Report Generation: Observation Extraction and Incorporation

3.1 Introduction

Radiology reports, which contain the textual description for a set of radiographs, are critical in the process of medical diagnosis and treatment. Nevertheless, the interpretation of radiographs is very time-consuming, even for experienced radiologists. Due to its large potential to alleviate the strain on the healthcare workforce, automated radiology report generation [4, 141] has attracted increasing research attention.

One significant challenge of this task is improving the clinical accuracy of the generated reports, ensuring that all relevant findings are accurately identified and described. Many previous works proposed to solve this through planning-based generation by first concluding the major observations identified in the radiographs before the word-level realization [73, 203, 126, 125]. Despite their progress, these methods still struggle to maintain the cross-modal consistency between radiographs and reports. A significant

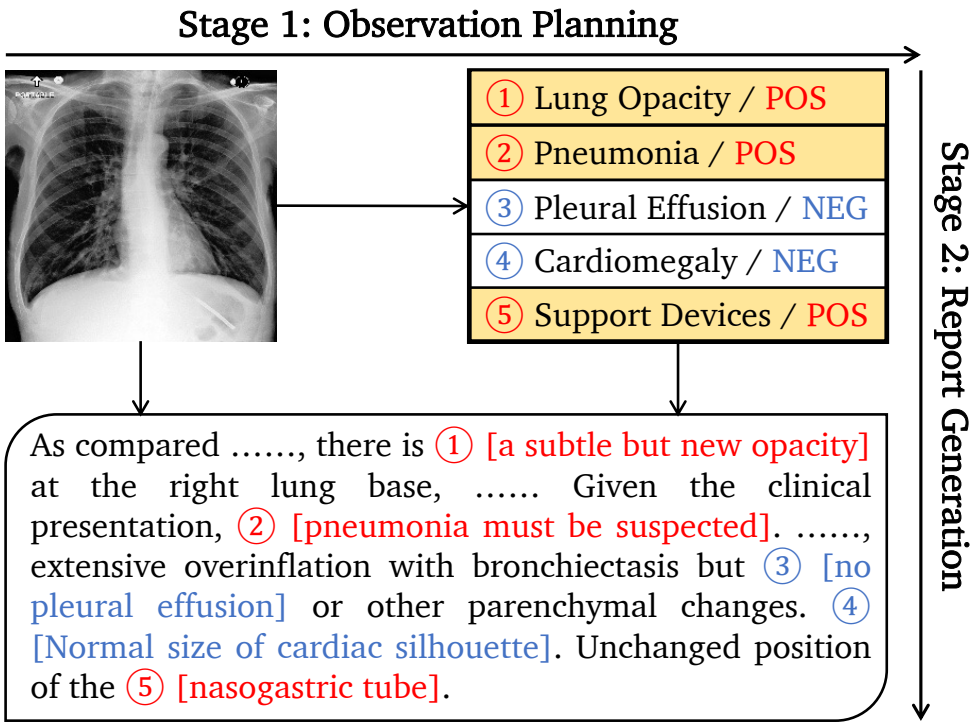


Figure 3.1: Our proposed framework contains two stages, including the observation planning stage and the report generation stage. Red color denotes positive observations, while Blue color denotes negative observations.

problem within these methods is that, in the stage of word-level generation, the semantic information of observations and radiographs is not fully utilized. They either generate the report only based on the high-level textual plan (i.e., major observations) or ignore the status of an observation (i.e., positive, negative, and uncertain), which is far from adequate. The observations contained in the high-level plan are extremely concise (e.g., *lung opacity*), while the final report needs to include more detailed information, such as the characteristics of the observation (e.g., *a subtle but new lung opacity*) and requires preliminary diagnosis inference based on the observation (e.g., lung infection must be suspected). In order to identify those detailed descriptions and clinical inferences about the observations, we need to further consider the image

information together with the textual plan, and stronger reasoning must be adopted during the generation process.

In this chapter, we propose ORGAN, an Observation-guided radiology Report GenerAtioN framework. Our framework mainly involves two stages, i.e., the observation planning and the report generation stages, as depicted in Figure 3.1. In the first stage, our framework produces the observation plan based on the given images, which includes the major findings from the radiographs and their statuses (i.e., positive, negative, and uncertain). In the second stage, we feed both images and the observation plan into a Transformer model to generate the report. Here, a tree reasoning mechanism is devised to enrich the concise observation plan precisely. Specifically, we construct a three-level observation graph, with the high-level observations as the first level, the observation-aware n-grams as the second level, and the specific tokens as the third level. These observation-aware n-grams capture different common descriptions of the observations and serve as the component of observation mentions. Then, we use the tree reasoning mechanism to capture observation-aware information by dynamically aggregating nodes in the graph.

Our main contributions can be summarized as follows:

- We propose an Observation-guided radiology Report GenerAtioN framework (ORGAN) that utilizes observation information to improve the clinical accuracy of generated free-text reports.

- To achieve better observation realization, we construct a three-level observation graph containing observations, n-grams, and tokens based on the training corpus. Then, we perform tree reasoning over the graph to dynamically select observation-relevant information.

- We conduct extensive experiments on two publicly available benchmarks, and experimental results demonstrate the effectiveness of our model. We also conduct

a detailed case analysis to illustrate the benefits of incorporating observation-related information.

3.2 Preliminary

3.2.1 Problem Formulation

Given an image X, the probability of a report $Y = \{y_1, \ldots, y_T\}$ is denoted as $p(Y|X)$. Our framework decomposes $p(Y|X)$ into two stages, where the first stage is observation planning, and the second stage is report generation. Specifically, observations of an image $Z = \{z_1, \ldots, z_L\}$ are firstly produced, modeled as $p(Z|X)$. Then, the report is generated based on the observation plan and the image, modeled as $p(Y|X, Z)$.

3.2.2 Observation Statistics

There are 14 categories of observations: *No Finding*, *Enlarged Cardiomediastinum*, *Cardiomegaly*, *Lung Lesion*, *Lung Opacity*, *Edema*, *Consolidation*, *Pneumonia*, *Atelectasis*, *Pneumothorax*, *Pleural Effusion*, *Pleural Other*, *Fracture*, and *Support Devices*. Table 3.1 lists the observation distributions annotated by CheXbert [157] in the train/valid/test split of the IU X-RAY and MIMIC-CXR datasets.

3.2.3 Observation Plan Extraction and Graph Construction

Observation Plan Extraction. There are two available tools for extracting observation labels from reports, which are CheXpert [67] and CheXbert [157]. We use CheXbert[1] instead of CheXpert because the former achieved better performance. To extract the observation plan of a given report, we first adopt the CheXbert to obtain

[1] https://github.com/stanfordmlgroup/CheXbert

#Observation	IU X-RAY		MIMIC-CXR	
	Positive	**Negative**	**Positive**	**Negative**
No Finding	744/108/318	1,325/188/272	64,677/514/229	206,133/1,616/3,629
Cardiomegaly	244/38/61	1,375/198/386	70,561/514/1,602	85,448/714/801
Pleural Effusion	60/13/15	1,559/230/452	56,972/477/1,379	170,989/1,310/1,763
Pneumothorax	9/2/5	1,528/231/449	8,707/62/106	190,356/1,495/2,338
Enlarged Card.	159/29/28	1,200/161/384	49,806/413/1,140	129,360/1,006/868
Consolidation	17/1/3	763/117/210	14,449/119/384	97,197/788/964
Lung Opacity	295/35/57	331/49/82	67,714/497/1,448	8,157/73/125
Fracture	84/6/15	137/22/50	11,070/59/232	9,632/72/53
Lung Lesion	85/14/17	92/10/30	11,717/123/300	1,972/21/11
Edema	28/2/7	119/17/31	33,034/257/899	51,639/409/669
Atelectasis	143/15/37	3/0/0	68,273/515/1,210	563/5/9
Support Devices	89/20/16	1/0/0	60,455/450/1,358	1,081/7/11
Pneumonia	20/2/1	68/9/25	23,945/184/503	21,976/165/411
Pleural Other	32/4/7	0/0/0	7,296/70/184	63/0/0

Table 3.1: Observation distribution across the train, validation, and test splits of the IU X-RAY and MIMIC-CXR datasets. Enlarged Card. denotes Enlarged Cardiomediastinum.

the observation labels within 14 categories $C = \{C_1, \ldots, C_{14}\}$ as indicated in [67]. More details about the distribution of observation can be found in Table 3.1. The label (or status) of each category belongs to *Present*, *Absent*, and *Uncertain*, except the *No Finding* category, which only belongs to *Present* and *Absent*. To simplify the observation plan and emphasize the abnormalities presented in a report, we regard *Present* and *Uncertain* as Positive and *Absent* as Negative. Then, observations are divided into a positive collection C/POS and a negative collection C/NEG, and each category with its corresponding label is then converted to its unique observation C_i/POS $\in C$/POS or C_i/NEG $\in C$/NEG, resulting in 28 observations. For example, as indicated in Figure 3.1, the report presents *Lung Opacity* while *Cardiomegaly* is

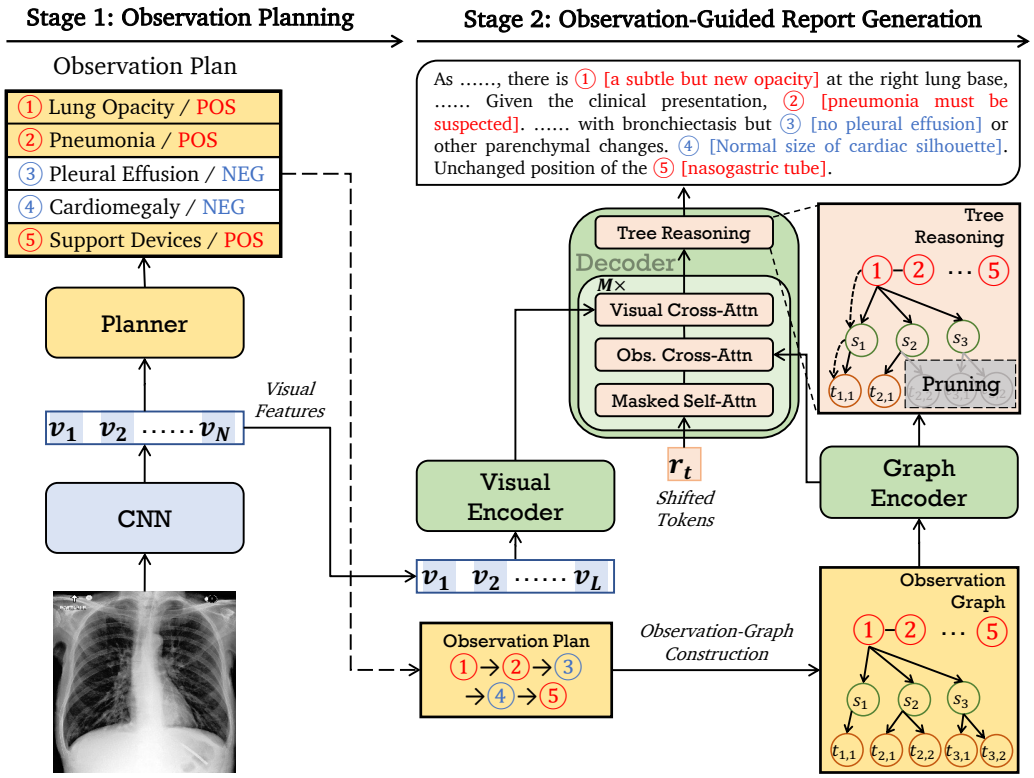


Figure 3.2: The overall framework of ORGAN, divided into two stages: Observation Planning and Observation-Guided Report Generation ("*Obs. Cross-Attn*" in the decoder refers to the observation-related cross-attention module).

absent in it. These categories are converted to two observations: *Lung Opacity*/POS and *Cardiomegaly*/NEG. Then, we locate each observation by matching mentions in the report and order them according to their positions. These mentions are either provided by [67][2] or extracted from the training corpus (i.e., n-grams), as will be illustrated in the following part. Finally, we can obtain the image's observation plan $Z = \{z_1, \ldots, z_L\}$.

Tree-Structured Observation Graph Construction. Since observations are

[2] https://github.com/stanfordmlgroup/chexpert-labeler

high-level concepts that are implicitly related to tokens in reports, it could be difficult for a model to realize these concepts in detailed reports without more comprehensive modeling. Thus, we propose to construct an observation graph by extracting observation-related n-grams as the connections between observations and tokens for better observation realization. Specifically, it involves two steps to construct such a graph: (1) n-grams extraction, where $n \in [1, 4]$ and (2) <observation, n-gram> association. Following previous research [31, 159], we adopt the pointwise mutual information (PMI) [24] to fulfill these two steps, where a higher PMI score implies two units with higher co-occurrence:

$$\text{PMI}(\bar{x}, \hat{x}) = log \frac{p(\bar{x}, \hat{x})}{p(\bar{x})p(\hat{x})}. \tag{3.1}$$

For the first step, we extract n-gram units $S = \{s_1, \ldots, s_{|s|}\}$ based on the training reports. Given two adjacent units $\bar{x}$ and $\hat{x}$ of a text sequence, a high PMI score indicates that they are good collection pairs to form a candidate n-gram s_*, while a low PMI score indicates that these two units should be separated. For the second step, given a predefined observation set $O = \{z_1, \ldots, z_{|O|}\}$, we extract the observation-related n-gram units with $\text{PMI}(z_i, s_j)$, where z_i is the i-th observation, s_j is the j-th n-gram, and $p(z_i, s_j)$ is the frequency that an n-gram s_j appears in a report with observation z_i in the training set. Then, we can obtain a set of observation-related n-grams $s^z = \{s_1^z, \ldots, s_k^z\}$, where $s_j^z = \{t_{j,1}^z, \ldots, t_{j,n}^z\}$, and tokens in n-grams form the token collection $T = \{t_1, \ldots, t_{|T|}\}$. Note that we remove all the stopwords in T, using the vocabulary provided by NLTK[3]. Finally, for each observation, we extract the top-K n-grams as the candidates to construct the graph, which contains three types of nodes $V = \{Z, S, T\}$.After extracting relevant information from the training reports, we construct an observation graph $G =< V, E >$ by introducing three types of edges $E = \{E_1, E_2, E_3\}$:

- E_1: This undirected edge connects two adjacent observations in an observation plan (i.e., $< z_i, z_{i+1} >$).

[3] https://www.nltk.org/

- E_2: This directed edge connects an observation and an n-gram (i.e., $< z_i, s_j >$).
- E_3: This directed edge connects an n-gram with its tokens (i.e., $< s_j, t_k >$).

3.3 Method

3.3.1 Visual Feature Extraction

Given an image X, a CNN and an MLP layer are first adopted to extract visual features $\boldsymbol{X}$:

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} = \text{MLP}(\text{CNN}(X)), \tag{3.2}$$

where $\boldsymbol{x}_i \in \mathbb{R}^h$ is the i-th visual feature.

3.3.2 Observation Planning

The output of observation planning is an observation sequence, which is the high-level summarization of the radiology report, as shown on the left side of Figure 3.2. While examining a radiograph, a radiologist should report positive observations. However, only part of the negative observations will be reported by the radiologist, depending on the overall conditions of the radiograph (e.g., co-occurrence of observations or the limited length of a report). Thus, it is difficult to plan without considering the observation dependencies (i.e., label dependencies). Here, we regard the planning problem as a generation task and use a Transformer [168] encoder-decoder for observation planning:

$$\boldsymbol{h}^v = \{\boldsymbol{h}_1^v, \ldots, \boldsymbol{h}_N^v\} = \text{Encoder}_p(\boldsymbol{X}), \tag{3.3}$$

$$\boldsymbol{z}_l = \text{Decoder}_p(\boldsymbol{h}^v, \boldsymbol{z}_{<l}), \tag{3.4}$$

$$p(z_l|X, Z_{<l}) = \text{Softmax}(\boldsymbol{W}_z \boldsymbol{z}_l + \boldsymbol{b}_z), \tag{3.5}$$

where $\boldsymbol{h}_i^v \in \mathbb{R}^h$ is the i-th visual hidden representation, Encoder_p is the visual encoder, Decoder_p is the observation decoder, $\boldsymbol{z}_* \in \mathbb{R}^h$ is the decoder hidden representation, $\boldsymbol{W}_z \in \mathbb{R}^{|O| \times h}$ is the weight matrix, and $\boldsymbol{b}_z \in \mathbb{R}^{|O|}$ is the bias vector. Then the planning loss $\mathcal{L}_p$ is formulated as:

$$\mathcal{L}_p = -\sum_{l=1}^{L} w_l \log p(z_l | X, Z_{<l}), \tag{3.6}$$

where

$$w_l = \begin{cases} 1+\alpha & \text{if } z_l \in C/\text{POS}, \\ 1 & \text{otherwise.} \end{cases} \tag{3.7}$$

By increasing α, the planner gives more attention to abnormalities. Note that the plugged α is applied to positive observations and *No Finding*/NEG instead of *No Finding*/POS.

3.3.3 Observation-Guided Report Generation

Observation Graph Encoding. We use a Transformer encoder to encode the observation graph constructed according to §3.2.3. To be specific, given the observation graph G with nodes $V = \{Z, S, T\}$ and edges $E = \{E_1, E_2, E_3\}$, we first construct the adjacency matrix $\hat{A} = A + I$ based on E. Then, V and $\hat{A}$ are fed into the Transformer for encoding. Now $\hat{A}$ serves as the self-attention mask in the Transformer, which only allows nodes in the graph to attend to connected neighbors and itself. To incorporate the node type information, we add a type embedding $\boldsymbol{P} \in \mathbb{R}^h$ for each node representation:

$$\boldsymbol{N} = \text{Embed}(V) + \boldsymbol{P}, \tag{3.8}$$

$$\boldsymbol{V} = \{\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{T}\} = \text{Encoder}_g(\boldsymbol{N}, \hat{A}), \tag{3.9}$$

where $\text{Embed}(\cdot)$ is the embedding function, and $\boldsymbol{N} \in \mathbb{R}^h$ represents node embeddings. For observation nodes, $\boldsymbol{P}$ denotes positional embeddings, and for n-gram and token

nodes, $\boldsymbol{P}$ represents type embeddings. $\boldsymbol{Z}$, $\boldsymbol{S}$, and $\boldsymbol{T} \in \mathbb{R}^h$ are encoded representations of observations, n-grams, and tokens, respectively.

Vision-Graph Alignment. As an observation graph may contain irrelevant information, it is necessary to align the graph with the visual features. Specifically, we jointly encode visual features $\boldsymbol{X}$ and token-level node representations $\boldsymbol{T}$ so that the node representations can fully interact with the visual features, and we prevent the visual features from attending the node representations by introducing a self-attention mask M:

$$[\boldsymbol{h}^v, \boldsymbol{T}^A] = \text{Encoder}_u([\boldsymbol{X}, \boldsymbol{T}], \text{M}), \tag{3.10}$$

where $\boldsymbol{h}^v, \boldsymbol{T}^A \in \mathbb{R}^h$ are the visual representation and the aligned token-level node representations, respectively.

Observation Graph Pruning. After aligning visual features and the observation graph, we prune the graph by filtering out irrelevant nodes. The probability of keeping a node is denoted as:

$$p(1|\boldsymbol{T}^A) = \text{Sigmoid}(\boldsymbol{W}_d \boldsymbol{T}^A + b_d), \tag{3.11}$$

where $\boldsymbol{W}_d \in \mathbb{R}^{1\times h}$ is the learnable weight and $b_d \in \mathbb{R}$ is the bias. We can optimize the pruning process with the following loss:

$$\mathcal{L}_d = [-\beta \cdot d \log p(1|\boldsymbol{T}^A) - (1-d)\log(1 - p(1|\boldsymbol{T}^A))], \tag{3.12}$$

where β is the weight to tackle the class imbalance issue, and d is the label indicating whether a token appears in the referential report. Finally, we prune the observation graph by masking out token-level nodes with $p(1|\boldsymbol{T}^A) < 0.5$ and masked token-level node representations denote as $\boldsymbol{T}^M = \text{Prune}(\boldsymbol{T})$.

Tree Reasoning over Observation Graph. We devise a tree reasoning (TrR) mechanism to aggregate observation-relevant information from the graph dynamically. The overall process is shown in Figure 3.3, where we aggregate node information from

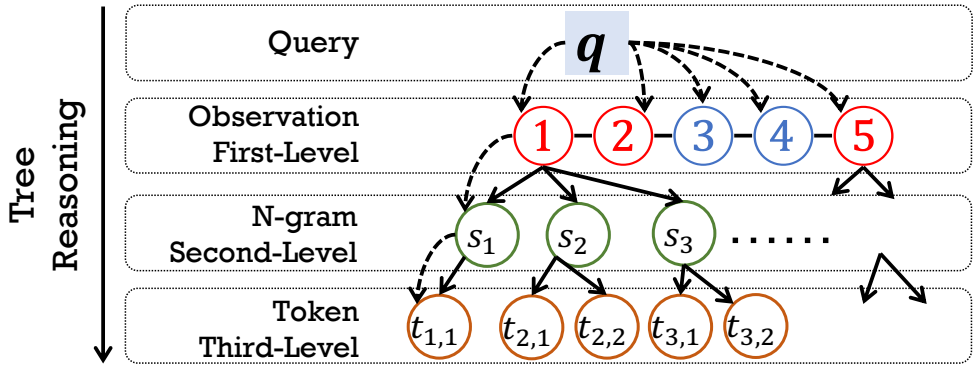


Figure 3.3: Illustration of the tree reasoning mechanism. It aggregates information from the observation level to the n-gram level and finally to the token level.

the observation level (i.e., first level) to the n-gram level (i.e., second level), then to the token level (i.e., third level). To be specific, given a query $\boldsymbol{q}^l$ and node representations at l-th level $\boldsymbol{k}^l \in \{\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{T}^M\}$, the tree reasoning path is $\boldsymbol{q}^0 \xrightarrow{\boldsymbol{Z}} \boldsymbol{q}^1 \xrightarrow{\boldsymbol{S}} \boldsymbol{q}^2 \xrightarrow{\boldsymbol{T}^M} \boldsymbol{q}^3$, and the overall process, is formulated as below:

$$\boldsymbol{v}^{l+1} = \text{MHA}(\boldsymbol{W}_q \boldsymbol{q}^l, \boldsymbol{W}_k \boldsymbol{k}^l, \boldsymbol{W}_v \boldsymbol{k}^l), \tag{3.13}$$

$$\boldsymbol{q}^{l+1} = \text{LayerNorm}(\boldsymbol{q}^l + \boldsymbol{v}^{l+1}), \tag{3.14}$$

where MHA and LayerNorm are the multi-head self-attention, and layer normalization modules [168], respectively. $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v \in \mathbb{R}^{h \times h}$ are weight metrics for query, key, and value vector, respectively. Finally, we can obtain the multi-level information $\boldsymbol{q}^3$, containing observation, n-gram, and token information.

Report Generation with Tree Reasoning. As shown in the right side of Figure 3.2, an observation-guided Transformer decoder is devised to incorporate the graph information, including (i) multiple observation-guided decoder blocks (i.e., Decoder$_g$), which aims to align observations with the visual representations, and (ii) a tree-reasoning block (i.e., TrR$_g$), which aims to aggregate observation-relevant information. For Decoder$_g$, we insert an observation-related cross-attention module before a visually-aware cross-attention module. By doing this, the model can correctly focus on regions

closely related to a specific observation. Given the visual representations $\boldsymbol{h}^v$, the node representations $\boldsymbol{V} = \{\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{T}^M\}$, and the hidden representation of the prefix $\boldsymbol{h}_*^w \in \mathbb{R}^h$, the t-th decoding step is formulated as:

$$\text{Decoder}_g = \begin{cases} \boldsymbol{h}_t^s = \text{Self-Attn}(\boldsymbol{h}_t^w, \boldsymbol{h}_{<t}^w, \boldsymbol{h}_{<t}^w), \\ \boldsymbol{h}_t^o = \text{Cross-Attn}(\boldsymbol{h}_t^s, \boldsymbol{Z}, \boldsymbol{Z}), \\ \boldsymbol{h}_t^p = \text{Cross-Attn}(\boldsymbol{h}_t^o, \boldsymbol{h}^v, \boldsymbol{h}^v), \end{cases} \tag{3.15}$$

$$\text{TrR}_g = \begin{cases} \boldsymbol{h}_t^d = \text{Self-Attn}(\boldsymbol{h}_t^p, \boldsymbol{h}_{<t}^p, \boldsymbol{h}_{<t}^p), \\ \boldsymbol{q}_t^3 = \text{TrR}(\boldsymbol{h}_t^d, [\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{T}^M]), \end{cases} \tag{3.16}$$

$$p(y_t | X, G, Y_{<t}) = \text{Softmax}(\boldsymbol{W}_g \boldsymbol{q}_t^3 + \boldsymbol{b}_g), \tag{3.17}$$

where Self-Attn is the self-attention module, Cross-Attn is the cross-attention module, $\boldsymbol{h}_t^s, \boldsymbol{h}_t^o, \boldsymbol{h}_t^p \in \mathbb{R}^h$ are self-attended hidden state, observation-related hidden state, visually-aware hidden state of Decoder_g, respectively. $\boldsymbol{h}_t^d \in \mathbb{R}^h$ is the self-attended hidden state of TrR_g, $\boldsymbol{W}_g \in \mathbb{R}^{|V| \times h}$ is the weight matrix, and $\boldsymbol{b}_g \in \mathbb{R}^{|V|}$ is the bias vector. We omit other modules (i.e., Layer Normalization and Feed-Forward Network) in the standard Transformer for simplicity. Note that we extend the observation plan Z to an observation graph G, so the probability of y_t conditions on G instead of Z. Then, we optimize the generation process using the negative log-likelihood loss:

$$\mathcal{L}_r = -\sum_{t=1}^{T} \log p(y_t | X, G, Y_{<t}). \tag{3.18}$$

Finally, the loss function of the generator is:

$$\mathcal{L}_g = \mathcal{L}_r + \mathcal{L}_d. \tag{3.19}$$

3.4 Experiments

3.4.1 Datasets

Following previous research [21, 20], we use two publicly available benchmarks to evaluate our method, which are IU X-RAY[4] [28] and MIMIC-CXR[5] [74]. Both datasets have been automatically de-identified, and we use the same preprocessing setup of [21].

- IU X-RAY is collected by Indiana University, containing 3,955 reports with two X-ray images per report, resulting in 7,470 images in total. We split the dataset into train/validation/test sets with a ratio of 7:1:2, which is the same data split as in [21].

- MIMIC-CXR consists of 377,110 chest X-ray images and 227,827 reports from 63,478 patients. We adopt the standard train/validation/test splits.

3.4.2 Evaluation Metrics and Baselines

Evaluation Metrics. We adopt NLG Metrics and CE Metrics to evaluate the performance of models. Specifically, BLEU-1/2/3/4 (B-1/2/3/4) [130], METEOR (MTR) [8], and ROUGE-L (R-L) [93] are selected as NLG Metrics, and we use the MS-COCO caption evaluation tool[6] to compute the results. For CE Metrics, we adopt CheXpert [67] for MIMIC-CXR dataset to label the generated reports compared with disease labels of the references and report the macro-weighted precision, recall, and F_1 score across 14 observations. Note that CheXpert is designed for MIMIC-CXR and we do not apply CE Metrics to IU X-RAY.

[4] https://openi.nlm.nih.gov/

[5] https://physionet.org/content/MIMIC-cxr-jpg/2.0.0/

[6] https://github.com/tylin/coco-caption

Baselines. To evaluate the performance of ORGan, we compare it with the following 10 state-of-the-art (SOTA) baselines: R2Gen [21], a memory-driven Transformer model; CA [102], which applies contrastive attention to highlight abnormal regions; CMCL [100], which employs competence-based multimodal curriculum learning for progressive training; PPKED [101], which captures and distills both posterior and prior knowledge; R2GenCMN [20], which utilizes a cross-modal memory network for better visual-text alignment; AlignTransformer [203], which hierarchically aligns visual features with disease-related tags; KnowMat [199], which incorporates external domain knowledge to enhance generation quality; $\mathcal{M}^2$Tr [126], which generates radiology reports in a coarse-to-fine manner; CMM-RL [133], which applies reinforcement learning to refine report generation; CMCA [158], which leverages contrastive attention to better model abnormal findings.

3.4.3 Implementation Details

We adopt the ResNet-101 [49] pretrained on ImageNet [29] as the visual extractor. For IU X-ray, we further fine-tune ResNet-101 on CheXpert [67]. The layer number of all the encoders and decoders is set to 3 except for the Graph Encoder, where the layer number is set to 2. The input dimension and the feed-forward network dimension of a Transformer block are set to 512, and each block contains 8 attention heads. The beam size for decoding is set to 4, and the maximum decoding step is set to 64/104 for IU X-ray and MIMIC-CXR, respectively.

