Visual Interpretation and Comprehension of Chest X-ray Anomalies: Toward Trustworthy AI in Medical Imaging

Sayeh Gholipour Picha

UNIVERSITÉ GRENOBLE ALPES

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Signal Image Parole Télécoms**

Arrêté ministériel : 7 août 2006

Présentée par
Sayeh GHOLIPOUR PICHA

Thèse dirigée par **Alice CAPLIER** et
codirigée par **Dawood AL CHANTI**

préparée au sein du
laboratoire GIPSA-LAB
dans **l'école doctorale Électronique, électrotechnique, automatique, traitement du signal (EEATS)**

Visual Interpretation and Comprehension of Chest X-ray Anomalies: Toward Trustworthy AI in Medical Imaging

Thèse soutenue publiquement le **16/12/2025**,
devant le jury composé de:

Harold MOUCHERE, Full Professor
Nantes Université, Rapporteur
Chantal MULLER, Assistant Professor
National Institute of Applied Science in Lyon, Rapporteur
Bertrand RIVET, Full Professor
Université Grenoble Alpes, Examinateur, Présidente du jury
Maria ZULUAGA, Assistant Professor
Eurecom, Examinateur
Alice CAPLIER, Full Professor
Université Grenoble Alpes, Directeur de thèse
Dawood AL CHANTI, Assistant Professor
Université Grenoble Alpes, Co-Encadrement de thèse

UNIVERSITÉ DE GRENOBLE ALPES

ÉCOLE DOCTORALE SIGLE ED

Description de complète de l'école doctorale

T H È S E

pour obtenir le titre de

docteur en sciences

de l'Université de Grenoble Alpes

Mention : Signal Image Parole Télécoms

Présentée et soutenue par

Sayeh GHOLIPOUR PICHA

Visual Interpretation and Comprehension of Chest X-ray Anomalies: Toward Trustworthy AI in Medical Imaging

Thèse dirigée par Alice CAPLIER et codirigée par Dawood ALCHANTI

préparée au laboratoire complet (GIPSA LAB)

soutenue 16/12/2025

Jury :

Rapporteurs :	Harold MOUCHERE	-	Nantes Université
	Chantal MULLER	-	National Institute of Applied Science in Lyon
Directeur :	Alice CAPLIER	-	Université Grenoble Alpes
Co-Encadrement :	Dawood AL CHANTI	-	Université Grenoble Alpes
Président :	Bertrand RIVET	-	Université Grenoble Alpes
Examinateur :	Bertrand RIVET	-	Université Grenoble Alpes
	Maria ZULUAGA	-	Eurecom

Acknowledgments

"What you seek is seeking you."
— Rumi

As I close the book on my PhD journey, I am filled with gratitude for the years of learning, growth, and discovery that have shaped me, made possible through the support, trust, and collaboration of all those who accompanied me along the way.

The research presented in this thesis was conducted at the GIPSA-Lab (Grenoble Images Parole Signal Automatique) laboratory, within Grenoble INP, Université Grenoble Alpes (UGA), and partially carried out at the Phelma engineering school of Grenoble INP, which kindly hosted me for several months and provided a focused and productive working environment. This work was supported by a doctoral contract from Grenoble INP and Université Grenoble Alpes, within GIPSA-Lab and the EEATS (Électronique, Électrotechnique, Automatique, Traitement du Signal) doctoral school.

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor Alice Caplier and Dr. Dawood Al Chanti, for their unwavering support, insightful guidance, and constant encouragement throughout these years. This work would not have been possible without their trust, belief in my ideas, and willingness to let me explore the directions that inspired me most. I am profoundly thankful for their mentorship, which taught me to think critically, write rigorously, and grow as a researcher. I will always cherish our stimulating discussions and the curiosity they inspired.

I am sincerely grateful to GIPSA-Lab and Université Grenoble Alpes for providing an inspiring research environment and technical resources. My special thanks go to the GRICAD team for maintaining the computing infrastructure that made my experiments possible, their continuous support played a major role in the realization of this work.

To my colleagues and friends in the ACTIV team, thank you for your openness, for always bringing fresh perspectives and state-of-the-art ideas to our discussions, and for inspiring me to think beyond the boundaries of my thesis. I am deeply grateful for your collaboration, your generosity, and our many insightful conversations.

My heartfelt thanks also go to my friends, both old and new, for their presence and understanding through the highs and lows of this journey. In particular, to Isabella Costa Maia, thank you for your amazing friendship, encouragement, and support when I needed them most. I am equally thankful to all my colleagues at GIPSA-Lab for their kindness, inspiration, and companionship.

Finally, I extend my appreciation to my family for their love, patience, and support from afar. As my favorite poet Ferdowsi wrote, “The wise build the world with knowledge and goodness.” I hope to continue along this path.

Sayeh Gholipour Picha
Grenoble, France

Contents

List of Figures

List of Tables

Use of AI Tools

During the preparation of this thesis, I used **ChatGPT**, a large language model developed by OpenAI, as a writing assistant for language refinement, specifically for grammar correction, rephrasing, and improving clarity. Its use was strictly limited to enhancing readability and presentation. **All research ideas, methodologies, experiments, and analyses presented in this thesis are entirely my own original work.**

Programming Languages

All experiments and implementations were conducted primarily in **Python**, with some modules developed in **C++**. The main editors used were **VSCode**, **Vim**, and **NeoVim**, with **Jupyter Notebook** employed for exploratory tasks.

Writing Tools

This manuscript was written entirely in LaTeX using the CNRS online LaTeX platform.

List of Acronyms

AI *Artificial Intelligence*

CNN *Convolution Neural Network*

CXR *Chest X-ray*

FID *Frechet Inception Distance*

HIPAA *Health Insurance Portability and Accountability Act*

IoU *Intersection over Union*

LLM *Large Language Model*

LSTM *Long Short-Term Memory*

MCSE *Medical Corpus Similarity Evaluation*

MS-SSIM *Multi-Scale Structural Similarity Index*

RNN *Recurrent Neural Network*

RRG *Radiology Report Generation*

RoI *Region of Interest*

VG *Visual Grounding*

VICCA *Visual Interpretation and Comprehension of Chest X-ray Anomalies in Generated Report*

ViT *Vision Transformer*

VLM *Visual Language Model*

VQA *Visual Question Answering*

Chapter 1

Introduction

Contents

1.1 Research Motivation, and Objectives

Medical imaging is a cornerstone of clinical diagnosis and decision-making across a wide range of healthcare applications. Among the various imaging modalities, *Chest X-ray* (CXR) is one of the most commonly used tools for screening and identifying thoracic diseases. However, interpreting these images requires specialized medical expertise. Radiologists must not only detect pathological abnormalities but also generate corresponding textual reports that summarize their findings in a clinically meaningful way.

In recent years, *Artificial Intelligence* (AI) has begun to support and alleviate expert workload in medicine. With the success of AI in computer vision tasks, medical AI research has flourished, especially in fields reliant on image interpretation, such as radiology, pathology, and ophthalmology [69]. This progress is driven by algorithmic innovation and the increasing availability of large-scale medical imaging datasets. In particular, the introduction of attention mechanisms and transformer architectures [105] has enabled the development of preliminary approaches to automatic radiology report generation [20, 76, 31]. While these works demonstrate the feasibility of generating clinically relevant text from images, they still suffer from limited accuracy and incomplete coverage of pathologies, highlighting the need for more robust and explainable solutions.

To address these gaps, the central contribution of this thesis is the development of a *dual-modality self-verification strategy* for CXR report generation. The key idea is to leverage multimodal AI models not only for producing outputs, but also for validating them by cross-checking consistency between text and image. This principle underpins the proposed VICCA

framework and guides the thesis contributions to reliability, interpretability, and semantic consistency in automated radiology reporting.

Building on this strategy, we identify a considerable gap between the strong performance of AI models on benchmark datasets and their clinical validation. In the case of CXR report generation, models often achieve promising results on established benchmarks, yet they may overlook critical aspects such as providing reliable predictions across diverse pathologies, ensuring interpretability of generated findings, and maintaining generalizability when applied to new patient populations. These dimensions are particularly important in radiology, where clinical decisions depend not only on accuracy but also on trustworthiness and clarity.

C1. *Our first contribution* addresses this gap by introducing evaluation strategies that explicitly target the *reliability*, *interpretability*, and *generalizability* of AI-generated CXR reports. Concretely, we (i) adapt visual grounding models to identify and localize disease-related regions in chest X-rays (Chapter 2), (ii) validate visual grounding reliability through anatomically and textually guided diffusion process and synthetic CXR generation (Chapter 3), and (iii) propose MCSE as a metric to assess textual semantic consistency (Chapter 5). Together, these strategies enable a more clinically relevant assessment of model outputs and form the basis for identifying limitations in current systems.

In this thesis, we address the challenges of reliability and interpretability in CXR report generation by relying on multimodal AI models that integrate both image and text information. Such models offer advantages over traditional architectures based solely on CNNs or RNNs, as they are inherently designed to capture cross-modal relationships: they can ground textual concepts in visual regions (e.g., MedRG [126], MIMO [18]), align medical findings with corresponding image features using generative fusion (e.g., X-Ray-CoT [80], Generate-to-Ground [82]), and provide richer contextual representations that enhance interpretability (e.g., ChEX [78]). These properties make multimodal models particularly suited to improving both the trustworthiness of predictions and the interpretability of generated reports.

This trend mirrors broader developments in natural image analysis, where large language models (LLMs) and vision-language models (VLMs) have achieved remarkable success in producing coherent, context-aware outputs for diverse applications, such as conversational AI and image captioning [117, 17]. However, while these advances demonstrate the potential of multimodal approaches, their direct transfer to medical imaging is limited. Adapting them to the clinical context remains essential to handle the specific requirements of radiology, such as domain-specific terminology, pathology coverage, and clinical validation.

Despite their popularity in natural image domains, LLMs and VLMs are still emerging in medical imaging. Medical image captioning, for instance, shows promise in improving diagnostic accuracy and enhancing contextual understanding [8, 78]. However, unlike captioning in natural images, *Radiology Report Generation* (RRG) requires deep domain expertise, nuanced interpretation of subtle pathological features, and strict adherence to structured clinical reporting standards [96, 99, 10, 45]. This makes radiology reports not only difficult to generate automatically, but also central to the question of interpretability: a VLM that takes an image and outputs a textual report must capture domain-specific terminology, clinical reasoning,

and long-form narrative structures. Previous studies have highlighted frequent challenges, including factual inconsistencies, omission of critical findings, and limited alignment with radiological conventions [10, 45, 96]. Addressing these gaps is therefore essential to ensure that RRG systems provide outputs that are both clinically accurate and interpretable.

In the specific case of CXR report generation, these challenges are compounded by the need to integrate text and image data in a way that maintains interpretability. Errors in generated reports may lead to clinical misjudgments or erode trust among practitioners [117]. To move beyond benchmark performance, robust validation methods are required to evaluate how well generated reports align with the underlying radiographic evidence. Such validation may take the form of specialized metrics or dedicated multimodal models capable of jointly assessing textual and visual consistency, thereby ensuring that reported findings are both clinically meaningful and grounded in the image.

C2. *Our second contribution* addresses this gap by proposing multimodal validation methods that combine both metrics and evaluation models to assess textual–visual consistency in generated CXR reports. Specifically, we design a scoring mechanism to evaluate interpretability in grounding models (Chapter 4) and introduce MCSE as a multimodal metric for report evaluation (Chapter 5). By explicitly linking reported findings to radiographic evidence, these methods enhance transparency and provide a more robust foundation for clinicians to evaluate and trust AI-generated reports.

In the context of Chest X-ray analysis, few studies [78, 103, 8] have focused on improving automated report generation by detecting pathologies in CXR images and producing structured, clinically relevant reports based on identified *Region of Interest* (RoI). However, these AI-generated reports are often not self-explanatory and do not include standardized metrics to evaluate their trustworthiness. Consequently, expert validation remains imperative, and the trustworthiness of these reports continues to be a significant concern. This issue is further complicated by the variability in current methods for CXR report generation, particularly in how differently they contextualize and structure content to address clinical terms. Differences in the choice of training datasets and model architectures can lead to significant variations in generated outputs [31, 103]. For instance, one model might generate a report indicating "No FINDINGs" for a given image, while another may identify and describe a pathology.

To address the broader limitations of existing AI-generated medical reports, this thesis develops a comprehensive framework that unifies the evaluation of reports with visual grounding and text-to-image synthesis. While prior works [78, 8] primarily focus on integrating image and text data for report generation, they often lack mechanisms to validate whether generated reports are consistent with the underlying medical images.

C3. *Our third contribution* is the design and implementation of the *Visual Interpretation and Comprehension of Chest X-ray Anomalies in Generated Report* (VICCA) framework, a novel multimodal pipeline that integrates visual grounding, text-guided diffusion, and report evaluation. VICCA enhances *visual interpretability*, *reliability*, and *semantic consistency* by introducing an additional evaluation layer that quantifies the alignment between generated reports and their corresponding medical images. This framework is motivated in Chapter 4,

and applied in Chapter 7 to benchmark state-of-the-art RRG models in an objective and reproducible way.

In this thesis, we adopt a task-specific notion of *reliability* tailored to CXR report generation. While prior works in medical AI often define reliability in terms of reproducibility across datasets and robustness to input noise or adversarial perturbations [22, 49], our focus is on the alignment between image-derived features and the semantic content of the generated report. Specifically, we quantify reliability through a scoring mechanism that evaluates whether the identified pathologies in the text are accurately reflected in the corresponding image regions, and vice versa. This definition is motivated by the clinical requirement that reported findings must be visually grounded in the radiograph, which is a prerequisite for interpretability and trust in automated reporting systems.

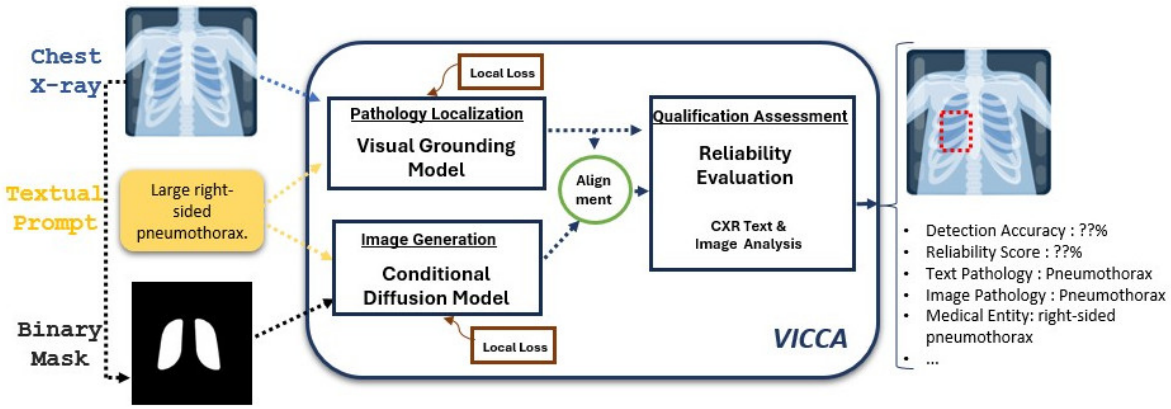


Figure 1.1: Overview of VICCA. Given a chest X-ray image and a text query, the model performs visual localization of the text input and generates a comprehensive set of information. This includes the localization accuracy, the alignment accuracy between the visual output and the text query, the presence of the specified pathology in both the text query and the chest X-ray image, the identification of the main entity within the text prompt, and a summarization of the text input into its key medical entities.

As illustrated in Figure 1.1, our framework operationalizes this dual-modality self-verification strategy. Specifically, it leverages the report content to localize pathological regions within the original CXR image using a visual grounding model. To validate this localization, a complementary AI module synthesizes a new CXR image conditioned on the same textual prompt. By comparing the regions of interest (ROIs) between the original and generated images, both semantically and spatially, we compute quantitative confidence scores that reflect the consistency and trustworthiness of the model's outputs. This general principle, using one modality to validate the other, forms the backbone of our thesis.

1.2 Thesis Approach

This thesis addresses two key challenges in medical AI: (1) the difficulty of aligning textual descriptions with image content in the absence of dense supervision, and (2) the lack of robust frameworks for validating the clinical reliability of generated content.

To address the first challenge, we adapt visual grounding methods to the medical imaging domain, enabling the localization of text-described anomalies in chest X-rays using weak supervision. This allows us to establish explicit links between clinical terminology in reports and corresponding image regions, thereby improving interpretability.

To address the second challenge, we develop a dual-modality self-verification strategy that combines visual grounding with text-guided image synthesis. By cross-checking consistency between original and generated images, we compute reliability scores that reflect the trustworthiness of model outputs. This approach is further extended to radiology report generation, where we benchmark state-of-the-art models and introduce the *Medical Corpus Similarity Evaluation (MCSE)* metric, tailored to clinical semantics.

Together, these steps form the basis of the VICCA framework, which unifies interpretability and reliability into a multimodal pipeline for evaluating and validating AI-generated medical reports.

Our multimodal pipeline operates as follows: To address interpretability, a visual grounding model processes a CXR image together with a textual prompt describing the pathologies. It outputs bounding boxes that localize the identified abnormalities within the image, thereby linking textual descriptions to specific visual regions.

To address reliability, a conditional diffusion model takes the same textual prompt together with a binary anatomical mask derived from the original CXR image. The mask serves as a structural constraint, guiding the generation process to remain anatomically plausible and reducing the risk of hallucinations. In parallel, conditioning the model on the textual prompt ensures that the synthesized image incorporates the pathology features described in the report. The outcome is not only a realistic synthetic CXR, but also a mechanism to cross-check whether textual descriptions and localized findings are consistent: if the grounded regions in the original and generated images align, we gain confidence in the accuracy and trustworthiness of the outputs.

Together, these modules establish a self-verification strategy in which one modality validates the other, providing both interpretable localization and quantifiable measures of reliability within the same pipeline.

Impact of Text Prompts

The clarity and richness of the text prompt significantly affect the performance of the conditional diffusion model. This can be illustrated in two distinct scenarios:

1. **Findings linked to anatomical structures.** Prompts such as *enlarged heart* (cardiomegaly) or *cardiomediastinal contour* are relatively easy for the model to interpret. Even though these prompts are not semantically rich, the model has learned the spatial location of the heart during training and can reliably associate the term with the correct anatomical region. As a result, the generated images remain plausible and well aligned with the prompt (see Table 1.1, row 2).

2. **Findings not tied to a fixed anatomical structure.** Pathologies such as pneumonia or pulmonary edema pose a greater challenge. These conditions can manifest in different parts of the lung fields and may appear visually similar, making semantic alignment between the text prompt and the image more difficult. In such cases, the level of detail in the report, for example, specifying laterality or affected lobe, directly influences the fidelity of the generated image and, by extension, the reliability of subsequent feature analyses (see Table 1.1, row 3). To guide interpretation, bounding boxes highlighting the relevant regions are included for each example.

Bounding Box Annotations in Table 1.1. Figures in Table 1.1 include bounding boxes highlighting pathological regions or anatomical structures. These annotations were originally provided in external CXR resources (e.g., educational radiology platforms). However, in several cases the bounding boxes were either not directly downloadable or not visually accessible in the provided images. To preserve the intent of the original annotation and ensure clarity for the reader, we manually redrew the bounding boxes based on the reference descriptions or accompanying metadata from the original source. These redrawings were used exclusively for illustrative purposes and were not used in any training, validation, or quantitative experiments in this thesis.

This analysis underscores the importance of prompt clarity and specificity: when findings are tied to well-defined anatomical structures, alignment is straightforward, whereas diffuse pathologies require more precise textual descriptions to ensure reliable multimodal consistency.

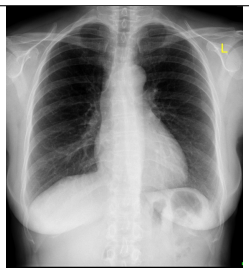
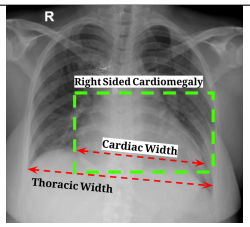
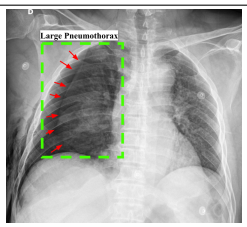
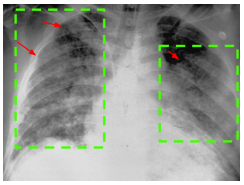
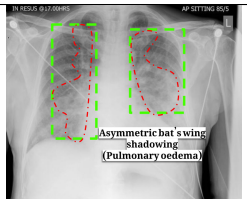
Evaluation Strategy

To evaluate the consistency and accuracy of the generated outputs, our approach produces two complementary scores:

1. **Detection Accuracy:** Quantifies how confidently the **visual grounding model** associates a reported pathology with its corresponding region in the image.