We use AdamW [113] as the optimizer and set the initial learning rate for the visual extractor as $5e-5$ and $1e-4$ for the rest of the parameters, with a linear schedule decreasing from the initial learning rate to 0. α is set to 0.5, the dropout rate is set to 0.1, and the batch size is set to 32. For IU X-ray, we train the planner/generator for 15/15 epochs, and β is set to 2. For MIMIC-CXR, the planner and generator are trained for 3 and 5 epochs, respectively, and β is set to 5. We select the best checkpoints

of the planner based on micro F_1 of all observations and select the generator based on the BLEU-4 on the validation set. Our model has 65.9M parameters, and the implementations are based on HuggingFace's Transformers [185]. We conduct all the experiments on an NVIDIA-3090 GTX GPU with mixed precision. The NLTK package version is 3.6.2.

3.5 Results and Analyses

3.5.1 Quantitative Analysis

Language Generation Results. The left part of Table 3.2 presents the language generation results. ORGAN outperforms most of the baselines (except CMCA on IU X-RAY) and achieves state-of-the-art performance. Specifically, our model achieves 0.195 BLEU-4 on the IU X-XRAY dataset, which is the second-best result, and 0.123 BLEU-4 on the MIMIC-CXR dataset, leading to a 5.1% increment of compared to the best baseline (i.e., CMCA). In terms of METEOR, ORGAN achieves competitive performance on both datasets. In addition, our model increases R-L by 0.6% on the MIMIC-CXR dataset compared to the best baseline and achieves the second-best result on the IU X-RAY dataset. This indicates that by introducing the guidance of observations, ORGAN can generate more coherent text than baselines. However, we notice that on the IU X-RAY dataset, there is still a performance gap between our model and the best baseline (i.e., CMCA). The reason may be that the overall data size of this dataset is small ($\sim$ 2,000 samples for training), in which positive observations are rare. It is difficult to train a good planner using a small training set, especially with cross-modal data. As we can see from Table 3.4, the planner only achieves 0.132 Macro-F_1 on the IU X-RAY dataset, which is relatively low compared to the performance of the planner on the MIMIC-CXR dataset. Thus, accumulation errors unavoidably propagate to the generator, which leads to lower performance.

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	**B-2**	**B-3**	**B-4**	**MTR**	**R-L**	**P**	**R**	**F_1**
IU X-RAY	R2GEN	0.470	0.304	0.219	0.165	–	0.371	–	–	–
	CA	0.492	0.314	0.222	0.169	0.193	0.381	–	–	–
	CMCL	0.473	0.305	0.217	0.162	0.186	0.378	–	–	–
	PPKED	0.483	0.315	0.224	0.168	-	0.376	–	–	–
	R2GENCMN	0.475	0.309	0.222	0.170	0.191	0.375	–	–	–
	$\mathcal{M}^2$TR	0.486	0.317	0.232	0.173	0.192	0.390	–	–	–
	ALIGNTRANSFOMER	0.484	0.313	0.225	0.173	–	0.379	–	–	–
	KNOWMAT	0.496	0.327	0.238	0.178	–	0.381	–	–	–
	CMM-RL	0.494	0.321	0.235	0.181	0.201	0.384	–	–	–
	CMCA	0.496	**0.349**	**0.268**	**0.215**	**0.209**	0.392	–	–	–
	ORGAN (Ours)	**0.510**	0.346	0.255	0.195	0.205	**0.399**	–	–	–
MIMIC -CXR	R2GEN	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
	CA	0.350	0.219	0.152	0.109	0.151	0.283	–	–	–
	CMCL	0.344	0.217	0.140	0.097	0.133	0.281	–	–	–
	PPKED	0.360	0.224	0.149	0.106	0.149	0.284	–	–	–
	R2GENCMN	0.353	0.218	0.148	0.106	0.142	0.278	0.344	0.275	0.278
	$\mathcal{M}^2$TR	0.378	0.232	0.154	0.107	0.145	0.272	0.240	**0.428**	0.308
	ALIGNTRANSFOMER	0.378	0.235	0.156	0.112	–	0.283	–	–	–
	KNOWMAT	0.363	0.228	0.156	0.115	–	0.284	**0.458**	0.348	0.371
	CMM-RL	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
	CMCA	0.360	0.227	0.156	0.117	0.148	0.287	0.444	0.297	0.356
	ORGAN (Ours)	**0.407**	**0.256**	**0.172**	**0.123**	**0.162**	**0.293**	0.416	0.418	**0.385**

Table 3.2: Experimental Results of our model and baselines on the IU X-RAY dataset and the MIMIC-CXR dataset, with the best scores shown in **boldface** and the second-best scores underlined.

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	**B-2**	**B-3**	**B-4**	**MTR**	**R-L**	**P**	**R**	**F_1**
IU X-RAY	ORGAN	0.510	0.346	0.255	0.195	0.205	0.399	–	–	–
	ORGAN *w/o* Plan	0.406	0.254	0.178	0.133	0.167	0.372	–	–	–
	ORGAN *w/o* Graph	0.461	0.302	0.218	0.164	0.186	0.383	–	–	–
	ORGAN *w/o* TrR	0.494	0.335	0.247	0.190	0.203	0.395	–	–	–
MIMIC -CXR	ORGAN	0.407	0.256	0.172	0.123	0.162	0.293	0.416	0.418	0.385
	ORGAN *w/o* Plan	0.334	0.211	0.145	0.107	0.136	0.282	0.384	0.239	0.252
	ORGAN *w/o* Graph	0.369	0.233	0.158	0.113	0.151	0.290	0.401	0.415	0.383
	ORGAN *w/o* TrR	0.405	0.254	0.170	0.121	0.161	0.291	0.411	0.419	0.386

Table 3.3: Ablation results of our model and its variants, where ORGAN *w/o* Plan is the standard Transformer model.

Clinical Efficacy Results. The clinical efficacy results are listed on the right side of Table 3.2. On the MIMIC-CXR dataset, our model outperforms previous SOTA results. Specifically, ORGAN reaches 0.385 CE F_1, increasing by 1.4% compared to the best baseline. In addition, 0.416 precision and 0.418 recall are achieved by our model, which are competitive results. This indicates that our model can successfully improve the clinical accuracy of the generated reports with observation information. In addition, as shown in Table 3.4, the planner trained on the MIMIC-CXR dataset achieves a Micro-F_1 score of 0.574 and a Macro-F_1 score of 0.397. Similarly, the generator attains a Macro-F_1 score of 0.385, which corresponds to 97% of the planner's performance. Despite the small gap between the planner and generator, ORGAN demonstrates strong observation realization capability. The performance on the IU X-RAY dataset is also reported for reference.

Ablation Results. To examine the effect of the observation plan and the TrR mechanism, we perform ablation tests, and the ablation results are listed in Table 3.3. There are three variants:

- ORGAN *w/o* Plan: This variant does not consider observation information, and

Dataset	Micro-F_1	Macro-F_1	B-2
IU X-RAY	0.507	0.132	0.499
MIMIC-CXR	0.574	0.397	0.357

Table 3.4: Experimental results of observation planning. Macro-F_1 and Micro-F_1 denote the macro F_1 and micro F_1 of abnormal observations, respectively.

Dataset	K	B-2	B-4	MTR	R-L
IU X-RAY	10	0.309	0.170	0.192	0.388
	20	0.333	0.180	0.202	0.393
	30	0.346	0.195	0.205	0.399
MIMIC-CXR	10	0.249	0.118	0.161	0.290
	20	0.252	0.120	0.159	0.292
	30	0.256	0.123	0.162	0.293

Table 3.5: Experimental results across different K (selected n-grams).

it is a standard Transformer model.

- ORGAN *w/o* Graph: This variant only considers observations but not the observation graph.

- ORGAN *w/o* TrR: This variant select information without using the TrR mechanism.

Compared to the full model, the performance of ORGAN *w/o* Plan drops significantly on both datasets. This indicates that observation information plays a vital role in generating reports. For ORGAN *w/o* Graph, the performance on NLG metrics decreases significantly, but the performance of clinical efficacy remains nearly the same as the full model. This is reasonable because the observation graph is designed to enrich the observation plan to achieve better word-level realization. On the performance of

ORGAN *w/o* TrR, a similar result of ORGAN *w/o* Graph is observed. This indicates that TrR can enrich the plan information, and stronger reasoning can help generate high-quality reports.

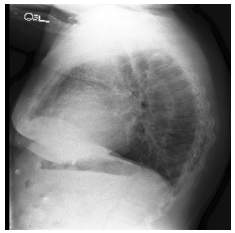
We also conduct experiments on the impact of the number (K) of selected n-grams, as shown in Table 3.5. There is a performance gain when increasing K from 10 to 20 and to 30 on both datasets. On the IU X-RAY dataset, B-2 increases by 2.4% and 3.7% and B-4 rises by 1.0% and 1.5%. A similar trend is also observed on the MIMIC-CXR dataset.

3.5.2 Qualitative Analysis

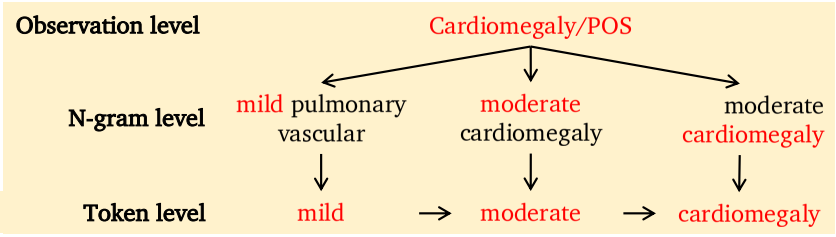
We conduct a case study and analyze several error cases produced by ORGAN on the MIMIC-CXR dataset to provide some insights.

Case Study. We conduct a case study to show how the observation and the tree reasoning mechanism guide the report generation process, as shown in Figure 3.4. We show the generated reports of ORGAN, ORGAN *w/o* TrR, and ORGAN *w/o* Plan, respectively. All three models successfully generate the first three negative observations and the last positive observation. However, variant *w/o* plan generates "*mild pulmonary vascular congestion without overt pulmonary edema*" which is not consistent with the radiograph. In terms of the output of variant *w/o* TrR, "*mediastinal silhouettes are unchanged*" is closely related to observation *Enlarged Cardiomediastinum* instead of *Cardiomegaly*. Only ORGAN can generate the *Cardiomegaly*/POS presented in the observation plan with a TrR path. This indicates that observations play a vital role in maintaining clinical accuracy. In addition, most of the tokens in the observation mention *mild to moderate cardiomegaly* can be found in the observation graph, which demonstrates that the graph can provide useful information in word-level realization.

Error Analysis. We present two error cases generated by ORGAN in Figure 3.5. We find that the primary error arises from the introduction of incorrect observations



Observation Plan:
① Consolidation/NEG
② Pleural Effusion/NEG
③ Pneumothorax/NEG
④ Cardiomegaly/POS



Tree Reasoning Path over Graph:

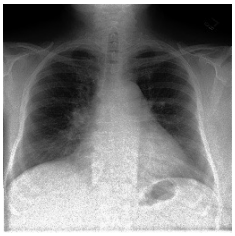
Report Generated by ORGAN
frontal and lateral views of the chest were obtained. the lungs are clear ①[without focal consolidation]. ②[no pleural effusion] or ③[pneumothorax] is seen. there is ④[mild to moderate cardiomegaly] unchanged. the aorta is calcified and tortuous.

Report Generated by ORGAN *w/o* TrR
frontal and lateral views of the chest were obtained. the lungs are clear ①[without focal consolidation]. ②[no pleural effusion] or ③[pneumothorax] is seen. ④[the cardiac silhouette is mildly enlarged.] mediastinal silhouettes are unchanged. there is mild to moderate degenerative change in the thoracic spine.

Report Generated by OGRAN *w/o* Plan
frontal and lateral views of the chest were obtained. there is mild pulmonary vascular congestion without overt pulmonary edema. ①[no focal consolidation] ②[pleural effusion] or ③[pneumothorax] is seen. the ④[cardiac and mediastinal silhouettes are stable].

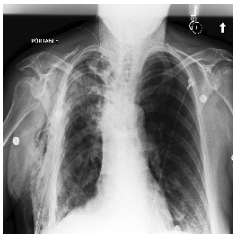
Reference
frontal and lateral views of the chest. the lungs are ①[clear of focal consolidation] ②[effusion] or ③[pneumothorax]. ④ [the heart is enlarged] similar to prior. right upper extremity vascular stent is partially visualized. multiple thoracic compression deformities are again seen.

Figure 3.4: Case study of our model with the tree reasoning path of the mention "*mild to moderate cardiomegaly.*"



Oracle Plan:
① Pneumonia/NEG ② Edema/NEG,
③ Cardiomegaly/POS ④ Enlarged Card./POS
⑤ Pleural Effusion/NEG ⑥ Pneumothorax/NEG
Reference: ..... there are ① [no lung opacities concerning for pneumonia or ② [pulmonary edema]. ③ [heart size is mildly enlarged and stable since]. ④ [mediastinal and hilar contours are unchanged]. there is ⑤ [no pleural effusion] or ⑥ [pneumothorax].

Generated Plan:
① Cardiomegaly/POS ② Edema/NEG
③ Pleural Effusion/NEG ④ Pneumonia/NEG
Generated Report: ..... ① [mild enlargement of the cardiac silhouette] ② [without pulmonary vascular congestion] ③ [pleural effusion] or ④ [acute focal pneumonia].



Oracle Plan:
① Support Devices/POS ② Pneumothorax/NEG
Reference: ..... ① [right chest tubes] appears to have been removed. ② [no definite pneumothorax is appreciated].
.....

Generated Plan:
① Support Devices/POS ② Pneumothorax /POS ③ Lung Opacity/POS ④ Atelectasis/POS ⑤ Pleural Effusion/POS
Generated Report: ..... ① [right-sided chest tube with the tip in the mid svc]. there is ② [a small right apical pneumothorax]. there is ③ [increased opacity] at the right lung base likely representing ④ [atelectasis]. there is ⑤ [a small left pleural effusion].

Figure 3.5: Examples of error cases. *Enlarged Card.* refers to *Enlarged Cardiomediastinum*. The upper case omits one positive observation and the bottom case contains false positive observations.

in the plans. Specifically, two types of errors defined by [205] are evident: omission of finding and false prediction of finding. The generated plan of the upper case omits one positive observation (i.e., *Enlarged Cardiomediastinum*/POS), resulting in false negative observations in its corresponding generated report (i.e., "*mediastinal and hilar contours are unchanged*"). Another type of error is the presence of false positive observations in the generated reports (e.g., the bottom case). There are four positive observations in the generated plans: *Pneumothorax*/POS, *Lung Opacity*/POS, *Atelectasis*/POS, and *Pleural Effusion*/POS, all of which are realized in the generated reports. For example, the inclusion of *Lung Opacity*/POS results in the mention of "*increased opacity*", while *Pleural Effusion*/POS leads to the phrase "*a small left pleural effusion*" in the generated report. Although these results demonstrate the strong surface realization capabilities of our model, errors made by the planner inevitably propagate to the generator in ORGan. Therefore, improving the performance of the planner remains a promising direction for future work to further enhance clinical accuracy.

3.6 Chapter Summary

In this chapter, we propose ORGan, an observation-guided radiology report generation framework, which first produces an observation plan and then generates the corresponding report based on the radiograph and the plan. To achieve better observation realization, we construct a three-level observation graph containing observations, observation-aware n-grams, and tokens, and we propose a tree reasoning mechanism to capture observation-related information by dynamically aggregating nodes in the graph. Experimental results demonstrate the effectiveness of our proposed framework in terms of improving the clinical accuracy of the generated reports.

There are several limitations to our framework. Specifically, since observations are introduced as guiding information, our framework requires observation extraction tools

to label the training set in advance. Then, the nodes contained in the observation graph are mined from the training data. Consequently, the mined n-grams can be biased when the training set is small and may not accurately capture the attributes of the observations. The temporal information inherent in follow-up studies, which reflects changes over time, is also overlooked, even though it plays a crucial role in generating clinically accurate and contextually coherent reports. In addition, our framework is a pipeline, and the report generation performance highly relies on the performance of observation planning. Thus, errors could accumulate through the pipeline, especially for small datasets. Finally, our framework is designed for radiology report generation targeting chest X-ray images. However, there are other types of medical images (e.g., Fundus Fluorescein Angiography images) that our framework needs to examine.

Chapter 4

Observation-aware Radiology Report Generation: Supplementary Knowledge Injection

4.1 Introduction

In Chapter 3, we discuss radiology report generation with a focus on observation extraction and incorporation. Building upon this foundation, in this chapter, we investigate observation-aware knowledge injection. Recent advances in foundation models [131, 22, 66], which leverage large language models (LLMs) for enhanced medical image analysis, have demonstrated remarkable potential in generating fluent and cohesive clinical text, aiding radiologists in their diagnostic workflow.

Despite their ability to generate highly readable and clinically plausible report content, LLMs still face persistent challenges in ensuring clinical accuracy. One major challenge lies in the knowledge gap between the medical and general domains. Many studies have attempted to bridge this disparity by augmenting models with retrieved domain-specific knowledge [199, 101, 89, 139, 160]. However, these approaches often overlook

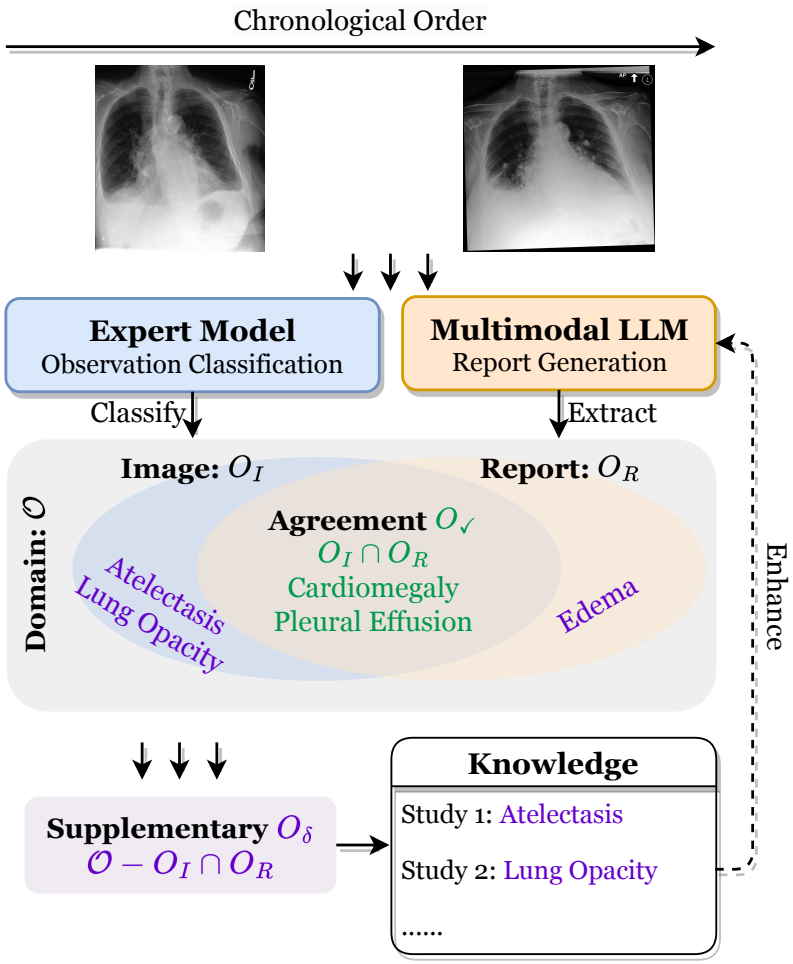


Figure 4.1: A motivating example. The report directly generated by the multimodal LLM showcases its knowledge regarding several findings (O_R) but can contain hallucinations and overlook some other findings. To address this, we regard the part that aligns with another expert model ($O_R \cap O_I$) as trustworthy and we incorporate supplementary knowledge for the remaining part ($\mathcal{O} - O_R \cap O_I$) to enhance the report generation.

the knowledge LLMs have already acquired. That is, much of the retrieved information is often duplicate knowledge already encoded within the model's parameters, leading to redundant information retrieval. Moreover, the knowledge learned by LLMs [99] is not always trustworthy, as hallucinations frequently occur [60]. For instance, in Figure 4.1, the LLM correctly identifies *Cardiomegaly*, making the retrieval of additional knowledge about this observation unnecessary. Additionally, the generated *Pleural Effusion* is highly credible, as it aligns with the expert model, whereas *Edema* remains uncertain. Thus, balancing learned and retrieved knowledge in radiology report generation is crucial to address these challenges.

In this chapter, we propose RADAR, a framework for RADiology report generation that integrates both the internal knowledge of LLMs and external supplementARy knowledge. Our framework primarily consists of two stages: preliminary findings generation and supplementary findings augmentation. In the first stage, RADAR generates an initial report from the input images. Subsequently, an expert model processes the images for observation classification. The overlapping information between the generated report and the classified observations is identified as high-confidence internal knowledge. In the second stage, RADAR additionally retrieves new knowledge to supplement the internal knowledge. Finally, both internal and supplementary knowledge sources are aggregated to enhance the report generation process. Our main contributions can be summarized as follows:

- We propose RADAR, a novel framework that enhances the clinical accuracy of radiology report generation by effectively integrating both the internal knowledge of LLMs and externally retrieved domain-specific knowledge.
- To optimize knowledge utilization, we introduce a knowledge extraction method that identifies and retains non-overlapping information from the model's learned knowledge, reducing redundancy and bridging the knowledge gap.
- We conduct extensive experiments on three benchmark datasets: MIMIC-CXR,

CheXpert-Plus, and IU X-ray, demonstrating the effectiveness of Radar.

4.2 Preliminary

4.2.1 Problem Formulation

A multimodal LLM (MLLM) generally consists of a vision encoder, a vision connector that transforms visual signals into the language space (e.g., MLP [104], Q-Former [84], or Perceiver Resampler [192]), and an LLM, as illustrated in the left part of Figure 6.2. For radiology report generation[1], the MLLM takes a radiograph X, its prior X_p (if available), and the clinical context C (e.g., *Indication* or *Prior Findings*) as input and generates the report $Y = \{y_1, \ldots, y_L\}$. The probability of the t-th token is computed as follows:

$$p(y_t) = \text{MLLM}(X, X_p, C, y_{<t}), \tag{4.1}$$

where the MLLM is optimized using the negative log-likelihood loss:

$$\mathcal{L} = -\sum_{t=1}^{L} \log p(y_t). \tag{4.2}$$

4.2.2 Semi-Structured Report as Knowledge

In this chapter, the training set of MIMIC-CXR serves as the knowledge source for radiology report generation. To effectively leverage the knowledge encoded in each report, we convert it into semi-structured data. Specifically, given a report consisting of N sentences, $Y = \{S_1, \ldots, S_N\}$, we annotate each sentence using the 14-category CheXpert observations [67] with the CheXbert model [157]. Each observation falls into one of four classes: *Positive*, *Negative*, *Uncertain*, or *Blank*. To ensure conciseness, we retain only sentences annotated with *Positive* observations. These selected sentences

[1] In this chapter, "report" typically refers to "findings," and we use these two terms interchangeably.

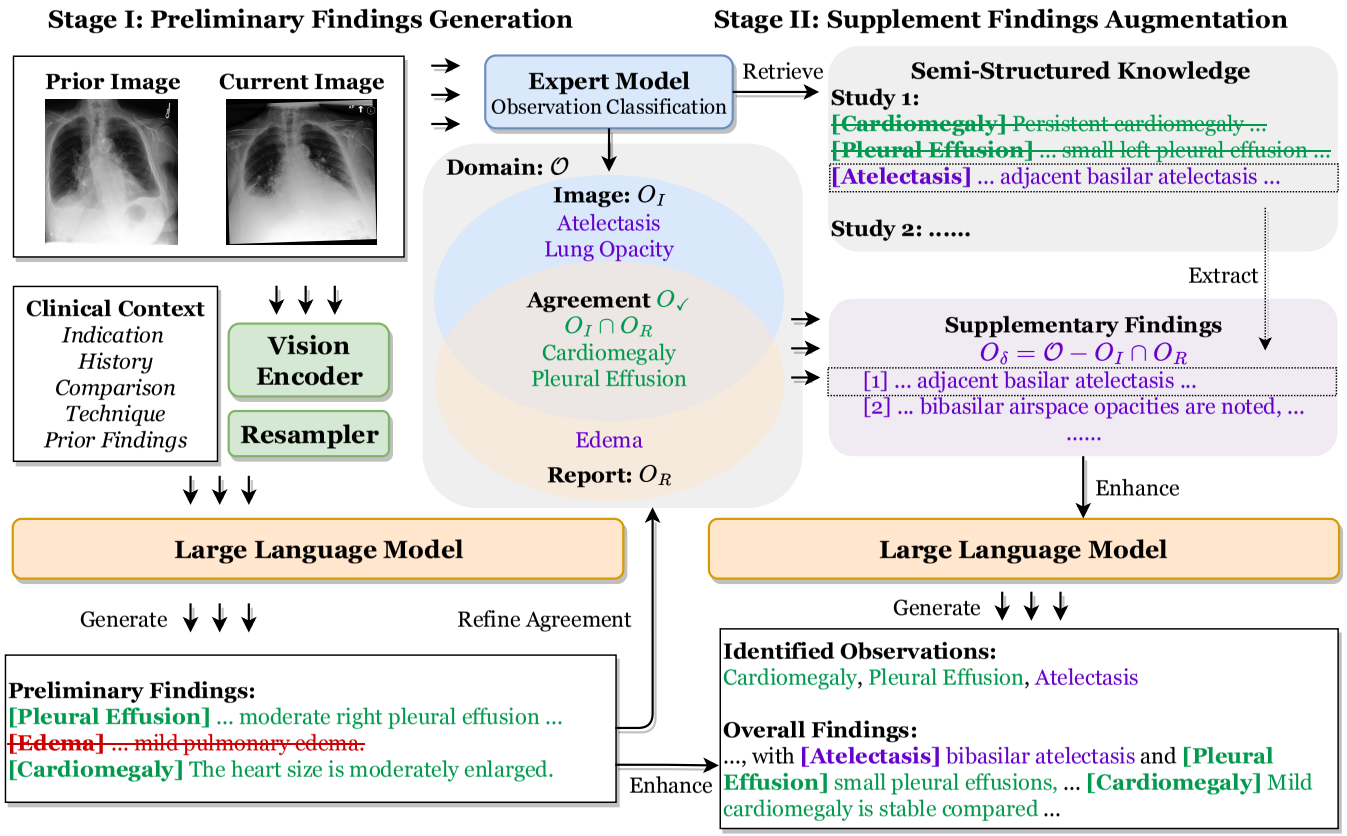


Figure 4.2: Overview of the RADAR. In Preliminary Findings, only sentences that reach agreement are retained, whereas in Supplementary Findings, only sentences that are relevant to Preliminary Findings are preserved.

collectively represent the knowledge extracted from the report, as illustrated in the top-right part of Figure 6.2. Note that we annotate and process Preliminary Findings (§4.3.1) and Supplementary Findings (§4.3.2) in the same manner.

4.3 Method

4.3.1 Preliminary Findings Generation

We illustrate the Stage I process in the left part of Figure 6.2. To assess the learned knowledge of an LLM, we first feed the input (X, X_p, and C) into RADAR to generate a report $\hat{Y}$:

$$\hat{Y} = \underset{\hat{Y} \in \mathcal{Y}}{\operatorname{argmax}} \prod_{t=1}^{T} \text{MLLM}(X, X_p, C, \hat{y}_{<t}),$$

where $\mathcal{Y}$ represents the set of possible reports. Note that exact maximization is intractable and we employ an approximate decoding algorithm for generation. Next, we convert the findings into semi-structured knowledge, as described in §4.2.2, and denote the observations of $\hat{Y}$ as O_R.

To extract credible knowledge from $\hat{Y}$ while filtering out untrustworthy information, we train an expert model that predicts observations for the image. Unlike previous works [57, 131], which consider only the image as input, we incorporate the clinical context to enhance performance. Specifically, the expert model $f(X)$ encodes X and C using an image encoder Encoder_v and a text encoder Encoder_t, respectively, and then processes their outputs through an MLP for observation classification:

$$\mathbf{h}_v = \text{Encoder}_v(X), \tag{4.3}$$

$$\mathbf{h}_t = \text{Encoder}_t(C), \tag{4.4}$$

$$p(O_i) = \sigma(\text{MLP}([\mathbf{h}_v; \mathbf{h}_t])), \tag{4.5}$$

where $[;]$ is the concatenation function, $\mathbf{h}_v$ and $\mathbf{h}_t$ are the pooled outputs of the image and text encoders, respectively, and $p(O_i)$ represents the probability of the i-th

observation. We denote the observations derived from $f(X)$ as O_I, and the credible and high-confidence observations, $O_\checkmark$, are then obtained by intersecting O_I and O_R, as follows:

$$O_\checkmark = O_I \cap O_R.$$

Finally, we refine $\hat{Y}$ by removing sentences that do not correspond to $O_\checkmark$, yielding the Preliminary Findings (PF).

To train the expert model, we collect observations from each report as image annotations and optimize the expert model using binary cross-entropy loss. Following [131], we address data imbalance by re-weighting the positive observations with a log-scale weight, defined as $\alpha_i = \log\left(1 + \frac{|\mathcal{D}_{\text{train}}|}{w_i}\right)$, where $|\mathcal{D}_{\text{train}}|$ is the total number of training samples and w_i denotes the frequency of observation O_i.

4.3.2 Supplementary Findings Augmentation

Supplementary Knowledge Retrieval. We follow the retrieval process of [199] to search for domain knowledge. Specifically, the expert model described in §4.3.1 produces probabilities for 14 observations, and we compute the similarity between different samples using KL-divergence:

$$\hat{z} = \text{Normalize}(f(X)), \tag{4.6}$$

$$\text{Sim}(X, X_i) = -\sum_{j=1}^{|\mathcal{O}|} \hat{z}_j \log \frac{\hat{z}_j}{\hat{z}_{i,j}}, \tag{4.7}$$

where Normalize$(\cdot)$ normalizes $f(X)$ to 1, $\hat{z}$ represents the normalized scores for (X), and $\hat{z}_{i,j}$ denotes the score of the j-th observation in the i-th sample from the database (i.e., the training samples of the MIMIC-CXR dataset). We then rank the samples based on their similarity scores, $\text{Sim}(X, X_i)$, and retrieve the top-K reports, denoted as $\mathcal{Y}^S = \{Y_1^S, \ldots, Y_K^S\}$.

Supplementary Knowledge Extraction. Since the retrieved information may

overlap with the knowledge learned by LLMs, we extract only supplementary knowledge based on two principles: (1) it should be concise and relevant, and (2) it should complement, rather than duplicate, the preliminary findings. Thus, for each supplementary report Y_i^S with its corresponding observations O^S, we retain only the following observations:

$$O_\delta = \mathcal{O} - O_\checkmark. \tag{4.8}$$

Next, we convert Y_i^S into semi-structured knowledge and remove sentences that do not correspond to O_δ, referring to these findings as Supplementary Findings (SF). Notably, all sentences corresponding to negative observations are removed, ensuring that SF remains concise and clinically relevant.

4.3.3 Enhanced Radiology Report Generation

We integrate both PF and SF into the clinical context C to form the augmented context C^A, from which the final report Y is generated as:

$$Y = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \, \text{MLLM}(X, X_p, C^A). \tag{4.9}$$

Since PF and SF contain information from various studies, summarizing high-level information before generating the report is necessary. Thus, we include the observations of Y as part of the training targets. Specifically, during training, Y is converted into a structured format:

$$Y^O = \{O_1, \ldots, O_N, y_1, \ldots, y_L\}, \tag{4.10}$$

where $\{O_1, \ldots, O_N\}$ represents the observations in Y, and $\{y_1, \ldots, y_L\}$ corresponds to the tokens of the report. We refer to this process as Observation Identification (OI). During inference, we extract the final report from the generated output for evaluation.

4.4 Experiments

4.4.1 Datasets

We evaluate our model using three publicly available radiology report generation datasets: MIMIC-CXR [74], CheXpert Plus[2] [15], and IU X-ray [28]:

- MIMIC-CXR contains 377,110 chest radiographs and 227,827 reports. We use this dataset for fine-tuning, including only frontal images in our experiments. The number of samples in the train/validation/test sets is 162,955/1,286/2,461.
- CheXpert Plus comprises 223,462 unique radiology report and chest X-ray pairs from 187,711 studies. We evaluate our model using only frontal images from the validation set, which includes 62 samples.
- IU X-ray is collected by Indiana University, and we use all frontal images for evaluation, with 3,199 studies in total, similar to [9].

4.4.2 Evaluation Metrics

NLG Metrics. Following previous research [21, 89], BLEU-1/4 [130], ROUGE-L [93], and METEOR [8] are adopted for evaluating the languages of generated outputs.

Clinical Metrics. We evaluate the factual accuracy using several metrics. Specifically, RG-F_1 and $RG_{\overline{ER}(ER)}$ [69] evaluate the entity-level factuality and $RadCliQ_0$ [205], denoted as $CliQ_0$, aligns with the preference of radiologists. For observation evaluation, ^{14}Macro-F_1 (^{14}Ma-F_1) and ^{14}Micro-F_1 (^{14}Mi-F_1) evaluate the macro and micro F_1 of 14 observations (refers to Table 4.8), respectively. In addition, ^{5}Macro-F_1 (^{5}Ma-F_1) and ^{5}Micro-F_1 (^{5}Mi-F_1) measure the performance of 5 common observations (*Atelectasis*,

[2] https://aimi.stanford.edu/datasets/chexpert-plus

Cardiomegaly, *Consolidation*, *Edema*, and *Pleural Effusion)*. Two lines of CheXpert results are reported, i.e., *Uncertain as Negative* and *Uncertain as Positive*.

4.4.3 Baselines

On the MIMIC-CXR dataset, we compare our models with the state-of-the-art (SOTA) MLLMs, including: RadFM [186], which is a radiology foundation model; XrayGPT [163], which is a fine-tuned vision-language model for report generation; LLaVA-Med [82], which is a general-purpose biomedical LLM; R2GenGPT [180], which employs Llama 2 for report generation; R2-LLM [99], which enhances visual-textual alignment via self-bootstrapping; RaDialog [131], which is a conversational assistant in radiology; CheXagent [22], which is a VLM with various CXR iterpretation abilites; GPT-4V [128], which is a general VLM and is able to interprate medical images; LLaVA-Rad [16], which is a fine-tuned VLM for radiology report generation; Med-PaLM [156], which is a medical LLM with strong performance on various medical tasks; MAIRA-1 [66], which employs a CXR image encoder and a Vicuna-7B LLM; MAIRA-2 [9], which can generate descriptions for specific regions in radiographs; MedVerse [222], which is a generalist model and supports multimodal inputs; and Libra [215], which leverages temporal images for radiographs. Other SOTA specialists are: R2Gen [21], a memory-driven Transformer model; R2GenCMN [20], which employs a cross-modal memory network; $\mathcal{M}^2$Tr [126], which generates radiology reports in a progressive manner; KnowMat [199], which integrates domain knowledge to enhance performance; CMM-RL [133], which leverages reinforcement learning to optimize report generation; CMCA [158], which incorporates contrastive attention to better capture abnormalities; KiUT [65], a knowledge-injected U-Transformer; DCL [89], which improves report generation through contrastive learning; METrans [179], RGRG [162], Recap [56], which leverages historical patient records to inform report generation; Controllable [25], which is a controllable radiology report generation framework and allows user to select the regions for reporting; and PromptMRG [71], which generates medical reports

guided by diagnosis-aware prompts. We also compare RADAR with LLaVA-Rad and MAIRA-2 on the IU X-RAY dataset. On the CHEXPERT-PLUS dataset, we compare RADAR with the baseline SWIN$_{v2}$-BERT [15] consisting of a Swin Transformer V2 [110] and a BERT decoder [30]. The SWIN$_{v2}$-BERT model includes three variants, each trained on a distinct dataset: the MIMIC-CXR dataset, the CHEXPERT PLUS dataset, and a combined version of both.

Hyperparameters	**Stage I**	**Stage II**
Trainable Module	Vision Encoder (LoRA) Perceiver Resampler (Full) LLM (LoRA)	LLM (LoRA)
Training Epoch	3	2
Learning Rate	$1e-4$	
Optimizer	AdamW	
LR Scheduler	Cosine	
Warmup Ratio	0.03	
LoRA Config	$r = 64, \alpha = 128$	
Batch Size	32	

Table 4.1: Detailed hyperparameters for training RADAR. LoRA is used to fine-tune both the vision encoder and the LLM, while the Perceiver Resampler is fully fine-tuned.

4.4.4 Implementation Details

Training and Inference. We implement RADAR using BLIP-3[3] [192] as the backbone, which comprises a SigLIP [210] vision encoder, a Perceiver Resampler, and a Phi-3-mini$_{3.8B}$ [1] language model. The expert model consists of a Swin Transformer V2[4] [110] and a BioClinicalBERT[5] [3]. Top-2 reports are selected as knowledge. The

[3]The model card is "Salesforce/xgen-mm-phi3-mini-instruct-interleave-r-v1.5."
[4]The model card is "microsoft/swinv2-large-patch4-window12to16-192to256-22kto1k-ft."
[5]The model card is "emilyalsentzer/Bio_ClinicalBERT."

Dataset: MIMIC-CXR (*Training and Evaluation*)											
Model	**NLG Metrics**				**Clinical Metrics** (*CheXpert: Uncertain as Negative / Positive*)						
	B-1	**B-4**	**MTR**	**R-L**	**RG-F_1**	**RG_{ER}**	**$CliQ_0$ (↓)**	**^{14}Ma-F_1**	**^{5}Ma-F_1**	**^{14}Mi-F_1**	**^{5}Mi-F_1**
RadFM	–	0.128	–	0.182	–	–	–	–	–	–	–
XrayGPT	0.128	0.004	0.079	0.111	–	–	–	–	–	–	–
R2GenGPT	0.411	0.134	0.160	0.297	–	–	–	0.389	–	–	–
R2-LLM	0.402	0.128	0.175	0.291	–	–	–	–	–	–	–
RaDialog	0.346	0.095	0.140	0.271	–	–	–	0.394	–	–	–
LLaVA-Med	0.354	0.149	0.353	0.276	0.191	0.238	3.30	0.269	0.363	0.427	0.439
CheXagent	0.169	0.047	–	0.215	–	0.205	–	0.247	0.345	0.393	0.412
GPT-4V	0.164	0.178	–	0.132	–	0.132	–	0.204	0.196	0.355	0.258
Med-PaLM	0.323	0.115	–	0.275	0.267	–	–	0.398	0.516	0.536	0.579
LLaVA-Rad	0.381	0.154	–	0.306	–	0.294	–	0.395	0.477	0.573	0.574
MAIRA-1	0.392	0.142	0.333	0.289	0.243	0.296	3.10	0.386 0.423	0.477 0.517	0.557 0.553	0.560 0.588
MAIRA-2	0.465	0.234	0.420	<u>0.384</u>	**0.346**	**0.396**	<u>2.64</u>	<u>0.416</u>	0.504	0.581	0.591
MedVerse	–	0.178	–	–	0.280	–	2.71	–	–	–	–
M4CXR	0.339	0.103	–	–	0.218	0.285	–	0.400	0.495	<u>0.606</u>	<u>0.618</u>
Libra	**0.513**	<u>0.245</u>	**0.489**	0.367	0.329	0.376	2.70	0.404	<u>0.538</u>	0.559	0.601
RADAR (Ours)	<u>0.509</u>	**0.262**	0.450	**0.397**	**0.346**	<u>0.393</u>	**2.61**	**0.460** 0.497	**0.567** 0.602	**0.627** 0.627	**0.653** 0.674

Table 4.2: Evaluation results of our model and baseline methods on the MIMIC-CXR dataset. Baseline results are cited from their respective literature. The best results are shown in **bold**, while <u>underlined</u> values indicate the second-best results. ↓ denotes that lower values are better. Results of CheXpert treat *Uncertain* labels as *Positive* when compared with MAIRA-1.

Dataset: IU X-RAY (*Evaluation Only*)						
Model	**NLG Metrics**		**Clinical Metrics**			
	B-4	**R-L**	**RG-F_1**	**CliQ$_0$** (↓)	**^{14}Ma-F_1**	**^{14}Mi-F_1**
LLaVA-Rad	–	0.253	–	–	–	0.535
MAIRA-2	0.117	0.274	0.271	2.68	0.319	0.525
RADAR (Ours)	0.116	0.276	0.237	2.78	0.325	0.546
BACKBONE	0.112	0.275	0.236	2.79	0.269	0.514

Table 4.3: Experimental results on the IU X-RAY dataset, with results for the models LLaVA-Rad and MAIRA-2 cited from [9].

hyperparameters used for training RADAR are provided in Table 4.1. During inference, we employ beam search with a beam width of 5 for report generation and set the length penalty to 2.0. As proposed by [192], BLIP-3 samples vision tokens using a Perceiver Resampler with learned queries and supports images of any resolution, resulting in significant performance gains across multiple tasks. In this chapter, we use only the base resolution (384×384) with 128 learned query tokens to ensure a fair comparison with other baselines. For training, in Stage I, we fine-tune all three components (i.e., the vision encoder, the Perceiver Resampler, and the LLM) in BLIP-3 since the model is not specifically designed for medical tasks. In Stage II, we further fine-tune only the LoRA of the LLM to enhance performance.

Data Preprocessing. Following previous research [66, 9, 215], we incorporate *Indication*, *History*, *Comparison*, *Technique*, and *Prior Findings* as clinical context for the MIMIC-CXR and CHEXPERT PLUS datasets, when available. Since the IU X-RAY dataset does not include follow-up studies, we extract only *Indication*, *Comparison*, and *Technique* as clinical context. For a better illustration, we provide the prompt template in Table 4.9.

Dataset: CheXpert Plus (*Evaluation Only*)						
Model	**Train**	**NLG Metrics**		**Clinical Metrics**		
		B-4	**R-L**	**RG**$_{\overline{\text{ER}}(\text{ER})}$	$^{14(5)}$**Ma-F**$_1$	$^{14(5)}$**Mi-F**$_1$
Swin$_{\text{v2}}$-BERT	M*	0.034	0.191	0.136 (0.198)	0.268 (0.383)	0.410 (0.423)
	C	0.057	0.228	0.183 (0.250)	0.331 (0.401)	0.508 (0.432)
	M&C	0.056	0.234	0.201 (0.277)	0.366 (0.495)	0.560 (0.532)
Radar (Ours)	M	0.076	0.203	0.143 (0.216)	0.362 (0.417) 0.401 (0.540)	0.541 (0.524) 0.554 (0.608)
Backbone	M	0.073	0.203	0.143 (0.206)	0.282 (0.437) 0.317 (0.502)	0.477 (0.466) 0.492 (0.552)

Table 4.4: Evaluation on the CheXpert Plus dataset. The results for Swin$_{\text{v2}}$-BERT are cited from [15], and we primarily compare Radar with its $\star$ variant. The "Train" column indicates the training datasets, where M and C denote the MIMIC-CXR and CheXpert Plus datasets, respectively.

4.5 Results and Analyses

4.5.1 Quantitative Analysis

Comparison with MLLMs. As shown in Table 4.2, Radar achieves SOTA performance compared to other MLLM baselines. In terms of lexical metrics, Radar outperforms the best baselines (i.e., Libra and MAIRA-2) with absolute improvements of 1.7% in BLEU-4 and 1.3% in ROUGE-L, while maintaining competitive performance of 0.509 in BLEU-1 and 0.450 in METEOR. Regarding entity-level clinical metrics, our model achieves the best performance on RG-F$_1$ and RadCliQ$_0$, attaining scores of 0.346 and 2.61, respectively. Additionally, Radar surpasses the top three baselines, achieving improvements across multiple observation-level clinical metrics, with ^{14}Macro-F$_1$ increasing to 0.460, ^{5}Macro-F$_1$ to 0.567, ^{14}Micro-F$_1$ to 0.627, and ^{5}Micro-F$_1$ to 0.653, respectively. Notably, the smallest gain over the second-best model is 2.1%,

Dataset: MIMIC-CXR *(Compared with SOTA Specialists)*									
Model	**NLG Metrics**						**CE (^{14}Macro) Metrics**		
	B-1	**B-2**	**B-3**	**B-4**	**MTR**	**R-L**	**P**	**R**	**F_1**
R2Gen	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
R2GenCMN	0.353	0.218	0.148	0.106	0.142	0.278	0.344	0.275	0.278
$\mathcal{M}^2$Tr	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308
KnowMat	0.363	0.228	0.156	0.115	–	0.284	0.458	0.348	0.371
CMM-RL	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
CMCA	0.360	0.227	0.156	0.117	0.148	0.287	0.444	0.297	0.356
KiUT	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
DCL	–	–	–	0.109	0.150	0.284	0.471	0.352	0.373
METrans	0.386	0.250	0.169	0.124	0.152	0.291	0.364	0.309	0.311
RGRG	0.373	0.249	0.175	0.126	0.168	0.264	0.380	0.319	0.305
ORGan	0.407	0.256	0.172	0.123	0.162	0.293	0.416	0.418	0.385
Recap	0.429	0.267	0.177	0.125	0.168	0.288	0.389	0.443	0.393
Controllable	0.486	0.366	0.295	0.246	0.216	**0.423**	**0.597**	**0.516**	**0.553**
PromptMRG	0.398	–	–	0.112	0.157	0.268	0.396	0.393	0.381
ICon	0.429	0.266	0.178	0.126	0.170	0.287	0.445	0.505	0.464
Radar (Ours)	**0.509**	**0.390**	**0.315**	**0.262**	**0.450**	0.397	**0.481** 0.523	**0.474** 0.500	**0.460** 0.497

Table 4.5: Experimental results of our model and SOTA specialists on the MIMIC-CXR dataset. Results denotes *Uncertain* as *Positive*.

underscoring RADAR's effectiveness. Furthermore, we provide an additional set of CheXpert results using the *Uncertain as Positive* policy and compare RADAR with MAIRA-1. We observe that the improvements under this setting follow a similar trend to those obtained with the *Uncertain as Negative* policy. These results collectively demonstrate the effectiveness of RADAR in generating coherent and clinically accurate radiology reports.

Comparison with SOTA Specialists. The results of other specialists on the MIMIC-CXR dataset are shown in Table 4.5. These specialists are mainly based on Transformers and demonstrate strong performance. We find that models incorporating clinical context (e.g., *Indication*) as input generally achieve better performance than those that do not, as the clinical context can provide the rationale for the study, including relevant clinical history and other pertinent information. For example, the Controllable model significantly outperforms other baselines across both lexical and clinical metrics, achieving a B-4 score of 0.246 and a ^{14}Macro-F$_1$ of 0.553. This trend also holds for MLLMs (e.g., MAIRA-1/2 and Libra), as shown in Table 4.2. Moreover, benefiting from the strong contextual comprehension and language generation capabilities of LLMs, RADAR further improves linguistic quality, which requires models to integrate diverse information sources. However, when evaluating under the *Uncertain as Positive* CheXpert setting, we observe that the ^{14}Macro-F$_1$ score of our model still lags behind that of the Controllable baseline (0.497 vs. 0.553). This discrepancy may stem from differences in learning objectives, as this baseline treats *Uncertain* cases as *Positive*.

Model Generalization. Following prior research [9], we further evaluate RADAR on the CHEXPERT PLUS and IU X-RAY datasets to assess its generalization capability, with the results presented in Table 4.3. On the IU X-RAY dataset, RADAR outperforms MAIRA-2 in the CheXpert metrics, achieving a ^{14}Macro-F$_1$ score of 0.325 and a ^{14}Micro-F$_1$ score of 0.546. However, a performance gap remains in RG-F$_1$ and RadCliQ$_0$, which may be attributed to differences in training data, as MAIRA-2 is trained with the

additional USMix dataset. Meanwhile, RADAR demonstrates comparable performance to the baselines in terms of lexical metrics. When compared to another baseline, namely LLaVA-Rad, our model outperforms it on two of the provided metrics (R-L and ^{14}Micro-F_1). On the CHEXPERT PLUS dataset, our model significantly outperforms SWIN$_{V2}$-BERT trained on the MIMIC-CXR dataset, across both lexical and clinical metrics. Furthermore, RADAR surpasses the baseline that is trained on CHEXPERT PLUS alone as well as the one trained on a combination of both datasets. These results demonstrate the strong generalization ability of RADAR across different datasets. Additionally, RADAR significantly outperforms the BACKBONE, underscoring the effectiveness of the integrated knowledge.

Analysis of PF, SF, and OI. We analyze the impact of PF, SF, and OI on the performance of RADAR, with results summarized in Table 4.6. RADAR$_{w/o\ F}$, which first identifies observations before report generation without incorporating knowledge, significantly improves the CheXpert metrics, particularly ^{14}Macro-F_1 and ^{5}Macro-F_1, as observation information captures high-level abstractions of reports and aligns closely with the objectives of these metrics. This highlights the crucial role of OI in enhancing clinical accuracy, independent of other components. When PF and SF are introduced individually with OI, introducing PF alone helps preserve the knowledge embedded in the LLM, resulting in comparable performance across both lexical and clinical metrics. In contrast, introducing SF alone substantially improves $^{14/5}$Macro-F1, but negatively impacts RG_{ER} and $RadCliQ_0$. Moreover, combining both PF and SF leverages the strengths of each, leading to further improvements in the clinical metrics while maintaining comparable performance across the other metrics. We notice that BACKBONE tends to retain easily acquired knowledge (i.e., PF) and that selectively supplementing it with external information (i.e., SF) is crucial for bridging the remaining knowledge gaps.

Analysis of RADAR versus RAG. To evaluate the effectiveness of knowledge integration in RADAR, we conduct experiments comparing our model against three

Model	Modules			Lexical Metrics				Clinical Metrics (*CheXpert: Uncertain as Negative*)						
	PF	**SF**	**OI**	**B-1**	**B-4**	**MTR**	**R-L**	**RG-F$_1$**	**RG$_{\text{ER}}$**	**CliQ$_0$ (↓)**	**^{14}Ma-F$_1$**	**^{5}Ma-F$_1$**	**^{14}Mi-F$_1$**	**^{5}Mi-F$_1$**
RADAR	✓	✓	✓	0.509	0.262	0.450	0.397	0.346	0.393	2.61	0.460	0.567	0.627	0.653
BACKBONE	✗	✗	✗	0.497	0.259	0.444	0.396	0.343	0.387	2.67	0.402	0.495	0.565	0.581
RADAR$_{w/o\ \text{F}}$	✗	✗	✓	0.506	0.260	0.448	0.396	0.343	0.391	2.63	0.442	0.545	0.624	0.651
RADAR$_{w/o\ \text{SF}}$	✓	✗	✓	0.508	0.262	0.451	0.398	0.346	0.394	2.62	0.447	0.543	0.626	0.650
RADAR$_{w/o\ \text{PF}}$	✗	✓	✓	0.508	0.261	0.450	0.396	0.344	0.389	2.63	0.456	0.559	0.623	0.652

Table 4.6: Ablation results of RADAR with different modules. Per-observation results of BACKBONE, Variant (a), Variant (b), and RADAR are provided in Appendix, Table 4.8.

Model	Modules			NLG Metrics				Clinical Metrics (*CheXpert: Uncertain as Negative*)						
	Vision	**Resampler**	**LLM**	**B-1**	**B-4**	**MTR**	**R-L**	**RG-F$_1$**	**RG$_{\text{ER}}$**	**CliQ$_0$ (↓)**	**^{14}Ma-F$_1$**	**^{5}Ma-F$_1$**	**^{14}Mi-F$_1$**	**^{5}Mi-F$_1$**
BACKBONE	✓	✓	✓	0.497	0.259	0.444	0.396	0.343	0.387	2.67	0.402	0.495	0.565	0.581
BACKBONE-V1	✗	✓	✗	0.430	0.183	0.359	0.318	0.245	0.296	3.15	0.284	0.415	0.476	0.508
BACKBONE-V2	✗	✓	✓	0.483	0.246	0.428	0.381	0.321	0.368	2.78	0.361	0.465	0.532	0.550

Table 4.7: Ablation results of fine-tuning different modules of BACKBONE.

Observation	P	R	F_1
Atelectasis	0.518	0.645	0.574
Cardiomegaly	0.656	0.783	0.713
Consolidation	0.370	0.174	0.237
Edema	0.518	0.610	0.560
Pleural Effusion	0.695	0.800	0.744
[5]Macro Average	0.551	0.602	0.567
[5]Micro Average	0.607	0.707	0.653
Enlarged Cardiomediastinum	0.277	0.204	0.235
Lung Opacity	0.644	0.496	0.561
Lung Lesion	0.492	0.207	0.291
Pneumonia	0.283	0.232	0.255
Pneumothorax	0.407	0.636	0.496
Pleural Other	0.333	0.173	0.228
Fracture	0.421	0.244	0.309
Support Devices	0.823	0.866	0.844
No Finding	0.302	0.569	0.395
[14]Macro Average	0.481	0.474	0.460
[14]Micro Average	0.614	0.640	0.627

Table 4.8: Experimental results of RADAR for each observation on the MIMIC-CXR dataset.

baselines: (1) BACKBONE PLUS, (2) BACKBONE+RAG, and (3) BACKBONE+PF+SF. The results are presented in Figure 4.3. Note that these baselines do not include the OI. Since RADAR undergoes two-stage training (i.e., two additional epochs), we apply the same extended training to BACKBONE, referring to this variant as BACKBONE PLUS. In addition, we introduce a standard RAG baseline (BACKBONE+RAG), which utilizes the same retrieved findings as RADAR. Building upon this baseline, BACKBONE+PF+SF further includes PF as context. Our findings reveal that while all four models achieve comparable performance on lexical metrics (e.g., 50%/26% B-1/4),

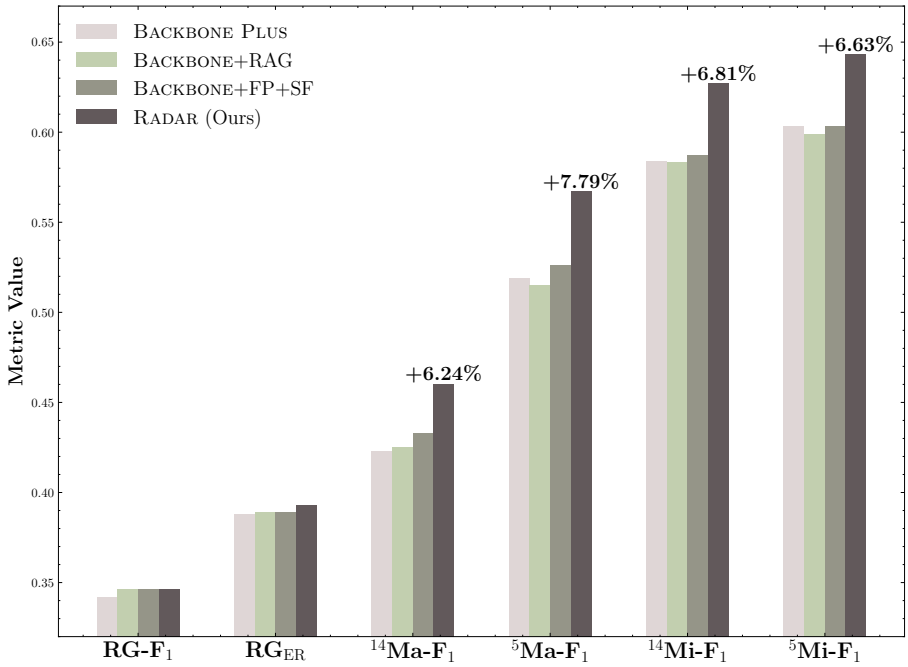


Figure 4.3: Comparisons among Backbone+RAG, Backbone+FP+SF, and Radar on six clinical metrics.

they differ in clinical metrics. Specifically, Backbone+RAG and Backbone Plus show similar performance, and Backbone+FP+SF outperforms these two baselines on CheXpert metrics and exhibits similar performance on RadGraph metrics. This demonstrates that incorporating credible knowledge can effectively enhance report generation even without OI. Moreover, Radar demonstrates a relative improvement of over 6% across four key CheXpert metrics. This suggests that structured integration of internal and external knowledge contributes to its enhanced clinical accuracy.

Analysis of Fine-tuning Different Modules in Backbone. To assess the contributions of different components in the base model (i.e., BLIP-3), we conduct an ablation study on the impact of fine-tuning the vision encoder, the Perceiver Resampler, and the LLM. The experimental results are presented in Table 4.7. By comparing Backbone and Backbone-V2, we find that fine-tuning the vision encoder

to incorporate domain-specific knowledge is crucial for achieving high clinical accuracy, even though both configurations exhibit strong language coverage in lexical metrics. In particular, ^{14}Macro-F$_1$ and ^{5}Macro-F$_1$ increase from 0.361 to 0.402 and from 0.465 to 0.495, respectively. In addition, there is a notable 3% improvement in both ^{14}Micro-F$_1$ and ^{5}Micro-F$_1$. Furthermore, fine-tuning the LLM (i.e., Phi-3) leads to substantial improvements in both lexical and clinical metrics, as demonstrated by the comparison between V1 and V2, where an improvement of nearly 5% across all metrics is observed. Notably, RADAR employs a lightweight LLM with 3.8B parameters as the language decoder and still outperforms many larger models (e.g., LLaVA-Med and MAIRA-1, both using 7B LLMs). This underscores both the effectiveness of the backbone and the importance of domain-specific adaptation for achieving optimal performance.