2. **Reliability Score:** A new interpretability measure that verifies whether the localized features of the pathology within the image align with the semantic content of the report. This score is derived by comparing the generated and original images in regions of interest (ROIs) that contain the most relevant pathological findings, thereby assessing feature consistency across modalities.

Table 1.1: Sample chest X-ray (CXR) images. The top row shows a normal CXR, while the remaining rows display CXRs with various pathologies, including right-sided cardiomegaly, large pneumothorax, viral pneumonia, and pulmonary edema. Bounding boxes highlight the relevant regions to guide interpretation.

Normal Chest X-ray Image	
	
Right-sided Cardiomegaly	Large Pneumothorax
	
Viral Pneumonia	Pulmonary Edema
	

The rationale behind this approach lies in detecting potential errors in generated reports. If the report inaccurately describes features not present in the original CXR image, these discrepancies will manifest in the synthetic image generated from the report's content. Consequently, the compatibility between the original and synthetic images in their respective ROIs will be low, signaling potential inaccuracies. This dual-scoring mechanism ensures a robust evaluation of both localization precision and semantic alignment, enhancing the trustworthiness of AI-generated reports.

To further address the lack of external radiologist inputs, we design an automated validation system for our pipeline that does not rely on expert feedback, thereby overcoming challenges such as limited accessibility, high costs, and the subjectivity of expert opinions. Instead, we leverage annotated datasets and established evaluation metrics to ensure an objective and reproducible validation process. This system is adaptable across diverse databases, enhancing the generalizability and robustness of our evaluation.

Pipeline Components

Our cross-validating multimodal pipeline can be summarized into the following parts, as shown in Figure 1.1:

1. **Pathology Localization Model:** A visual grounding model is adapted from natural image tasks to chest X-ray applications. To enhance performance, we integrate a BERT-like encoder [28, 10] pre-trained on CXR reports.

   Output and impact: Bounding boxes localize the detected findings, explicitly linking text to image regions. Unlike prior works that predicted text without spatial grounding, this module provides interpretable evidence, a prerequisite for trustworthy clinical adoption.

2. **Image Generation:** A stable diffusion model [29, 125] generates CXRs conditioned on input reports, guided by a binary anatomical mask from the original CXR.

   Output and impact: The generated image preserves anatomical plausibility while integrating pathology features described in the text. In contrast to generic text-to-image approaches, this guided generation enables controlled variation of pathologies and provides a means to validate text–image consistency.

3. **Medical Validation:** Validation methods assess the anatomical structure and pathological plausibility of generated CXRs, using either existing models (TorchXRayVision [21]) or customized AI architectures [11].

   Output and impact: This step ensures that synthetic images remain clinically meaningful, adding a safeguard against hallucinations and strengthening the reliability of the pipeline.

This cross-validating analysis enhances both interpretability and reliability by combining visual and textual evidence into a unified pipeline.

1.3 Research Questions

This work is guided by the following research questions (RQs). For each question, we indicate what we expect to learn, how it is addressed in the thesis, and the related contributions and publications.

- **RQ1:** How can visual-language models trained on natural datasets be adapted for grounding disease-related phrases in CXRs? *Expectation:* To determine whether visual grounding can provide interpretable localization of medical findings in chest X-rays. *Addressed in:* Chapter 2, where we adapt visual grounding models using weak supervision. *Contribution:* Forms part of **C1** (reliability, interpretability, and generalizability). *Publication:* Visual interpretation and comprehension of chest X-ray anomalies in the generated report without human feedback.[34].

- **RQ2:** Can synthetic CXR images generated from textual prompts preserve both anatomical structure and pathology? *Expectation:* To test whether diffusion-based generation can produce realistic CXRs that faithfully represent the conditions described in reports. *Addressed in:* Chapter 3, where we design a conditional diffusion model guided by anatomical masks. *Contribution:* Part of **C1** and **C3**, by enabling cross-modal validation in VICCA. *Publication:* Visual interpretation and comprehension of chest X-ray anomalies in the generated report without human feedback.[34].

- **RQ3:** How can the reliability of grounded regions be quantified using multimodal signals? *Expectation:* To establish scoring mechanisms that verify whether textual descriptions are visually consistent with radiographic evidence. *Addressed in:* Chapters 4 and 7, where we propose multimodal validation methods combining grounding and image synthesis. *Contribution:* Directly corresponds to **C2** (multimodal validation). *Publication:* Currently being prepared for submission.

- **RQ4:** What evaluation metrics best capture semantic correctness in radiology report generation? *Expectation:* To design metrics that reflect clinical semantics rather than surface-level similarity. *Addressed in:* Chapter 5, where we propose the Medical Corpus Similarity Evaluation (MCSE) metric. *Contribution:* Forms part of **C1** and **C2**. *Publication:* Semantic Textual Similarity Assessment in Chest X-ray Reports Using a Domain-Specific Cosine-Based Metric. [86].

1.4 Contributions

1. **Adaptation of Visual Grounding Models:** We adapt a visual grounding model to localize text-described anomalies in CXRs using weak supervision.

2. **CXR Image Generation with Text-Guided Diffusion:** We develop a conditional diffusion model that generates CXRs conditioned on radiology reports and anatomical masks.

3. **Reliability Score for Grounded ROIs:** We introduce a scoring mechanism to evaluate consistency between grounded ROIs in original and generated images.

4. **Radiology Report Generation and Assessment:** We evaluate state-of-the-art RRG models and propose a new metric, MCSE, for measuring clinical relevance.

5. **Validation of our Pipeline:** Our pipeline is applied to assess the interpretability and accuracy of our *Visual Interpretation and Comprehension of Chest X-ray Anomalies in Generated Report* (VICCA), a multimodal model for CXR analysis.

Table 1.2 provides an overview of how these detailed contributions align with the broader high-level objectives of the thesis.

Table 1.2: Mapping between high-level contributions (C1–C3) and detailed technical contributions (#1–#5).

High-level Contribution	**Scope / Description**	**Detailed Items**
C1. Reliability, Interpretability, Generalizability	Evaluation strategies for AI-generated CXR reports, covering reliability across pathologies, interpretability via visual–language grounding, and generalizability to new cohorts.	#3, #4
C2. Multimodal Validation Methods	Metrics and/or dedicated evaluation models (e.g., MCSE) that jointly assess textual–visual consistency to ensure clinical validity of generated reports.	#3, #4
C3. VICCA Framework	End-to-end pipeline integrating visual grounding, text-to-image diffusion, and report evaluation to enhance semantic alignment and clinical trustworthiness.	#1, #2, #5

1.5 Organization of the Thesis Manuscript

This thesis is structured as follows:

- **Chapter 1: Introduction.** Presents the background, research objectives, and a summary of the contributions made throughout this work.
- **Chapter 2: Visual Grounding for Chest X-Rays.** Explores visual grounding models, their adaptation to the medical domain, and the training protocols used for localizing disease-related regions in chest X-ray images.

- **Chapter 3: CXR Generation and Pipeline Validation.** Introduces a text-guided diffusion model for generating anatomically consistent synthetic chest X-rays and describes the proposed cross-modal validation pipeline used to assess localization reliability.
- **Chapter 4: Assessing VICCA Pipeline.** Proposes a novel scoring framework for evaluating the alignment and interpretability of visual grounding models without requiring expert feedback.
- **Chapter 5: Radiology Textual Report Generation Assessment.** Introduces the Medical Corpus Similarity Evaluation (MCSE) metric, which quantitatively evaluates generated medical reports based on clinical entity alignment and semantic coherence.
- **Chapter 6: Radiology Report Generation.** Reviews state-of-the-art methods for automated report generation from medical images and analyzes their respective strengths and limitations.
- **Chapter 7: Objective Evaluation of RRGs using VICCA.** Applies the VICCA framework to benchmark various radiology report generation (RRG) models, offering an objective evaluation strategy based on multi-modal consistency.
- **Chapter 8: Conclusion and Future Directions.** Summarizes the key findings, discusses the limitations of the current work, and outlines potential directions for future research.

Figure 1.2 provides an overview of how the chapters of this thesis map onto the three high-level contributions (C1–C3). It visually summarizes how each part of the manuscript supports the progression from foundational components, to multimodal validation tools, and finally to the fully integrated VICCA framework.

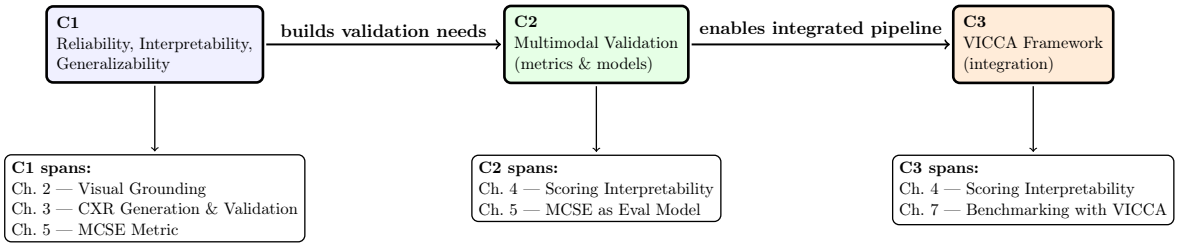


Figure 1.2: Thesis roadmap linking high-level contributions (C1–C3) to chapters. Each contribution spans specific chapters and flows toward the integrated VICCA framework.

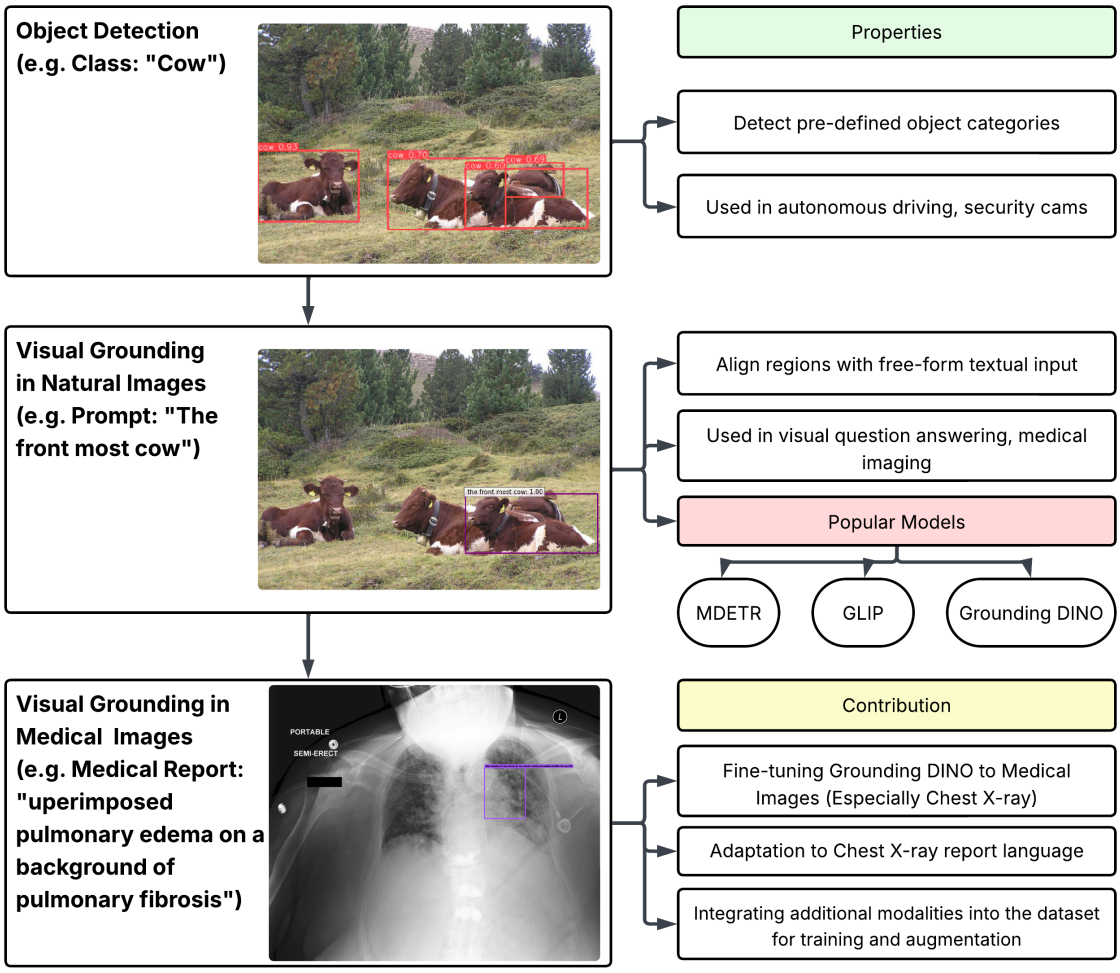
Chapter 2

From Report to Region: Visual Grounding in Medical Imaging

Contents

This chapter introduces the fundamental concepts of object detection and traces its evolution toward more interpretable visual grounding techniques that incorporate natural language prompts. We begin by reviewing the core principles and technical progression of object detection models, from early region proposal-based systems to transformer-based architectures capable of learning complex visual-textual associations. Particular emphasis is placed on how these models transition from generic region identification to semantically guided visual grounding, where textual input guides the localization of specific concepts within the image.

To contextualize these developments, we examine representative models applied to natural image domains, discussing their design choices, training strategies, and interpretability capabilities. Building upon this foundation, the chapter then shifts focus to the medical domain, with a particular emphasis on chest X-rays. We explore the challenges and adaptations



Chapter overview: from object detection to visual grounding and domain-specific adaptation in chest X-rays.

required to apply visual grounding techniques to medical images, including the scarcity of annotated localization data, the complexity of clinical language, and the need for robust semantic alignment.

By the end of this chapter, the reader will gain an understanding of the technical advancements from object detection to visual grounding, the potential benefits of phrase-level localization in clinical settings, and the limitations that must be addressed for successful integration into medical AI systems.

Terminology

Before proceeding, this section defines several terms that will be used throughout the chapter but will not be explained in detail, in order to avoid redundancy and maintain the overall flow of the narrative.

Attention Mechanisms *(see page 29)*
Attention mechanisms [105] are components in neural networks that dynamically focus on different parts of the input when generating an output. Originally introduced in the context of machine translation, attention allows the model to weigh the relevance of each input token when predicting each output token. In vision tasks, attention enables the model to emphasize salient image regions by learning dependencies between spatial locations. Variants include self-attention (used in transformers), cross-attention (between modalities like image and text), and multi-head attention.

Caption Tokens *(see page 22)*
Caption tokens are the individual words or subword units derived from a caption, typically produced through tokenization during preprocessing in vision-language models. These tokens serve as the basic input units for models that align or condition image representations on textual descriptions. The granularity of tokens (e.g., word-level or byte-pair encodings) can impact the model's ability to semantically ground visual elements.

Contrastive Learning *(see page 19)*
A training paradigm where a model learns to bring semantically similar inputs (e.g., an image and its corresponding caption) closer in the embedding space while pushing dissimilar ones apart. It is commonly used in vision-language pretraining to align image and text representations [15].

Contrastive Loss *(see page 22)*
A loss function used to learn a representation space where semantically similar pairs (e.g., an image and its corresponding caption) are pulled closer together, while dissimilar pairs are pushed apart. This is typically implemented in multimodal settings (such as image-text models) to encourage alignment across modalities by maximizing similarity for positive pairs and minimizing it for negative pairs in the embedding space [15].

Fine-grained Alignment *(see page 19)*
Fine-grained alignment refers to the precise association between elements of different modalities, such as specific words or phrases in a text and localized regions within an image. Unlike coarse-level alignment, which may link an entire sentence to a broad image area, fine-grained alignment ensures that each linguistic component (e.g., "right lower lobe opacity") is accurately mapped to its corresponding visual evidence in the image, enabling more detailed interpretation and improved model interpretability [16].

Hungarian Matching Loss *(see page 20)*
A permutation-invariant loss function used to assign predicted object detections to ground truth annotations in a one-to-one fashion. It relies on the Hungarian algorithm to find the

optimal bipartite matching between predicted and actual bounding boxes based on a composite cost (e.g., box coordinates, classification score, objectness) [59]. This technique is commonly used in transformer-based object detection models like DETR [11] and its derivatives.

Medical Report *(see page 21)*
A medical report is a formal, structured document generated by a healthcare professional, often a radiologist or physician, that summarizes observations, interpretations, and diagnostic conclusions based on a patient's clinical data, such as imaging results. In the context of chest X-rays, the report typically includes findings about the lungs, heart, bones, and other thoracic structures, often described in precise medical terminology to communicate pathological conditions or the absence of them [50].

Open-Vocabulary *(see page 21)*
Open-vocabulary refers to a model's ability to recognize, understand, or generate words or phrases beyond a predefined set of fixed labels. In the context of vision-language tasks, open-vocabulary models can dynamically interpret arbitrary textual queries and associate them with relevant visual content, even if those terms were not explicitly seen during training. This capability is critical for handling real-world variability in language and supporting generalization to unseen concepts.

Post-hoc Matching Task *(see page 20)*
A downstream evaluation step where the model is not trained to generate bounding boxes directly but rather selects the most relevant regions after processing both the image and the text. This task is typically used in retrieval-based visual grounding, where alignment is inferred from pretrained representations instead of learned end-to-end [98].

Spatial Information *(see page 17)*
In the context of visual data, spatial information refers to the geometric and positional attributes of objects or features within an image. It includes the location, size, orientation, and relative arrangement of visual elements, enabling models to understand where something appears in an image. In medical imaging, preserving spatial information is crucial for accurately identifying anatomical regions and pathological findings.

Swin Transformer *(see page 21)*
A hierarchical vision transformer architecture that introduces a shifted windowing mechanism to enable efficient and scalable image modeling [74]. Unlike traditional vision transformers that operate globally, the Swin Transformer computes self-attention within local non-overlapping windows and shifts the window partitioning between layers to enable cross-window communication. This design makes it well-suited for dense prediction tasks such as object detection, segmentation, and visual grounding.

Vision Transformer (ViT) *(see page 23)*
The Vision Transformer (ViT) [30] is a deep learning architecture that applies transformer models, originally developed for natural language processing, to image data. It splits an image into fixed-size patches, linearly embeds them, and processes the sequence of patch embeddings using standard transformer blocks with self-attention. ViT enables global context modeling

across the image and has demonstrated strong performance in various vision tasks without relying on convolutional operations.

Zero-shot Learning *(see page 21)*
Zero-shot learning [60] refers to the ability of a model to perform tasks or make predictions on previously unseen classes or concepts without having been explicitly trained on labeled examples of those classes. This is typically achieved by leveraging a shared embedding space between modalities (e.g., vision and language), where the model can infer semantic relationships and align novel inputs with learned concepts. In visual grounding, zero-shot capability allows the model to localize or recognize objects described by free-form textual queries that were not present during training.

2.1 Introduction

Artificial intelligence has made remarkable progress in interpreting visual data, particularly through advancements in object detection, image captioning, and *Visual Grounding* (VG). While object detection provides spatial information about prominent regions within an image, it lacks semantic specificity, often identifying only the presence and location of general classes such as "car", "animal" or in the case of medical domain "lung opacity" or "nodule." In contrast, visual grounding, the process of linking natural language phrases to specific image regions, offers a more comprehensive understanding by aligning visual features with fine-grained textual descriptions. This capability becomes especially critical in medical imaging, where precise interpretation can significantly impact clinical decision-making.

The motivation for this chapter stems from the need to bridge the semantic gap between the radiology reports and visual evidence in the image. In sensitive domains like healthcare, this lack of alignment can undermine trust and limit adoption. Visual grounding offers a way forward by enabling localized visual explanations for each phrase in a report, effectively answering the question: "Where in the image is this mentioned condition or observation located?"