4.5.2 Qualitative Analysis

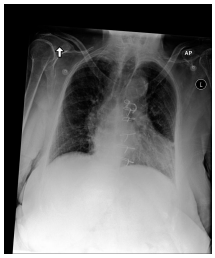
Case Study. We conduct a case study to illustrate the advantages of incorporating both internal knowledge and retrieved information, as shown in Figure 4.4. In Case A, RADAR initially generates a report that includes the finding *Atelectasis*. However, expert assessment indicates the image shows no abnormal findings. As a result, the intersection between the preliminary report and the expert model's outputs is $\emptyset$, and by removing this incorrect observation, RADAR ultimately produces an accurate report. This example highlights the model's ability to refine its predictions when guided by expert constraints, effectively eliminating unnecessary or incorrect findings. Another more complex case is presented on the right side (Case B) of this Figure. Specifically, RADAR initially identifies findings related to *Edema* and *Cardiomegaly*, which the expert model also notes. However, the observation of *Atelectasis* is omitted from the preliminary findings, while indicated by the supplementary findings. By incorporating retrieved evidence such as *"... linear atelectasis ..."* and *"Mild areas of atelectasis ..."*, RADAR successfully corrects the omission and generates a complete and accurate report that includes all the relevant observations. This case demonstrates the

Figure 4.4: Two cases generated by RADAR, where false positive observation appears in the PF of case A, and false negative observation shows in the PF of case B.

Indication: History of chest pain x20 minutes, please evaluate for pneumonia or other acute process.

Prior

Current



Preliminary Findings: O_R **[Cardiomegaly]**
[Cardiomegaly] Mild cardiomegaly has been stable compared to exams dated back to at least ____.
Supplementary Findings:
[Edema] There is mild-to-moderate interstitial edema.
[Lung Opacity] ....., there is an ill-defined somewhat rounded opacity .....
[Fracture] Left lateral rib fractures are again noted.
Expert Observations: O_I **[Edema, Fracture, Cardiomegaly]**
Agreement: $O_\checkmark$ **[Cardiomegaly]**
Overall Findings:
[Cardiomegaly] Mild cardiomegaly is stable compared to exams dated back to at least ____.
~~[Edema] There is mild pulmonary vascular congestion with mild pulmonary edema.~~ .....
Reference:
[Cardiomegaly] There is stable mild cardiomegaly.
[Lung Opacity] There appears to be a subtle increase in opacification in the retrocardiac region, ...

Figure 4.5: Error case generated by RADAR, where ~~spans~~ and spans indicate incorrect and correct findings, respectively.

model's capability to leverage external knowledge to recover missing findings, thereby improving factual completeness.

Error Analysis. We conduct an error analysis to gain deeper insights on the report generation process, as shown in Figure 4.5. RADAR initially generates a report containing the observation *Cardiomegaly*, which is also present in the expert model's output. In this case, the observation reflects credible knowledge possessed by the LLM and should be preserved. Subsequently, RADAR produces a false prediction of finding, *Edema*, which aligns with the retrieved supplementary findings. This error may result

from the model's overreliance on external knowledge. Moreover, since *Edema* is clinically associated with *Cardiomegaly*, it is possible that RADAR has learned only superficial correlations between them. To address these issues, potential solutions include refining the expert model and expanding the training dataset. Additionally, another common error is the omission of finding [205], where the model misses a case of *Lung Opacity*. Despite relevant information in the supplementary findings, RADAR fails to leverage it. This could be mitigated with relevance-aware integration to promote the utilization of such information.

4.6 Chapter Summary

In this chapter, we introduce RADAR, a novel approach designed to enhance radiology report generation by leveraging both the internal knowledge of an LLM and externally retrieved information. Our model first generates a report and subsequently classifies the image based on observations, with their shared components regarded as internal knowledge. It then retrieves supplementary information to further refine and complement this knowledge. Extensive experiments on three public datasets demonstrate that RADAR achieves SOTA performance in both language quality and clinical accuracy, highlighting the effectiveness of integrating internal and external knowledge for more accurate and coherent radiology report generation.

Our experiments are conducted using a single backbone architecture. While this choice provides a controlled evaluation, the performance of alternative architectures remains unexplored. Future work should investigate whether different model architectures can achieve comparable or better results. In addition, our study focuses exclusively on a single imaging modality (e.g., Chest X-ray). The model's effectiveness in other imaging modalities, such as CT scans or MRI, has not been evaluated. Extending our approach to multiple imaging modalities would be an important direction for future research to enhance its clinical utility and generalizability.

Role	Prompt
SYSTEM	<|system|> You are an assistant in radiology, responsible for analyzing medical imaging studies and generating detailed, structured, and accurate radiology reports. <|end|>
USER	<|user|> <prior image> (*If prior available*) <current image> *Indication*: ..... *History*: ..... *Comparison*: ..... *Technique*: ..... *Prior Findings*: ..... (*If prior available*) *Preliminary Findings*: ..... (*If available*) *Supplementary Findings*: ..... (*If available*) Generate a comprehensive and detailed description of findings based on this chest X-ray image. Include a thorough comparison with a prior chest X-ray, emphasizing any significant changes, progression, or improvement. (*If prior available*) Before this, systematically identify all observations. <|end|>
ASSISTANT	<|assisitant|> *Identified Observations*: ..... *Overall Findings*: *(e.g., the target)* ..... <|end|>

Table 4.9: The prompt template for RADAR and its variants, consisting of three roles: System, User, and Assistant.

Chapter 5

Spatiotemporally Precise Radiology Report Generation

5.1 Introduction

Aiming to generate clinically coherent and factually accurate free-text reports, many research works [126, 125, 26, 10, 162] have made significant efforts in improving the clinical factuality of generated reports. Despite their progress, these methods still struggle to produce precise and accurate free-text reports. One significant problem within these methods is that although they successfully captured the semantic information of observations, their attributes still remain imprecise. They either ignored historical records (i.e., temporal information) that are required for assessing patients' current conditions or omitted the fine-grained attributes of observations (i.e., spatial information) that are crucial in quantifying the severity of diseases, which are far from adequate and often lead to imprecise reports. Incorporating both temporal and spatial information are important for generating precise and accurate reports. For instance, as illustrated in Figure 5.1, the patient's conditions can change from time to time, and the observations become `Better` and `Stable`. Only if accessing the historical records,

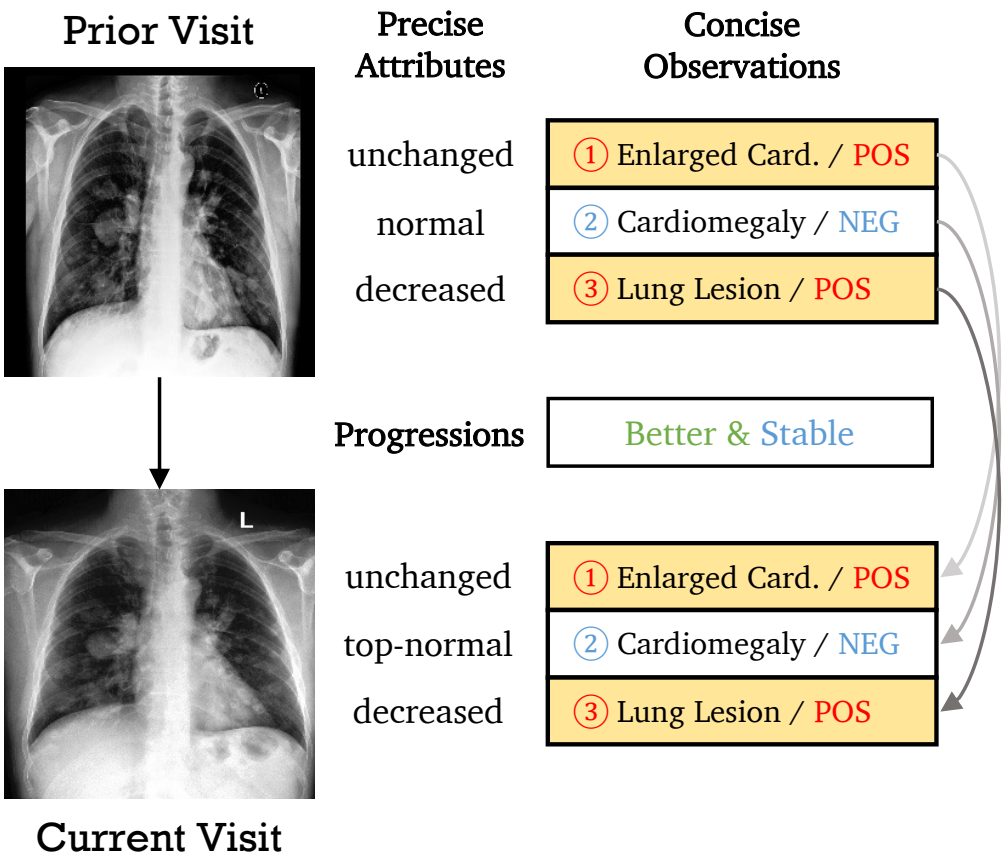


Figure 5.1: An example of a follow-up visit record with its prior visit record. Part of their observations are listed with their precise attributes. *Enlarged Card.* denotes *Enlarged Cardiomediastinum*.

the overall conditions could be estimated. In addition, different attributes reflect the severity of an observation, such as *normal* and *top-normal* for *Cardiomegaly*. In order to produce precise and accurate free-text reports, we must consider both kinds of information and apply stronger reasoning to strengthen the generation process with precise attribute modeling.

In this chapter, we propose RECAP, which captures both temporal and spatial information for radiology REport Generation via DynamiC DiseAse Progression Reasoning. Specifically, RECAP first predicts observations and progressions given two consecutive radiographs. It then combines them with the historical records and the current radiograph for report generation. To achieve precise attribute modeling, we construct a disease progression graph, which contains the prior and current observations, the progressions, and the precise attributes. We then devise a dynamic progression reasoning (PrR) mechanism that aggregates information in the graph to select observation-relevant attributes.

Our contributions can be summarized as follows:

- We propose RECAP, which can capture both spatial and temporal information for generating precise and accurate free-text reports.

- To achieve precise attribute modeling, we construct a disease progression graph containing both observations and fine-grained attributes that quantify the severity of diseases. Then, we devise a dynamic disease progression reasoning (PrR) mechanism to select observation/progression-relevant attributes.

- We conduct extensive experiments on two publicly available benchmarks, and experimental results demonstrate the effectiveness of our model in generating precise and accurate radiology reports.

5.2 Preliminary

5.2.1 Problem Formulation

Given a radiograph-report pair $D^c = \{X^c, Y^c\}$, with its record of last visit being either $D^p = \{X^p, Y^p\}$ or $D^p = \emptyset$ if the historical record is missing, the task of radiology report generation aims to maximize $p(Y^c|X^c, D^p)$. Note that there are two kinds of records (i.e., first-visit and follow-up-visit). If it is the first visit of a patient, the historical record does not exist. To learn the spatiotemporal information, observations O (i.e., spatial information) [67] and progressions P (i.e., temporal information) [187] are introduced. Then, the report generation process is divided into two stages in our framework, i.e., observation and progression prediction (i.e., Stage 1) and spatiotemporal-aware report generation (i.e., Stage 2). Specifically, the probability of observations and progressions are denoted as $p(O|X^c)$ and $p(P|X^c, X^p)$, respectively, and then the generation process is modeled as $p(Y^c|X^c, D^p, O, P)$.

5.2.2 Progression Graph Construction

Observation and Progression Extraction. For each report, we first label its observations $O = \{o_1, \ldots, o_{|o|}\}$ with CheXbert [157]. Similar to [57], each observation is further labeled with its status (i.e., *Positive*, *Negative*, *Uncertain*, and *Blank*). We convert *Positive* and *Uncertain* as POS, *Negative* as NEG, and remove *Blank*, as shown in Figure 5.1. Then, we extract progression information P of a patient with Chest ImaGenome [187] which provides progression labels (i.e., `Better`, `Stable`, or `Worse`) between two regions of interest (ROIs) in X^p and X^c, respectively. However, extracting ROIs could be difficult, and adopting such ROI-level labels may not generalize well across different datasets. Thus, we use image-level labels, which only indicate whether there are any progressions between X^p and X^c. As a result, a patient may have different progressions (e.g., both `Better` and `Worse`). The statistics of observations

#Observation	MIMIC-ABN	
	Positive	**Negative**
No Finding	5002/32/22	66,784/514/784
Cardiomegaly	16,312/118/244	804/4/8
Pleural Effusion	10,502/80/186	1,948/18/24
Pneumothorax	1,452/24/4	1,792/10/26
Enlarged Card.	5,202/40/90	1,194/10/14
Consolidation	4,104/36/96	3,334/20/34
Lung Opacity	22,598/166/356	748/10/4
Fracture	4,458/32/76	330/0/0
Lung Lesion	5,612/54/112	120/2/2
Edema	8,704/76/168	1,898/16/32
Atelectasis	19,132/134/220	116/2/0
Support Devices	9,886/58/196	394/0/10
Pneumonia	17,826/138/260	3,226/22/34
Pleural Other	2,850/30/62	8/0/0

Table 5.1: Observation distribution in the train/valid/test split of the MIMIC-ABN dataset, and the distribution of the MIMIC-CXR dataset is provided in Table 3.1.

and progressions can be found in Table 5.1 and Table 5.2.

Spatial/Temporal Entity (Attribute) Collection. To model spatial and temporal information, we collect a set of entities to represent it. For temporal entities, we adopt the entities provided by [10], denoted as E^T. For spatial entities E^S, we adopt the entities with a relation *modify* or *located_at* in RadGraph [69], and we also filter out stopwords and temporal entities from them. Attributes are included in the entity set as provided by [69]. For simplicity, we use "attribute" and "entity" interchangeably in this chapter. Part of the temporal and spatial entities are listed here: healed, fractured, healing, nondisplaced, top, size, heart, normal, mediastinum,

#Progression	MIMIC-ABN	MIMIC-CXR
`Better`	929/2/19	14,790/110/345
`Worse`	1,219/6/30	18,083/163/431
`Stable`	4,114/31/99	41,721/334/1,085
Total	6,440/48/137	64,498/535/1,566
Ratio	9%/8.8%/17%	24%/25.1%/40.6%

Table 5.2: Progression distribution in train/valid/test split of the MIMIC-ABN and MIMIC-CXR datasets.

widening, contour, widened, consolidative, collapse, underlying, developing, fibrosis, thickening, biapical, blunting, indistinctness, asymmetrical, haziness, asymmetric, layering, subpulmonic, thoracentesis, trace, small, adjacent, tiny, atypical, developing, supervening, multifocal, correct, superimposed, patchy, and borderline. For temporal entities, we use the same settings of [10], which are: bigger, change, cleared, constant, decrease, decreased, decreasing, elevated, elevation, enlarged, enlargement, enlarging, expanded, greater, growing, improved, improvement, improving, increase, increased, increasing, larger, new, persistence, persistent, persisting, progression, progressive, reduced, removal, resolution, resolved, resolving, smaller, stability, stable, stably, unchanged, unfolded, worse, worsen, worsened, worsening and unaltered. We list top-5 attributes for each observation in Table 5.3.

Progression Graph Construction. Our progression graph $G =< V, R >$ is constructed based purely on the training corpus in an unsupervised manner. Specificially, $V = \{O, E^T, E^S\}$ is the node-set, and $R = \{S, B, W, R_S, R_O\}$ is the edge set, where S, B, and W denote three progressions `Stable`, `Better`, and `Worse`, connecting an observation with an temporal entity. In addition, R_s and R_o are additional relations connecting current observations with spatial entities and prior/current observations, respectively. To extract spatial/temporal triples automatically, we use the proven-efficient statistical tool, i.e., pointwise mutual information (PMI) [24], where a higher

Observation	**Top-5 Attributes**
Cardiomegaly	cardiomegaly, borderline, moderately, severely, mildly
Pleural Effusion	layering, subpulmonic, thoracentesis, trace, small
Pneumothorax	hydropneumothorax, apical, tiny, tension, component
Enlarged Card.	mediastinum, widening, contour, widened, lymphadenopathy
Consolidation	consolidative, collapse, underlying, developing, consolidations
Lung Opacity	opacification, opacifications, patchy, heterogeneous, scarring
Fracture	healed, fractured, healing, nondisplaced, posterolateral
Lung Lesion	nodular, nodule, mass, nodules, mm
Edema	indistinctness, asymmetrical, haziness, asymmetric, interstitial
Atelectasis	atelectatic, atelectasis, collapsed, subsegmental, collapse
Support Devices	sidehole, carina, coiled, tunneled, duodenum
Pneumonia	infectious, infection, atypical, supervening, developing
Pleural Other	fibrosis, thickening, biapical, blunting, scarring

Table 5.3: Top-5 attributes for each observation.

PMI score implies two units with higher co-occurrence, similar to [57]:

$$\text{PMI}(\bar{x}, \hat{x}) = \log \frac{p(\bar{x}, \hat{x})}{p(\bar{x})p(\hat{x})} = \log \frac{p(\hat{x}|\bar{x})}{p(\hat{x})}, \tag{5.1}$$

Specifically, we set $\bar{x}$ to (o_i, r_j) where $r_j \in R$ and set $\hat{x}$ to e_k^* where $e_k^* \in \{E^T, E^S\}$. Then, we rank these triples using $\text{PMI}((o_i, r_j),\ e_k^*)$ and select top-K of them as candidates for each (o_i, r_j). Finally, we use observations as the query to retrieve relevant triples. We consider edges in the graph: $e_i^* \xrightarrow{r_j} o_k^p \xrightarrow{R_O} o_l^c \xrightarrow{r_m} e_n^*$, as shown in the top-right of Figure 5.2, consistent with the progression direction.

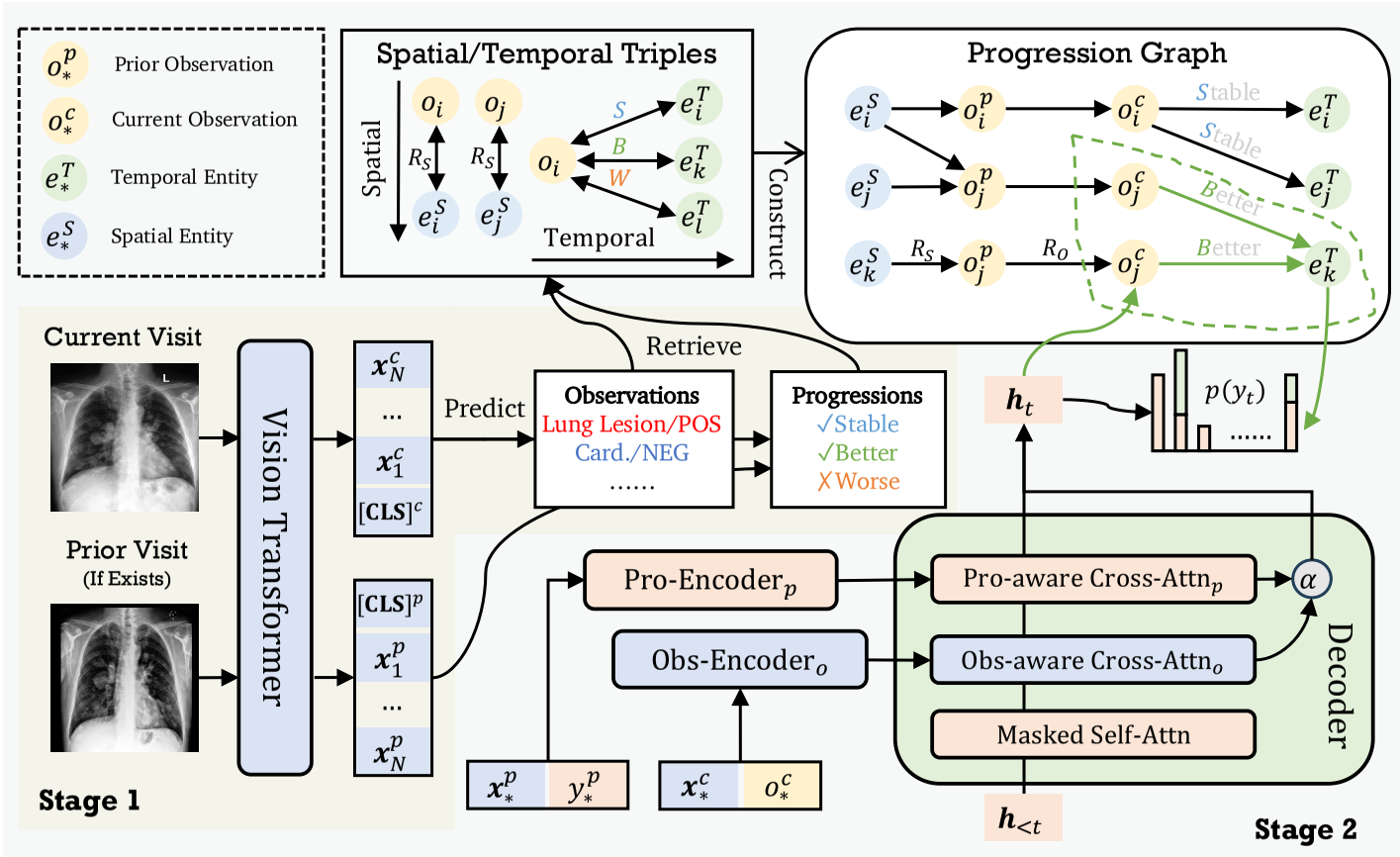


Figure 5.2: Overview of the RECAP framework. *Pro-Encoder*$_p$ is the progression-related encoder and *Obs-Encoder*$_o$ is the observation-related encoder, respectively. Other modules in the decoder are omitted for simplicity.

5.3 Method

5.3.1 Visual Representation Extraction

Given an image X^c, an image processor is first to split it into N patches, and then a visual encoder (i.e., Vision Transformer [32]) is adopted to extract visual representations $\boldsymbol{X}^c$:

$$\boldsymbol{X}^c = \{[\mathbf{CLS}]^c, \boldsymbol{x}_1^c, \ldots, \boldsymbol{x}_N^c\} = \mathrm{ViT}(X^c), \tag{5.2}$$

where $[\mathbf{CLS}]^c \in \mathbb{R}^h$ is the representation of the class token `[CLS]` prepended in the patch sequence, $\boldsymbol{x}_i^c \in \mathbb{R}^h$ is the i-th visual representation. Similarly, the visual representation of image X^p is extracted using the same ViT model and represented as $\boldsymbol{X}^p = \{[\mathbf{CLS}]^p, \boldsymbol{x}_1^p, \ldots, \boldsymbol{x}_N^p\}$.

5.3.2 Observation and Progression Prediction

Observation Prediction. As observations can be measured from a single image solely, we only use the pooler output $[\mathbf{CLS}]^c$ of X^c for observation prediction. Inspired by [162], we divide it into two steps, i.e., detection and then classification. Specifically, the detection probability $p_d(o_i)$ of the i-th observation presented in a report and the probability of this observation $p_c(o_i)$ being classified as abnormal are modeled as:

$$p_d(o_i) = \sigma(\boldsymbol{W}_{d_i}[\mathbf{CLS}]^c + b_{d_i}), \tag{5.3}$$

$$p_c(o_i) = \sigma(\boldsymbol{W}_{c_i}[\mathbf{CLS}]^c + b_{c_i}), \tag{5.4}$$

where σ is the Sigmoid function, $\boldsymbol{W}_{d_i}, \boldsymbol{W}_{c_i} \in \mathbb{R}^h$ are the weight matrices and $b_{d_i}, b_{c_i} \in \mathbb{R}$ are the biases. Finally, the probability of the i-th observation is denoted as $p(o_i) = p_d(o_i) \cdot p_c(o_i)$. Note that for observation *No Finding* o_n is presented in every sample, i.e., $p_d(o_n) = 1$ and $p(o_n) = p_c(o_n)$.

Progression Prediction. Similar to observation prediction, the pooler outputs $[\mathbf{CLS}]^p$ of X^p and $[\mathbf{CLS}]^c$ of X^c are adopted for progression prediction, and the probability of the j-th progression $p(p_j)$ is modeled as:

$$[\mathbf{CLS}] = [[\mathbf{CLS}]^p; [\mathbf{CLS}]^c], \tag{5.5}$$

$$p(p_j) = \sigma(\boldsymbol{W}_j[\mathbf{CLS}] + b_j), \tag{5.6}$$

where $[;]$ is the concatenation operation, $\boldsymbol{W}_j \in \mathbb{R}^{2h}$ is the weight matrix, and $b_j \in \mathbb{R}$ are the bias. As we found that learning sparse signals from image-level progression labels is difficult and has side effects on the performance of observation prediction, we detach [**CLS**] from the computational graph while training.

Training. We optimize these two prediction tasks by minimizing the binary cross-entropy loss. Specifically, the loss of observation detection $\mathcal{L}_d$ is denoted as:

$$\mathcal{L}_d = -\frac{1}{|O|}\sum[\alpha_d \cdot l_{d_i} \cdot \log p_d(o_i) + (1 - l_{d_i}) \cdot \log(1 - p_d(o_i))], \tag{5.7}$$

where α_d is the weight to tackle the class imbalance issue, l_{d_i} denotes the label of i-th observation d_i. Similarly, the loss of observation classification $\mathcal{L}_c$ and progression prediction $\mathcal{L}_p$ can be calculated using the above equation. Note that $\mathcal{L}_c$ and $\mathcal{L}_p$ are unweighted loss. Finally, the overall loss of Stage 1 is:

$$\mathcal{L}_{\text{S1}} = \mathcal{L}_d + \mathcal{L}_c + \mathcal{L}_p. \tag{5.8}$$

5.3.3 SpatioTemporal-aware Radiology Report Generation

Observation-aware Visual Encoding. To learn the observation-aware visual representations, we jointly encode $\boldsymbol{X}^c$ and its observations O^c using a Transformer encoder [168]. Additionally, a special token `[FiV]` for first-visit records or `[FoV]` for follow-up-visit records is appended to distinguish them, represented as `[F*V]`:

$$\boldsymbol{h}^c = [\boldsymbol{h}^c_X; \boldsymbol{h}^c_o] = \text{Encoder}_o([\boldsymbol{X}^c; \texttt{[F*V]}; O^c]), \tag{5.9}$$

where $\boldsymbol{h}_X^c, \boldsymbol{h}_o^c \in \mathbb{R}^h$ are the visual hidden representations and observation hidden representations of the current radiograph and observations.

Progression-aware Information Encoding. We use another encoder to encode the progression information (i.e., temporal information). Specifically, given $\boldsymbol{X}^p$ and Y^p, the hidden states of the prior record are represented as:

$$\boldsymbol{h}^p = [\boldsymbol{h}_X^p; \boldsymbol{h}_Y^p] = \text{Encoder}_p([\boldsymbol{X}^p; Y^p]), \tag{5.10}$$

where $\boldsymbol{h}_X^p, \boldsymbol{h}_Y^p \in \mathbb{R}^h$ are the visual hidden representations and textual hidden representations of prior records, respectively.

Concise Report Decoding. Given $\boldsymbol{h}^p$ and $\boldsymbol{h}^c$, a Transformer decoder is adopted for report generation. Since not every sample has a prior record and follow-up records may include new observations, controlling the progression information is necessary. Thus, we include a soft gate α to fuse the observation-related and progression-related information, as shown in Figure 5.2:

$$\text{Decoder} = \begin{cases} \boldsymbol{h}_t^s = \text{Self-Attn}(\boldsymbol{h}_t^w, \boldsymbol{h}_{<t}^w, \boldsymbol{h}_{<t}^w), \\ \tilde{\boldsymbol{h}}_t^c = \text{Cross-Attn}_o(\boldsymbol{h}_t^s, \boldsymbol{h}^c, \boldsymbol{h}^c), \\ \tilde{\boldsymbol{h}}_t^p = \text{Cross-Attn}_p(\tilde{\boldsymbol{h}}_t^c, \boldsymbol{h}^p, \boldsymbol{h}^p), \\ \alpha = \sigma(\boldsymbol{W}_\alpha \tilde{\boldsymbol{h}}_t^c + b_\alpha), \\ \boldsymbol{h}_t = \alpha \cdot \tilde{\boldsymbol{h}}_t^p + (1 - \alpha) \cdot \tilde{\boldsymbol{h}}_t^c, \end{cases} \tag{5.11}$$

$$p_{\mathcal{V}}(y_t) = \text{Softmax}(\boldsymbol{W}_{\mathcal{V}} \boldsymbol{h}_t + \boldsymbol{b}_{\mathcal{V}}), \tag{5.12}$$

where Self-Attn is the self-attention module, Cross-Attn is the cross-attention module, $\boldsymbol{h}_t^s, \tilde{\boldsymbol{h}}_t^c, \tilde{\boldsymbol{h}}_t^p, \boldsymbol{h}_t \in \mathbb{R}^h$ are self-attended hidden state, observation-related hidden state, progression-related hidden state, and spatiotemporal-aware hidden state, respectively, $\boldsymbol{W}_\alpha \in \mathbb{R}^h, \boldsymbol{W}_{\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}| \times h}$ are weight matrices and $b_\alpha \in \mathbb{R}, \boldsymbol{b}_{\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|}$ are the biases.