In this chapter, we advance beyond traditional object detection by exploring visual grounding as a crucial step toward more interpretable and reliable medical AI systems. We begin with an overview of visual grounding techniques developed for natural image domains, such as those used in Visual Genome [58] and RefCOCO [55] benchmarks. We review prominent state-of-the-art architectures like MDETR [53], GLIP [124], and Grounding DINO [71], highlighting how they map textual phrases to bounding boxes in a weakly supervised or end-to-end fashion. These architectures establish a strong technical foundation for aligning visual and textual modalities.

We then transition to the medical domain, where applying visual grounding poses unique challenges:

- The absence of ground-truth spatial annotations for phrases in most radiology datasets;

- The highly structured and domain-specific nature of medical language;
- The clinical requirement for high precision in both localization and semantic interpretation.

To address these challenges, we present a pipeline for visual grounding in chest X-rays that leverages both the image and its associated radiology report for supervised report-to-region alignment. While contrastive approaches such as MedCLIP [111] have demonstrated the ability to align images and reports without explicit region annotations, they only provide global similarity scores rather than precise localizations. In contrast, our focus in this study is on supervised grounding, which produces explicit bounding boxes for phrases,an essential requirement for enabling explainability and spatial interpretability in clinical contexts. This chapter contributes to the overall thesis by providing the localization backbone that enables visual explainability of text in generated reports. It serves as a conceptual and technical prerequisite for the subsequent chapters, where we explore how this alignment can be used to:

- Enhance the interpretability of generative models for chest X-rays;
- Evaluate the coherence between generated and reference reports through visual-textual consistency;
- Assess the trustworthiness and reliability of clinical findings derived by AI systems.

Ultimately, this chapter lays the foundation for understanding where in the image each reported finding is visually grounded, enabling a more transparent, accountable, and clinically relevant use of AI in medical imaging.

2.2 From Object Detection to Visual Grounding

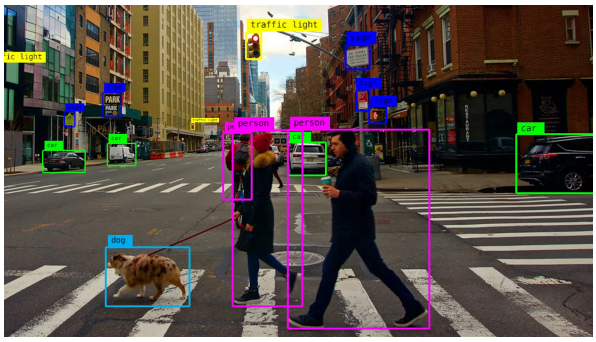
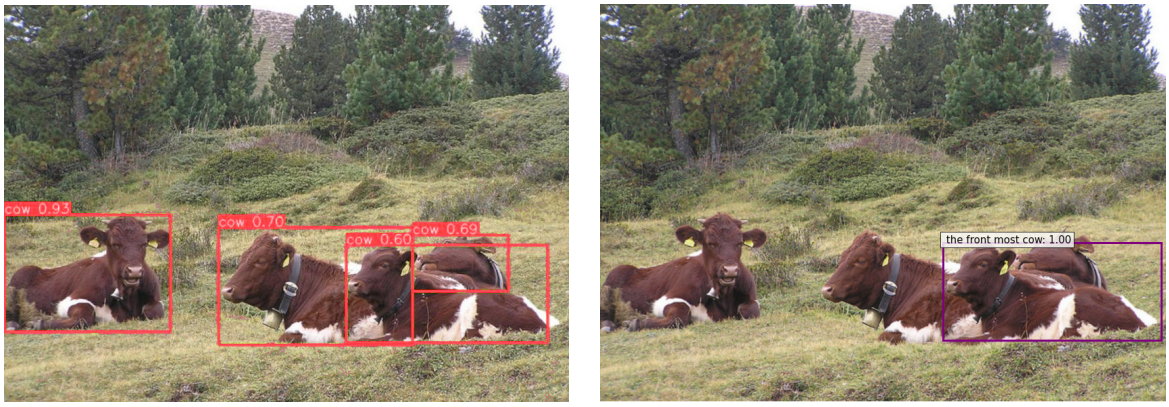


Figure 2.1: Object Detection in Real World Application using Deep Learning Approaches. Image from Medium using YOLO Model.

Traditional object detection methods in computer vision aim to identify and localize objects within an image by drawing bounding boxes around them. While highly effective for

standard visual tasks, such as detecting cars, people, or animals, as shown in figure 2.1, these approaches are inherently limited in their ability to capture complex semantic relationships or achieve fine-grained alignment between visual content and natural language descriptions [6]. Models such as Faster R-CNN [93], YOLO [92], and DETR [11] have played a foundational role in object detection by delivering impressive performance on fixed-category recognition tasks. However, these architectures rely on predefined label vocabularies and are not designed to handle arbitrary linguistic inputs [6]. As a result, they fall short in scenarios that require deeper semantic understanding or flexible interaction with free-form text. Visual grounding models address this limitation by enabling dynamic alignment between textual expressions and visual regions. Unlike conventional detectors, these models are capable of localizing image regions corresponding to diverse, open-ended textual queries, thereby offering a more adaptable and expressive interface for multimodal reasoning.

While *localization* broadly refers to identifying regions of interest (e.g., where a lesion is), *visual grounding* extends this idea by establishing an explicit mapping between a given text prompt and its corresponding region. More generally, *visual grounding* include any task that aligns visual and textual modalities, whether at the level of words, phrases, or full sentences. Figure 2.2 demonstrate the distinction using an example of cows in a field. In conventional object detection, localization is limited to identifying object classes (e.g., "cow"), without integrating textual context. In contrast, visual grounding models enable nuanced understanding by linking free-form text, such as "the front most cow," to its precise location in the image.



Object Detection using YOLO Model [92] Visual Grounding using MDETR Model [53]

Figure 2.2: Comparison between traditional object detection and visual grounding applied to an image. While object detection localizes predefined objects, visual grounding uses natural language prompts to flexibly and semantically guide localization, enabling more semantically meaningful and flexible image-text alignment.

2.3 Visual Grounding in Natural Images

Recent breakthroughs in visual grounding for natural images have been driven by multimodal transformer architectures and contrastive learning. State-of-the-art models such as MDETR

[53], GLIP [124], and Grounding DINO [71] demonstrate high performance by learning cross-modal representations that align image regions and text embedding in a shared space.

- **MDETR** reformulates grounding as a direct end-to-end detection problem by conditioning a DETR-style model on textual input, enabling it to predict bounding boxes based on phrases.
- **GLIP** leverages large-scale pretraining on grounded captions and open-vocabulary labels to generalize across domains.
- **Grounding DINO** introduces an efficient vision transformer for zero-shot object grounding, making it highly adaptable to unseen phrases.

These models are typically trained on large-scale datasets such as Visual Genome [58] or COCO Captions [68], where detailed textual annotations are explicitly associated with image regions. While these models demonstrate high accuracy and robust alignment in natural image domains, their direct application to medical imaging introduces several unique challenges. These include domain shift, the lack of richly annotated datasets, and the complexity of clinical language.

In the following sections, we provide a more detailed technical overview of each of these models and critically assess their potential for adaptation to medical imaging tasks, particularly in chest X-ray interpretation.

2.3.1 MDETR: Modulated Detection for End-to-End Multi-Modal Understanding

MDETR [53] introduces a unified architecture for visual grounding by directly conditioning a Transformer-based object detector on textual input. Unlike traditional models that treat visual grounding as a post-hoc matching task, which first generate region proposals and then associate them with language [24, 122], MDETR reformulates it as a set prediction problem where each prediction is influenced by both image features and linguistic content.

The model extends the DETR framework [11] by introducing a cross-modal encoder that fuses multi-scale visual features with text embeddings. Specifically, the image is encoded using a standard CNN backbone (e.g., ResNet [40]), and the text is embedded using a pretrained language model (such as RoBERTa [73]). These two modalities are concatenated and passed through a stack of Transformer layers, enabling bi-directional interaction between vision and language.

During training, MDETR receives a full image and a set of referring expressions (phrases or sentences) as input and learns to predict a set of aligned bounding boxes and their corresponding textual spans. This is supervised via bipartite matching loss based on Hungarian assignment, along with standard box regression and classification losses. Notably, MDETR

does not require explicit region-phrase alignment at training time; instead, it leverages weak supervision via matching at the set level, making it highly scalable.

MDETR's ability to perform zero-shot visual grounding derives from its end-to-end learning of a shared multimodal representation space, where visual and linguistic features are aligned through cross-modal attention. This design allows the model to generalize to novel phrases or concepts not encountered during training, an essential capability for tasks involving open-ended language queries.

Application to Chest X-ray Images. While MDETR demonstrates strong generalization in natural image domains, its direct application to chest X-rays is limited by several domain-specific challenges: the visual domain shift between photographic and radiographic data, the need for fine-grained clinical semantics, and the scarcity of paired region-level annotations in medical datasets.

Nevertheless, MDETR's architecture presents promising traits for adaptation. Its ability to ground free-form text enables alignment with descriptive medical reports , and its end-to-end training scheme can be adapted for weak supervision using image-report pairs. However, the model's reliance on general visual features and natural image pretraining may reduce its sensitivity to subtle or low-contrast abnormalities typical in chest X-rays.

These constraints suggest that domain adaptation strategies or hybrid learning protocols are necessary for effective deployment in medical imaging. While MDETR establishes a robust foundation, the need for greater modularity and domain-specific flexibility motivates the adoption of alternatives.

2.3.2 GLIPv2: Unifying Localization and Vision-Language Understanding

GLIPv2 [124] builds upon the original GLIP framework to create a unified model capable of both object localization and broader vision-language (VL) understanding tasks. Unlike traditional detection models constrained to predefined labels, GLIPv2 formulates detection as a grounding task where textual descriptions guide region identification. This allows the model to operate in an open-vocabulary setting, effectively bridging the gap between object detection and vision-language understanding.

The core architecture of GLIPv2 is a two-tower design comprising a vision encoder and a text encoder. The visual backbone is a modified Swin Transformer that produces multi-scale visual feature maps, while the text encoder is a large pretrained language model (e.g., RoBERTa [73]) that outputs contextualized phrase embeddings. The key innovation lies in the contrastive alignment and region-text matching objectives that simultaneously train the model to detect objects and understand multimodal relationships.

Training involves a hybrid supervision scheme that includes:

1. *Region-level grounding*: associating image regions with corresponding textual phrases via soft label assignment and contrastive loss.

2. *Caption grounding*: aligning dense region proposals with caption tokens.

3. *Classification supervision*: using standard detection labels for strong supervision where available.

To enhance generalization, GLIPv2 is trained on a large mixture of datasets, including COCO [68], Visual Genome [58], OpenImages [57], and custom image-text pairs. This results in improved performance on both detection (mAP) and VL tasks (e.g., referring expression comprehension).

Adaptation to Medical Imaging GLIPv2's ability to handle open-vocabulary detection and align phrases with regions makes it an attractive candidate for grounding radiological descriptions in chest X-rays. However, applying GLIPv2 to medical images introduces several challenges:

- The domain shift between natural and medical images reduces performance without domain-specific finetuning.
- The scarcity of paired image-text data in medical settings limits the effectiveness of its large-scale pretraining strategy.
- Complex and nuanced medical phrases (e.g., "mild bibasilar opacities") require specialized textual embeddings and visual representations that GLIPv2 does not natively support.

Nonetheless, the underlying cross-modal alignment capability remains highly relevant. With appropriate adaptation, such as training on curated radiology datasets and incorporating clinical text encoders, GLIPv2 could offer meaningful improvements in visual grounding for chest X-rays. That said, compared to newer architectures like Grounding DINO, which are specifically optimized for zero-shot and long-text grounding scenarios, GLIPv2 is less efficient and requires heavier pretraining. As such, while technically sound and generalizable, GLIPv2 is not the primary choice in our pipeline.

2.3.3 Grounding DINO

Grounding DINO [71] introduces a unified and scalable framework for open-set object detection and visual grounding by integrating the transformer-based DINO architecture with a grounded pretraining method. Unlike conventional object detectors that rely on closed vocabulary classification, Grounding DINO adopts a text-conditioned detection strategy, enabling the

model to handle open-vocabulary inputs and localize arbitrary phrases without additional fine-tuning.

The model builds upon the encoder-decoder architecture of DINO, where a backbone network, typically a ResNet [40] or a Vision Transformer (ViT) variant such as Swin Transformer, extracts multiscale image features. These features are then processed by a transformer encoder, employing multi-head self-attention to capture global contextual relationships across the image. A separate text encoder, such as BERT [28] or CLIP [90], maps the textual input into the same feature space aligned with the visual representations. A transformer decoder takes the encoded features and object queries, which are learnable embeddings, to decode information for predicting object regions and their alignment with textual descriptions.

The model employs a multimodal alignment to ensure that the embeddings of corresponding text and image elements share a unified feature space, typically using a contrastive learning approach. Predictions are made through specialized heads: one for bounding box localization and another for scoring the alignment between textual queries and detected objects. The training process uses a Hungarian matching loss for object-query assignments, a box regression loss for refining bounding boxes, and a grounding loss to enforce accurate text-object associations. This architecture enables robust performance in tasks requiring simultaneous visual and textual reasoning. This dual-objective encourages both spatial localization accuracy and semantic understanding, making it especially powerful in zero-shot settings. Importantly, the model supports multi-phrase queries and produces bounding boxes along with confidence scores for each grounded phrase.

Figure 2.3 demonstrates an example of using the Grounding DINO model for the visual grounding task. The model processes a natural language prompt (“A dog resting in the grass with a ball") and localizes the described objects (one dog, one patch of grass, and one ball) within the associated image, generating corresponding bounding boxes for each identified region.

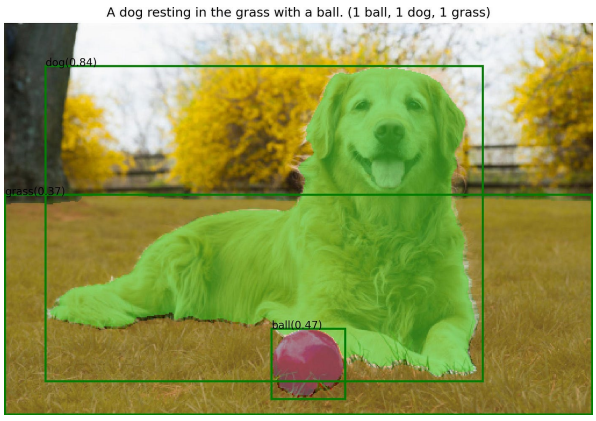


Figure 2.3: An example of visual grounding using the Grounding DINO model [71] on a natural image. The output includes localization bounding boxes for all objects mentioned in the text that are present in the image, along with their associated detection probabilities.

Applied to datasets such as Visual Genome [58], RefCOCO [55], and OpenImages, Grounding DINO achieves state-of-the-art performance in both phrase localization and general object detection tasks, especially under zero-shot and open-vocabulary conditions. Its ability to localize unseen classes and phrases during inference makes it highly adaptable and practical for real-world applications.

Application to Medical Chest X-rays. Grounding DINO's capacity to handle free-form medical phrases and localize relevant anatomical or pathological regions in a zero-shot manner is particularly valuable in radiology. In the context of chest X-rays, where detailed phrase-level annotations are scarce and domain-specific terminology is critical, Grounding DINO can be leveraged to ground textual findings (e.g., "right basilar opacity" or "mild cardiomegaly") without requiring fully supervised training on medical bounding boxes.

Its transformer backbone and language-conditioned detection pipeline allow better integration to clinical language, especially when coupled with prompt engineering and domain adaptation. Thus, among the evaluated visual grounding models, Grounding DINO stands out as the most viable for integration into medical imaging pipelines. It balances flexibility, accuracy, and scalability, offering a robust foundation for phrase-level visual-textual alignment in chest X-ray interpretation.

Among the models reviewed, Grounding DINO emerges as the most adaptable and clinically viable choice for visual grounding in chest X-rays. While MDETR and GLIP offer strong performance on natural images, their limitations in annotation dependence and domain generalization prevent their utility in medical applications. In contrast, Grounding DINO's zero-shot capacity, architectural efficiency, and pretraining flexibility make it a well-suited foundation for exploring report-driven region localization in medical AI.

2.4 Adapting Visual Grounding to Medical Imaging

Translating visual grounding models to the medical domain, particularly in radiology, presents a set of domain-specific challenges that diverge significantly from natural image tasks. Medical reports are inherently more complex and nuanced than generic image captions; they often include uncertain language, implicit reasoning, and negation, all of which require deeper semantic understanding. Furthermore, large-scale annotated datasets with fine-grained region-to-phrase alignments remain scarce. While datasets like MS-COCO [68] offer thousands of grounded captions, medical datasets such as MIMIC-CXR, the largest CXR dataset, rarely provide explicit mappings between report phrases and image regions.

Moreover, medical abnormalities often present subtly in chest X-rays, making reliable region detection significantly more difficult than in natural images. Conditions such as "subtle cardiomegaly" or "faint right basal opacity" demand both high-resolution feature extraction and precise linguistic modeling to detect them. These abnormalities often lack salient features, making them challenging to localize even for human experts, visual grounding in this context

must not only identify the correct region but also accurately interpret the clinical semantics of the accompanying text.

2.4.1 Related Works in Medical VG

Recent efforts in medical visual grounding focus on linking textual findings to corresponding regions in radiological images. Several approaches have demonstrated the potential of using cross-modal models for localizing disease-related content [10, 19, 107]. For instance, Boecking et al. [10] propose a zero-shot method that maps textual findings to RoIs in CXRs. Similarly, Chen et al. [19] highlight challenges in locating small or overlapping findings, particularly where visual and textual cues may conflict. Despite their progress, these models primarily serve as localization tools and often lack mechanisms to validate their outputs without expert feedback. This limits their reliability in practical clinical settings. Additionally, their generalizability is constrained due to limited medical training data compared to natural image datasets. Addressing these limitations, we introduce an automated reliability evaluation score that reduces dependence on expert validation by assessing the semantic alignment between the generated report and the image.

Alternatively, Grounded reporting methods [46, 103, 78, 126] extend visual grounding by combining localization with automated report generation. These methods aim to improve the accuracy and clinical relevance of generated reports by first detecting regions of interest and then generating textual descriptions. For example, Tanida et al. [103] demonstrate that grounded approaches outperform direct image-to-text generation in identifying multiple anomalies. While grounded reporting improves report precision, the generated reports still face challenges in terms of interpretability and validation. Moreover, these methods primarily focus on generating new reports rather than enhancing existing ones.

In contrast to previous approaches, our work leverages medical visual grounding to localize findings described in a given CXR report, while focusing on enhancing the interpretability and trustworthiness of existing AI-generated reports. This method facilitates visual validation and bridges the gap between textual and visual modalities. Furthermore, we enhance the interpretability by designing a new dual-scoring system that transforms trust from a binary agree/disagree metric into a continuous, quantifiable measure of performance, thereby enabling greater confidence in the model's outputs.

Unlike models trained predominantly on natural image datasets [71], our architecture is fine-tuned using medical data, ensuring it is specifically adapted to the unique complexities of CXR images, such as subtle anatomical structures and overlapping pathologies. To assess model performance without expert feedback, we integrate a validation pipeline that utilizes annotated datasets and standardized metrics. This provides a robust and objective evaluation framework, ensuring both reliability and clinical relevance in the assessment process.

2.4.2 Adapting Grounding DINO to Medical Contexts

Following our comparative analysis of visual grounding models in natural image domains, we identified Grounding DINO as the most suitable candidate for our project, due to its ability to perform open-vocabulary visual grounding, an essential requirement for aligning medical language with image regions. While Grounding DINO is highly effective for natural images, its direct application to medical imaging is challenging due to the unique demands of the domain. Adapting Grounding DINO model necessitates fine-tuning with annotated medical data, including bounding boxes for pathologies and detailed text descriptions. Figure 2.4 presents a chest X-ray image grounded to the phrase "Cardiomegaly with mild pulmonary vascular congestion" using the Grounding DINO model trained on natural data. Despite the simplicity of the medical prompt and the accessibility of the text features for even non-experts, the model incorrectly identifies the entire image as the object. This highlights the challenge of capturing subtle pathological features in medical imaging, underscoring the need for domain-specific adaptations.

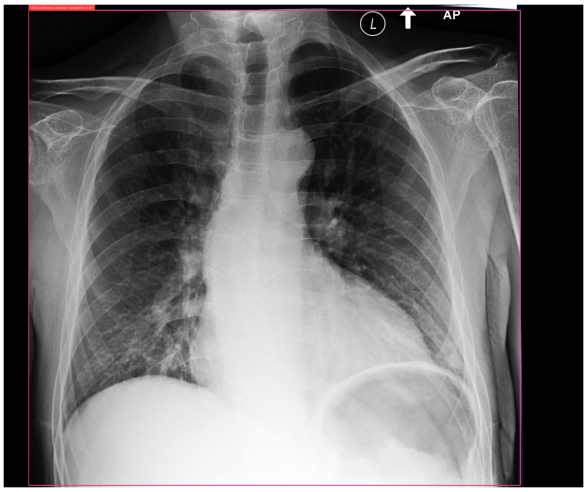


Figure 2.4: An example of visual grounding using the Grounding DINO model trained exclusively on natural data with the phrase "Cardiomegaly with mild pulmonary vascular congestion". The model inaccurately localizes the entire image as the detected object.

We fine-tuned Grounding DINO on CXR datasets to overcome these challenges, customizing it for medical applications. For this process, we used two key datasets:

1. MS-CXR Dataset [10] A specialized dataset designed for visual grounding tasks, as illustrated in Figure 2.5. It contains 1,162 annotated CXR images, each paired with detailed bounding box annotations and corresponding pathology descriptions, making it a valuable resource for developing and evaluating visual grounding models. However, the dataset's limited size poses a challenge, as it is insufficient for training a conventional object detection model. This necessitates strategies such as few-shot learning or the incorporation of additional datasets to ensure robust model performance.

2. VinDr-CXR Dataset [81] A larger dataset with over 18,000 subjects' annotated images,

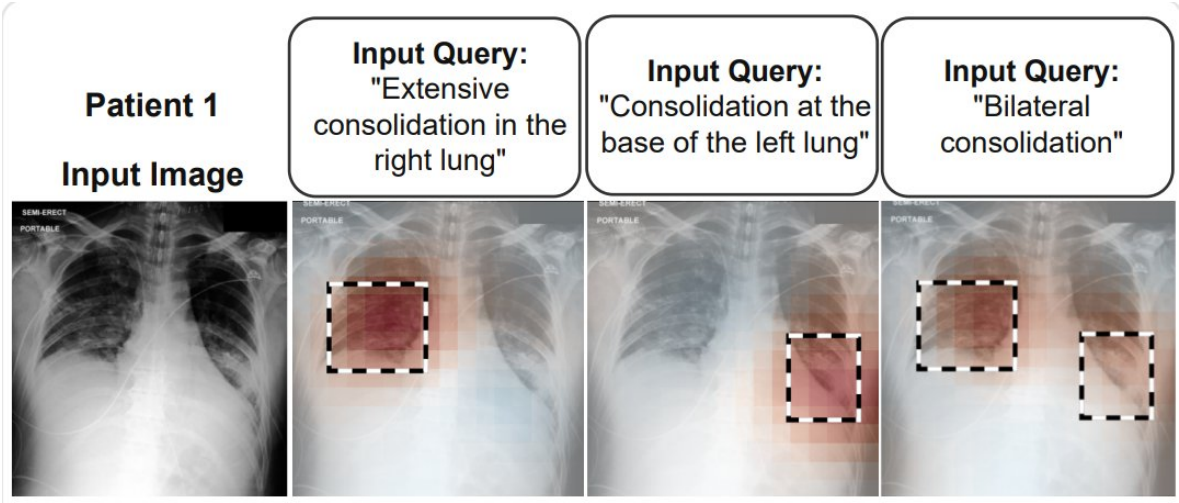


Figure 2.5: A well-annotated dataset for visual grounding, featuring bounding box locations, associated text queries, and corresponding pathology descriptions [10].

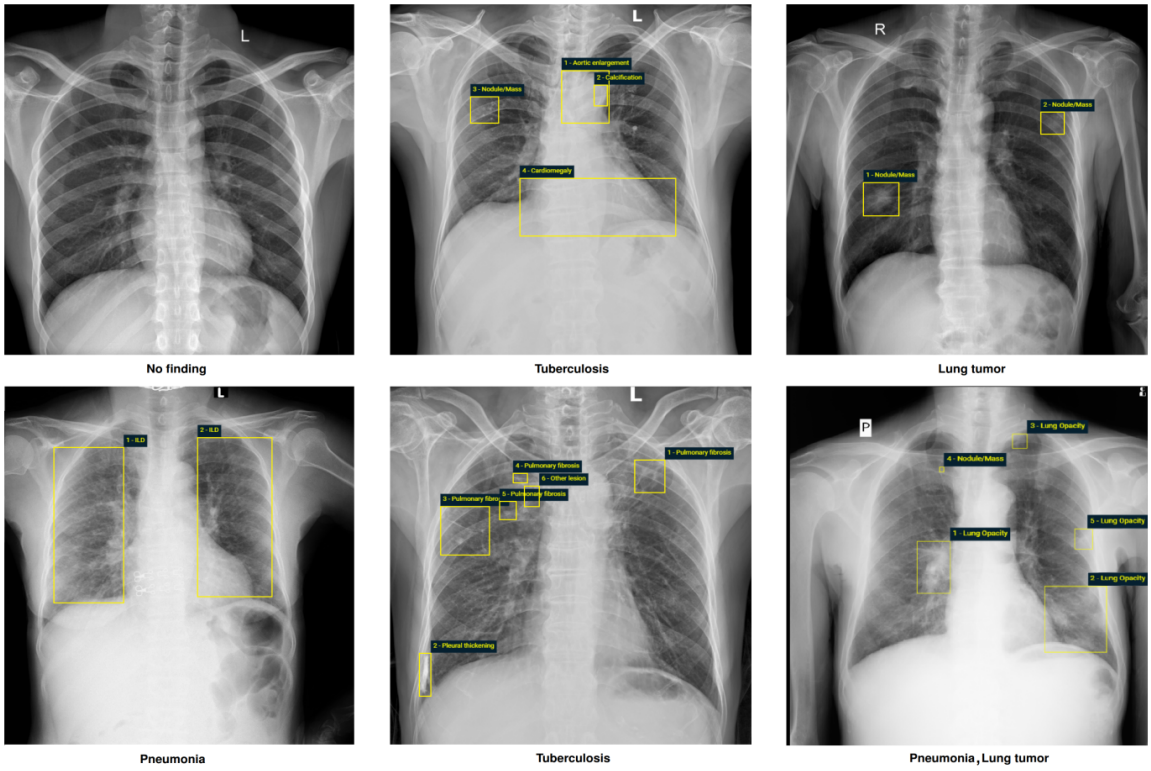
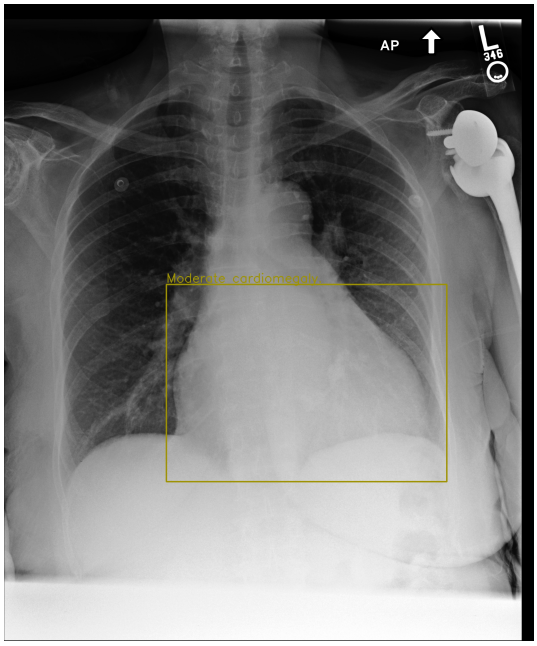


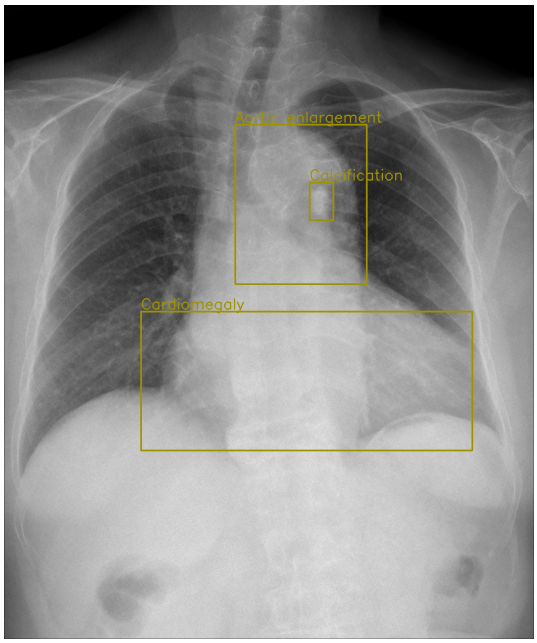
Figure 2.6: Examples of chest X-rays with expert radiologist annotations in VinDr-CXR dataset. Local findings are highlighted with bounding boxes overlaid on the original images. Global diagnostic labels are shown below each example [81].

providing bounding box locations for 22 common thoracic diseases, as shown in Figure 2.6. However, VinDR-CXR lacks detailed written reports associated with each subject, which is

critical for our visual grounding purposes.



Example from the MS-CXR dataset. Images are heterogeneous, and annotations primarily cover anatomical regions derived from MIMIC-CXR.



Example from the VinDr-CXR dataset. Images are consistent, and annotations include bounding boxes for 22 thoracic pathologies, but lack textual reports or anatomical labels.

Figure 2.7: Comparison between MS-CXR (based on MIMIC-CXR) and VinDr-CXR datasets.

Figure 2.7 presents a sample comparison between the MS-CXR and VinDr-CXR datasets. The VinDr-CXR images are relatively clean and consistent in quality, as they are collected from hospitals in Vietnam. In contrast, MS-CXR, derived from MIMIC-CXR, is more heterogeneous, with greater variation in contrast, resolution, and acquisition settings. From an annotation perspective, MS-CXR provides rich anatomical region annotations (e.g., lung zones, atrium) but only limited coverage of pathologies, whereas VinDr-CXR focuses on detailed bounding boxes for 22 common thoracic pathologies, without corresponding textual reports or anatomical descriptors.

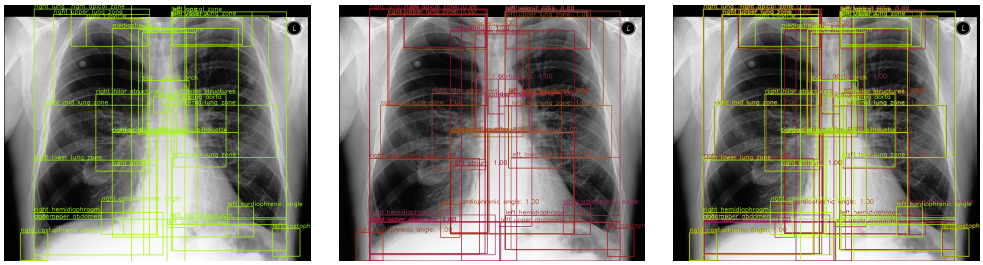
One potential solution to this issue is the use of automated report generation models to create descriptive reports. However, as one of our objectives is to enhance the interpretability of such models, we opted not to include them during training to avoid introducing potential biases. Instead, we enhanced the dataset by associating pathologies with their anatomical regions to enrich the spatial information.

2.4.3 Data Protocol for Establishing Phrase-Region Pairs

To enrich VinDr-CXR with fine-grained anatomical context, we developed a protocol that automatically associates pathology labels with anatomical regions. This enables the creation of phrase-region pairs that are essential for supervised visual grounding. The protocol consists of five main steps.

2.4.3.1 Automated Anatomical Detection

We first leveraged the Chest ImaGenome dataset [115], which provides detailed scene graphs for chest X-rays, linking 36 anatomical regions to bounding boxes and attributes. ImaGenome also includes over 1,200 relation types and 670,000 localized comparisons across sequential exams (e.g., improved, worsened, or stable), and a manually curated gold-standard scene graph set for 500 patients, making it a high-quality source of spatial annotations.



Ground-truth annotation from Chest ImaGenome

Automatic anatomical detection (DETR, trained on ImaGenome)

Overlap between ground truth and DETR predictions

Figure 2.8: Comparison between ground-truth annotations and automatic anatomical detection. Left: manual annotations from the Chest ImaGenome dataset. Middle: bounding boxes predicted by our trained DETR model. Right: overlap of ground truth and predictions.

On this dataset, we trained the DETR (**Detection Transformer**) model [11] from scratch to automatically detect anatomical regions in chest X-rays. DETR is a state-of-the-art object detection framework with an end-to-end transformer-based architecture that removes the need for post-processing heuristics such as non-maximum suppression. This makes it particularly effective in medical imaging scenarios where anatomical structures are often overlapping or closely spaced. Its global attention mechanisms further enable the robust identification of subtle and partially overlapping regions. Figure 2.8 illustrates a representative example: red bounding boxes correspond to anatomical regions detected by our trained DETR model, while green boxes denote the ground-truth annotations from the Chest ImaGenome dataset. As shown, the detections closely overlap with the annotations, confirming the model's ability to capture anatomical structures accurately. The model was further validated on the MIMIC-CXR test set.

Training converges rapidly, reaching a DETR loss of 3.50 and a generalized **IoU loss** of 0.29 after only 14 epochs. On the test set, it achieves a mean IoU (mIoU) of 76%. By

comparison, a DETR model pretrained on COCO and fine-tuned on the same task achieves only 62% mIoU. This performance gap underscores the importance of domain-specific training: unlike natural images, anatomical regions in CXRs appear in relatively consistent positions across patients, which enables a model trained from scratch on medical data to learn these spatial regularities more effectively. Together, these results demonstrate both the effectiveness and efficiency of our approach for anatomical region detection in CXRs.

2.4.3.2 Application to VinDr-CXR

We applied the trained DETR model to the VinDr-CXR dataset [81], obtaining anatomical bounding boxes for each image (e.g., "right upper lung zone," "left atrium"). These predictions enriched VinDr, which originally provided only pathology bounding boxes, by adding the missing anatomical context necessary for phrase–region alignment. Figure 2.9 illustrates this process: the pathology bounding box ("Lung Opacity") is shown in blue, while the anatomical regions automatically detected by our trained DETR model are shown in red. This augmentation provides a critical spatial link between pathologies and their anatomical locations, which was not originally available in the VinDr-CXR dataset.

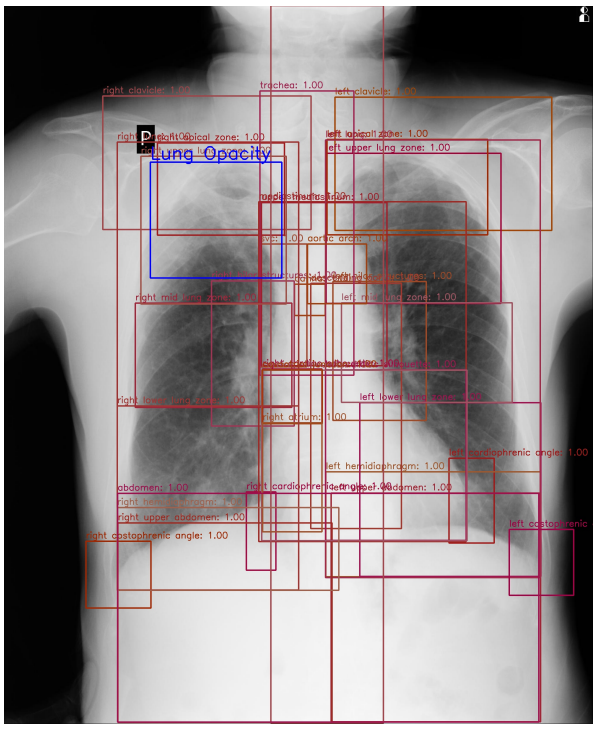


Figure 2.9: An example CXR image from the VinDr-CXR dataset. The blue bounding box represents the dataset-provided annotation for the pathology present in the image. The red bounding boxes indicate the anatomical regions automatically detected by our trained model.

2.4.3.3 Pathology-Anatomy Alignment

Since VinDr-CXR provides pathology labels and bounding boxes but no anatomical descriptors, we associated each pathology box with the closest anatomical region box predicted by DETR. Matching was performed using empirical distance and intersection-over-union (IoU) metrics. This step allowed us to anchor each pathology to a specific anatomical location.

2.4.3.4 Phrase-Region Pair Generation

We then generated enriched textual prompts by combining pathology labels with anatomical regions. For example, a VinDr pathology box labeled "Lung Opacity" that overlaps most with the "right upper lung zone" was relabeled as:

"Lung opacity in right upper lung zone".

This step transforms generic pathology annotations into clinically meaningful phrase-region pairs, significantly improving dataset granularity and suitability for grounding tasks.

2.4.3.5 Refinements and Validation

To ensure reliability, we manually verified 200 enriched samples against radiology descriptors [1]. Based on this analysis, we refined the protocol to handle edge cases. In particular, for bilateral pathologies, we duplicated associations to both left and right anatomical regions to avoid information loss.

Overall, this approach provides a frugal yet effective solution: by reusing ImaGenome annotations and leveraging DETR, we enrich VinDr-CXR without requiring costly new labels. The resulting dataset supports supervised visual grounding with phrase-region pairs, offering explicit spatial supervision rather than global image-text similarity alone. The fine-tuning process involved few-shot training, using a subset of annotated images to adapt the model's weights to the medical domain while preserving its pretrained capabilities for natural data. This step is critical to balance generalization and domain-specific accuracy.

Limitations of Phrase–Region Enrichment. While the enrichment protocol provides structured phrase–region pairs that are effective for training supervised grounding, this approach diverges from the nature of full radiology reports. Real-world reports are written in free text and often include multiple findings per sentence, contextual modifiers, and frequent use of negation (e.g., "no pleural effusion"). By contrast, our structured phrases (e.g., "left pleural effusion") impose a simplified and compositional form that may not fully capture the linguistic complexity of radiology reporting.

[1] www.radiologymasterclass.co.uk

Algorithm 1: Data Protocol for Enriching VinDr-CXR with Anatomical Phrase–Region Pairs

Input:

- VinDr-CXR: CXR images with pathology bounding boxes and labels (22 diseases), but no anatomical descriptors.
- Chest ImaGenome: CXR images with bounding boxes for 36 anatomical regions and a radiologist-constructed ontology.

Output: Extended VinDr-CXR annotations containing (*pathology + anatomical phrase*, *pathology bbox*, *anatomical label*).

Step 1: Train Anatomical Region Detector (DETR).
Train a Detection Transformer (DETR) from scratch on Chest ImaGenome to detect 36 anatomical regions.
Rationale: create a robust anatomical detector independent of pathology labels.

Step 2: Predict Anatomy on VinDr-CXR.
Apply the trained DETR to each VinDr-CXR image, producing anatomical region bounding boxes (e.g., *right upper lung zone*, *left atrium*).
Rationale: enrich VinDr with anatomical structure absent from the original dataset.

Step 3: Align Pathology with Anatomy.
For each pathology box B_p (from VinDr) and anatomical box B_a (from DETR), compute Intersection-over-Union (IoU) and center distance. Assign B_p to the best-matching B_a using IoU as primary criterion and distance as tie-breaker.
Edge cases: if B_p overlaps multiple anatomical regions (e.g., bilateral findings), duplicate the association to preserve both sides.

Step 4: Generate Phrase–Region Pairs.
Compose enriched textual phrases by combining pathology and anatomical region, e.g.:

```
"Lung opacity in right upper lung zone"
```

Store triplets: (*phrase*, *pathology bbox* B_p, *anatomical label*).
Rationale: enable supervised visual grounding with explicit phrase-to-region alignment.

Step 5: Quality Refinement and Validation.
Manually verify a subset of samples (e.g., 200 images) against radiology references. Refine thresholds for IoU matching and adjust handling of bilateral/ambiguous cases.

Output. The enriched VinDr-CXR dataset now provides phrase–region pairs, supporting visual grounding tasks with explicit bounding boxes. This goes beyond global image–text similarity and contributes a key resource for interpretable medical AI.