Disease Progression Encoding. As there are different relations between nodes, we adopt an L-layer Relational Graph Convolutional Network (R-GCN) [145] to encode

the disease progression graph, similar to [70]:

$$\boldsymbol{h}_{v_i}^{l+1} = \text{ReLU}\left(\frac{1}{c_i}\sum_{v_k \in V}^{r_j \in R} \boldsymbol{W}_{r_j}^{l} \boldsymbol{h}_{v_k}^{l} + \boldsymbol{W}_0^{l} \boldsymbol{h}_{v_i}^{l}\right), \tag{5.13}$$

where c_i is the number of neighbors connected to the i-th node, $\boldsymbol{W}_{r_j}^{l}, \boldsymbol{W}_0^{l} \in \mathbb{R}^{h \times h}$ are learnable weight metrics, and $\boldsymbol{h}_{v_i}^{l}, \boldsymbol{h}_{v_i}^{l+1}, \boldsymbol{h}_{v_k}^{l} \in \mathbb{R}^{h}$ are the hidden representations.

Precise Report Decoding via Progression Reasoning. Inspired by [70] and [120], we devise a dynamic disease progression reasoning (PrR) mechanism to select observation-relevant attributes from the progression graph. The reasoning path of PrR is $o_i^c \xrightarrow{r_j} e_k$, where r_j belongs to either three kinds of progression or R_s. Specifically, given t-th hidden representation $\boldsymbol{h}_t$, the observation representation $\boldsymbol{h}_{o_i}^{L}$, and the entity representation $\boldsymbol{h}_{e_k}^{L}$ of e_k, the progression score $\hat{ps}_t(e_k)$ of node e_k is calculated as:

$$ps_t(e_k) = \frac{1}{|\mathcal{N}_{e_k}|}\sum_{(o_i, r_j) \in \mathcal{N}_{e_k}} \phi(\boldsymbol{h}_t^{\top} \boldsymbol{W}_{r_i}[\boldsymbol{h}_{o_i}^{L}; \boldsymbol{h}_{e_k}^{L}]), \tag{5.14}$$

$$\hat{ps}_t(e_k) = \gamma \cdot ps_t(e_k) + \phi(\boldsymbol{h}_t \boldsymbol{W}_s \boldsymbol{h}_{e_k}^{L}), \tag{5.15}$$

where ϕ is the Tangent function, γ is the scale factor, $\mathcal{N}_{e_k}$ is the neighbor collection of e_k, and $\boldsymbol{W}_{r_i} \in \mathbb{R}^{h \times 2h}$ and $\boldsymbol{W}_s \in \mathbb{R}^{h \times h}$ are weight matrices for learning relation r_i and self-connection, respectively. In the PrR mechanism, the relevant scores (i.e., $ps_t(e_k)$) of their connected observations are also included in $\hat{ps}_t(e_k)$ since $\boldsymbol{h}_t$ contains observation information, and higher relevant scores of these connected observations indicate a higher relevant score of e_k. Then, the distribution over all entities in G is denoted as:

$$p_G(y_t) = \text{Softmax}(\hat{ps}_t(e_k)). \tag{5.16}$$

Finally, a soft gate $g_t = \sigma(\boldsymbol{W}_g \boldsymbol{h}_t + b_g)$ is adopted to combine $p_{\mathcal{V}}(y_t)$ and $p_G(y_t)$ into $p(y_t)$:

$$p(y_t) = g_t \cdot p_{\mathcal{V}}(y_t) + (1 - g_t) \cdot p_G(y_t), \tag{5.17}$$

where $\boldsymbol{W}_g \in \mathbb{R}^{h}$ and $b_g \in \mathbb{R}$ are the weight matrix and bias, respectively.

Training. The generation process is optimized using the negative log-likelihood loss, given each token's probability $p(y_t)$ and the probability of g_t:

$$\mathcal{L}_{\text{NLL}} = -\sum_{t=1}^{T} \log p(y_t), \tag{5.18}$$

$$\mathcal{L}_g = -\sum_{t=1}^{T} [l_{g_t} \log g_t + (1 - l_{g_t}) \log(1 - g_t)], \tag{5.19}$$

where l_{g_t} indicates t-th token appears in G. Finally, the loss of Stage 2 is:

$$\mathcal{L}_{\text{S2}} = \mathcal{L}_{\text{NLL}} + \lambda \mathcal{L}_g. \tag{5.20}$$

Hyperparameter	**MIMIC-ABN**	**MIMIC-CXR**
Training Epoch	10	5
Dropout Rate	0.1	0.1
Learning Rate	$1e-4$	$1e-4$
Batch Size	$\{64, \mathbf{128}\}$	$\{64, \mathbf{128}\}$
Sample Weight (α_d)	$\{1, 2, \mathbf{3}\}$	$\{1, 2, \mathbf{3}\}$

Table 5.4: Selected hyperparameters of Stage 1 training. The final hyperparameters in **boldface** are tuned on the validation set and others are set empirically.

5.4 Experiments

5.4.1 Datasets

We use two datasets to evaluate both the baselines and our models: MIMIC-ABN[1] [124] and MIMIC-CXR [74]. Since the IU X-RAY does not contain any temporal information, we exclude it from this chapter.

[1] https://github.com/zzxslp/WCL

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	**B-2**	**B-3**	**B-4**	**MTR**	**R-L**	**P**	**R**	**F_1**
MIMIC -ABN	R2Gen	0.290	0.157	0.093	0.061	0.105	0.208	0.266	0.320	0.272
	R2GenCMN	0.264	0.140	0.085	0.056	0.098	0.212	0.290	0.319	0.280
	ORGan	0.314	0.180	0.114	0.078	**0.120**	**0.234**	0.271	0.342	0.293
	Recap (Ours)	**0.321**	**0.182**	**0.116**	**0.080**	**0.120**	0.223	**0.300**	**0.363**	**0.305**
MIMIC -CXR	R2Gen	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
	R2GenCMN	0.353	0.218	0.148	0.106	0.142	0.278	0.344	0.275	0.278
	$\mathcal{M}^2$Tr	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308
	KnowMat	0.363	0.228	0.156	0.115	–	0.284	**0.458**	0.348	0.371
	CMM-RL	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
	CMCA	0.360	0.227	0.156	0.117	0.148	0.287	0.444	0.297	0.356
	KiUT	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
	DCL	–	–	–	0.109	0.150	0.284	0.471	0.352	0.373
	METrans	0.386	0.250	0.169	0.124	0.152	0.291	0.364	0.309	0.311
	ORGan	0.407	0.256	0.172	0.123	0.162	**0.293**	0.416	0.418	0.385
	Recap (Ours)	**0.429**	**0.267**	**0.177**	**0.125**	**0.168**	0.288	0.389	**0.443**	**0.393**

Table 5.5: Experimental Results of our model and baselines on the MIMIC-ABN and MIMIC-CXR datasets, with the best scores shown in **boldface** and the second-best scores underlined. The experimental results on the MIMIC-ABN dataset are replicated based on their corresponding repositories.

- MIMIC-CXR consists of 377,110 chest X-ray images and 227,827 reports from 63,478 patients. We adopt the settings of [21].

- MIMIC-ABN is a modified version of MIMIC-CXR and only contains abnormal sentences. The original train/validation/test split of [124] is 26,946/3,801/7,804 samples, respectively. To collect patients' historical information and avoid information leakage, we recover the data-split used in MIMIC-CXR according to the *subject_ id*[2]. Finally, the data-split used in our experiments is 71,786/546/806 for train/validation/test sets, respectively.

[2] *subject_ id* is the anonymized identifier of a patient.

Model	Sections	B-2	CE-F_1
R2Gen	*Find. & Imp.*	0.212	0.148
IFCC	*Findings*	0.217	0.270
CXR-RePaiR-Sel	*Impressions*	0.050	0.274
BioViL-T	*Impressions*	0.159	0.348
BioViL-T	*Find. & Imp.*	0.213	0.359
ORGan	*Findings*	0.256	0.385
Recap (Ours)	*Findings*	**0.265**	**0.393**

Table 5.6: BLEU score and CheXbert score of our model and baselines on the MIMIC-CXR dataset. Results of baselines are cited from [10] and [57].

5.4.2 Evaluation Metrics and Baselines

NLG Metrics. BLEU [130], METEOR [8], and ROUGE [93] are selected as the natural language generation metrics (NLG Metrics), and we use the MS-COCO evaluation tool to compute the results.

CE Metrics. For Clinical Efficacy (CE Metrics), CheXbert [157] is adopted to label the generated reports compared with disease labels of the references. Macro-weighted precision, recall, and F_1 score are employed as evaluation metrics. Besides, we use the temporal entity matching scores (TEM), proposed by [10], to evaluate how well the models generate progression-related information.

Baselines. For performance evaluation, we compare our model with the following state-of-the-art (SOTA) baselines: R2Gen [21], a memory-driven Transformer model; R2GenCMN [20], which employs a cross-modal memory network; KnowMat [199], which integrates domain knowledge to enhance performance; $\mathcal{M}^2$Tr [126], which generates radiology reports progressively; CMM-RL [133], which leverages reinforcement learning to optimize report generation; CMCA [158], which incorporates contrastive attention to better capture abnormalities; CXR-RePaiR-Sel/2 [35],

Model	B-4	R-L	CE-F_1	TEM
CXR-RePaiR-2	0.021	0.143	0.281	0.125
BioViL-NN	0.037	0.200	0.283	0.111
BioViL-T-NN	0.045	0.205	0.290	0.130
BioViL-AR	0.075	0.279	0.293	0.138
BioViL-T-AR	0.092	**0.296**	0.317	0.175
RECAP (Ours)	**0.118**	0.279	0.400	**0.304**
RECAP *w/o* OP	0.093	0.260	0.256	0.203
RECAP *w/o* Obs	0.104	0.270	0.307	0.240
RECAP *w/o* Pro	0.103	0.266	0.395	0.269
RECAP *w/o* PrR	0.115	0.279	**0.403**	0.296

Table 5.7: Progression modeling performance of our model and baselines on the MIMIC-CXR dataset. The *-NN models use nearest neighbor search for report generation, and the *-AR models use autoregressive decoding, as indicated in [10].

which uses the CLIP model [134] for report generation; BioViL-T [10], which models temporal information in CXR images; DCL [89], which improves report generation through contrastive learning; METrans [179], a Transformer model with learnable expert tokens; KiUT [65], a knowledge-injected U-Transformer; and ORGAN [57], which incorporates observation-level information for report generation.

5.4.3 Implementation Details

We use the ViT [32], a vision transformer pretrained on ImageNet [29], as the visual encoder[3]. The maximum decoding step is set to 64/104 for MIMIC-ABN and MIMIC-CXR, respectively. γ is set to 2 and K is set to 30 for both datasets.

We use AdamW [113] as the optimizer, and the learning rate is set to $5e-5$ and $1e-4$

[3]The model card is "google/vit-base-patch16-224-in21k."

Model	$\mathbf{RG}_{\mathrm{E}}$	$\mathbf{RG}_{\mathrm{ER}}$	$\mathbf{RG}_{\overline{\mathrm{ER}}}$
$\mathcal{T}_{\mathrm{NLL}}$	0.230	0.202	0.153
ORGAN	0.303	0.275	0.199
RECAP (Ours)	**0.307**	**0.276**	**0.205**

Table 5.8: Radgraph evaluation results on the MIMIC-CXR dataset. Results of $\mathcal{T}_{\mathrm{NLL}}$ are cited from [26].

for the pretrained ViT and the rest of the parameters, respectively. The layer number of the Transformer encoder and decoder are both set to 3, and the dimension of the hidden state is set to 768, which is the same as the one of ViT. The layer number L of the R-GCN is set to 3. The learning rate decreases from the initial learning rate to 0 with a linear scheduler. The dropout rate is set to 0.1, the batch size is set to 32, and λ is set to 0.5. We select the best checkpoints based on the BLEU-4 score on the validation set. Our model has 160.05M trainable parameters, and the implementations are based on HuggingFace's *Transformers* [185]. We implement our models on an NVIDIA-3090 GTX GPU with mixed precision. We adopt the preprocessing setup used in [21], and the minimum count of each token is set to 3/10 for MIMIC-ABN/MIMIC-CXR, respectively. Other tokens are replaced with a special token `[UNK]`. Table 5.4 shows the hyperparameters used in Stage 1 training for two datasets. Note that l_{d_i} is the weight for observation detection, and the weights of observation classification and progression classification are both set to 1. In addition, two data augmentation methods are used during training. Specifically, we first resize an input image to 256×256, and then the image is randomly cropped to 224×224, and finally, we flip the image horizontally with a probability of 0.5. We select the best checkpoint based on the Macro-F_1 of abnormal observations at this stage. As the variant *w/o* OP and the variant *w/o* Obs in Table 5.9 are not trained in Stage 1, they are trained with more epochs (i.e., 10 epochs).

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	**B-2**	**B-3**	**B-4**	**MTR**	**R-L**	**P**	**R**	**F_1**
MIMIC -ABN	RECAP	0.321	0.182	0.116	0.080	0.120	0.223	0.300	0.363	0.305
	RECAP *w/o* OP	0.303	0.170	0.109	0.074	0.113	0.227	0.289	0.300	0.280
	RECAP *w/o* Obs	0.302	0.174	0.114	0.079	0.114	0.231	0.341	0.314	0.282
	RECAP *w/o* Pro	0.306	0.169	0.107	0.072	0.114	0.220	0.298	0.361	0.298
	RECAP *w/o* PrR	0.320	0.180	0.115	0.079	0.120	0.224	0.295	0.365	0.301
MIMIC -CXR	RECAP	0.429	0.267	0.177	0.125	0.168	0.288	0.389	0.443	0.393
	RECAP *w/o* OP	0.350	0.219	0.150	0.109	0.140	0.278	0.356	0.259	0.266
	RECAP *w/o* Obs	0.356	0.224	0.153	0.113	0.144	0.283	0.464	0.281	0.296
	RECAP *w/o* Pro	0.402	0.245	0.161	0.112	0.157	0.278	0.379	0.433	0.386
	RECAP *w/o* PrR	0.415	0.257	0.171	0.119	0.164	0.285	0.381	0.443	0.391

Table 5.9: Ablation results of our model and its variants. RECAP *w/o* OP is the standard Transformer model, *w/o* Obs stands for without observation, and *w/o* Pro stands for without progression.

5.5 Results and Analyses

5.5.1 Quantitative Analysis

Language Generation Results. The language generation results of two datasets are listed on the left side of Table 5.5 and Table 5.6. As we can see from Table 5.5, RECAP achieves the best performance compared with other SoTA models, with substantially improvements on both datasets. Specifically, as shown in Table 5.5, on the MIMIC-ABN dataset, our model improves the BLEU-1/2/3/4 scores from 0.314/0.180/0.114/0.078 to 0.321/0.182/0.116/0.080, while achieving comparable METEOR and ROUGE-L performance. On the MIMIC-CXR dataset, the improvements are more pronounced, with BLEU-1/2 scores increasing from 0.407/0.256 to 0.429/0.267. We further compare our model with baselines trained on different report sections (i.e., *Findings* or *Impression*), and RECAP consistently outperforms all of them. Furthermore, it shows

Dataset	Better	Worse	Stable	Macro
MIMIC-ABN	0.286	0.468	0.934	0.563
MIMIC-CXR	0.389	0.455	0.896	0.580

Table 5.10: Experimental results of progression prediction (F_1) after Stage 1 training.

a comparable improvement of 5.2% over a temporally-aware baseline (e.g., BioViL-T), as presented in Table 5.6.

Clinical Efficacy Results. The clinical efficacy results are shown on the right side of Table 5.5 and Table 5.6. RECAP achieves SOTA performance on F_1 score, leading to a 1.2% improvement over the best baseline (i.e., ORGAN) on the MIMIC-ABN dataset. Similarly, on the MIMIC-CXR dataset, our model achieves a score of 0.393, increasing by 0.8% compared with the second-best. Furthermore, compared with the best temporally-aware baseline trained on the Findings section of MIMIC-CXR, as shown in Table 5.6, RECAP outperforms BioViL-T by more than 3%, demonstrating its effectiveness in generating clinically accurate radiology reports. In terms of entity-level performance, RECAP achieves the best results among all baselines, as shown in Table 5.8. Compared to ORGAN, our model raises RG_E from 0.303 to 0.307 and $RG_{\overline{ER}}$ from 0.199 to 0.205, respectively. The improvements are even more pronounced when compared to $\mathcal{T}_{NLL}$. These results demonstrate the effectiveness of RECAP in spatiotemporal modeling for precise radiology report generation. Regarding per-observation performance of RECAP, we present the experimental results in Table 5.12. Three observations achieve a score above 0.65, including Cardiomegaly, Pleural Effusion, and Support Devices, all of which are common findings in the MIMIC-CXR dataset. This highlights the importance of having sufficient training samples to achieve strong performance on specific observations.

Temporal Modeling Results. Since the MIMIC-ABN dataset contains relatively few follow-up visit records, we mainly focus on analyzing the MIMIC-CXR dataset, as shown in Table 5.7 and Table 5.11. RECAP achieves the best performance on BLEU-4.

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	**B-2**	**B-3**	**B-4**	**MTR**	**R-L**	**P**	**R**	**F_1**
w. Historical Record D^p										
MIMIC -ABN	RECAP	0.327	0.183	0.117	0.081	0.124	0.227	0.274	0.372	0.297
	RECAP *w/o* OP	0.300	0.164	0.106	0.072	0.110	0.217	0.281	0.274	0.257
	RECAP *w/o* Obs	0.306	0.173	0.110	0.076	0.114	0.233	0.270	0.288	0.259
	RECAP *w/o* Pro	0.295	0.158	0.099	0.070	0.109	0.209	0.249	0.361	0.278
	RECAP *w/o* PrR	0.320	0.177	0.112	0.076	0.121	0.218	0.266	0.377	0.292
MIMIC -CXR	RECAP	0.423	0.260	0.170	0.118	0.169	0.279	0.387	0.457	0.400
	RECAP *w/o* OP	0.321	0.196	0.131	0.093	0.130	0.260	0.350	0.238	0.256
	RECAP *w/o* Obs	0.347	0.213	0.144	0.104	0.141	0.270	0.465	0.293	0.307
	RECAP *w/o* Pro	0.396	0.236	0.151	0.103	0.153	0.266	0.383	0.447	0.395
	RECAP *w/o* PrR	0.420	0.257	0.168	0.115	0.166	0.279	0.386	0.459	0.403
w/o Historical Record D^p										
MIMIC -ABN	RECAP	0.319	0.182	0.116	0.080	0.120	0.223	0.306	0.360	0.306
	RECAP *w/o* OP	0.303	0.171	0.109	0.074	0.110	0.217	0.299	0.302	0.283
	RECAP *w/o* Obs	0.301	0.174	0.114	0.079	0.114	0.231	0.353	0.313	0.282
	RECAP *w/o* Pro	0.309	0.171	0.109	0.073	0.115	0.222	0.314	0.360	0.302
	RECAP *w/o* PrR	0.320	0.181	0.116	0.079	0.120	0.225	0.299	0.362	0.302
MIMIC -CXR	RECAP	0.427	0.268	0.180	0.128	0.168	0.294	0.378	0.417	0.374
	RECAP *w/o* OP	0.371	0.236	0.164	0.121	0.130	0.260	0.357	0.259	0.268
	RECAP *w/o* Obs	0.363	0.231	0.161	0.119	0.146	0.291	0.415	0.262	0.277
	RECAP *w/o* Pro	0.406	0.251	0.151	0.103	0.153	0.266	0.364	0.405	0.365
	RECAP *w/o* PrR	0.412	0.257	0.172	0.122	0.163	0.289	0.364	0.414	0.368

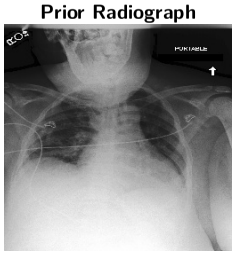
Table 5.11: Ablation results of our model and its variants on progression modeling. RECAP *w/o* OP is the standard Transformer model, *w/o* Obs stands for without observation, and *w/o* Pro stands for without progression.

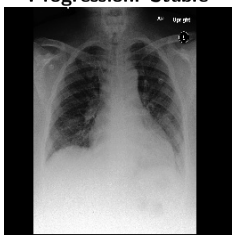
In terms of clinical F_1, RECAP *w/o* PrR outperforms all baselines by a substantial margin, achieving an improvement of over 8% compared to the best-performing baseline, BioViL-T-AR. Additionally, our model yields a notable 12.9% increase in the TEM score relative to this baseline. These results underscore the importance of incorporating historical records to generate accurate follow-up reports. Furthermore, the progression prediction results are presented in Table 5.10. We observe that the model performs best on the `Stable` class, while its performance on the `Better` class remains relatively poor.

Ablation Results. To evaluate the contributions of each module, we perform an ablation analysis, and the ablation results are listed in Table 5.9. We also list the ablation results on progression modeling in Table 5.11. There are four variants in the ablation study:

- RECAP *w/o* OP: This is a standard Transformer model with spatiotemporal information removed.
- RECAP *w/o* Obs: This variant excludes observation information and considers only progression information.
- RECAP *w/o* Pro: This variant removes progression information and includes only observation information.
- RECAP *w/o* PrR: This variant does not utilize the disease progression reasoning mechanism.

As we can see from Table 5.9 and Table 5.11, without the spatiotemporal information (i.e., RADAR *w/o* OP), the performances of language generation drop significantly on both datasets, which indicates the necessity of spatiotemporal modeling in free-text report generation. In addition, compared with RADAR *w/o* OP, the performance of RECAP *w/o* Obs improves substantially on the MIMIC-CXR dataset, with the TEM score increasing from 0.203 to 0.240. This result demonstrates the importance of





① Enlarged Card./NEG
② Cardiomegaly/POS
③ Edema/NEG
④ Consolidation/NEG
⑤ Pneumothorax/NEG
⑥ Pleural Effusion/NEG

Prior Report: low lung volumes are present. ② this accentuates the size of the cardiac silhouette which is likely mildly enlarged. ① mediastinal and hilar contours are likely within normal limits. a right brachiocephalic venous stent is re-demonstrated. there is crowding of the bronchovascular structures with probable ③ mild pulmonary vascular congestion. ⑥ no pleural effusion or ⑤ pneumothorax is identified.

Reference: ap and lateral views of the chest. the lungs are ④ **clear** of consolidation or ⑥ effusion. ② the cardiac silhouette is **enlarged** but **unchanged**. no acute osseous abnormality is detected. right brachiocephalic venous stent is again noted.

Ours: the lungs are **clear** without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac silhouette is **top-normal** to **mildly enlarged**. mediastinal contours are **stable** and **unremarkable**. there is no pulmonary edema.

Ours *w/o* OP: ap upright and lateral views of the chest provided. lung volumes are **low** limiting assessment. allowing for this there is ④ no focal consolidation ⑥ effusion or ⑤ pneumothorax. ① the cardiomediastinal silhouette is **normal**. imaged osseous structures are **intact**. no free air below the right hemidiaphragm is seen.

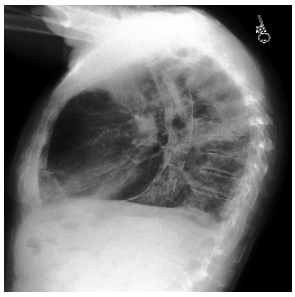
Figure 5.3: Case study of a follow-up-visit sample, given its prior radiograph and prior report. Attributes of observations in reports are highlighted in **bold**, and spans with colors in reports indicate mentions of observations.

incorporating historical records when assessing the current condition of patients. In terms of observation metrics, learning from the observation information boosts the performance of RECAP drastically, with an improvement of 12%. Furthermore, the performance of RECAP increases compared with variant *w/o* PrR. This indicates that PrR can help generate precise and accurate reports.

In terms of ablation of first-visit and follow-up-visit samples, as shown in Table 5.11, our model achieves better performance on samples with historical records compared to those without in the MIMIC-CXR dataset, as historical information contains rich semantic information. These results demonstrate that incorporating patients' historical information can effectively enhance the quality of generated reports. RECAP learns to integrate prior studies with the current CXR, enabling the generation of more temporally grounded outputs, while preserving its ability to produce accurate reports for first-visit cases. Additionally, we observe that RECAP maintains similar performance across samples in the MIMIC-ABN dataset. This may be attributed to the limited number of follow-up visit records available in this dataset, which potentially constrains the model's ability to fully leverage temporal information. Nevertheless, each module within the model continues to contribute to its overall performance, highlighting the robustness of the architecture even under data-sparse conditions.

5.5.2 Qualitative Analysis

Case Study. We conduct a detailed case study on how RECAP generates precise and accurate attributes of a given radiograph in Figure 5.3. Specifically, RECAP successfully generates six observations, including five negative observations and one positive observation. Regarding attribute modeling, our model can generate the precise description "*the lungs are clear without focal consolidation*", which also appears in the reference, while RECAP *w/o* OP can not generate relevant descriptions. Additionally, RECAP can learn to compare with the historical records so as to precisely measure the



① Enlarged Card./NEG
② Cardiomegaly/F.NEG
③ Lung Opacity/F.POS
④ Pneumothorax/NEG
⑤ Effusion/F.POS

Reference: there is no new consolidation. right lower lobe pneumonia that was present in prior exams has significantly improved. esophageal stent is in unchanged position. there is no pneumomediastinum or pneumothorax. there is no pleural effusion . mediastinal and cardiac contours are stable.

Ours: ② the heart size is normal. the hilar and mediastinal contours are within normal limits. there is no pneumothorax. again seen is a ⑤ small right pleural effusion. the visualized osseous structures are unremarkable. there has been interval improvement of the ③ right basilar opacity.

Figure 5.4: Error case generated by RECAP. The span and the spans denote false negative observation and false positive observation, respectively.

observations. For example, our model produces "*cardiac silhouette is top-normal to mildly enlarged*" for *Cardiomegaly*/POS, whereas RECAP *w/o* OP fails to capture the positive observations. This indicates that spatiotemporal information plays a vital role in the generation process.

Error Analysis. We depict error analysis to provide more insights, as shown in Figure 5.4. There are two major errors, which are false-positive observations (i.e., Positive *Lung Opacity* and *Pleural Effusion*) and false-negative observations (i.e., Negative *Cardiomegaly*). Since RECAP relies on predicted observations for accurate and precise report generation, these errors are mainly propagated from Stage 1. Specifically, *Cardiomegaly*/NEG, *Pleural Effusion*/POS, and *Lung Opacity*/POS are rendered as "*the heart size is normal*", "*small right pleural effusion*", and "*right basilar opacity*", respectively. Thus, improving the performance of observation prediction could be an important direction in enhancing the quality of generated reports. In addition, although RECAP aims to model precise attributes of observations presented in the

Observation	**P**	**R**	**F_1**
Enlarged Cardiomediastinum	0.323	0.589	0.417
Cardiomegaly	0.585	0.836	0.689
Lung Opacity	0.489	0.499	0.494
Lung Lesion	0.265	0.044	0.075
Edema	0.562	0.587	0.574
Consolidation	0.285	0.233	0.256
Pneumonia	0.242	0.444	0.313
Atelectasis	0.426	0.800	0.556
Pneumothorax	0.265	0.167	0.205
Pleural Effusion	0.691	0.781	0.728
Pleural Other	0.184	0.050	0.078
Fracture	0.155	0.081	0.107
Support Devices	0.720	0.660	0.689
No Finding	0.265	0.429	0.327
Macro Average	0.389	0.443	0.393

Table 5.12: Per-observation performance results after Stage 2 training on the MIMIC-CXR dataset.

radiograph, it still can not cover all the cases, especially those rare ones. This might be alleviated by incorporating external knowledge.