This simplification introduces a trade-off: it improves the alignment between phrases and spatial regions, but reduces linguistic variability. Consequently, the trained model may be less robust when exposed to longer, composite, or negated statements. To mitigate this, we comple-

ment VinDr-CXR with MS-CXR, which offers short phrase-level annotations closer in spirit to our enriched pairs, while acknowledging that both remain relatively constrained compared to full free-text reports. This limitation is important for understanding the scope of our future pipeline (VICCA): while visual grounding achieves reliable localization for well-structured prompts, extending it to free-text report interpretation requires additional mechanisms for handling complex expressions, negations, and broader semantic contexts.

2.4.4 Visual Grounding Model Architecture

As shown in Figure 2.10, we fine-tuned the Grounding DINO's architecture in the following components:

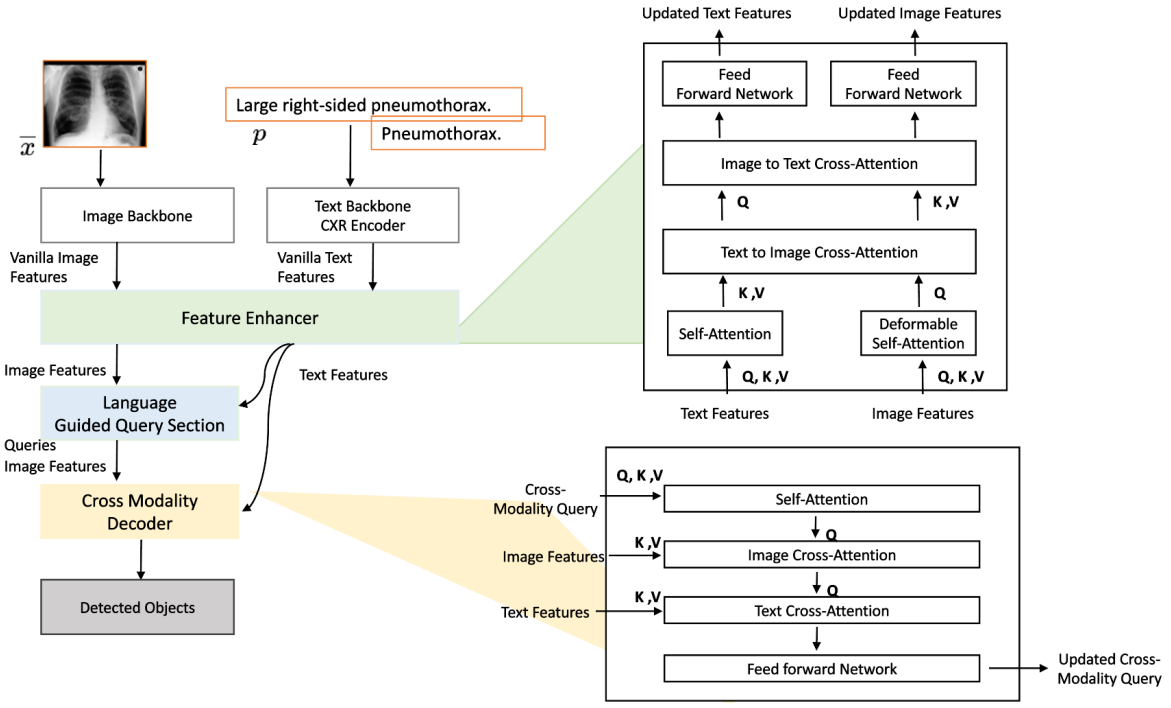


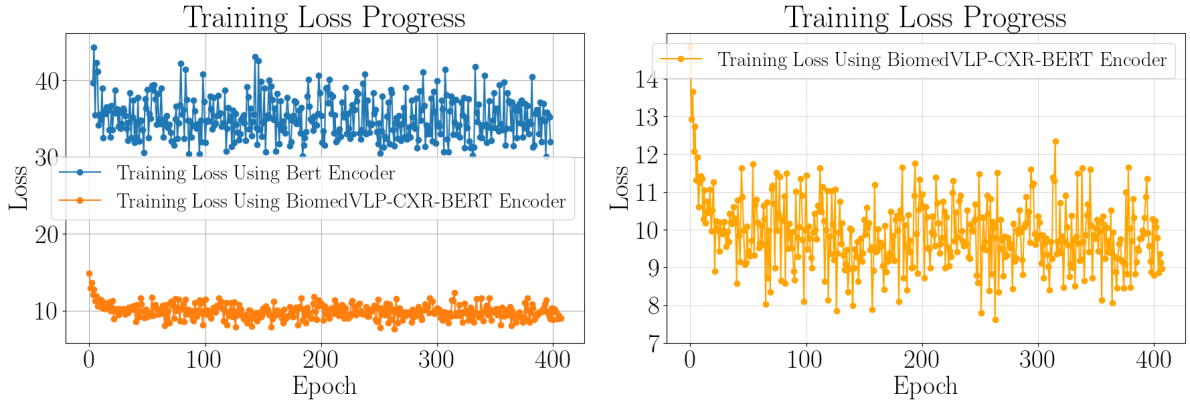
Figure 2.10: The architecture of the Grounding DINO model integrates the BiomedVLP-CXR-BERT as the text encoder for medical text attention. This model grounds the provided text input onto the image using cross-attention maps between the image and text.

Language Encoder tokenizes and embeds textual prompts for semantic representation. Grounding DINO uses BERT encoder [28], a text encoder pretrained on the English language using a masked language modeling (MLM) objective, empirically proven powerful for natural language tasks, equipped with a large token vocabulary. However, numerous tokens in this vocabulary are not relevant to the medical domain. Despite efforts to integrate models such as ClinicalBERT [44], specifically trained for medical applications, into Grounding DINO's training process, this did not result in notable enhancements in feature extraction. To overcome these limitations, we integrated BiomedVLP-CXR-BERT [10], a text encoder specifically

trained on chest X-ray reports. By leveraging BiomedVLP-CXR-BERT in the fine-tuning process, we aligned the text encoder's outputs more closely with the medical context, significantly enhancing localization outcomes.

By integrating the BiomedVLP-CXR-BERT encoder into the Grounding DINO model, we effectively transformed the previously natural language outputs into medically contextualized ones. This adjustment resulted in a substantial performance enhancement. Specifically, the general DETR-like loss during training decreased from 35.85 when using the BERT encoder to 8.79 when using the BiomedVLP-CXR-BERT encoder. Figure 2.11a compares the training loss progression between the two encoders, showing that the specialized encoder leads to both lower loss and faster convergence. The model using the BERT encoder shows significant fluctuations in training loss, ranging from 30.08 to 44.29. This high variation is caused from unstable feature matching in the visual grounding model, indicating that the BERT encoder struggles to extract clinically relevant features from the text. As a result, it fails to consistently align textual and visual representations, which disrupts the learning process.

To ensure that overfitting is not occurring, Figure 2.11b provides a zoomed-out view of the training curve with BiomedVLP-CXR-BERT. Despite fluctuations, the loss remains within a stable range [7.6, 11.8], which is consistent with the typical behavior of Grounding DINO models [32], before converging to 8.79 by epoch 407. This pattern suggests normal variability during training rather than evidence of overfitting. Validation loss curves are omitted, as the objective of Figure 2.11 is to contrast the training dynamics of the different encoders rather than to evaluate generalization performance directly. Model generalization is more appropriately assessed through the metrics reported in section 4.4 and chapter 7.



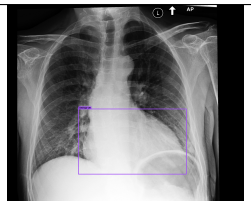
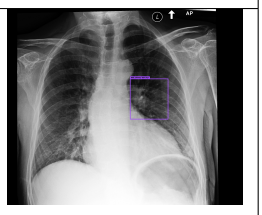
(a) (b)

Figure 2.11: (a) Training loss progression for the visual grounding model using two text encoders (BERT and BiomedVLP-CXR-BERT). The specialized BiomedVLP-CXR-BERT encoder (orange curve), designed for CXR text, significantly reduces the loss and achieves faster convergence compared to using the BERT encoder (blue curve). (b) A zoomed-out view of the loss curve for the BiomedVLP-CXR-BERT encoder shows a stable yet non-monotonic pattern, suggesting variability in learning but no clear signs of overfitting.

Image encoder processes input images to extract relevant visual features. Grounding DINO supports both ResNet [40] and Swin Transformer [74] architectures as backbones. For our fine-tuning, we integrated only the Swin Transformer due to its superior performance on high-resolution images and its patch-merging strategy, which effectively aligns text and image features both globally and locally. This choice proved particularly advantageous for the intricate spatial details present in chest X-ray analysis.

Loss Function: We adopted the original loss configuration of Grounding DINO, which includes L1 loss and Generalized Intersection over Union (GIoU) for bounding box regression [124]. The model also employs a contrastive loss to associate predicted objects with their corresponding classifications. This process involves computing the dot product between each query and text feature, generating logits for each text token, and applying focal loss [67] to these logits. For fine-tuning, a multi-task loss function combining localization loss (L1 and GIoU) and classification loss (focal loss) was utilized. Training was conducted over 400 epochs, with hyperparameters such as learning rate and batch size optimized via grid search. The final model exhibited a stable loss convergence with the loss stabilizing around 8.7, ensuring both precise localization and robust classification capabilities.

Table 2.1: The results of the visual grounding model using two different text prompts. The ground truth report, “Cardiomegaly with mild pulmonary vascular congestion," got the bounding box detection with 76% accuracy. The last column shows the results with a random prompt, “Left-sided Pulmonary Edema," which achieved a bounding box accuracy of 20%.

Input	Cardiomegaly with mild pulmonary vascular congestion.	Left-sided Pulmonary Edema.
Output		

Success in the visual grounding task provides the first score in our dual-scoring mechanism, offering a detection accuracy metric for each bounding box. Table 2.1 presents a case study using the optimized Grounding DINO model with two different textual inputs. The CXR image was annotated with the accurate diagnosis of “Cardiomegaly with mild pulmonary vascular congestion," achieving a bounding box detection accuracy of 76%. Conversely when exposed to an unrelated prompt, “Left-sided Pulmonary Edema," the accuracy dropped significantly to 20%, localizing an anomaly in the left lung but failing to identify edema. This result highlights the model’s capability to differentiate and accurately localize the correct pathology when provided with a relevant text prompt, while demonstrating reduced performance for unrelated or

incorrect prompts.

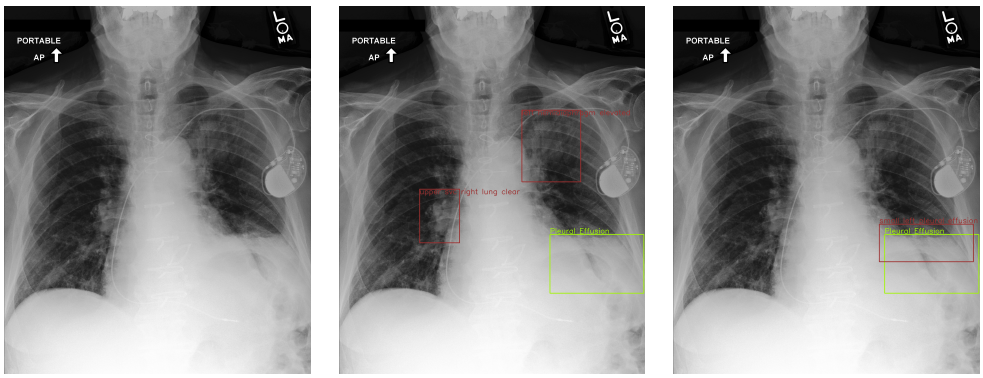
2.4.5 Training Protocols and Data Augmentation Assessment

Table 2.2 summarizes the datasets used for training and validation, including the number of images and annotations. Only chest X-rays in the Posterior-Anterior (PA) or Anterior-Posterior (AP) views are included, while lateral views are excluded. In addition, VinDr-CXR images labeled as "No FINDING" are removed from training to ensure pathology-relevant supervision.

Table 2.2: Dataset distribution used for training and validation across the models, including the number of images and paired annotations.

visual Grounding Model			
Dataset	Train	Validation	Test
MS-CXR [10]	817	169	176
VinDr-CXR [81]	34,367	3,000	2,697
Automatic Lung Segmentation			
Chest ImaGenome [115]	166,521	23,593	47,393

To assess the impact of our data augmentation protocol, we conducted an ablation study. Specifically, we compared the visual grounding model trained (i) only on the MS-CXR dataset and (ii) on MS-CXR combined with VinDr-CXR enriched by our automatic anatomical detection model (trained DETR).



Ground-truth MS-CXR image

VG Model trained only on MS-CXR

VG Model trained on MS-CXR + enriched VinDr-CXR

Figure 2.12: Effect of dataset augmentation on visual grounding. Adding VinDr-CXR enriched with anatomical regions improves the model's ability to correctly localize the critical pathology (*Pleural Effusion*) instead of being distracted by irrelevant findings.

Figure 2.12 illustrates a representative case from the MS-CXR test set. The original report is long and contains many irrelevant findings, e.g.:

"AP chest reviewed in the absence of prior chest radiographs: Right PICC line ends in the upper SVC. Transvenous pacer lead projects over the right ventricular apex. Right lung clear. 5 cm left suprahilar mass. Left hemidiaphragm elevated. Small left pleural effusion present. Heart size normal. House office and I discussed the findings by telephone at 9:40 a.m. as soon as the findings were recognized."

The ground-truth category is **Pleural Effusion** (shown in green). When trained only on MS-CXR, the model fails to capture the critical finding, grounding instead on irrelevant phrases such as *"upper SVC"* or *"left hemidiaphragm elevated"*. By contrast, when trained with our augmented dataset (MS-CXR VinDr-CXR enriched with anatomical regions), the model correctly grounds the pathology and aligns the textual description with the corresponding image abnormal region (shown in red).

2.4.6 Results

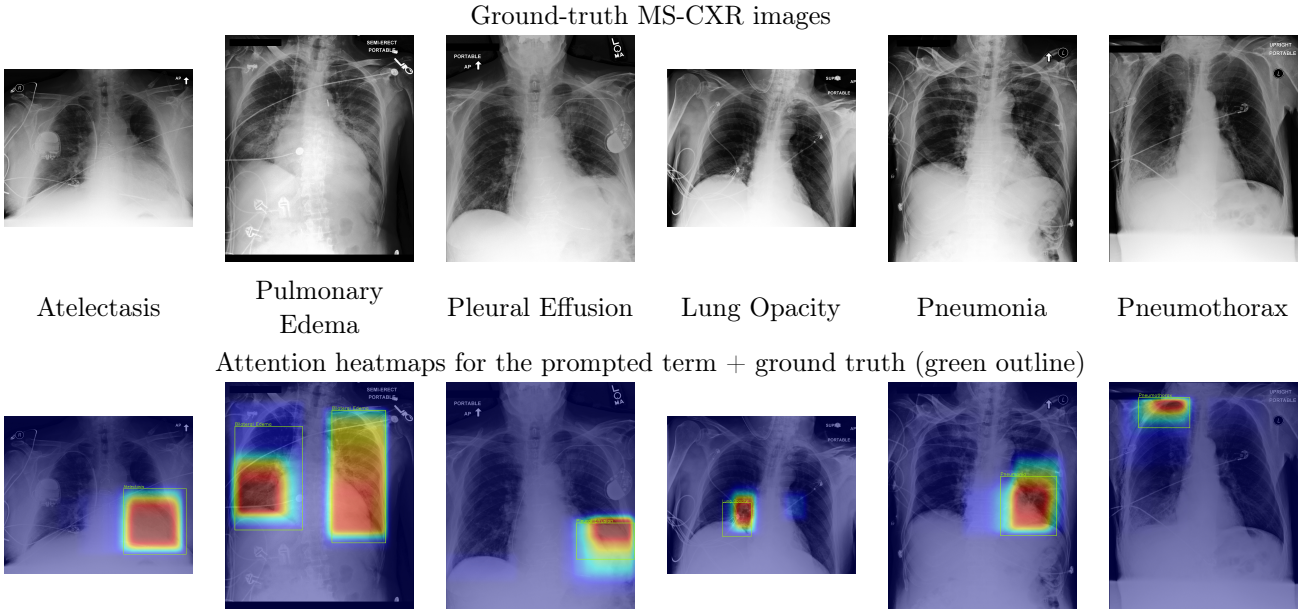


Figure 2.13: Qualitative grounding results on MS-CXR. Top: input images from test set. Bottom: token-conditioned attention heatmaps (for concise, anatomy-aware prompts) +the ground truth (green boxes).

To qualitatively assess the Visual Grounding model, we visualize token–region cross-attention as a heatmap for six common thoracic findings: *Atelectasis*, *Pulmonary Edema*, *Pleural Effusion*, *Lung Opacity*, *Pneumonia*, and *Pneumothorax*. Figure 2.13 shows, for each case, the MS-CXR test image (top row) with its ground-truth annotation (green box, bottom row) and the corresponding attention heatmap produced by our model (bottom row).

For clarity, we employ concise, anatomy-aware prompts that directly correspond to each label (e.g., "*left pleural effusion*", "*bilateral pulmonary edema*"). The resulting heatmaps in-

dicate the image regions most attended by the model for the prompted term. In the case of *bilateral pulmonary edema*, where the pathology involves both lungs, the model appropriately distributes attention across both fields. For more localized findings (e.g., *pleural effusion* or *pneumothorax*), the peak attention is concentrated on the affected side. Across several examples, the attention maxima align closely with the ground-truth annotations, suggesting that the model can reliably associate clinical phrases with their corresponding anatomical locations.

2.4.7 Evaluation and Analysis

The standard approach for evaluating a visual grounding model involves mean Average Precision (mAP) for token understanding and mean Intersection over Union (mIoU) for object detection. To this end, we evaluate the model using the test sets of two datasets, with detailed information provided in Table 2.2. For the VinDr-CXR dataset, captions are generated based on the pathology and the closest associated anatomical region. The mAP is calculated using the probability scores for each detection, and the results are summarized in Table 2.3.

The results for the VinDr-CXR dataset, presented in Table 2.3, reflect the performance of benchmark models on the object detection task. Since our method of generating captions based on anatomical regions and pathologies is a novel contribution, direct comparisons with other models for the visual grounding task are not possible. Nevertheless, our approach outperforms the benchmark models on the test set for object detection in this dataset. For the MS-CXR dataset, the results pertain to the visual grounding task. Our model achieves comparable performance to the benchmarks in terms of mAP and surpasses them in mIoU.

Table 2.3: Visual Grounding model performance against VinDR-CXR and MS-CXR datasets.

VinDr-CXR		
Model	mAP ↑	mIoU ↑
ChEX [78]	14.12 ± 0.95	-
BioVIL [10]	2.82 ± 0.25	-
VICCA (Ours)	**34.36 ± 1.79**	**38.32 ± 2.7**
MS-CXR		
ChEX [78]	**44.47 ± 2.21**	47.52 ± 1.45
TransVG [27]	44.05 ± 2.63	53.51 ± 1.53
BioVIL [10]	18.62 ± 1.37	28.57 ± 1.31
VICCA (Ours)	41.67 ± 0.69	**55.27 ± 2.36**

The results highlight a dataset-dependent correlation. On VinDr-CXR, where pathology labels lack phrase-level supervision and must be aligned with automatically inferred anatomical regions, our model significantly outperforms baselines in both detection accuracy and localization (mAP and mIoU). This demonstrates the strength of our enrichment protocol in scenarios with weak or noisy supervision.

By contrast, on MS-CXR, where phrase-level annotations are explicitly provided, the performance gap narrows. While our model achieves the best mIoU, it performs slightly below

some baselines in raw detection mAP. This suggests that when richer textual supervision is available, models trained directly on phrase-level annotations may benefit more from the explicit alignment, reducing the relative advantage of our approach.

Overall, these results indicate that our model is particularly effective in low-supervision settings (VinDr), where bridging the gap between pathology and anatomy is critical, while remaining competitive in settings with stronger phrase-level supervision (MS-CXR). This supports our hypothesis that limited diversity in training text supervision constrains generalization, and highlights the value of our approach for datasets where phrase-to-region alignment is less explicit.

While our enhancements improve the visual interpretability of the generated text reports, the outputs still require validation by radiologists to ensure diagnostic reliability and completeness. To mitigate this dependency and move toward automated assessment, we propose an auxiliary model designed to evaluate the accuracy and reliability of the predicted bounding box localizations.

To quantify the semantic alignment between the text prompt and the image content, ensuring that the text accurately describes the image and the image visually supports the textual description, we introduce a reliability score. This score serves as an objective metric to assess the trustworthiness of the model's outputs.

In the following chapter, we delve into this approach by first introducing diffusion models for image generation and then presenting our auxiliary model, which generates CXR images from textual prompts while preserving anatomical structures. This enables a direct visual comparison between the synthesized anomalies, guided by textual descriptions, and the original pathology, providing a novel strategy for evaluating alignment and reliability in medical image-to-report generation.

Chapter 3

Towards Reliable Synthetic Chest X-rays: Guided Generation Using Diffusion

---

Contents

This chapter explores diffusion-based methods for image generation and examines the specific challenges of applying them to chest X-rays.

Problem: The central problem is to generate realistic synthetic CXR images from textual descriptions while preserving anatomical fidelity and pathological relevance.

Objective: The objective is to adapt these models to the medical domain with spatial control, by conditioning generation on a binary lung mask that guides synthesis within the parenchyma while distinguishing it from peripheral regions.

Within the pipeline, the role of this chapter is twofold. First, the diffusion generator renders textual descriptions into visual evidence under mask constraints. Second, and crucially, the diffusion model itself serves as the auxiliary validator, by synthesizing an image from the report (and mask) and measuring its visual alignment with the original *Chest X-ray* via the visual grounding mechanism introduced in Chapter 2, we derive a reliability score on the

similarity of generation with the original that provides a complementary perspective on the trustworthiness of model predictions.

3.1 Introduction to Diffusion Models

Diffusion models [113] are a recent and powerful class of generative algorithms designed to synthesize data, such as images or audio, by learning from a large collection of training examples. The objective of generative modeling is to produce diverse outputs that capture the underlying distribution of the training data without simply memorizing or replicating it. At their core, diffusion models operate by gradually corrupting input data through the iterative addition of Gaussian noise, effectively destroying the structure of the original data over a series of steps. During training, the model learns to reverse this degradation process: starting from pure noise, it progressively denoises the input to recover realistic samples. Once trained, the model can generate entirely new samples by initiating this reverse process from random noise (Figure 3.1).

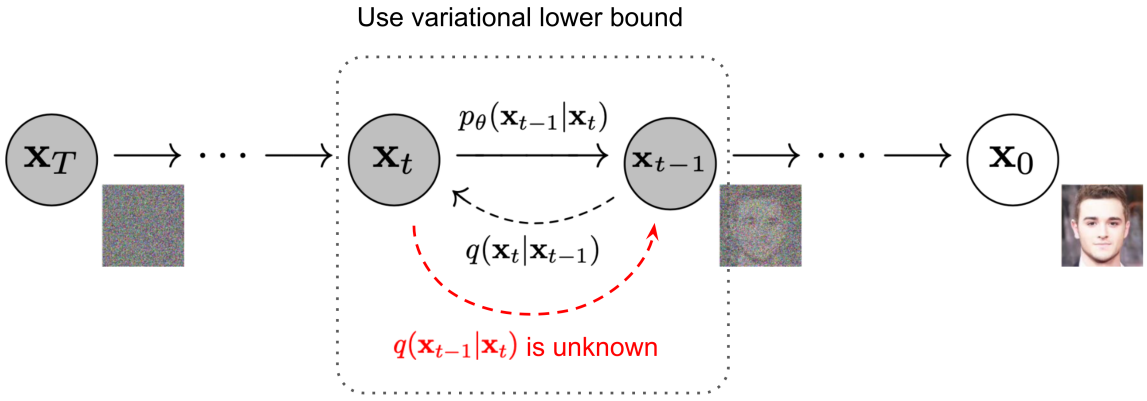


Figure 3.1: The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. (Image source: Ho et al. [41])

Inspired by non-equilibrium thermodynamics, diffusion models implement a Markov chain framework that simulates a forward noising process and a learned reverse denoising process. Unlike other generative approaches such as Variational Autoencoders (VAEs) [56] or normalizing flows [94], diffusion models do not require an explicit likelihood parameterization or an invertible architecture. Instead, they gradually construct samples through iterative denoising steps. Recent advancements, such as SDXL [88], further improve latent diffusion models by enhancing architecture depth, conditioning mechanisms, and high-resolution synthesis capabilities, without relying on normalizing-flow formulations. Instead, they use a fixed, stochastic forward process and learn only the reverse trajectory, operating directly in the high-dimensional data space. This property contributes to their ability to generate high-fidelity and diverse outputs, making them particularly suitable for tasks requiring fine-grained detail.

Figure 3.2 offers a comparative visual summary of four major classes of generative models,

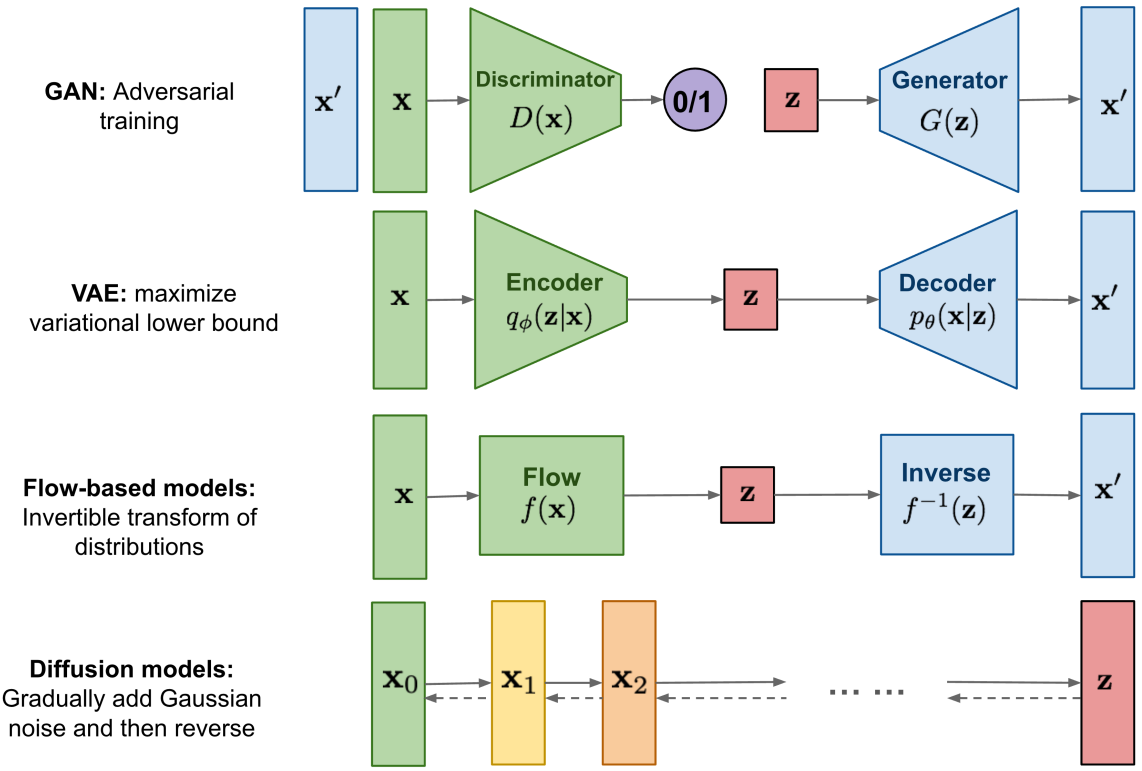


Figure 3.2: Overview of different types of generative models. Image source [114]

GANs [36], VAEs, flow-based models, and diffusion models, highlighting their core architectures and learning principles. In GANs, generation is achieved through adversarial training, where a generator learns to produce realistic samples that a discriminator cannot distinguish from true data. VAEs, in contrast, rely on variational inference to encode data into a latent space and then decode it back into the original domain. Flow-based models apply invertible transformations between data and latent distributions, enabling exact likelihood estimation. Diffusion models, shown at the bottom, take a fundamentally different approach: they gradually corrupt the data with Gaussian noise through a forward process and learn to reverse this process to generate new data samples. Unlike the other architectures, diffusion models do not require an explicit latent representation or adversarial loss, instead relying on a learned denoising trajectory in the data space. This unique process contributes to their robustness and ability to generate high-quality outputs, making them particularly well-suited for applications in medical imaging where fine-grained detail and reliability are crucial.

3.2 Mathematical Foundations of Diffusion Models

Diffusion models are built upon a forward noising process and a reverse denoising process, both governed by stochastic differential equations. The forward process incrementally adds

Gaussian noise to a data sample x_0 over T discrete time steps, producing a sequence of latent variables $x_1, x_2, \ldots, x_T$:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}), \tag{3.1}$$

where $\beta_t \in (0, 1)$ controls the noise variance at time step t. The entire forward process can be compactly written as:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{3.2}$$

with $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. This means that we can directly sample x_t from x_0 without iterating through each step, which is useful for training.

The reverse process attempts to learn the denoising trajectory from noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ back to a data sample x_0, parameterized by a neural network:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{3.3}$$

In practice, many implementations fix the variance Σ_θ and train the model to predict the noise ϵ added at each step using a simplified loss:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right], \tag{3.4}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$. This objective encourages the model to accurately denoise the corrupted image at each time step.

Generation. At inference time, we sample $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively apply the learned reverse process $p_\theta(x_{t-1} \mid x_t)$ to obtain a synthetic image x_0.

Key Strengths. Diffusion models are especially effective at generating high-resolution, detail-preserving images because they avoid the mode-collapse issues common in GANs and provide a more stable training objective grounded in maximum likelihood estimation.

In the context of medical imaging, particularly chest X-rays, these properties are advantageous due to the subtlety of radiological features. Preserving anatomical structure and pathological integrity is essential for clinical utility, motivating the use of diffusion models in our framework.

3.3 Conditional Generation

While standard diffusion models generate samples by reversing a noise process in an unconditional manner, real-world tasks often require outputs to be conditioned on external information. This leads to the framework of *conditional diffusion models*, where generation is guided by a conditioning input, such as class labels, text, or images, enabling the production of outputs that conform to specific constraints or semantics.

In conditional diffusion, the reverse process becomes:

$$p_\theta(x_{t-1} \mid x_t, c), \tag{3.5}$$

where c represents the conditioning signal. This conditioning can be incorporated into the model in various ways, such as:

- Concatenating c to the input x_t.
- Modifying the architecture (e.g., cross-attention layers or adaptive normalization).
- Using classifier guidance, where a separately trained model guides the generation toward desired attributes.

This conditional setup enables more controllable and task-specific generation, and it has been successfully used in domains like class-conditional image synthesis, image inpainting, and text-to-image generation.

Relevance to Medical Imaging. In our work, we adopt a conditional diffusion model to generate synthetic chest X-ray (CXR) images conditioned on two inputs:

1. A radiology report or textual description of the pathology.
2. A binary anatomical mask derived from the original CXR image.

This setup ensures that the synthesized image not only reflects the pathology described in the text but also maintains the anatomical structure of the original image. The anatomical mask acts as a spatial prior that constrains the generation process, allowing the model to respect the underlying geometry of the lungs, heart, and surrounding tissues.

Motivation. This strategy is particularly crucial in medical imaging applications, where anatomical inconsistencies between images can lead to invalid or misleading pathology comparisons. By integrating semantic conditioning from text (which encodes pathology descriptions) with spatial conditioning from anatomical masks (which preserve structural integrity), our model generates outputs that are both clinically meaningful and visually coherent.

In summary, conditional diffusion provides a powerful framework for guided image synthesis. When extended with anatomical constraints, it enables high-fidelity and interpretable generation of medical images—supporting use cases such as anomaly verification, data augmentation, and cross-modal reliability assessment.

3.4 Related works

Text-to-image diffusion models represent a novel approach in medical imaging by generating synthetic CXR images from textual reports. These models operate by mapping text descriptions to visual features, effectively synthesizing images that align with the described pathologies and anatomical context. This process can enhance medical image analysis by addressing issues such as data scarcity, enabling augmentation [87], and supporting research validation [91]. Despite their promise, current methods face challenges in maintaining anatomical consistency, a critical requirement for clinical applicability.

Pioneering works like RoentGen [12] and Cheff [112] have demonstrated the potential of generating CXRs directly from free-form text reports. While these methods can produce plausible images, they often lack control over critical spatial features, resulting in inconsistencies in the anatomical structure. For example, a generated image describing "left lung consolidation" might misplace the pathology or alter unrelated anatomical regions, thereby reducing clinical reliability [13, 97].

Class-conditional X-ray generation methods are often motivated by privacy, data scarcity, or dataset balancing. However, not all works addressing privacy in CXR data involve generative modeling. For example, Packhäuser et al. [83] demonstrate that chest X-rays contain strong biometric signatures, enabling near-perfect patient re-identification even across institutions and over long time intervals. Their findings underline the importance of privacy-preserving data release, but they do not investigate synthetic or class-conditional image generation for augmentation.

In contrast, approaches that genuinely target synthetic augmentation often emphasize sampling diversity more than anatomical precision, making them unsuitable for tasks requiring spatially accurate, pathology-guided synthesis. In our earlier work [33], we highlighted the importance of using synthetic data to balance datasets effectively, underscoring that, for our research, preserving anatomical structure in chest X-rays is essential, particularly when incorporating spatial information like pathology.

A more recent study, XReal [39], takes a significant step forward by employing anatomy and pathology masks to guide X-ray generation. This ensures the precise placement of pathologies while maintaining organ-level accuracy. However, the reliance on detailed pathology masks introduces a dependency on expert input, which may limit scalability when using only textual CXR reports. In contrast, our guided image generation employs binary lung segmentation masks together with medical text prompts to validate the contextual fidelity of prompts by ensuring anatomical accuracy and consistent localization of pathologies. While this choice

provides a lightweight and scalable solution, more granular or multi-class segmentation masks could be incorporated in future work to offer finer control over anatomical regions and pathology placement.

3.5 Methods

We introduce an auxiliary model to enhance the reliability of the visual interpretation of the text report. Specifically, we utilize a CXR generation model to produce an alternative CXR image that is anatomically similar to the original but guided by the input text report. This approach provides an additional set of CXR images that can be compared to the localized regions in the original image. The comparison allows us to evaluate the report's accuracy and its alignment with the image by determining whether the report provides sufficient spatial details, such as pathology and anatomical regions. If accurate, the localized regions in the original CXR image should closely align with the corresponding areas in the generated images.

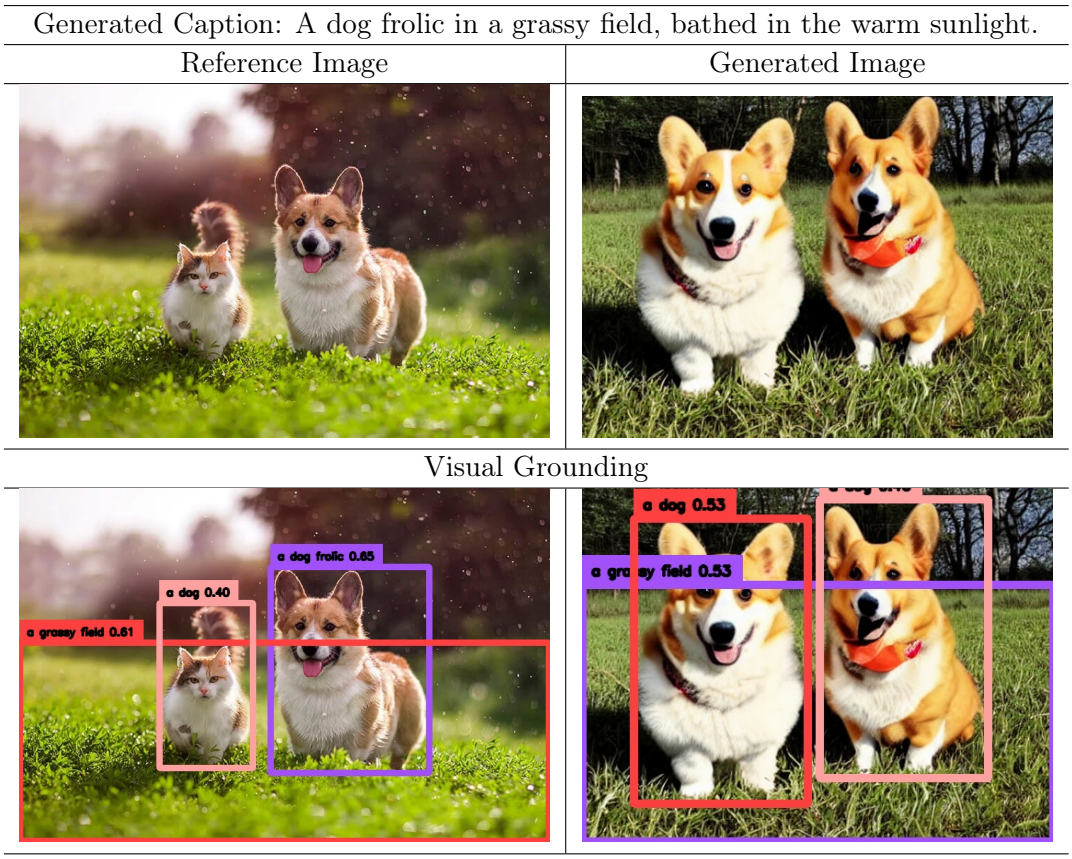


Table 3.1: Comparison between the reference image containing both a cat and a dog (left) and a synthetic image of two dogs generated based on the caption (right). The second row illustrates object localization results using the generated caption, highlighting how the model aligns text references with visual regions in both real and synthesized images.

To contextualize this approach, consider an analogous example from the natural image domain. Table 3.1 presents a simple scenario involving an image containing both a cat and a dog. An AI model generates the caption, "A dog frolics in a grassy field, bathed in warm sunlight." Notably, the caption omits any mention of the cat. In such cases, detection-based models like Grounding DINO may misclassify or completely overlook the secondary object (in this case, the cat), focusing instead on the explicitly referenced dog.

To resolve this ambiguity, a text-to-image diffusion model can synthesize an image based solely on the given caption. When comparing regions of interest between the original image and the generated one, the dog in the reference image shows a much higher similarity score (64%) with its synthesized counterpart than the cat does (13%). This gap suggests that the cat's presence is not semantically supported by the caption, indicating the description may be incomplete or misleading.

Applying this concept to the medical domain, particularly chest X-rays, the 'cat' could represent a missed pathology such as 'pneumonia'. If the report omits pneumonia but includes another condition like 'pleural effusion', the localization model might incorrectly attribute a region indicative of pneumonia with low detection accuracy to pleural effusion.

Although such discrepancies are relatively easy to detect in natural images, they can be far more subtle and clinically significant in medical domains. In such contexts, auxiliary models like ours can play a critical role in revealing omissions, ambiguities, or inconsistencies in machine-generated reports.

This example parallels our medical imaging approach and demonstrates how multimodal validation, linking text and image, can support automated fact-checking, semantic alignment, and trust calibration in AI systems.

Given that most thoracic diseases occur within or near the lung region as visible on chest X-rays, and to ensure the preservation of anatomical structures similar to the original image, we extract the lung segmentation from the original image. This is achieved using an existing model [1].

Using the binary lung segmentation image and the input text report, we generate a chest X-ray image through a conditional diffusion model originally designed for text-to-image synthesis. The binary lung mask provides spatial guidance by constraining the synthesis to the lung fields and separating them from peripheral regions. This design ensures anatomical consistency, but it also introduces certain limitations. For instance, because the mask is limited to lung boundaries, the model is not explicitly guided to represent structures outside this region, such as humerus arm bone or the clavicle (collarbone). In practice, these elements can still be synthesized through the richness of the text encoder and the model's learned distribution, but not with the same degree of spatial control as for the lungs.

Furthermore, as the segmentation masks are automatically extracted from the original CXRs, they are not perfect and may contain small artifacts. Examples include residual letters

[1] https://github.com/IlliaOvcharenko/lung-segmentation.git

(R or L) at the image borders, syringe-like shapes, or regions affected by medical devices. The model often learns to associate such artifacts with its training distribution, for example in the case of text artifacts, it tends to generate random pixel patterns resembling letters in the upper image corners; in the case of devices, it may occasionally introduce tubes or implants based on report content, even when these are absent in the original image. While these behaviors illustrate the robustness of the model, they also highlight areas where our binary mask guidance may fall short, and they underscore the potential benefit of extending the approach to more granular or multi-class segmentation masks in future work.

For this task, we build upon the stable diffusion model (SD v1.5 [97]), leveraging its pretrained weights to retain its understanding of general tokens. We then adopt the ControlNet model architecture [125] as the backbone and integrate a BiomedVLP-CXR-BERT encoder for tokenizing chest X-ray entities.