5.6 Chapter Summary

In this chapter, we propose RECAP, which can capture both spatial and temporal information for generating precise and accurate radiology reports. To achieve precise attribute modeling in the generation process, we construct a disease progression graph containing both observations and fined-grained attributes which quantify the severity of

diseases and devise a dynamic disease progression reasoning (PrR) mechanism to select observation-relevant attributes. Experimental results demonstrate the effectiveness of our proposed model in terms of generating precise and accurate radiology reports. This chapter addresses several limitations of ORGan introduced in Chapter 3. While ORGan enhances clinical accuracy through the use of observation graphs, Recap incorporates temporal information to more effectively model disease progression. Moreover, Recap leverages a knowledge graph, rather than mined n-grams from training reports, for explicit attribute modeling, thereby further improving the quality of the generated outputs.

Our proposed two-stage framework requires pre-defined observations and progressions for training, which may not be available for other types of radiographs. Despite incorporating temporal information, the images within each study (e.g., AP or lateral views) are still treated as separate samples, which can lead to hallucinations about non-existent inputs. In addition, the outputs of Stage 1 are the prerequisite inputs of Stage 2, and thus, our framework may suffer from error propagation. Finally, although prior information is important in generating precise and accurate free-text reports, historical records are not always available, even in the two benchmark datasets. Our framework will still generate misleading free-text reports, conditioning on non-existent priors, as indicated in [138]. This might be mitigated through rule-based removal operations.

Chapter 6

Consistent Radiology Report Generation

6.1 Introduction

In Chapters 3 and 4, we investigate observation-aware radiology report generation, while in Chapter 5, we address clinical accuracy at both the observation and entity levels. Additionally, recent studies [125, 162] have made noteworthy progress in enhancing the clinical accuracy of generated reports.

However, constructing a credible report generation system goes beyond the overall accuracy. There is another crucial quality for report generation systems that has been largely overlooked in the existing literature of medical report generation, which is, *inter-report consistency* [33]. To illustrate the disparity between accuracy and inter-report consistency, we exemplify two semantically equivalent cases as shown in Figure 6.1, where they share similar observations and reports. Specifically, System α demonstrates the ability to maintain both inter-report consistency and factual accuracy for two similar cases (i.e., *"small bilateral pleural effusions"* for positive *Pleural Effusion*), whereas other systems (i.e., β and γ) fail to meet these criteria. These systems

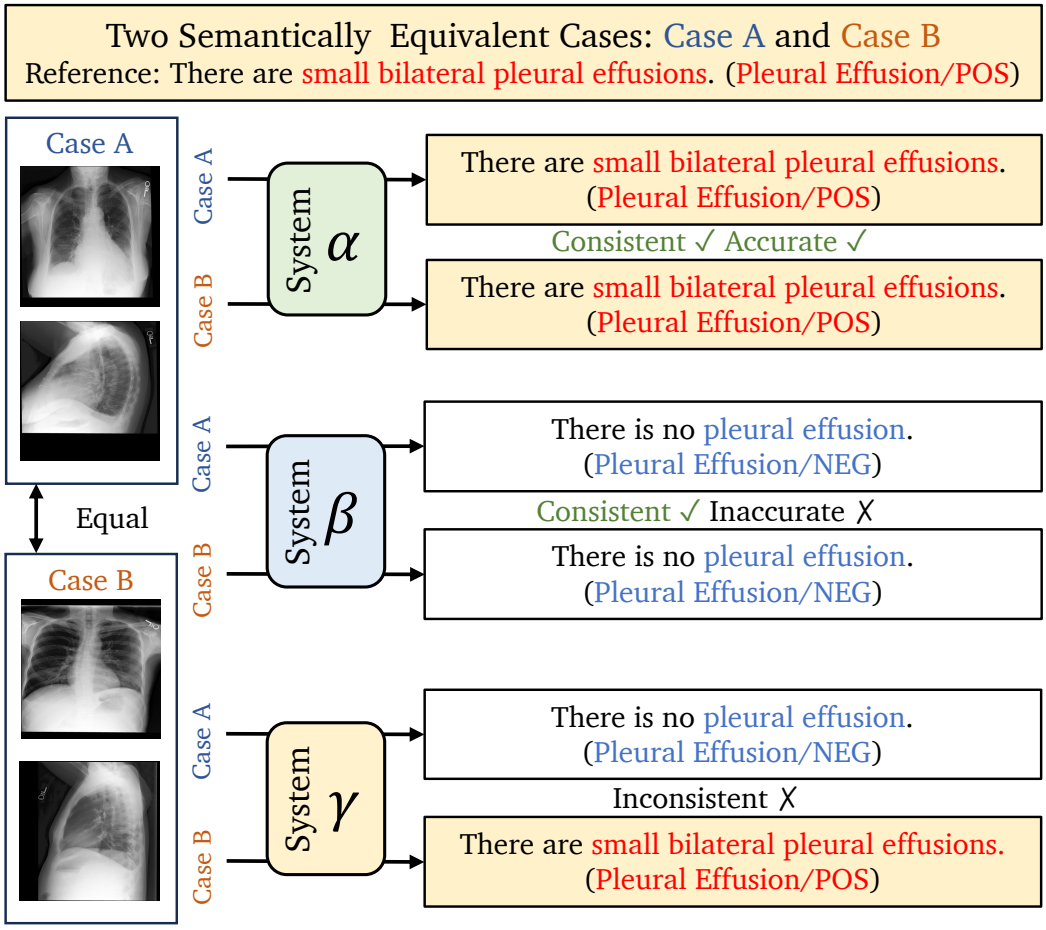


Figure 6.1: Given two semantically equivalent cases (i.e., Case A and Case B), an example to illustrate the difference between three radiology report generation systems: a consistent and accurate system (i.e., System α) and a consistently inaccurate system (i.e., System β), and an inconsistent system (i.e., System γ).

might have overfitted to ordinary cases and could be vulnerable to noise or attack. In terms of enhancing the system's credibility, inter-report consistency might even hold greater significance than the overall accuracy, since a system prone to providing conflicting results would severely undermine users' trust [132, 5]. Regrettably, existing report generation systems struggle to maintain this important quality. They tend to exhibit biases towards common patterns, primarily describing normal observations and are susceptible to lesion variants and context noise [21, 133, 116, 75]. We argue that this is largely due to their limited capability of capturing shared attributes of similar patterns, which arises from the data scarcity of distributed lesions and their semantically equivalent variants, rendering it challenging for neural models to accurately locate and describe abnormalities.

In this chapter, we propose ICon, which aims to Improves inter-report Consistency of radiology report generation. Our proposed method involves first extracting lesions from given input images, followed by examining the attributes of these lesions. Subsequently, both the radiographs and their associated attributes are utilized as inputs for report generation. To further enhance the inter-report consistency, we introduce a lesion-aware mixup technique by learning from linearly combined lesions and synthesized attributes that belong to the same observation. In summary, the contributions of this paper are as follows:

- To the best of our knowledge, we are the first to introduce *inter-report consistency* in radiology report generation. To this end, we devise two metrics (Con and R-Con) to measure such consistency.

- We propose ICon, which improves both the *consistency* and *accuracy* in radiology report generation by capturing abnormalities at the region level. ICon only requires coarse-grained labels (i.e., image-level labels) for training to extract lesions[1].

[1]In this context, the term "lesion" generally refers to a specific abnormality. It encompasses most

- Extensive experiments are conducted on three publicly available datasets, and the results demonstrate the effectiveness of ICon in terms of improving both the consistency and accuracy of the generated reports.

6.2 Preliminaries

6.2.1 Problem Formulation

Given a set of radiographs $\mathcal{X} = \{X_1, \ldots, X_L\}$ in one study, along with its historical records $\mathcal{X}^p = \{X_1^p, \ldots, X_{|p|}^p\}$ (or $\mathcal{X}^p = \emptyset$ if no historical records are available), and its report $\mathcal{Y}$, the task of radiology report generation (RRG) is to generate the report $\mathcal{Y}$ based on $\mathcal{X}$ and $\mathcal{X}^p$. Our proposed method, denoted as ICon, decomposes the RRG task into two stages: Lesion Extraction (Stage 1) and Report Generation (Stage 2). Specifically, given the input images $\mathcal{X}$, ICon first extracts M region candidates $\mathcal{R} = \{R_1, \ldots, R_M\}$ and then classifies regions as lesions $\mathcal{Z} = \{Z_1, \ldots, Z_{|O|}\}$, where O denotes the observations. Subsequently, in Stage 2, ICon generates a report based on the input images $\mathcal{X}$, historical records $\mathcal{X}^p$, and the extracted lesions $\mathcal{Z}$.

6.2.2 Observation and Attribute Annotation

Observations for Lesion Extraction. Lesion extraction requires report-level labels, and we adopt CheXbert [157] for this purpose. Specifically, CheXbert annotates a report with 14 observation categories $O = \{o_1, \ldots, o_{14}\}$. Each observation is assigned one of four statuses: *Present*, *Absent*, *Uncertain*, and *Blank*. During training and evaluation, *Present* and *Uncertain* are merged into the *Positive* category, which represents abnormal observations. Note that for the observation *No Finding*, only

observation categories, excluding *Support Devices*, *Cardiomegaly*, and *Enlarged Cardiomediastinum*. For simplicity, we consider all corresponding regions as lesions.

two statuses, *Present* or *Absent*, are applicable. Finally, observation information is utilized for lesion extraction as described in §6.3.2.

Attributes for Lesion-Attribute Alignment. After extracting observations, we further extract entities that represent their characteristics. Specifically, we adopt the attributes released by [56][2], which are entities (with a relation *modify* or *located_at*) extracted from RadGraph [69] using PMI [24]. Part of the attributes are listed in Table 5.3. These attributes are then utilized for lesion-attribute alignment as will be described in §6.3.3.

6.2.3 Inter-Report Consistency Metrics

To assess the inter-report consistency of a model, we introduce two metrics, CON and R-CON, inspired by [33]. Semantically equivalent samples should have high observation and entity similarity, and we identify these samples using the Overlap Coefficient [155]:

$$\text{Overlap}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}. \tag{6.1}$$

For a report Q_i and its relevant reports $\mathcal{K}_i = \{K_{i,1}, \ldots, K_{i,N}\}$, when the observation similarity satisfies $\text{Overlap}(O_{Q_i}, O_{K_{i,j}}) \geq 0.75$ and the entity similarity satisfies $\text{Overlap}(Q_i, K_{i,j}) \geq 0.5$, we regard them as semantically equivalent samples. We then collect the corresponding outputs of $\mathcal{K}_i$ from a model, denoted as $\widehat{\mathcal{K}}_i = \{\widehat{K}_{i,1}, \ldots, \widehat{K}_{i,N}\}$. The similarity between two outputs $\widehat{Q}_i$ and $\widehat{K}_{i,j}$ is:

$$\text{Overlap}(\widehat{Q}_i, \widehat{K}_{i,j}) = \frac{|\widehat{e}_i \cap \widehat{e}_j|}{\min(|\widehat{e}_i|, |\widehat{e}_j|)}, \tag{6.2}$$

where $\widehat{e}_i$ and $\widehat{e}_j$ are entities and attributes in $\widehat{Q}_i$ and $\widehat{K}_{i,j}$, respectively. The inter-report consistency is then defined as:

$$\text{CON}(\widehat{Q}_i, \widehat{\mathcal{K}}_i) = \frac{1}{N} \sum_{j=1}^{N} \text{Overlap}(\widehat{Q}_i, \widehat{K}_{i,j}). \tag{6.3}$$

[2]The attributes are available at https://github.com/wjhou/Recap.

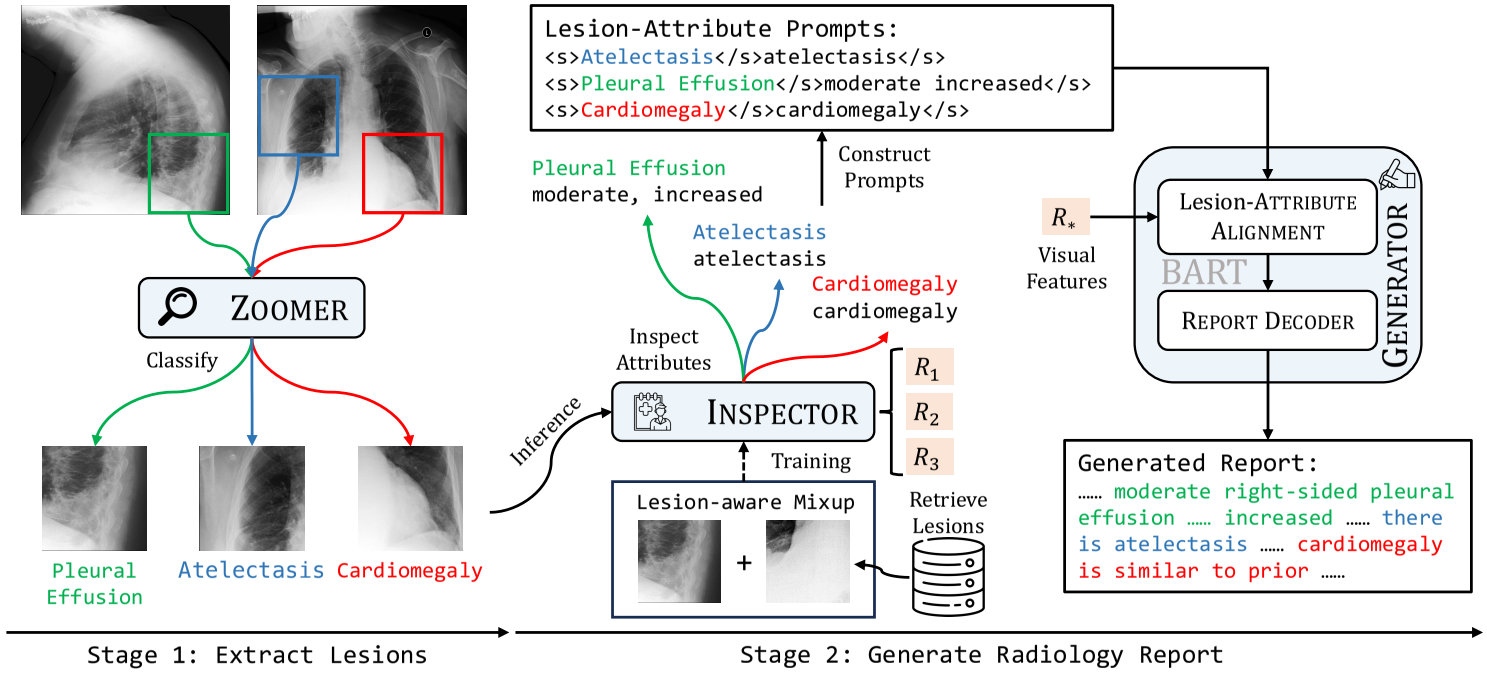


Figure 6.2: Overview of the ICon framework, which first extracts lesions, then aligns these lesions with corresponding attributes, and finally generates comprehensive reports. The attributes are extracted using RadGraph [69].

Since CON only considers inter-report consistency without accounting for reference quality, we introduce R-CON, which considers both consistency and accuracy:

$$\text{R-CON}(\widehat{Q}_i, \widehat{\mathcal{K}}_i) = \tau_i \cdot \text{CON}(\widehat{Q}_i, \widehat{\mathcal{K}}_i), \tag{6.4}$$

where $\tau_i = \text{Overlap}(\widehat{Q}_i, Q_i)$ is the similarity between the hypothesis and its reference.

6.3 Method

6.3.1 Visual Representation Extraction

Given an image X_l, an image processor is first utilized to split X_l into N patches. Then, a visual encoder f_θ, e.g., Swin Transformer [111], is employed to extract visual representations $\boldsymbol{X}_l$ and the pooler output $\boldsymbol{P}_l \in \mathbb{R}^h$:

$$[\boldsymbol{P}_l, \boldsymbol{X}_l] = f_\theta(X_l), \tag{6.5}$$

where $\boldsymbol{X}_l = \{\boldsymbol{x}_{l,i}, \ldots, \boldsymbol{x}_{l,N}\}$ and $\boldsymbol{x}_{l,i} \in \mathbb{R}^h$ is the i-th visual representation.

6.3.2 Lesion Extraction via Observation Classification

Observation Classification. A ZOOMER is a visual encoder parameterized by θ_Z and trained to classify a given input $\mathcal{X}$ into abnormal observations as mentioned in §6.2.2:

$$p(o_i) = \text{ZOOMER}(\mathcal{X}). \tag{6.6}$$

Specifically, ZOOMER first encodes images $\mathcal{X} = \{X_1, \ldots, X_L\}$ as outlined in §6.3.1, and then takes the averaged pooler output for classification, following these steps:

$$[\boldsymbol{P}_l, \boldsymbol{X}_l] = f_{\theta_\text{Z}}(X_l), \tag{6.7}$$

$$\boldsymbol{P} = \frac{1}{L}\sum \boldsymbol{P}_l, \tag{6.8}$$

$$p(o_i) = \sigma(\boldsymbol{W}_i \boldsymbol{P} + b_i), \tag{6.9}$$

where $\boldsymbol{W}_i \in \mathbb{R}^h$ is the weight for the i-th observation, $b_i \in \mathbb{R}$ is its bias, and σ is the Sigmoid function.

Zooming In for Lesion Extraction. Upon completing training ZOOMER, we can use it to extract lesions without the need for object detectors [140]. It is worth noting that our method does not require fine-grained labels, such as bounding boxes [162].

For an image X_l, a sliding window with a 0.375 ratio of X_l is applied to extract M region candidates $\mathcal{R}_l = \{R_{l,1}, \ldots, R_{l,M}\}$ from X_l, as shown in the left side of Figure 6.2. These regions are then sequentially fed into ZOOMER for classification. Specifically, there are two steps in extraction lesions: candidate generation and candidate classification. Given an image with a resolution of 1024×1024, padding if needed, we apply a sliding window of 384×384, with a step size of 128 to extract candidates for classification. This operation results in 36 regions. Then, each region is fed into the ZOOMER for classification, and only the top-1 region is selected for each observation. Note that before extracting lesions, each input case is first assigned with their observations by ZOOMER, and as a result, the number of lesions corresponds to the number of observations. The probability of a region $R_{l,j}$ being classified as an abnormal observation o_i is:

$$p_{l,j}(o_i) = \text{ZOOMER}(R_{l,j}). \tag{6.10}$$

For each study, all images in $\mathcal{X}$ are iterated, and only the region with the highest $p_{l,j}(o_i)$ is chosen as a lesion Z_i corresponding to the observation o_i. Finally, the set of lesions is denoted as $\mathcal{Z} = \{Z_1, \ldots, Z_{|O|}\}$.

Training ZOOMER. ZOOMER is optimized using the binary cross-entropy (BCE) loss. To handle the class-imbalanced issue (refer to Table 3.1 and Table 5.1 for details), a weight factor α_j is applied for each abnormal observation, and the loss function $\mathcal{L}_{\text{S1}}$ is:

$$\text{BCE}(p(O), O) = -\frac{1}{|O|} \sum_j \left[\alpha_j \cdot o_j \cdot \log p(o_j) \; + (1 - o_j) \cdot \log(1 - p(o_j))\right], \tag{6.11}$$

where $o_j \in \{0, 1\}$ is the observation label, $\alpha_j = 1 + \log\left(\frac{|\mathcal{D}_{\text{train}}| - w_j}{w_j}\right)$, and $|\mathcal{D}_{\text{train}}|$ and w_j are the number of samples and the number of j-th observations in the training set, respectively.

6.3.3 Lesion Inspection

Inspecting Lesions with Attributes. Given that lesions of the same observation can exhibit different characteristics, it is crucial to inspect each lesion and match it with corresponding attributes (§6.2.2) to differentiate it from other variations. Specifically, an INSPECTOR is a visual encoder parameterized by θ_I, similar to §6.3.2. INSPECTOR$(\boldsymbol{P}^p, \boldsymbol{P}, Z_j)$ takes prior and current visit chest X-rays as context, along with a lesion region as input:

$$[\boldsymbol{P}_{Z_j}, \boldsymbol{Z}_j] = f_{\theta_I}(Z_j), \tag{6.12}$$

$$p_j(a_k) = \sigma(\text{MLP}(\boldsymbol{P}^p, \boldsymbol{P}, \boldsymbol{P}_{Z_j})), \tag{6.13}$$

where MLP is a two-layer perceptron with non-linear activation, and $\boldsymbol{P}^p, \boldsymbol{P}, \boldsymbol{P}_{Z_j} \in \mathbb{R}^h$ are pooler outputs of prior images, current images, and the lesion, respectively. Concurrently, the lesion features $\boldsymbol{\mathcal{Z}} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{|O|}\}$ are collected for report generation. For image encoding, we use another visual encoder f_{θ_V} to encode $\mathcal{X}$ into $\boldsymbol{\mathcal{X}}$ and $\mathcal{X}^p$ into $\boldsymbol{\mathcal{X}}^p$. By inspecting lesion-level features, ICON can capture fine-grained details which are beneficial for generating consistent outputs.

Lesion-aware Mixup. To further improve the consistency of the generated outputs, we adopt the mixup augmentation method [211] and devise a lesion-aware mixup during the training phase. Specifically, for a lesion-attribute pair (Z_j, A_j), we retrieve a similar pair (Z_k, A_k) with the same observation from the training data based on report similarity. These lesions are synthesized by a linear combination, as illustrated in Figure 6.3:

$$Z_j^* = \lambda Z_j + (1 - \lambda) Z_k, \tag{6.14}$$

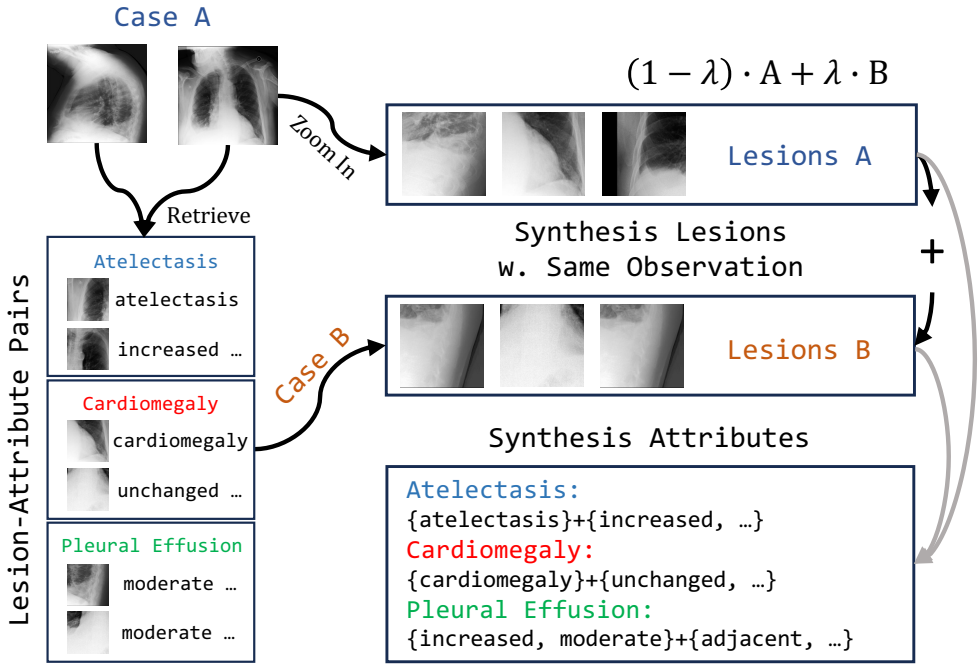


Figure 6.3: Overview of our proposed lesion-aware mixup augmentation.

where λ is set to 0.75. Note that during training, Z_j^* is used for both INSPECTOR and GENERATOR.

Training INSPECTOR. Similar to §6.3.2, we adopt a linearly combined BCE loss to optimize INSPECTOR:

$$\mathcal{L}_{\mathrm{I}} = \lambda \mathrm{BCE}_j + (1-\lambda)\mathrm{BCE}_k, \tag{6.15}$$

where BCE_j and BCE_k take A_j and A_k as their respective labels. Notably, only the attributes that are shared between Z_j and Z_k are fully optimized. Consequently, our lesion-aware mixup technique facilitates the improvement of output consistency for two semantically equivalent lesions.

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	B-2	B-3	B-4	MTR	R-L	P	R	F_1
MIMIC -ABN	R2Gen	0.290	0.157	0.093	0.061	0.105	0.208	0.266	0.320	0.272
	R2GenCMN	0.264	0.140	0.085	0.056	0.098	0.212	0.290	0.319	0.280
	ORGan	0.314	0.180	0.114	0.078	0.120	0.234	0.271	0.342	0.293
	Recap	0.321	0.182	0.116	0.080	0.120	0.223	0.300	0.363	0.305
	ICon (Ours)	**0.337**	**0.195**	**0.126**	**0.086**	**0.129**	**0.236**	**0.332**	**0.430**	**0.360**
MIMIC -CXR	R2Gen	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
	R2GenCMN	0.353	0.218	0.148	0.106	0.142	0.278	0.344	0.275	0.278
	$\mathcal{M}^2$Tr	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308
	KnowMat	0.363	0.228	0.156	0.115	–	0.284	**0.458**	0.348	0.371
	CMM-RL	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
	CMCA	0.360	0.227	0.156	0.117	0.148	0.287	0.444	0.297	0.356
	KiUT	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
	DCL	–	–	–	0.109	0.150	0.284	0.471	0.352	0.373
	METrans	0.386	0.250	0.169	0.124	0.152	0.291	0.364	0.309	0.311
	RGRG	0.373	0.249	0.175	**0.126**	0.168	0.264	0.380	0.319	0.305
	ORGan	0.407	0.256	0.172	0.123	0.162	**0.293**	0.416	0.418	0.385
	Recap	**0.429**	**0.267**	0.177	0.125	0.168	0.288	0.389	0.443	0.393
	ICon (Ours)	**0.429**	0.266	**0.178**	**0.126**	**0.170**	0.287	0.445	**0.505**	**0.464**

Table 6.1: Experimental results of our model and the baselines on the MIMIC-ABN and MIMIC-CXR datasets, with the best scores shown in **bold** and the second-best scores underlined.

6.3.4 Radiology Report Generation

Lesion-Attribute Alignment. To bridge the modality gap between lesion representations and attributes, we leverage a BART [79] encoder to extract attribute representations. The attributes associated with each lesion are formulated as a prompt: `<s>` o_j `</s>` A_j `</s>`, as depicted in Figure 6.2. Then, a cross-attention module [168] is inserted after every self-attention module. This module aligns the lesion representations with the attribute representations by querying visual representations using attribute representations, similar to Q-Former [85]:

$$\boldsymbol{H}_j^a = \text{Cross-Attn}(\boldsymbol{H}_j^s, \boldsymbol{Z}_j, \boldsymbol{Z}_j), \tag{6.16}$$

where $\boldsymbol{H}_j^a, \boldsymbol{H}_j^s \in \mathbb{R}^h$ are the aligned attribute representation and the self-attended representation of A_j, respectively. All prompts are encoded, and the attribute representations of $\boldsymbol{\mathcal{Z}}$ are denoted as $\boldsymbol{\mathcal{H}}^a$.

Report Generation. Given the input images $\boldsymbol{\mathcal{X}}$, images of prior visits $\boldsymbol{\mathcal{X}}^p$, the lesions $\boldsymbol{\mathcal{Z}}$, and attribute $\boldsymbol{\mathcal{H}}^a$, we utilize a BART decoder in conjunction with the Fusion-in-Decoder (FiD; [68]) that simply concatenates multiple context sequences for report generation. Then, the probability of the t-th step is expressed as:

$$\boldsymbol{h}_t = \text{FiD}([\boldsymbol{\mathcal{X}}; \boldsymbol{\mathcal{X}}^p; \boldsymbol{\mathcal{Z}}; \boldsymbol{\mathcal{H}}^a], \boldsymbol{h}_{<t}), \tag{6.17}$$

$$p(y_t|\mathcal{X}, \mathcal{X}^p, \mathcal{Z}, \mathcal{Y}_{<t}) = \text{Softmax}(\boldsymbol{W}_g \boldsymbol{h}_t + \boldsymbol{b}_g), \tag{6.18}$$

where $\boldsymbol{h}_t \in \mathbb{R}^h$ is the t-th hidden representation, $\boldsymbol{W}_g \in \mathbb{R}^{|\mathcal{V}| \times h}$ is the weight matrix, $\boldsymbol{b}_g \in \mathbb{R}^{|\mathcal{V}|}$ is the bias vector, and $\mathcal{V}$ is the vocabulary.