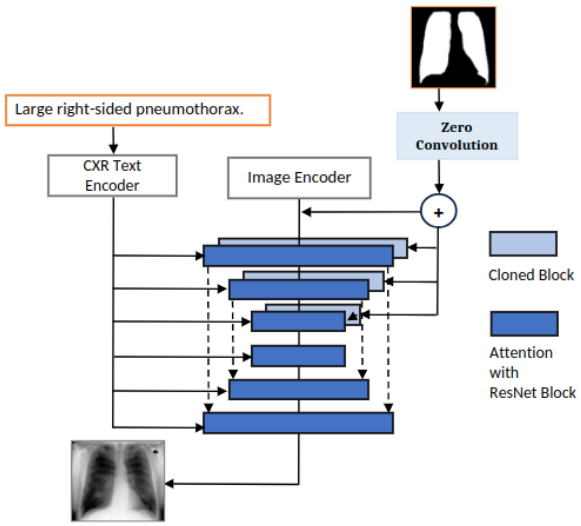


Figure 3.3: Overview of the Diffusion model using ControlNet architecture: The output of the Diffusion Model should closely approximate the original input image. The "zero convolution" is 1×1 convolution with both weight and bias initialized to zeros. It serves as a way to integrate the encoded text with the encoded guided binary mask, ensuring the output maintains anatomical structural fidelity.

The final architecture of the model is illustrated in Figure 3.3 diffusion model. Similar to the stable diffusion model, ControlNet uses an autoencoder architecture for text-to-image generation. However, in ControlNet, additional cloned encoder blocks are introduced to process the binary image, providing control over the inpainting process. While the original encoder blocks learn, the cloned blocks are frozen, and vice versa. A zero-convolution layer is then used to link the cloned blocks with the locked model, enabling smooth integration and enhanced control.

During training, we first fine-tuned the ControlNet architecture on a subset of CXR images from the MIMIC-CXR dataset, allowing it to adapt to the specific visual characteristics of chest radiographs. Once stabilized, we jointly trained the entire architecture, ControlNet and

the BiomedVLP-CXR-BERT encoder, so that the text encoder could effectively align with the visual features learned by ControlNet. This joint training phase ensures that both modalities are harmonized for the downstream generation task.

For training, We used the MIMIC-CXR v2.0.0 dataset [35, 50], a large-scale and publicly available collection of chest radiographs paired with corresponding free-text radiology reports. The dataset comprises 227,835 studies from 65,079 patients collected at Beth Israel Deaconess Medical Center. In addition to paired images and reports, it includes thoracic disease labels, facilitating supervised learning for various diagnostic tasks.

Our conditional diffusion framework was trained over 25 epochs. The training process achieved a final training loss of 0.078 and a validation loss of 0.125, indicating stable convergence and generalization. These results demonstrate the model's capacity to adapt to domain-specific image-text patterns while maintaining anatomical structure and semantic consistency across modalities.

Table 3.2 presents four samples generated using our method. In the third row, the generated images result from feeding the binary lung masks and their paired text reports (from the MIMIC-CXR dataset) into the model. The final row illustrates generated images created using the same binary masks but paired with random, unrelated reports. One challenge observed in these generations is the presence of artifacts around the periphery of the images. While these artifacts are random and irrelevant to the study's focus, they do not affect our main objective, as we concentrate on the lung regions and their associated pathologies. The trained model demonstrates strong capabilities in preserving anatomical structures. For example, in the fourth sample, the right lung in the original image is completely opaque. However, when paired with a random report that does not describe this opacity, the model still synthesizes it accurately, highlighting its robustness in anatomical reconstruction.

Limitation. It is important to note, a key limitation of this approach, within our VICCA framework, we assume that radiology reports are accurate and truthful reflections of image content. If a report contains hallucinated findings, the synthesized image will reproduce them. Conversely, if a report misses a finding, the synthetic image will also omit it. This constitutes a failure mode that leaves our pipeline dependent on report completeness. Nevertheless, since our reliability scoring compares the synthesized image against the original through visual grounding, any missed finding would still appear in the original image grounding but not in the synthesized alignment, allowing us to capture such discrepancies as part of the evaluation process.

For a clearer demonstration of the model's performance with varying reports, Table 3.3 presents an additional example. The middle column displays an image generated using the ground truth report, while the last column shows an image generated using a random, unrelated report. Images generated with the ground truth report closely resemble the original input image, while those generated with incorrect pathology descriptions do not. Notably, when provided with a random report mentioning "cardiomegaly," the model expanded the heart region in the generated image, reflecting its ability to interpret and map text input to

Table 3.2: The results of the CXR generative model using two different text prompts. The ground truth report is in the third row, and a random CXR report is in the fourth row. Feature similarity is used to calculate the generation's similarity with the Ground Truth (GT) image.

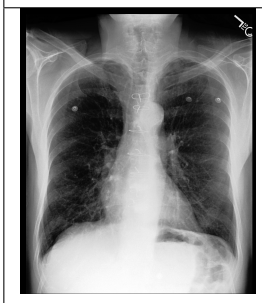
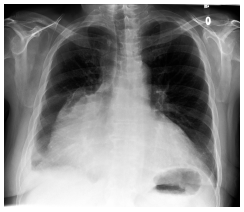
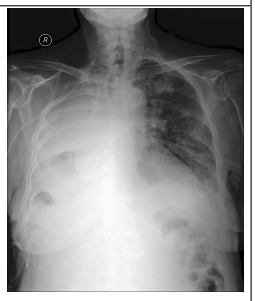
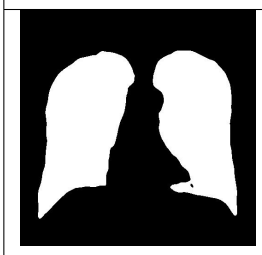
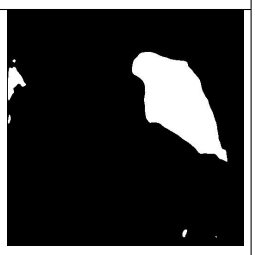
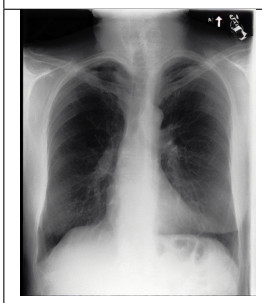
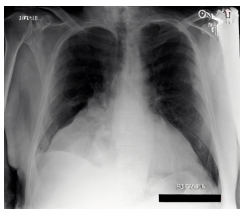
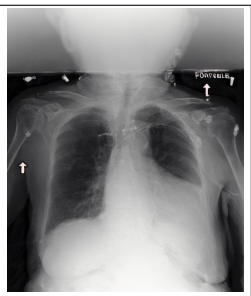
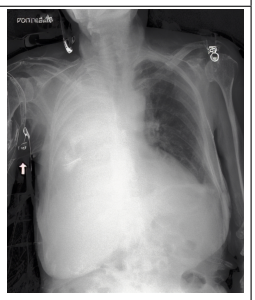
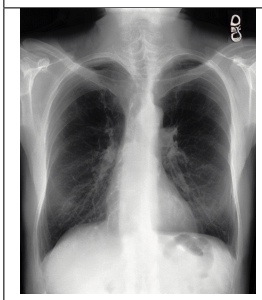
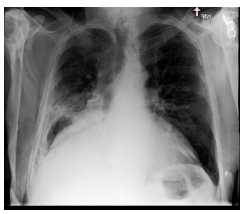
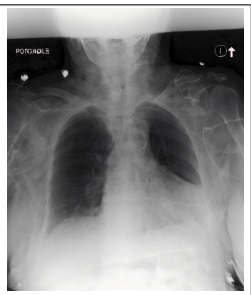
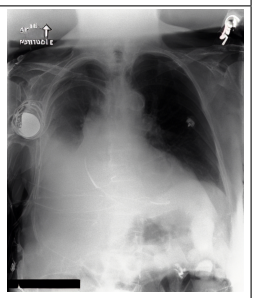
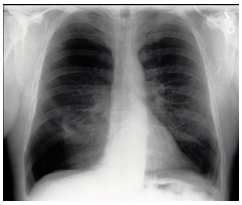
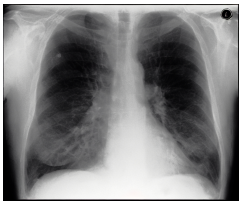
Original Image			
			
Binary Image Lung Segmentation			
			
Generated with the Original Report			
			
Similarity with the GT: 88.6%	Similarity with the GT: 85.3%	Similarity with the GT: 78.6%	Similarity with the GT: 81.9%
Generated with A False Report			
			
Similarity with the GT: 84.2%	Similarity with the GT: 79.6%	Similarity with the GT: 72.7%	Similarity with the GT: 67.3%

Table 3.3: The results of the CXR generative model using two different text prompts. The ground truth report is in the middle column, and a random CXR report is in the last column.

		
Original Image	Generated Image by *Ground Truth Report:* Large right-sided pneumothorax with mild leftward shift of mediastinal structures indicative of tension.	Generated Image by *Random Report:* Cardiomegaly with mild pulmonary vascular congestion.

corresponding anatomical features effectively.

Accurately generating features from text enables a robust comparison of feature-level similarity within regions of interest, facilitating the evaluation of the reliability score and validating the alignment between the report and the image.

Before continuing, we address several questions that naturally arise from this methodology of generating CXR images.

Q1: How does the model ensure generation of pathology-related features? The denoising U-Net in Stable Diffusion is augmented with (i) **textual embeddings** from BiomedVLP-CXR-BERT, injected via cross-attention into the U-Net blocks, and (ii) **spatial embeddings** delivered by ControlNet from the binary lung mask (Figure 3.3). This dual conditioning, text tokens specifying *what* abnormality to draw and the mask specifying *where* synthesis should be constrained, is the component responsible for injecting pathology-specific features while preserving anatomy.

Q2: Which part of the model is responsible for this capability? Cross-attention layers in the U-Net fuse text features with image latents, ensuring that pathology attributes from the report modulate the denoising trajectory at multiple resolutions. ControlNet processes the binary lung mask into mask-conditioned feature maps that are coupled to the main U-Net through zero-convolution layers, enforcing spatial fidelity to lung regions during synthesis. Together, these components provide content control (via text) and location control (via mask).

Q3: How is the right mapping from report context to image enforced? During each denoising step, noisy latent patches (image tokens) attend to text tokens via cross-attention, aligning visual features with clinical entities (e.g., `right lower lobe opacity`). In parallel, ControlNet guides the U-Net with mask-conditioned feature maps coupled through zero-convolution, biasing the network so that synthesized evidence emerges within the lung fields rather than arbitrarily across the image. Thus, phrase semantics drive content and mask conditioning drives location.

3.6 Training Protocols and Datasets

Table 3.4: Dataset distribution used for training and validation, including the number of images and their paired annotations.

Dataset	Train	Validation	Test
MIMIC-CXR [51]	207827	2991	2189

Table 3.4 summarizes the distribution of training and validation samples in the MIMIC-CXR dataset, including the number of images and their corresponding annotations. Only chest X-ray images captured in the Posterior-Anterior (PA) or Anterior-Posterior (AP) views are retained for this study; lateral view images are excluded to maintain consistency in anatomical presentation.

The original MIMIC-CXR dataset includes approximately 53,000 images labeled as "No FINDING". To address class imbalance and improve the representational diversity of pathological cases, we randomly exclude 20,000 of these normal cases from the training set. This preprocessing step ensures a more balanced distribution across different pathology labels and prevents the model from being biased toward the overrepresented "No FINDING" class during training.

3.7 Results and Validation

To assess both visual quality and clinical validity, we employ a multi-step validation protocol:

1. **Visual fidelity:** FID for realism; MS-SSIM for perceptual similarity.
2. **Anatomical consistency:** Dice and IoU against anatomical regions detected by a DETR model trained on Chest ImaGenome bounding boxes (Section 2.4.3).
3. **Clinical plausibility:** Pathology classification of generated CXRs using pre-trained clinical detectors (*e.g.*, TorchXrayVision) compared against the conditioning phrases/reports (details in Medical Validation 3.8).

These layers jointly evaluate whether the conditional diffusion model (i) generates realistic images, (ii) preserves anatomy, and (iii) faithfully maps report context to visual evidence in the correct locations.

To evaluate the generative model, we compare its performance against benchmark models using three metrics: *Multi-Scale Structural Similarity Index* (MS-SSIM), Dice score, and *Frechet Inception Distance* (FID). These metrics collectively assess the quality and anatomical fidelity of the generated images.

For evaluation, we generated 10 samples for each of the 2,065 reports in the test set, resulting in a total of 20,650 generated images. For the FID evaluation, we use all generated

samples. For the MS-SSIM and Dice evaluations, we employ an automated selection approach to identify the most representative sample among the 10 generated images. The selected sample is the one with the highest *Intersection over Union* (IoU) score and the closest pathology classification match to the original CXR image. The results, presented in Table 3.5, indicate that our model achieved the second-best FID score while outperforming all benchmarks in MS-SSIM and Dice. These findings highlight the superior quality of our model's generated images, particularly in preserving anatomical structures and ensuring alignment with pathology information.

The results highlight the effectiveness of our inpainting technique during training, demonstrating that a model originally trained on natural images can be successfully adapted for CXR images by utilizing binary lung segmentation to maintain anatomical consistency and a specialized text encoder for medical contexts. In comparison, the Cheff model with the best FID score generates images solely based on text input and lacks this anatomical precision. Furthermore, our model surpasses the performance of the XReal model, which also focuses on preserving anatomical accuracy, further validating our approach.

Table 3.5: The comparison of FID scores among three models.

Model	MS-SSIM ↑	Dice ↑	FID ↓
Cheff [112]	0.415	0.5	**24.64**
RoentGen [12]	0.386	0.631	82.14
XReal [39]	0.701	0.838	55.12
VICCA (Ours)	**0.71**	**0.841**	35.76

While the quantitative metrics used are valuable for validating the generative model's performance, the sensitivity of the medical domain necessitates a deeper focus on medical validation. In this study, we design a validation framework that does not rely on direct radiologist feedback, aiming to develop an expert-independent assessment pipeline. To ensure the model's medical reliability, we leverage well-annotated datasets and standardized evaluation metrics, establishing a structured and objective validation process that enhances reproducibility and clinical applicability.

Before the integration of this auxiliary model in the pipeline, it is crucial to validate the generative model specifically through medical criteria, as it serves as a reference for assessing localization accuracy. To ensure the robustness of the generated outputs, we add a novel system to the framework. This system is designed to evaluate the anatomical and pathological aspects of the model. This additional validation ensures that the generated images align with the medical standards required for accurate interpretation and assessment.

3.8 Medical Validation

The validation process for the generative model in this research is inspired by the way experts acquire knowledge. It centers on evaluating the fundamental aspects of anatomical regions

and correlating the anomalous parts of the image with the corresponding pathology. Consequently, our first step is to establish a method for validating the anatomical structures of the generated images. This ensures that the generated outputs maintain consistency with real CXR anatomical accuracy.

3.8.1 Anatomical Validation

A CXR image consists of approximately 36 distinct anatomical regions, such as the "right upper lung zone," "right mid lung zone," "right atrium," "descending aorta," "carina," "left upper abdomen," and others. Identifying these anatomical regions is a straightforward task, thanks to the availability of the Chest ImaGenome dataset [35, 115], which provides radiologist-annotated bounding boxes for these 36 regions within the MIMIC-CXR dataset. Leveraging this dataset, we successfully train a robust region detector using the DETR model [11]. A notable advantage of the DETR model is its capacity to detect all class objects simultaneously, which is particularly useful given the close proximity of anatomical features in CXR images. The model demonstrates rapid convergence, achieving a DETR-like loss of 3.5030 and a generalized **IoU loss** of 0.29 after just 14 epochs. On the test set, the IoU accuracy reaches an impressive 76%.

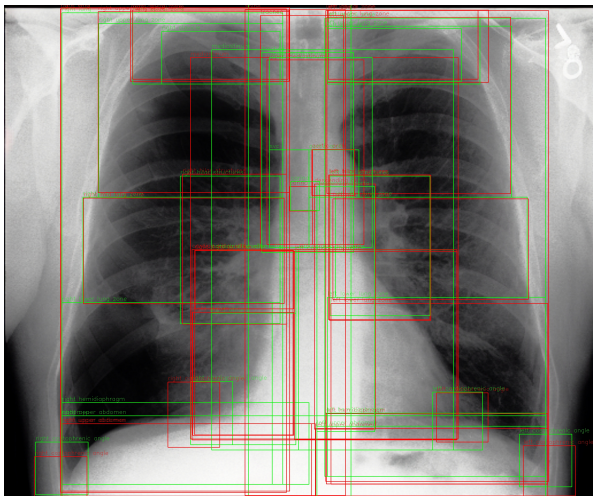


Figure 3.4: Ground truth anatomical regions in green and automatical detection of the anatomical regions on generated CXR in red. The mIoU of the bounding boxes is 76%.

Subsequently, we apply the trained DETR model to the generated CXRs to obtain bounding boxes and evaluate their IoU against the ground truth. The IoU score for the generated CXR set is 68%, validating the generative model's effectiveness in preserving anatomical structures. Figure 3.4 demonstrates the performance of the trained DETR model on a generated CXR image. The red bounding boxes represent the automatically detected anatomical regions, while the green bounding boxes depict the ground truth annotations from the Chest ImaGenome dataset. Since generated images often contain slight spatial shifts compared to ground truth images, we address this issue by aligning the bounding boxes based on the su-

perior vena cava (SVC) region, which is typically central in CXR images. By detecting the positional shift of the SVC in both images, we adjust the bounding boxes accordingly. The mIoU score between the detected and ground truth bounding boxes is 76.2%, indicating a high degree of overlap and alignment for the majority of anatomical regions.

3.8.2 Pathological Validation

To validate the pathological accuracy of the generated CXR images, we utilize the TorchXrayVision library for pathology classification [21]. While various datasets provide annotations for 14 to 15 thoracic diseases, many existing classification models are either outdated or limited to a subset of these classes. TorchXrayVision overcomes these limitations by leveraging multiple datasets for comprehensive pathology classification. Using this library, we classify the pathologies in the generated CXRs and compare the detected pathologies' accuracy with those in the original images. The generative model achieves an average classification accuracy of 88.53% on the generated test set. These results highlight our generative model's capability to produce CXR images that accurately reflect pathological features from text descriptions while maintaining anatomical fidelity.

3.9 Perspectives and Limitations

While our validation strategy focuses on phrase–level alignment between textual descriptions and localized image regions, it does not yet extend to the full report level. In clinical practice, radiology reports often contain multiple findings, differential diagnoses, or negations (e.g., "no evidence of pneumonia"). Our model is capable of generating images that reflect several pathologies from a single report, but handling such cases still requires a more advanced alignment system capable of reasoning over composite findings. For negated statements, the corresponding pathologies are removed from the prompt during preprocessing through our entity extraction mechanism, which will be discussed in detail in Chapter 5.

Beyond its role in reliability scoring, our guided diffusion generator also offers potential for dataset augmentation. By combining text prompts with anatomical masks, the model can synthesize rare or underrepresented pathologies. For instance, starting from a normal CXR, one could input a prompt such as "pulmonary nodule in the left lower lung" and generate a synthetic image with the specified abnormality. This capability provides a means to address class imbalance in training diagnostic models, particularly for rare conditions where annotated examples are limited.

3.10 Conclusion

In pursuit of the main objective of this thesis, enhancing the interpretability of medical reports both visually and quantitatively, we have presented a two-stage approach across two chapters.

First, we introduced a visual grounding model that localizes textual findings within chest X-ray images, assigning each localized region a *localization score* to reflect its alignment with the report content. Second, we proposed an auxiliary model that generates anatomically consistent synthetic CXR images from textual descriptions. By comparing the localized regions across the original and generated images, we compute a *reliability score* that quantifies the consistency and trustworthiness of the alignment between text and image. While this pipeline outlines our complete framework, the previous two chapters have focused on presenting and evaluating each model independently. In the next chapter, we shift our attention to the integration of these components, unifying the visual grounding and generation modules to form the full **VICCA** framework. In the next chapter, We conduct a comprehensive evaluation of VICCA across multiple approaches, using both conventional and clinically-informed metrics to validate its interpretability and diagnostic reliability.