Training Generator. The generation process is optimized using the negative log-likelihood loss, given each token's probability $p(y_t|\mathcal{X}, \mathcal{X}^p, \mathcal{Z}, \mathcal{Y}_{<t})$:

$$\mathcal{L}_{\text{G}} = -\sum_{t=1}^{T} \log p(y_t|\mathcal{X}, \mathcal{X}^p, \mathcal{Z}, \mathcal{Y}_{<t}). \tag{6.19}$$

The loss function of Stage 2 is: $\mathcal{L}_{\text{S2}} = \mathcal{L}_{\text{I}} + \mathcal{L}_{\text{G}}$.

6.4 Experiments

6.4.1 Datasets

Three public datasets are used to evaluate our models, i.e., IU X-RAY [28], MIMIC-CXR [74], and MIMIC-ABN [124]. We follow previous research [21] to preprocess these datasets.

- IU X-RAY consists of 3,955 samples where each report corresponds to a frontal and lateral CXR. We follow previous research [21] and split the dataset into train/validation/test sets with a ratio of 7:1:2.
- MIMIC-CXR consists of 377,110 chest X-ray images and 227,827 reports from 63,478 patients. We adopt the standard train/validation/test splits.
- MIMIC-ABN is modified from the MIMIC-CXR dataset and its reports only contain abnormal part. We adopt the data-split as used in [56], and the data-split is 71,786/546/806 for train/validation/test sets.

Unlike previous research [21] which only used one view for report generation on the MIMIC-CXR and MIMIC-ABN datasets, we collect all views for each visit in experiments.

6.4.2 Evaluation Metrics and Baselines

NLG Metrics. To assess the quality of generated reports, we adopt several NLG metrics for evaluation. BLEU-1/2/3/4 (B-1/2/3/4) [130], METEOR (MTR) [8], and ROUGE-L (R-L) [93] are selected as NLG Metrics, and we use the MS-COCO caption evaluation tool to compute the results.

CE Metrics. Following previous research [21, 20], we adopt clinical efficacy (CE) metrics to evaluate the observation-level factual accuracy, and CheXbert [157] is used

Model	B-4	R-L	CE-F_1	TEM
CXR-RePaiR-2	0.021	0.143	0.281	0.125
BioViL-NN	0.037	0.200	0.283	0.111
BioViL-T-NN	0.045	0.205	0.290	0.130
BioViL-AR	0.075	0.279	0.293	0.138
BioViL-T-AR	0.092	**0.296**	0.317	0.175
RECAP	0.118	0.279	0.400	0.304
ICON (Ours)	**0.120**	0.279	**0.468**	**0.335**

Table 6.2: Progression modeling results on the MIMIC-CXR dataset. Results of BioViL-* are cited from [10].

in this chapter. To measure the entity-level factual accuracy, we leverage a knowledge graph built from radiology reports, i.e., RadGraph [69] to evaluate the performance, following [26]. Specifically, we include three metrics: RG_E, which focuses on entities; RG_{ER}, which evaluates both entities and their relations; and $RG_{\overline{ER}}$, which considers entities along with detailed relations. In addition, we adopt the temporal entity matching (TEM) score proposed by [10] for further evaluation.

Consistency Metrics. CON and R-CON (§6.2.3) are utilized to measure the inter-report consistency. Note that entities used in measuring consistency are adopted from RadGraph [69]. A MAJORITY baseline which outputs the same report for all inputs, is included in Table 6.4.

Baselines. We compare our models with the following state-of-the-art (SoTA) baselines: R2GEN [21], a memory-driven Transformer model; R2GENCMN [20], which employs a cross-modal memory network; KNOWMAT [199], which integrates domain knowledge to enhance performance; $\mathcal{M}^2$TR [126], which generates radiology reports in a progressive manner; CMM-RL [133], which leverages reinforcement learning to optimize report generation; CMCA [158], which incorporates contrastive attention to better capture abnormalities; CXR-RePaiR-Sel/2 [35], which utilizes

Dataset	Model	NLG Metrics		RadGraph		
		B-4	R-L	RG_E	RG_{ER}	$RG_{\overline{ER}}$
IU X-RAY	R2GEN	0.120	0.298	–	–	–
	$\mathcal{M}^2$TR	**0.121**	0.288	–	–	–
	$\mathcal{T}_{NLL}$	0.114	–	0.230	0.202	0.153
	ICON	0.098	**0.320**	**0.342**	**0.312**	**0.246**
MIMIC-CXR	$\mathcal{T}_{NLL}$	0.105	0.253	0.230	0.202	0.153
	ORGAN	0.123	**0.293**	0.303	0.275	0.199
	RECAP	0.125	0.288	0.307	0.276	**0.205**
	ICON	**0.126**	0.287	**0.312**	**0.278**	0.197

Table 6.3: Radgraph evaluation results on the IU X-RAY and MIMIC-CXR datasets. Results of $\mathcal{T}_{NLL}$ are cited from [26].

the CLIP model [134] for report generation; BioViL-T [10], which models temporal information in CXRs; DCL [89], which improves report generation through contrastive learning; METrans [179], a Transformer model enhanced with learnable expert tokens; KiUT [65], a knowledge-injected U-Transformer; ORGAN [57], which incorporates observation-level information into report generation; and RECAP [56], which leverages historical patient records to inform report generation.

6.4.3 Implementation Details

The small[3] and tiny[4] versions of Swin Transformer V2 [110] are used as the visual backbone for ZOOMER and INSPECTOR, respectively. The GENERATOR is initialized with the base version of BART [79] pretrained on biomedical corpus[5] [207]. Other parameters are randomly initialized. For Stage 2 training, the learning rate is $5e-5$

[3]The model card is "microsoft/swinv2-small-patch4-window8-256."
[4]The model card is "microsoft/swinv2-tiny-patch4-window8-256."
[5]The model card is "GanjinZero/biobart-v2-base."

Model	MIMIC-ABN		MIMIC-CXR	
	CON	R-CON	CON	R-CON
MAJORITY	1.000	–	1.000	–
R2GEN	0.280	0.072	0.137	0.042
R2GENCMN	0.302	0.091	0.155	0.049
ORGAN	**0.338**	0.127	0.345	0.126
RECAP	0.311	0.108	0.345	0.114
ICON (Ours)	0.316	**0.140**	**0.351**	**0.163**
ICON *w/o* ZOOM	0.183	0.073	0.175	0.066
ICON *w/o* INSPECT	0.253	0.100	0.245	0.090
ICON *w/o* MIXUP	0.286	0.119	0.334	0.156

Table 6.4: The CON score and the R-CON score. MAJORITY: outputs the same report for all inputs.

with linear decay, the batch size is 32, and the models are trained for 20 and 5 epochs on MIMIC-ABN and MIMIC-CXR with early stopping, respectively. Since the number of samples in IU X-RAY is too small to train a multimodal model, we only provide results produced by models trained on MIMIC-CXR as a reference, similar to [26]. For Stage 1, all three datasets use the same hyper-parameters for training ZOOMER, with a learning rate of $1e-4$, batch size of 128, and dropout rate of 0.1, and the number of training epochs is adjusted accordingly. We train ZOOMER for 5, 10, and 15 epochs on MIMIC-CXR, MIMIC-ABN, and IU X-RAY, respectively. During training, several data augmentation methods are applied. The input resolution of Swin Transformer is 256×256, and we first resize an image to 288×288, and then randomly crop it to 256×256 with random horizontal flip. For Stage 2, no data augmentation is applied, and we conduct experiments on MIMIC-ABN and IU X-RAY using two NVIDIA-3090 GTX GPUs, and on MIMIC-CXR using four NVIDIA-V100 GPUs, both with half precision. Our model has 328.38M trainable parameters, and

Model	MIMIC-ABN			MIMIC-CXR		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**
R2Gen	0.340	0.413	0.348	0.390	0.336	0.337
R2GenCMN	0.360	0.363	0.336	0.358	0.276	0.290
RGRG	–	–	–	0.461	0.475	0.447
ORGan	0.418	0.471	0.412	0.493	0.560	0.493
Recap	0.366	0.468	0.382	0.447	0.558	0.464
ICon	0.512	0.428	0.436	0.513	0.597	0.522
ICon *w/o* Zoom	0.397	0.406	0.372	0.440	0.362	0.373
ICon *w/o* Inspect	0.430	0.479	0.424	0.506	0.553	0.500
ICon *w/o* Mix-up	0.433	0.509	0.438	0.507	0.590	0.517

Table 6.5: Example-based CE results on the MIMIC-ABN and MIMIC-CXR datasets.

the implementations are based on HuggingFace's Transformers [185]. In terms of data preprocessing, we adopt the same preprocessing setup used in [21], and the minimum count of each token is set to 3/3/10 for IU X-ray/MIMIC-ABN/MIMIC-CXR, respectively. Other tokens are replaced with a special token <unk>.

As stated in [10, 56], without historical information, it is unreasonable to generate reports with comparisons between two consecutive visits and will lead to hallucinations [138]. As a result, we include historical records as context information for report generation. Since we collect all views of a study for report generation on the MIMIC-ABN and MIMIC-CXR datasets, each generated output for a study with L images is duplicated L times. This ensures that the number of samples in the evaluation is consistent with previous research, enabling a fair comparison.

Dataset	Model	Components			NLG Metrics						CE Metrics		
		Zoom	Inspect	Mixup	B-1	B-2	B-3	B-4	MTR	R-L	P	R	F_1
MIMIC -ABN	ICon	✓	✓	✓	0.337	0.195	0.126	0.086	0.129	0.236	0.332	0.430	0.360
	ICon *w/o* Zoom	✗	✗	✗	0.310	0.181	0.119	0.084	0.120	0.243	0.306	0.353	0.306
	ICon *w/o* Inspect	✓	✗	✗	0.315	0.182	0.117	0.081	0.121	0.236	0.338	0.401	0.352
	ICon *w/o* Mixup	✓	✓	✗	0.335	0.192	0.124	0.085	0.129	0.239	0.332	0.413	0.356
MIMIC -CXR	ICon	✓	✓	✓	0.429	0.266	0.178	0.126	0.170	0.287	0.445	0.505	0.464
	ICon *w/o* Zoom	✗	✗	✗	0.377	0.237	0.162	0.119	0.149	0.288	0.363	0.280	0.278
	ICon *w/o* Inspect	✓	✗	✗	0.399	0.248	0.168	0.122	0.157	0.287	0.444	0.447	0.423
	ICon *w/o* Mixup	✓	✓	✗	0.427	0.264	0.176	0.124	0.169	0.285	0.444	0.502	0.462

Table 6.6: Ablation results of our model and its variants on the MIMIC-ABN and MIMIC-CXR datasets. A ✓indicates that the component is included, while an ✗denotes that it is removed.

6.5 Results and Analyses

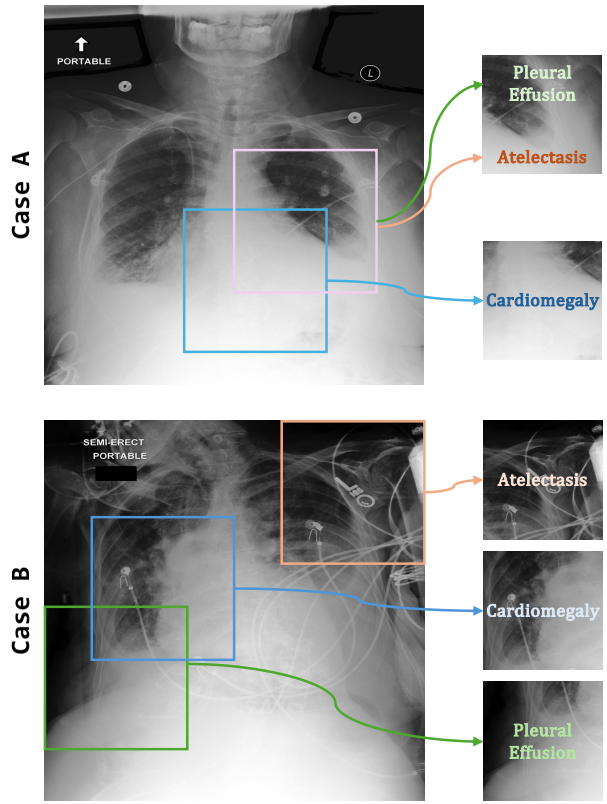
6.5.1 Quantitative Analysis

Inter-Report Consistency Analysis. Table 6.4 provides CON and R-CON scores of baselines, our model, and its ablated variants. ICON achieves the highest R-CON scores on both datasets, increasing from 0.127 to 0.140 and from 0.126 to 0.163, respectively, thereby demonstrating the best inter-report consistency. In terms of the CON score, ICON demonstrates competitive performance on the MIMIC-ABN dataset compared to the best baseline, ORGAN, and achieves the highest performance on the MIMIC-CXR dataset (0.351). Furthermore, we observe that all three components (ZOOMER, INSPECTOR, and MIXUP) contribute to improved inter-report consistency, highlighting the effectiveness of these modules.

Language Generation and Temporal Modeling Results. The language generation results are presented in Table 6.1 and the temporal modeling results are listed in Table 6.2. Among all models, ICON achieves SoTA performance on the NLG and temporal metrics. As shown in Table 6.1, our model demonstrates significant improvements on the MIMIC-ABN dataset and achieves competitive performance on the MIMIC-CXR dataset. Specifically, ICON achieves BLEU-1/2/3/4 scores of 0.337/0.195/0.126/0.086, a METEOR score of 0.129, and a ROUGE-L score of 0.236 on the MIMIC-ABN dataset, while demonstrating performance comparable to RECAP, with BLEU-1/2/3/4 scores of 0.429/0.266/0.178/0.126, a METEOR score of 0.170, and a ROUGE-L score of 0.287. These results demonstrate that our model is capable of producing highly readable and coherent radiology reports. Additionally, we provide experimental results on the IU X-RAY dataset as a reference in Table 6.3. Our model achieves competitive performance in terms of ROUGE-L score, demonstrating a 2% improvement over the R2GEN baseline, despite not being trained on the IU X-RAY dataset. This result further highlights the strong generalization ability of ICON.

Regarding temporal modeling, as shown in Table 6.2, ICon demonstrates significant improvements over other baselines in terms of BLEU and TEM scores, achieving 0.120 and 0.335, respectively. At the same time, it maintains competitive performance on the ROUGE-L score at 0.296, indicating an enhanced ability to effectively leverage historical records.

Clinical Efficacy Results. In the right section of Table 6.1, we observe that ICon achieves SoTA clinical efficacy, increasing the macro CE F_1 score from 0.393 to 0.464 on the MIMIC-CXR dataset and rising by 5.5% on the MIMIC-ABN dataset. These results indicate that our model is capable of generating accurate radiology reports when provided with region-level information. When compared to the best baseline (Recap), ICon demonstrates significant improvements on several observations, including *Lung Opacity*, *Lung Lesion*, *Pneumonia*, *Pleural Effusion*, and *Fracture* comparing the per-observation results in Table 5.12 and Table 6.7. These improvements may be attributed to a better vision model, specifically the Swin Transformer used in ICon compared to the ViT used in Recap, as well as to the fine-grained regional information extracted by Zoomer. Furthermore, Table 6.3 presents the RadGraph F_1 score on both the IU X-ray and MIMIC-CXR datasets. Our model achieves competitive performance compared to the non-RL-optimized baselines. Specifically, ICon attains RG_E and RG_{ER} scores of 0.312 and 0.278 on the MIMIC-CXR dataset, respectively. On the IU X-ray dataset, our model achieves improvements of 0.112, 0.110, and 0.093 on the three RadGraph metrics, respectively. These results indicate that ICon can generate radiology reports that are accurate at both the observation and entity levels. We also provide example-based CE results in Table 6.5 and the performance of Zoomer in Table 6.8 for reference. As shown in Table 6.5, our model achieves the highest example-based CE scores compared to all baseline methods on both datasets. Regarding the per-observation results, we find that ICon achieves strong performance on rich-resource observations (e.g., *Cardiomegaly* and *Support Devices*). This is consistent with the results presented in previous chapters. Furthermore, we observe



Reference A: in comparison with the study of there is little overall change. continued enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis. central catheter remains in place.

ICon *w/o* Zoom: as compared to the previous radiograph the patient has received a right-sided picc line. the course of the line is unremarkable the tip of the line projects over the mid svc. there is no evidence of complications notably no pneumothorax. otherwise the radiograph is unchanged.

ICon: in comparison with the study of there is little overall change. continued enlargement of the cardiac silhouette with pulmonary vascular congestion and mild to moderate cardiomegaly. small bilateral pleural effusions with areas of compressive atelectasis. the right picc line has been removed. nasogastric tube remains in place.

Reference B: in comparison with the study of there is again enlargement of the cardiac silhouette with extensive bilateral pleural effusions and compressive atelectasis combined with pulmonary vascular congestion.

ICon *w/o* Zoom: as compared to the previous radiograph there is no relevant change. moderate cardiomegaly with bilateral pleural effusions and subsequent areas of atelectasis. moderate pulmonary edema. no newly appeared focal parenchymal opacity suggesting pneumonia.

ICon: as compared to the previous radiograph there is unchanged evidence of moderate cardiomegaly with pulmonary vascular congestion and moderate pulmonary edema. unchanged moderate bilateral pleural effusions with areas of compressive atelectasis at the lung bases. no new parenchymal opacities. no pneumothorax.

Figure 6.4: A case study of ICon on two semantically equivalent cases (i.e., Case A and Case B), given their radiographs and lesions. Spans with the same color (*Cardiomegaly*, *Pleural Effusion*, *Atelectasis*, and *Edema*) represent the same positive observation. Consistent and accurate outputs are highlighted with underline.

a performance gap of 2.8% between image classification and report classification results on the MIMIC-CXR dataset, suggesting that the predicted observations are not fully realized in the generated reports. This discrepancy may be attributed to the coarse-grained regional features extracted by Zoomer, which is trained using image-level labels. Additionally, Zoomer, when trained on the IU X-Ray dataset, exhibits weaker performance (0.225 F_1) than when trained on the MIMIC-ABN (0.411 F_1) and MIMIC-CXR (0.491 F_1) datasets, as presented in Table 6.8. These results suggest that the number of training samples plays a critical role in achieving better performance on this task.

Ablation Results. The ablation results for MIMIC-ABN and MIMIC-CXR are listed in Table 6.4 and Table 6.6. We study three variants to investigate the effectiveness of Zoomer, Inspector, and Mixup:

- ICon *w/o* Zoomer: This variant is a standard Transformer-based encoder-decoder model where Zoomer, Inspector, and Mixup are all removed.
- ICon *w/o* Inspector: This variant includes only Zoomer, with both Inspector and Mixup removed.
- ICon *w/o* Mixup: This variant includes both Zoomer and Inspector, with only Mixup removed.

We observe that the performance of ICon *w/o* Zoomer drops significantly on the NLG and CE metrics across both datasets, with notable decreases in CE F_1 score of 5.4% and 18.6%, respectively. In contrast, the variant *w/o* Inspector still achieves competitive performance, with F_1 scores of 0.352 and 0.423 on the MIMIC-ABN and MIMIC-CXR datasets, respectively. This suggests that Zoomer effectively extracts lesion-related features and provides relevant abnormality information that benefits clinical accurate report generation. Secondly, we assess the role of Inspector. The variant *w/o* Mixup yields further performance improvements, particularly on

the MIMIC-CXR dataset, with an F_1 score increase of 3.9%. This highlights the effectiveness of INSPECTOR in transforming concise lesion information into precise diagnostic reports. Finally, the introduction of lesion-aware mixup (ICON) leads to slight improvements in both NLG and CE metrics compared to ICON *w/o* MIXUP. These results underscore the overall effectiveness of ICON in accurate radiology report generation. We observe similar trends in the example-based CE scores, as presented in Table 6.5. Regarding consistency metrics, we find that both INSPECTOR and MIXUP play significant roles in enhancing inter-report consistency, as shown in Table 6.4. Specifically, we observe that INSPECTOR primarily contributes to improving the CON score, while MIXUP leads to an increase in the R-CON score. Introducing INSPECTOR raises CON from 0.183/0.175 to 0.253/0.245, and including MIXUP improves R-CON from 0.119/0.156 to 0.140/0.163 on the two datasets, respectively. These results demonstrate that the introduced modules collectively contribute to the improvement of inter-report consistency.

6.5.2 Qualitative Analysis

Case Study. Figure 6.4 showcases the outputs of ICON on two semantically equivalent cases, i.e., Case A and Case B, extracted from the test set of MIMIC-CXR. In both instances, ICON successfully identifies abnormal observations (e.g., *Cardiomegaly*, *Pleural Effusion*, *Atelectasis*, *Edema*). Then, by extracting and incorporating lesions from the given radiograph, ICON generates consistent phrases including "*pulmonary vascular congestion*", "*bilateral pleural effusions*", and "*compressive atelectasis*" for these two cases. Conversely, the variant *w/o* ZOOM fails to produce these descriptions in Case A. This demonstrates that ZOOMER plays a crucial role in identifying lesions and highlights the ability of the mixup augmentation to ensure the alignment of lesions with their corresponding attributes.

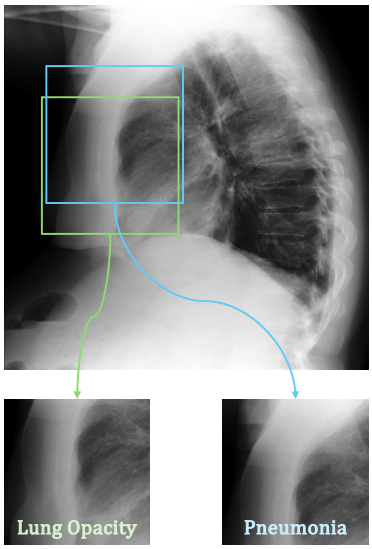
Error Analysis. We conduct an error analysis to provide more insights, and Figure

Observation	Image Classification			Report Classification		
	P	**R**	**F_1**	**P**	**R**	**F_1**
Enlarged Cardiomediastinum	0.426	0.540	0.476	0.442	0.525	0.428
Cardiomegaly	0.635	0.838	0.722	0.630	0.822	0.714
Lung Opacity	0.535	0.725	0.616	0.542	0.563	0.552
Lung Lesion	0.318	0.187	0.235	0.321	0.177	0.228
Edema	0.471	0.851	0.607	0.464	0.784	0.583
Consolidation	0.283	0.227	0.251	0.275	0.162	0.204
Pneumonia	0.367	0.396	0.381	0.341	0.350	0.345
Atelectasis	0.541	0.660	0.595	0.539	0.620	0.577
Pneumothorax	0.392	0.481	0.432	0.400	0.444	0.421
Pleural Effusion	0.719	0.842	0.776	0.721	0.827	0.770
Pleural Other	0.289	0.440	0.349	0.295	0.315	0.304
Fracture	0.266	0.198	0.227	0.225	0.164	0.190
Support Devices	0.747	0.850	0.795	0.785	0.784	0.785
No Finding	0.366	0.459	0.407	0.263	0.535	0.352
Macro Average	0.454	0.550	0.491	0.445	0.505	0.464

Table 6.7: Experimental results of each observation on the MIMIC-CXR dataset. Image classification denotes the results of Zoomer, and report classification refers to the results of CheXbert.

Dataset	**P**	**R**	**F_1**
IU X-ray	0.223	0.243	0.225
MIMIC-ABN	0.379	0.472	0.411
MIMIC-CXR	0.454	0.550	0.491

Table 6.8: Abnormal observation prediction results of Zoomer at Stage 1. Results on the IU X-ray dataset are only provided for reference.



① Pneumonia/False NEG
② Lung Opacity/False POS

Reference: pa and lateral views of the chest. there are new opacities in the superior segment of the left lower lobe and in the right lower lobe most consistent with ① multifocal pneumonia. no pleural effusion or pneumothorax. cardiomediastinal and hilar contours are normal.

ICon: .....the heart size remains unchanged and is within normal limits. ...the pulmonary vasculature is not congested. no signs of acute or chronic parenchymal infiltrates are present and the lateral and posterior pleural sinuses are free. no pneumothorax in the apical area on frontal view. when comparison is made with the next preceding study there is a new area of ② increased opacity in the left.

Figure 6.5: An error case produced by ICon, with the its reference and extracted regions provided. The span and span denote false negative and false positive observations, respectively.

6.5 presents an error case produced by ICon. Although Zoomer successfully identifies *Pneumonia* in the given radiographs, the Generator fails to realize it into descriptions like "*multifocal pneumonia*" (i.e., a false negative observation). We note that the region associated with this observation may not be precisely identified, and the regional information available in this lateral CXR may be insufficient for accurate report generation. Additionally, Zoomer outputs a false positive observation *Lung Opacity*, leading to an inaccurate phrase "*increased opacity*". To mitigate these issues, a better Zoomer trained with larger datasets could be beneficial.

6.6 Chapter Summary

In this chapter, we propose ICon, comprising three components to improve both accuracy and inter-report consistency. ICon first extracts lesions and then matches fine-grained attributes for report generation. A lesion-aware mixup method is devised for attribute alignment. Experimental results on three datasets demonstrate the effectiveness of ICon. In the future, we plan to explore incorporating large language models (LLMs) into our framework, given their advanced capabilities in planning and generation, to further enhance the performance of radiology report generation. Leveraging the strengths of LLMs could provide more refined signals to enhance the performance of ICon. It also addresses the limitation of non-existent inputs mentioned in Chapter 5, because ICon collects all images from a study, as well as the prior-visit study, as inputs.

Although ICon can improve the consistency of radiology report generation, it still exhibits some limitations. Since our lesion extraction method is based on image labels, training such a model requires annotations for images. However, obtaining these annotations can be challenging in some medical settings. Recent advances in foundation vision models [76] and open-set learning [209] could be a potential direction to address this issue. Additionally, image labels are coarse-grained, so the overall accuracy is likely to be lower than when using fine-grained labels (e.g., bounding boxes). Moreover, since our framework consists of two stages, prediction errors can propagate through the pipeline, making the final performance of our framework largely dependent on Stage 1. Reinforcement learning [125] that takes factual improvement as a reward could be a solution to optimize the framework in an end-to-end manner.

Chapter 7

Conclusions and Future Work

This thesis makes significant strides in enhancing radiology report generation by extracting and incorporating clinical information and knowledge from various sources. Our research addresses three primary research problems: (1) How to improve the disease/observation accuracy of generated reports given CXR images, especially when LLMs can produce highly readable and coherent clinical texts? (2) How to properly model the attributes of diseases/observations that reflect both spatial characteristics and temporal progression, given sequential CXRs? (3) How to regulate a radiology report generation model to produce consistent reports at the attribute-level when semantically equivalent radiological studies are provided as input? By exploring these key aspects of radiology report generation, this work tackles fundamental challenges in improving the accuracy and reliability of automatically generated reports.

Through our exploration, we demonstrated the effectiveness of the proposed approaches, highlighting their ability to improve the accuracy, consistency, and clinical relevance of generated radiology reports. Our findings indicate the great potential of this work in advancing automated radiology report generation, paving the way for more reliable AI-assisted diagnostic tools. By addressing key challenges in report generation, our work serves as a foundation for future advancements in medical AI, bringing us closer to

a healthcare system where automated tools seamlessly support clinicians in delivering accurate and timely diagnoses.

7.1 Summary of Contributions

The following sections summarize the main contributions of this thesis.

7.1.1 Observation-aware Radiology Report Generation

- We propose an observation-guided radiology report generation framework (ORGan) that can maintain the clinical consistency between radiographs and generated free-text reports. To achieve better surface realization for observations, we construct a three-level observation graph containing observations, n-grams, and tokens based on the training corpus. Then, we perform tree reasoning over the graph to dynamically select observation-relevant information.

- To further enhance clinical accuracy leveraging LLMs, we propose Radar, a novel framework that effectively integrates both the internal knowledge of LLMs and externally retrieved domain-specific knowledge. To optimize knowledge utilization, we introduce a knowledge extraction method that identifies and retains non-overlapping information from the model's learned knowledge, reducing redundancy and bridging the knowledge gap.

- We conduct extensive experiments on three benchmark datasets: MIMIC-CXR, CheXpert-Plus, and IU X-ray, demonstrating the effectiveness of ORGan and Radar in terms of language quality and clinical accuracy.

7.1.2 Spatiotemporally Precise Radiology Report Generation

- We propose RECAP, which can capture both spatial and temporal information for generating precise and accurate free-text reports.

- To achieve precise attribute modeling, we construct a disease progression graph containing both observations and fine-grained attributes that quantify the severity of diseases. Then, we devise a dynamic disease progression reasoning (PrR) mechanism to select observation/progression-relevant attributes.

- We conduct extensive experiments on two publicly available benchmarks, and experimental results demonstrate the effectiveness of our model in generating precise and accurate radiology reports.

7.1.3 Consistent Radiology Report Generation

- To the best of our knowledge, we are the first to introduce inter-report consistency in radiology report generation. To this end, we devise two metrics (CON and R-CON) to measure such consistency.

- We propose ICON, which improves both the consistency and accuracy in radiology report generation by capturing abnormalities at the region level. ICON only requires coarse-grained labels (i.e., image labels) for training to extract lesions, in contrast to previous methods that require fine-grained labels (i.e., bounding boxes).

- Extensive experiments are conducted on three publicly available datasets, and the results demonstrate the effectiveness of ICON in terms of improving both the consistency and accuracy of the generated reports.

7.2 Future Work

While this thesis has laid the foundation for enhanced radiology report generation, several areas for future research remain open to further refinement and expansion of our findings:

- **Multimodal LLMs Adaptation:** Radiology report generation is a medical image captioning task and thus could be addressed with multimodal LLMs. Given the complex input structures in radiographic studies, such as single images, multiple images, and images accompanied by clinical context, multimodal LLMs offer a flexible and effective way to handle heterogeneous inputs. These models can convert inputs into sequences of representations via appropriate encoders, enabling them to process inputs of varying types and lengths. Future research could explore adapting LLMs for this task along two axes: input formulation and training strategies. For input formulation, promising directions include more principled integration of multimodal information (e.g., pretraining stronger multimodal encoders or pruning redundant content). For training strategies, LLMs could be adapted to improve performance by injecting domain knowledge and introducing new capabilities (e.g., segmentation).

- **Enhanced Image Understanding Capability:** Incorporating high-level structured or semi-structured clinical information has the potential to significantly improve radiology report generation by guiding models to accurately describe findings in medical images. Future research could explore the integration of advanced image understanding models, which may provide improved visual representations or extract more accurate clinical information (e.g., diseases) from images. By incorporating enhanced representations and clinical information, models could generate more accurate and structured reports, ultimately improving diagnostic accuracy and clinical decision-making.

- **Better Sequential Studies Modeling:** Effectively capturing differences between sequential studies of the same patient is crucial for generating accurate and precise radiology reports. Previous approaches have primarily relied on a single prior study for context in progression modeling, potentially overlooking important variations across multiple studies. Future research could focus on developing more advanced methods to analyze longer and more complex sequences of structured studies, enabling models to better capture nuanced changes over time and improve the consistency and clinical relevance of generated reports.

- **Introducing Expert Intervention:** While artificial intelligence holds great promise for medical applications, directly deploying AI-driven diagnostic tools carries risks, including potential misdiagnoses and ethical concerns. Future research could explore the integration of expert intervention during report generation and other stages of the diagnostic workflow. By incorporating real-time feedback from experts, AI systems could dynamically adjust their outputs, correct errors, and refine their interpretations. This interactive approach would not only enhance the reliability and safety of automated radiology reporting but also foster a collaborative AI-human diagnostic process, ultimately improving patient outcomes.

- **Customizing Reports for Patients:** As medical report generation datasets are mostly collected from clinical practice, the reports are typically written by radiologists. Despite being rigorous, the terminologies used in these reports are usually not patient-friendly and often require additional interpretation from experts. This can pose challenges for patients trying to understand their CXRs. Future work could explore converting radiology reports into plain language or utilizing LLMs to explain the findings. By doing this, radiology reports become more accessible and understandable to patients and other non-experts.

By exploring these future research directions, we can further advance the field of

radiology report generation, enabling medical AI systems to integrate clinical information more effectively. This will lead to more accurate, detailed, and context-aware interpretations, ultimately enhancing diagnostic precision and improving patient care.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu,

Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In Proceedings of the IEEE International Conference on Computer Vision, pages 8948–8957, 2019.

[3] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019.

[4] 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 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society, 2018.

[5] Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. Artificial intelligence and human trust in healthcare: Focus on clinicians. J Med Internet Res, 22(6):e15154, Jun 2020.

[6] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[7] Fan Bai, Yuxin Du, Tiejun Huang, Max Q. H. Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models, 2024.

[8] 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, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[9] Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024.

[10] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing, 2023.

[11] Siddharth Biswal, Cao Xiao, M Brandon Westover, and Jimeng Sun. Eegtotext: learning to write medical reports from eeg recordings. In Machine learning for healthcare conference, pages 513–531. PMLR, 2019.

[12] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing, page 1–21. Springer Nature Switzerland, 2022.

[13] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis, 66:101797, December 2020.

[14] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation, 2022.

[15] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats, 2024.

[16] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation, 2024.

[17] Hao Chen, Wei Zhao, Yingli Li, Tianyang Zhong, Yisong Wang, Youlan Shang, Lei Guo, Junwei Han, Tianming Liu, Jun Liu, and Tuo Zhang. 3d-ct-gpt: Generating 3d radiology reports through integration of large vision-language models, 2024.

[18] Pingyi Chen, Honglin Li, Chenglu Zhu, Sunyi Zheng, Zhongyi Shui, and Lin Yang. Wsicaption: Multiple instance generation of pathology reports for gigapixel whole-slide images, 2024.

[19] Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9494–9509, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[20] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 5904–5914. Association for Computational Linguistics, 2021.

[21] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1439–1449, Online, November 2020. Association for Computational Linguistics.

[22] 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, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation, 2024.

[23] Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large language model with region-guided referring and grounding for ct report generation, 2024.

[24] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22–29, 1990.

[25] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. Controllable chest X-ray report generation from longitudinal representations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4891–4904, Singapore, December 2023. Association for Computational Linguistics.

[26] 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 Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[27] Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 12902–12915, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[28] 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. Journal of the American Medical Informatics Association, 23(2):304–310, 2016.

[29] 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, pages 248–255, 2009.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[31] Shizhe Diao, Ruijia Xu, Hongjin Su, Yilei Jiang, Yan Song, and Tong Zhang. Taming pre-trained language models with n-gram representations for low-resource domain adaptation. 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), pages 3336–3349, Online, August 2021. Association for Computational Linguistics.

[32] 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.

[33] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. Transactions of the Association for Computational Linguistics, 9:1012–1031, 2021.

[34] 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 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, editors, Proceedings of Machine Learning for Health, volume 158 of Proceedings of Machine Learning Research, pages 209–219. PMLR, 04 Dec 2021.

[35] 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 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, editors, Proceedings of Machine Learning for Health, volume 158 of Proceedings of Machine Learning Research, pages 209–219. PMLR, 04 Dec 2021.

[36] Mehrdad Eshraghi Dehaghani, Amirhossein Sabour, Amarachi B. Madu, Ismini Lourentzou, and Mehdi Moradi. Representation Learning with a Transformer-Based Detection Model for Localized Chest X-Ray Disease and Progression Detection . In proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, volume LNCS 15001. Springer Nature Switzerland, October 2024.

[37] ROGERS FB. Medical subject headings. Bulletin of the Medical Library Association, 51:114–116, 1963.

[38] Alex Foo, Wynne Hsu, Mong Li Lee, and Gavin SW Tan. Dp-gat: A framework for image-based disease progression prediction. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 2903–2912, 2022.

[39] Gusztáv Gaál, Balázs Maga, and András Lukács. Attention u-net based adversarial architectures for chest x-ray lung segmentation. arXiv preprint arXiv:2003.10304, 2020.

[40] Yijian Gao, Dominic Marshall, Xiaodan Xing, Junzhi Ning, Giorgos Papanastasiou, Guang Yang, and Matthieu Komorowski. Anatomy-guided radiology report generation with pathology-aware regional prompts, 2024.

[41] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 7995–8004, 2024.

[42] Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction, 2024.

[43] Deepak Gupta, Russell Loane, Soumya Gayen, and Dina Demner-Fushman. Medical image retrieval via nearest neighbor search on pre-trained image features, 2022.

[44] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography, 2024.

[45] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging . In proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, volume LNCS 15012. Springer Nature Switzerland, October 2024.

[46] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Doga, Omer Faruk Durugol, Weicheng Dai, Murong Xu, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz

Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography, 2025.

[47] Michael P Hartung, Ian C Bickle, Frank Gaillard, and Jeffrey P Kanne. How to create a great radiology report. Radiographics, 40(6):1658–1670, 2020.

[48] P Harzig, YY Chen, F Chen, and R Lienhart. Addressing data bias problems for chest x-ray image report generation. arxiv 2019. arXiv preprint arXiv:1908.02123.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.

[50] Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models. arXiv preprint arXiv:2411.18672, 2024.

[51] Dennis Hein, Zhihong Chen, Sophie Ostmeier, Justin Xu, Maya Varma, Eduardo Pontes Reis, Arne Edward Michalson, Christian Bluethgen, Hyun Joo Shin, Curtis Langlotz, and Akshay S Chaudhari. Chexalign: Preference fine-tuning in chest x-ray interpretation models without human feedback, 2025.

[52] Yukina Hirata, Kenya Kusunose, Takumasa Tsuji, Kohei Fujimori, Jun'ichi Kotoku, and Masataka Sata. Deep learning for detection of elevated pulmonary artery wedge pressure using standard chest x-ray. Canadian Journal of Cardiology, 37(8):1198–1206, 2021.

[53] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[54] Wenjun Hou, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiang Liu. ICON: Improving inter-report consistency in radiology report generation via lesion-aware mixup augmentation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9043–9056, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[55] Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. Radar: Enhancing radiology report generation with supplementary knowledge injection, 2025.

[56] Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. Recap: Towards precise radiology report generation via dynamic disease progression reasoning, 2023.

[57] Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. ORGAN: Observation-guided radiology report generation via tree reasoning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8108–8122, Toronto, Canada, July 2023. Association for Computational Linguistics.

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

[59] Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I-Hung Lin, Kang Wang, Hiromasa Morikawa, Hernghua Chang, Jesper Tegner, and Marcel Worring. Deepopht: medical report generation for retinal images via deep models and visual explanation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2442–2452, 2021.

[60] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2):1–55, January 2025.

[61] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3942–3951, 2021.

[62] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 3942–3951, October 2021.

[63] Xiaofei Huang, Wenting Chen, Jie Liu, Qisheng Lu, Xiaoling Luo, and Linlin Shen. Damper: A dual-stage medical report generation framework with coarse-grained mesh alignment and fine-grained hypergraph matching, 2024.

[64] Xiyang Huang, Yingjie Han, Yx L, Runzhi Li, Pengcheng Wu, and Kunli Zhang. CmEAA: Cross-modal enhancement and alignment adapter for radiology report generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, Proceedings of the 31st International Conference on Computational Linguistics, pages 8546–8556, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[65] 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), pages 19809–19818, June 2023.

[66] Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. Maira-1: A specialised large multimodal model for radiology report generation, 2024.

[67] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In The Thirty-Third AAAI Conference on Artificial Intelligence, pages 590–597, 2019.

[68] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 874–880, Online, April 2021. Association for Computational Linguistics.

[69] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. CoRR, abs/2106.14463, 2021.

[70] Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. Language generation with multi-hop reasoning on commonsense knowledge graph. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 725–736, Online, November 2020. Association for Computational Linguistics.

[71] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation, 2024.

[72] Baoyu Jing, Zeya Wang, and Eric P. Xing. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 6570–6580. Association for Computational Linguistics, 2019.

[73] Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. In Iryna Gurevych and Yusuke Miyao, editors, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pages 2577–2586. Association for Computational Linguistics, 2018.

[74] 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. arXiv preprint arXiv:1901.07042, 2019.

[75] Sara Kaviani, Ki Jin Han, and Insoo Sohn. Adversarial attacks and defenses on ai in medical imaging informatics: A survey. Expert Systems with Applications, 198:116815, 2022.

[76] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 4015–4026, October 2023.

[77] Kleanthis Konstantinidis. The shortage of radiographers: A global crisis in healthcare. Journal of medical imaging and radiation sciences, 55(4):101333, 2024.

[78] Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation, 2024.

[79] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[80] Bingchuan Li, Guixia Kang, Kai Cheng, and Ningbo Zhang. Attention-guided convolutional neural network for detecting pneumonia on chest x-rays. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 4851–4854, 2019.

[81] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation, 2019.

[82] 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, 2023.

[83] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024.

[84] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[85] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[86] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[87] Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, yingfeng zheng, Yizhi Liu, Flora D. Salim, Karin Verspoor, Xiaodan Liang, and Xiaojun Chang. 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.

[88] Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. Cross-modal clinical graph transformer for ophthalmic report generation, 2022.

[89] 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), pages 3334–3343, June 2023.

[90] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. World Wide Web, 26(1):253–270, 2023.

[91] Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 2863–2874, 2023.

[92] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 1537–1547, 2018.

[93] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[94] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[95] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[96] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. Medical image analysis, 42:60–88, 2017.

[97] Aohan Liu, Yuchen Guo, Jun-hai Yong, and Feng Xu. Multi-grained radiology report generation with sentence-level image-language contrastive learning. IEEE Transactions on Medical Imaging, 2024.

[98] Bo Liu, Donghuan Lu, Dong Wei, Xian Wu, Yan Wang, Yu Zhang, and Yefeng Zheng. Improving medical vision-language contrastive pretraining with semantics-aware triage. IEEE Transactions on Medical Imaging, 42(12):3579–3589, 2023.

[99] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, AAAI, pages 18635–18643, 2024.

[100] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 3001–3012. Association for Computational Linguistics, 2021.

[101] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 13753–13762. Computer Vision Foundation / IEEE, 2021.

[102] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of Findings of ACL, pages 269–280. Association for Computational Linguistics, 2021.

[103] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. CoRR, abs/1904.02633, 2019.

[104] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[105] Hui Liu, Ning Ding, Xinying Li, Yunli Chen, Hao Sun, Yuanyuan Huang, Chen Liu, Pengpeng Ye, Zhengyu Jin, Heling Bao, and Huadan Xue. Artificial intelligence and radiologist burnout. JAMA Network Open, 7(11):e2448714–e2448714, 11 2024.

[106] Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. In-context learning for zero-shot medical report generation. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 8721–8730, 2024.

[107] Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. Hc-llm: Historical-constrained large language models for radiology report generation, 2024.

[108] Wufeng Liu, Jiaxin Luo, Yan Yang, Wenlian Wang, Junkui Deng, and Liang Yu. Automatic lung segmentation in chest x-ray images using improved u-net. Scientific Reports, 12(1):8649, 2022.

[109] Yunyi Liu, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. Mrscore: Evaluating radiology report generation with llm-based reward system, 2024.

[110] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12009–12019, June 2022.

[111] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 10012–10022, October 2021.

[112] Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. From observation to concept: A flexible multi-view paradigm for medical report generation. IEEE Transactions on Multimedia, 26:5987–5995, 2023.

[113] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.

[114] Justin Lovelace and Bobak Mortazavi. Learning to generate clinically coherent chest X-ray reports. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1235–1243, Online, November 2020. Association for Computational Linguistics.

[115] Andrew Lukaszewicz, Joseph Uricchio, and Grygori Gerasymchuk. The art of the radiology report: practical and stylistic guidelines for perfecting the conveyance of imaging findings. Canadian Association of Radiologists Journal, 67(4):318–321, 2016.

[116] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition, 110:107332, 2021.

[117] Fred A Mettler Jr, Walter Huda, Terry T Yoshizumi, and Mahadevappa Mahesh. Effective doses in radiology and diagnostic nuclear medicine: a catalog. Radiology, 248(1):254–263, 2008.

[118] 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, pages 5288–5304, Online, June 2021. Association for Computational Linguistics.

[119] Maram Mahmoud A Monshi, Josiah Poon, and Vera Chung. Deep learning in generating radiology reports: A survey. Artificial Intelligence in Medicine, 106:101878, 2020.

[120] Feiteng Mu and Wenjie Li. Enhancing text generation via multi-level knowledge aware reasoning. In Lud De Raedt, editor, Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, pages 4310–4316. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[121] Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, Can Zhao, Ziyue Xu, Yufan He, Greg Heinrich, Yee Man Law, Benjamin Simon, Stephanie Harmon, Stephen Aylward, Marc Edgar, Michael Zephyr, Song Han, Pavlo Molchanov, Baris Turkbey, Holger Roth, and Daguang Xu. Vila-m3: Enhancing vision-language models with medical expert knowledge, 2025.

[122] Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. Pragmatic radiology report generation. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, Proceedings of the 3rd Machine Learning for Health Symposium, volume 225 of Proceedings of Machine Learning Research, pages 385–402. PMLR, 10 Dec 2023.

[123] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2022.

[124] Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian McAuley. Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1954–1960, Online, November 2020. Association for Computational Linguistics.

[125] Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. Factual accuracy is not enough: Planning consistent description order for radiology report generation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Online, December 2022. Association for Computational Linguistics.

[126] 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, pages 2824–2832, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[127] Gabriel Iluebe Okolo, Stamos Katsigiannis, and Naeem Ramzan. Ievit: An enhanced vision transformer architecture for chest x-ray image classification. Computer Methods and Programs in Biomedicine, 226:107141, 2022.

[128] OpenAI. Gpt-4v(ision) system card. 2023.

[129] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, pages 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[130] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[131] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance, 2023.

[132] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. Secure and robust machine learning for healthcare: A survey. IEEE Reviews in Biomedical Engineering, 14:156–180, 2020.

[133] Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 448–458. Association for Computational Linguistics, 2022.

[134] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[135] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[136] Anand Rajaraman and Jeffrey David Ullman. Data Mining, page 1–17. Cambridge University Press, 2011.

[137] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017.

[138] Vignav Ramesh, Nathan Andrew Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors, 2022.

[139] Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. Retrieval augmented chest x-ray report generation using openai gpt models, 2023.

[140] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, page 91–99, Cambridge, MA, USA, 2015. MIT Press.

[141] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 1179–1195. IEEE Computer Society, 2017.

[142] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

[143] Andrew B Rosenkrantz, Danny R Hughes, and Richard Duszak Jr. The us radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. Radiology, 279(1):175–184, 2016.

[144] David A Rosman, Judith Bamporiki, Rebecca Stein-Wexler, and Robert D Harris. Developing diagnostic radiology training in low resource countries. Current Radiology Reports, 7:1–7, 2019.

[145] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In European semantic web conference, pages 593–607. Springer, 2018.

[146] Manuel Schultheiss, Philipp Schmette, Jannis Bodden, Juliane Aichele, Christina Müller-Leisse, Felix G Gassert, Florian T Gassert, Joshua F Gawlitza, Felix C Hofmann, Daniel Sasse, et al. Lung nodule detection in chest x-rays using synthetic ground-truth data comparing cnn-based diagnosis to human performance. Scientific Reports, 11(1):15857, 2021.

[147] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11):2673–2681, 1997.

[148] 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. International Journal of Computer Vision, 128(2):336–359, October 2019.

[149] Harshita Sharma, Valentina Salvatelli, Shaury Srivastav, Kenza Bouzid, Shruthi Bannur, Daniel C Castro, Maximilian Ilse, Sam Bond-Taylor, Mercy Prasanna Ranjit, Fabian Falck, et al. Maira-seg: Enhancing radiology report generation with segmentation-aware multimodal large language models. arXiv preprint arXiv:2411.11362, 2024.

[150] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. Annual review of biomedical engineering, 19(1):221–248, 2017.

[151] Junjie Shentu and Noura Al Moubayed. Cxr-irgen: an integrated vision and language model for the generation of clinically accurate chest x-ray image-report

pairs. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 5212–5221, 2024.

[152] George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence, 1(1):e180041, 2019. PMID: 33937785.

[153] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2497–2506, 2016.

[154] Wilson Silva, Alexander Poellinger, Jaime S Cardoso, and Mauricio Reyes. Interpretability-guided content-based medical image retrieval. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, pages 305–314. Springer, 2020.

[155] George Gaylord Simpson. Mammals and the nature of continents. American Journal of Science, 241(1):1–31, 1943.

[156] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan

Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022.

[157] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1500–1519, Online, November 2020. Association for Computational Linguistics.

[158] Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. Cross-modal contrastive attention model for medical report generation. In Proceedings of the 29th International Conference on Computational Linguistics, pages 2388–2397, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[159] Yixuan Su, Yan Wang, Deng Cai, Simon Baker, Anna Korhonen, and Nigel Collier. Prototype-to-style: Dialogue generation with style-aware editing on retrieval memory. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:2152–2161, 2021.

[160] Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 643–655, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[161] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, XINHENG LYU, Tao Lin, and Lin Yang.

Pathgen-1.6m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In The Thirteenth International Conference on Learning Representations, 2025.

[162] 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 (CVPR), pages 7433–7442, June 2023.

[163] Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. XrayGPT: Chest radiographs summarization using large medical vision-language models. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pages 440–448, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[164] Adrian Tousignant, Paul Lemaître, Doina Precup, Douglas L. Arnold, and Tal Arbel. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data. In M. Jorge Cardoso, Aasa Feragen, Ben Glocker, Ender Konukoglu, Ipek Oguz, Gozde Unal, and Tom Vercauteren, editors, Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, volume 102 of Proceedings of Machine Learning Research, pages 483–492. PMLR, 08–10 Jul 2019.

[165] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[166] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil

Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai, 2023.

[167] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine, 5(1):48, 2022.

[168] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[169] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.

[170] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In CVPR, pages 3156–3164. IEEE Computer Society, 2015.

[171] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5784–5789, Hong Kong, China, November 2019. Association for Computational Linguistics.

[172] Jun Wang, Abhir Bhalerao, Terry Yin, Simon See, and Yulan He. Camanet: class activation map guided attention network for radiology report generation. IEEE Journal of Biomedical and Health Informatics, 28(4):2199–2210, 2024.

[173] Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng. Fine-grained medical vision-language representation learning for radiology report generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15949–15956, Singapore, December 2023. Association for Computational Linguistics.

[174] Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, Chuanfu Li, and Jin Tang. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset, 2024.

[175] Xiao Wang, Fuling Wang, Haowen Wang, Bo Jiang, Chuanfu Li, Yaowei Wang, Yonghong Tian, and Jin Tang. Activating associative disease-aware vision token memory for llm-based x-ray report generation, 2025.

[176] 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), page 3462–3471. IEEE, July 2017.

[177] Xinyi Wang, Grazziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. A survey of deep learning-based radiology report generation using multimodal data. arXiv preprint arXiv:2405.12833, 2024.

[178] Zhanyu Wang, Hongwei Han, Lei Wang, Xiu Li, and Luping Zhou. Automated radiographic report generation purely on transformer: A multicriteria supervised approach. IEEE Transactions on Medical Imaging, 41(10):2803–2813, 2022.

[179] 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), pages 11558–11567, June 2023.

[180] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. Meta-Radiology, 1(3):100033, 2023.

[181] Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. A medical semantic-assisted transformer for radiographic report generation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 655–664. Springer, 2022.

[182] 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, pages 2433–2442, 2021.

[183] Zhuhao Wang, Yihua Sun, Zihan Li, Xuan Yang, Fang Chen, and Hongen Liao. Llm-rg4: Flexible and factual radiology report generation across diverse input contexts, 2024.

[184] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3876–3887, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[185] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin

Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[186] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data, 2023.

[187] 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, Leo Anthony Celi, and Mehdi Moradi. Chest imagenome dataset for clinical reasoning. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[188] Joy T. Wu, Ken C. L. Wong, Yaniv Gur, Nadeem Ansari, Alexandros Karargyris, Arjun Sharma, Michael Morris, Babak Saboury, Hassan Ahmad, Orest Boyko, Ali Syed, Ashutosh Jadhav, Hongzhi Wang, Anup Pillai, Satyananda Kashyap, Mehdi Moradi, and Tanveer Syeda-Mahmood. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. JAMA Network Open, 3(10):e2022779–e2022779, 10 2020.

[189] Ruiqi Wu, Chenran Zhang, Jianle Zhang, Yi Zhou, Tao Zhou, and Huazhu Fu. MM-Retinal: Knowledge-Enhanced Foundational Pretraining with Fundus Image-Text Expertise . In proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, volume LNCS 15001. Springer Nature Switzerland, October 2024.

[190] Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. DocLens: Multi-aspect fine-grained medical text evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Sriku-

mar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 649–679, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[191] Jing Xu, Haojie Ren, Shenzhou Cai, and Xiaoping Zhang. An improved faster r-cnn algorithm for assisted detection of lung nodules. Computers In Biology And Medicine, 153:106470, 2023.

[192] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models, 2024.

[193] 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, pages 4009–4015, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[194] Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. Style-aware radiology report generation with RadGraph and few-shot prompting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 14676–14688, Singapore, December 2023. Association for Computational Linguistics.

[195] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2982–2990, 2022.

[196] Sixing Yan, William K Cheung, Keith Chiu, Terence M Tong, Ka Chun Cheung, and Simon See. Attributed abnormality graph embedding for clinically accurate x-ray report generation. IEEE Transactions on Medical Imaging, 42(8):2211–2222, 2023.

[197] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, Eric Wang, Ellery Wulczyn, Fayaz Jamil, Theo Guidroz, Chuck Lau, Siyuan Qiao, Yun Liu, Akshay Goel, Kendall Park, Arnav Agharwal, Nick George, Yang Wang, Ryutaro Tanno, David G. T. Barrett, Wei-Hung Weng, S. Sara Mahdavi, Khaled Saab, Tao Tu, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Jorge Cuadros, Gregory Sorensen, Yossi Matias, Katherine Chou, Greg Corrado, Joelle Barral, Shravya Shetty, David Fleet, S. M. Ali Eslami, Daniel Tse, Shruthi Prabhakara, Cory McLean, Dave Steiner, Rory Pilgrim, Christopher Kelly, Shekoofeh Azizi, and Daniel Golden. Advancing multimodal medical capabilities of gemini, 2024.

[198] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis, 86:102798, 2023.

[199] Shuxin Yang, Xian Wu, Shen Ge, Shaohua Kevin Zhou, and Li Xiao. Knowledge matters: Radiology report generation with general and specific knowledge. CoRR, abs/2112.15009, 2021.

[200] Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. Writing by memorizing: Hierarchical retrieval-based medical report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, 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), pages 5000–5009, Online, August 2021. Association for Computational Linguistics.

[201] Yan Yang, Jun Yu, Jian Zhang, Weidong Han, Hanliang Jiang, and Qingming Huang. Joint embedding of deep visual and semantic features for medical image report generation. IEEE Transactions on Multimedia, 25:167–178, 2021.

[202] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In 2019 IEEE international conference on data mining (ICDM), pages 728–737. IEEE, 2019.

[203] 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 Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part III, volume 12903 of Lecture Notes in Computer Science, pages 72–82. Springer, 2021.

[204] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 101–111. Springer, 2023.

[205] 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, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. medRxiv, 2022.

[206] Ke Yu, Shantanu Ghosh, Zhexiong Liu, Christopher Deible, Clare B Poynton, and Kayhan Batmanghelich. Anatomy-specific progression classification in chest radiographs via weakly supervised learning. Radiology: Artificial Intelligence, page e230277, 2024.

[207] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. BioBART: Pretraining and evaluation of a biomedical generative language model. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 97–109, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[208] 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, pages 721–729. Springer, 2019.

[209] Giacomo Zara, Subhankar Roy, Paolo Rota, and Elisa Ricci. Autolabel: Clip-based framework for open-set video domain adaptation, 2023.

[210] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

[211] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In International Conference on Learning Representations, 2018.

[212] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine, 30(11):3129–3141, August 2024.

[213] Ke Zhang, Yan Yang, Jun Yu, Jianping Fan, Hanliang Jiang, Qingming Huang, and Weidong Han. Attribute prototype-guided iterative scene graph for explainable radiology report generation. IEEE Transactions on Medical Imaging, 2024.

[214] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[215] Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. Libra: Leveraging temporal images for biomedical radiology analysis, 2024.

[216] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In Proceedings of the Thirty-Fourth Conference on Association for the Advancement of Artificial Intelligence (AAAI), pages 12910–12917, 2020.

[217] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2022.

[218] Zizhao Zhang, Pingjun Chen, Manish Sapkota, and Lin Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references, 2017.

[219] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3549–3557, 2017.

[220] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[221] Zhusi Zhong, Yuli Wang, Lulu Bi, Zhuoqi Ma, Sun Ho Ahn, Christopher J. Mullin, Colin F. Greineder, Michael K. Atalay, Scott Collins, Grayson L. Baird, Cheng Ting Lin, Webster Stayman, Todd M. Kolb, Ihab Kamel, Harrison X. Bai, and Zhicheng Jiao. Abn-blip: Abnormality-aligned bootstrapping language-image pre-training for pulmonary embolism diagnosis and report generation from ctpa, 2025.

[222] Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J. Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation, 2024.

[223] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nature Machine Intelligence, 4(1):32–40, 2022.

[224] Yuanpin Zhou and Huogen Wang. Divide and conquer radiology report generation via observation level fine-grained pretraining and prompt tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7597–7610, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[225] Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M Summers, and Zhiyong Lu. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 189–198. Springer, 2023.

[226] Xun Zhu, Ying Hu, Fanbin Mo, Miao Li, and Ji Wu. Uni-med: A unified medical generalist foundation model for multi-task learning via connector-moe. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[227] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. Medical Image Analysis, 72:102125, 2021.