Chapter 4

VICCA: Visual Interpretation and Comprehension of Chest X-ray Anomalies in Generated Report

Contents

4.1 Introduction

Our architecture of *Visual Interpretation and Comprehension of Chest X-ray Anomalies in Generated Report* (VICCA), illustrated in Figure 4.1, is a unified framework designed to assess the reliability and interpretability of AI-generated chest X-ray (CXR) reports. By integrating visual grounding, image generation, and semantic analysis modules, VICCA establishes an

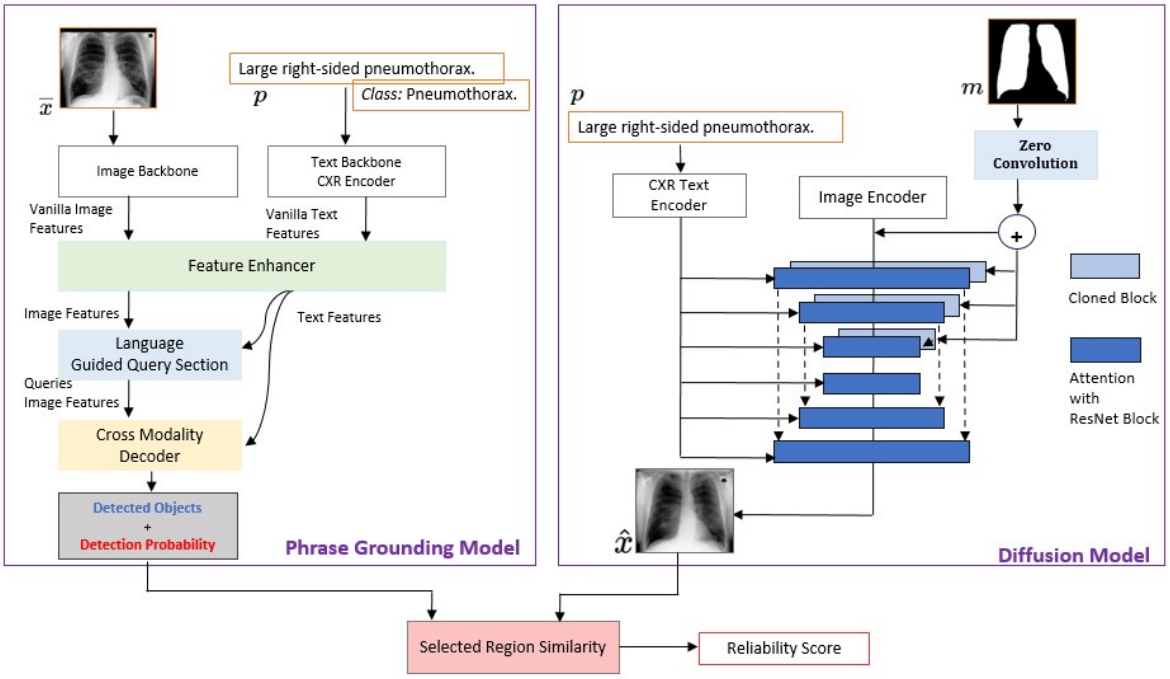


Figure 4.1: Overview of the VICCA model pipeline

objective, feedback-free scoring system that quantifies how well a report aligns with actual image content.

The ultimate goal of VICCA is to move beyond traditional black-box evaluation methods by providing explainable, cross-modal assessments, evaluating a report from both visual and textual perspectives. The pipeline operates as follows:

1. **Visual Grounding:** Given a CXR image and a text input (e.g., a radiology report or clinical prompt), VICCA first performs visual grounding. This step localizes image regions associated with each described abnormality using a vision-language grounding model (see Chapter 2).

2. **Synthetic Image Generation:** The same textual input is then passed to a conditional diffusion model (introduced in Chapter 3) to generate a synthetic CXR image. This model ensures anatomical plausibility while reflecting the content of the report.

3. **Region Alignment Check:** The localized regions from the original and synthetic images are compared to assess cross-modal consistency. This comparison provides insight into whether the text-based generation aligns with actual findings.

4. **Pathology Presence Consistency:** The pipeline also verifies the presence or absence of specific pathologies. For example, if the report mentions a "nodule," VICCA evaluates whether such a region is evident in both the real and generated images.

5. **Entity Identification:** To support interpretability, the main pathology or entity described in the report is identified and linked to visual regions.
6. **Medical Entity Summarization:** Finally, the input report is summarized into a structured set of medical entities using a concept extractor (discussed in Chapter 5).

These components produce a comprehensive set of outputs:

- **Localization Accuracy:** Evaluates whether the grounding model correctly identifies the visual regions for each described abnormality.
- **Reliability Accuracy:** Measures the overlap and consistency between grounded regions in the real image and those inferred from the generated image.
- **Presence/Absence Checks:** Validates whether each described pathology is visually supported.
- **Entity-Level Summarization:** Extracts and aligns clinical concepts across modalities.

Together, these outputs establish VICCA as a scalable and generalizable evaluation framework, fully independent of human annotation. This chapter presents a comprehensive assessment of VICCA as an integrated pipeline, analyzing both its quantitative performance and qualitative outcomes. We begin with an overview of the system's structure and outputs, followed by detailed evaluations and case studies. Before proceeding, however, it is essential to clearly define the **reliability score**, a key component of this framework, and to explain its formulation and operational principles in detail.

4.2 Reliability Score: Structural and Visual Consistency

The **reliability score** is introduced as a quantitative interpretability measure designed to assess the degree of visual–textual alignment between the generated and original chest X-ray (CXR) images. Specifically, it evaluates whether the pathological regions inferred from the textual report are consistently represented in the corresponding visual regions of the original image. By comparing localized features across modalities, the reliability score captures the extent to which generated visual evidence supports the textual description.

4.2.1 Mathematical Formulation

Let I_{orig} denote the original CXR image and I_{gen} the corresponding image synthesized by the conditional diffusion model guided by the report and binary lung mask. The evaluation is performed on localized regions of interest (ROIs) $\{R_1, R_2, \ldots, R_N\}$ obtained from the visual grounding model.

For each ROI R_i, we extract the corresponding subregions from both images:

$$I_{\text{orig}}^{(i)} = I_{\text{orig}} \odot M_i, \qquad I_{\text{gen}}^{(i)} = I_{\text{gen}} \odot M_i,$$

where M_i is the binary mask corresponding to the i^{th} ROI and $\odot$ denotes the element-wise product.

To measure visual similarity between these paired subregions, we compute the **Structural Similarity Index Measure (SSIM)** [110] defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{4.1}$$

where:

- μ_x and μ_y are the local means of image patches x and y;
- σ_x^2 and σ_y^2 are the local variances;
- σ_{xy} is the local covariance between x and y;
- $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$ are small constants for numerical stability, with L being the dynamic range of pixel values.

The global reliability score $\mathcal{R}$ is computed as the mean SSIM over all N regions of interest:

$$\mathcal{R} = \frac{1}{N} \sum_{i=1}^{N} \text{SSIM}(I_{\text{orig}}^{(i)}, I_{\text{gen}}^{(i)}). \tag{4.2}$$

4.2.2 Interpretation

A higher $\mathcal{R}$ value indicates greater structural consistency between real and generated CXRs within clinically relevant regions, implying that the visual content generated from the report accurately reflects the original image features. Conversely, lower values may indicate discrepancies due to missing pathologies, spatial misalignment, or semantic inconsistencies between the text and the image. It provides a quantitative bridge between textual reasoning and visual evidence, serving as a key metric in the VICCA framework.

Figure 4.2 illustrates a region of interest (ROI) from the original image alongside its corresponding region in the generated image (left). For clarity, the ROI is also displayed in an enlarged inset highlighted in red. The reliability score is computed by applying the SSIM metric to these paired regions, quantifying their local structural similarity. By aggregating the ROI-level SSIM values, VICCA derives the global reliability score $\mathcal{R}$, which provides an intuitive measure of how well the visual content aligns with the textual description across modalities, reflecting the generated image of focus on the region emphasized in the report.

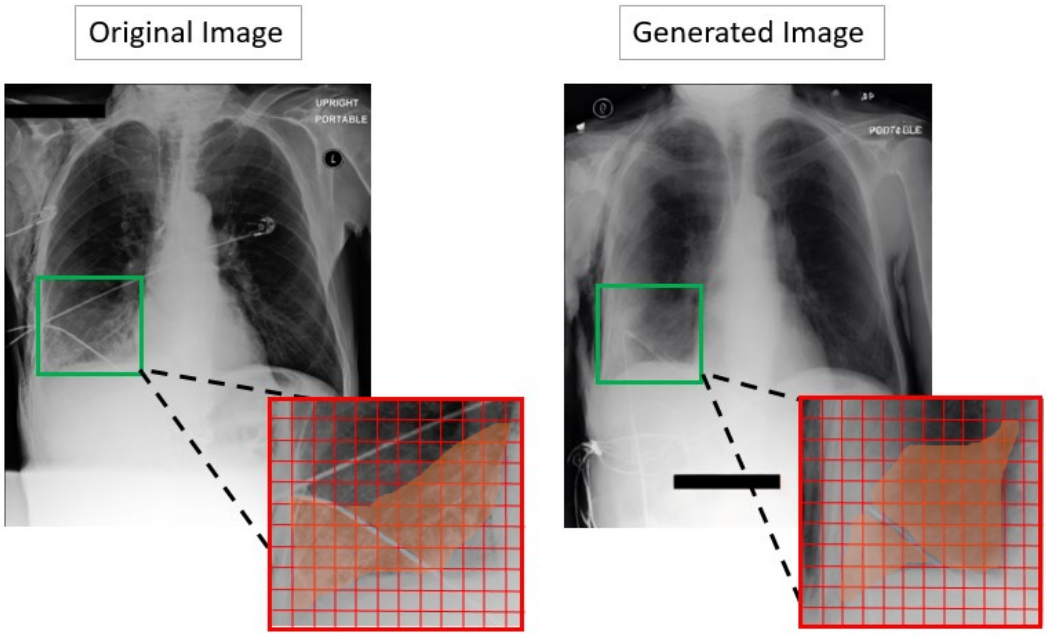


Figure 4.2: Illustration of reliability score computation: ROIs localized in the original image (right) and their corresponding generated regions (left) are compared using SSIM to derive region-wise and global reliability scores.

4.3 VICCA Model Result

An example of the VICCA pipeline output is illustrated in Figure 4.3. In addition to the two primary scoring metrics, localization accuracy and the reliability score, the pipeline also provides spatial information about the detected pathology, extracted from the input text. This extraction is performed using our previously introduced method [86] (described in detail in Chapter 5), which identifies clinically relevant medical entities within the report.

To categorize the pathology present in the image, we use the TorchXRayVision model [21], a pre-trained classifier that predicts 14 common thoracic pathologies from CXR images. Since both the original image and the associated prompt originate from the MIMIC-CXR dataset, we additionally incorporate the CheXpert labels [47] provided by the dataset for reference and comparison. This multi-faceted output enables a more comprehensive understanding of how well the textual and visual elements align across the VICCA pipeline.

4.4 VICCA Pipeline Evaluation

In the previous chapters, we evaluated the individual components of our framework, namely the visual grounding and diffusion-based generation models, using standard validation techniques. Both modules demonstrated strong performance separately. We now shift our focus to the central objective of this thesis: quantifying a reliability score for the automatically localized

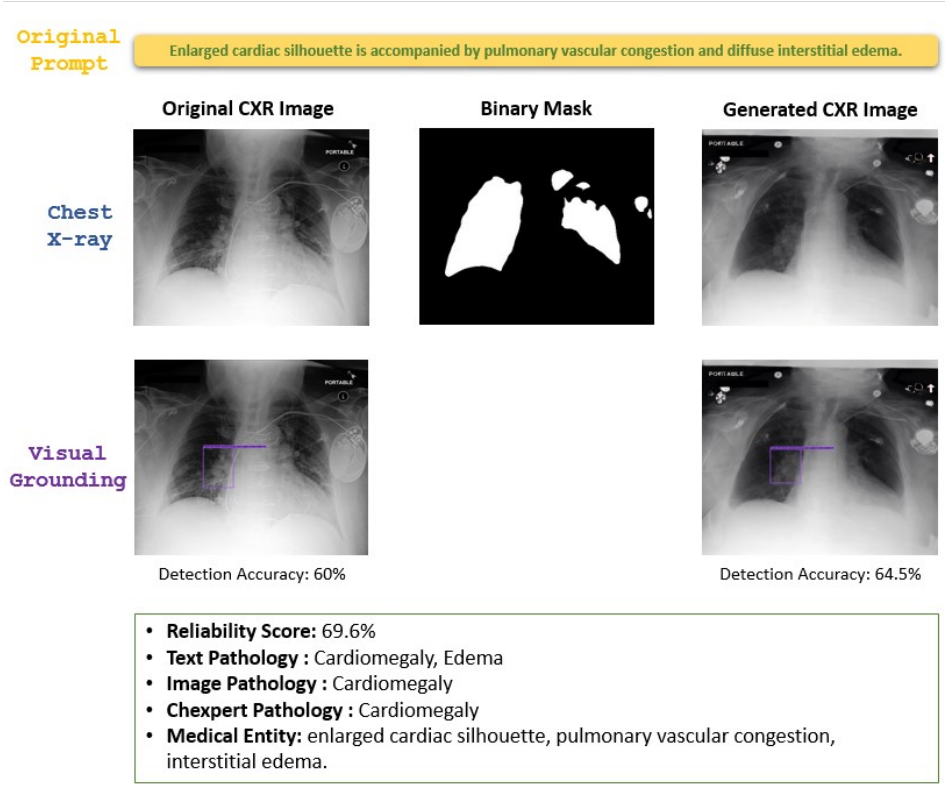


Figure 4.3: An example output of our VICCA model.

regions in AI-generated medical reports through a unified evaluation pipeline.

To accomplish this, we project the bounding boxes generated by the visual grounding model onto the corresponding synthetic images produced by our diffusion model, ensuring a region-level alignment between both modalities. A chest X-ray–specific image encoder is then employed to extract features from these matched regions in both the original and generated images.

To measure the similarity between corresponding regions, we adopt a two-fold evaluation strategy. First, we assess pixel-level fidelity using the *Multi-Scale Structural Similarity Index* (MS-SSIM). Second, we extract high-level semantic features using a dedicated CXR image encoder and compute the χ^2 distance between the resulting feature vectors as a statistical measure of correspondence.

Given the lack of direct ground truth for this form of evaluation, we introduce a two-step interpretability framework to guide the interpretation of the MS-SSIM and χ^2 scores and to establish a consistent, quantifiable reliability score for each localized region.

1. **Using the Original Report:**

   - Bounding boxes are detected based on the original report, and both MS-SSIM and χ^2 values are calculated by comparing the corresponding regions in the original and

generated images.

- A χ^2 value closer to zero indicates a high feature similarity, while an MS-SSIM value closer to one signifies a strong structural resemblance between the regions.

2. **Using a Random Report:**

- The same process is repeated using a randomly selected report from the MIMIC-CXR dataset. The random report is explicitly chosen to be unrelated to the original report and to share no overlapping thoracic disease classes with it.
- The expectation is that the MS-SSIM values would be significantly lower, and the χ^2 values significantly higher, confirming a reduced similarity between the localized and generated regions when the report does not correspond to the actual content.

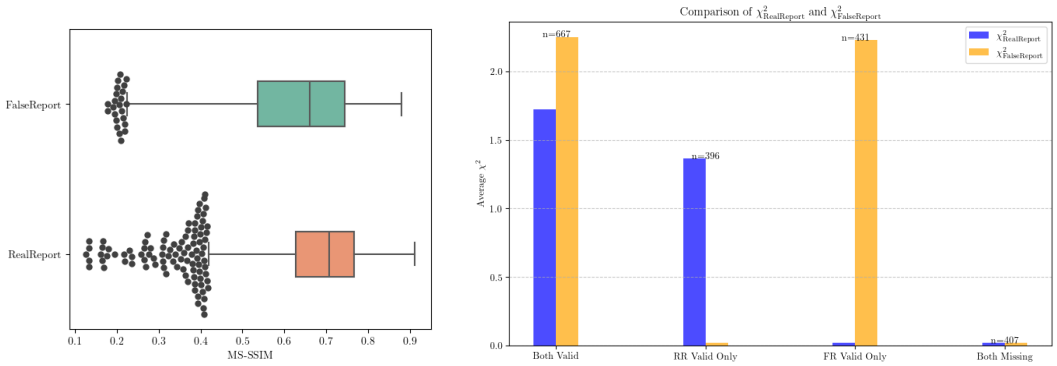


Figure 4.4: The similarity comparison between the localization and the generated CXR images is evaluated using MS-SSIM and χ^2 calculations based on two text inputs: the Real Report (RR), which represents the ground truth associated with the CXR image, and a False Report (FR), which has no correlation with the RR. The χ^2 plot on the right categorizes the results into four scenarios based on the output of the visual grounding model: Both RR and FR produce valid localizations, only RR produces a valid localization, only FR produces a valid localization, neither RR nor FR yields a valid localization. The MS-SSIM boxplot on the left illustrates the distribution of similarity scores for each scenario, showing the density of evaluation for both the real and false reports.

Figure 4.4 illustrates the results of this comparison based on the average MS-SSIM and the χ^2 values. For MS-SSIM, the similarity scores based on the Real Report indicate a narrower range, with outliers remaining close to the margins, indicating a consistent and high degree of similarity. In contrast, the scores based on the False Report show a wider distribution, with a noticeably lower average similarity, reflecting the reduced alignment between the generated and original images in this scenario. The higher similarity scores observed can be attributed to the method used for selecting random reports, which is based on pathology annotations from the MIMIC-CXR dataset. Random reports are considered as acceptable if they lack any explicit similarity to the original reports. However, in certain cases, overlapping visual traits between pathologies, for instance, “lung opacity” and “nodule”, led to unintentional

correlations. These shared visual characteristics in the selected random reports contribute to the detection of higher similarity scores.

Regarding feature similarity using χ^2, analysis across 1901 samples shows the following observations:

- 407 samples have low localization probabilities for both the original and random reports, rendering the results inconclusive.
- 396 samples show a valid localization only when using the original report.
- 431 samples produce more acceptable localization results with the random report compared to the original report.
- 667 samples demonstrate significantly lower χ^2 values for the original report compared to the random report, indicating better feature correspondence.

Although there are 431 cases where the localization model fails to detect regions accurately based on the original report, the pipeline's performance is evident. Even in these cases, the average χ^2 value for regions localized using the random report is relatively high, reflecting the system's ability to identify and highlight dissimilarities between false report segments in generated and original images. This underscores the robustness and sensitivity of the pipeline, even when faced with similarities in the pathology.

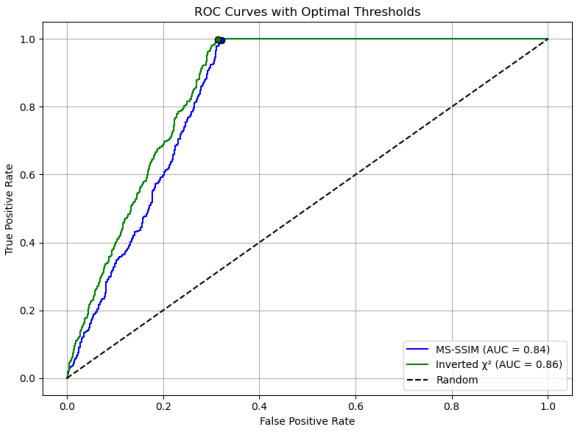


Figure 4.5: ROC curves for the dual-scoring interpretability system using MS-SSIM and inverted χ^2 metrics.

To support the system's utility in clinical decision-making, Figure 4.5 presents the ROC curves for the MS-SSIM and χ^2 scoring metrics. As explicit interpretability labels are unavailable, we assign a ground truth label of 1 to cases where the visual grounding model detects regions in both the original and synthesized images, and 0 otherwise. To determine optimal thresholds, we used Youden's J statistic [118] (J = True Positive Rate (TPR) – False Positive Rate (FPR)), which identifies the point on the ROC curve that maximizes the difference between the true positive rate and the false positive rate. Based on this criterion, the optimal

thresholds were found to be 0.299 for MS-SSIM and 0.123 for χ^2, with both metrics achieving a true positive rate of 1.00 and a false positive rate of approximately 0.31. This means that while all interpretable cases are correctly identified, about one-third of non-interpretable cases are misclassified.

These results suggest that the dual-scoring system prioritizes sensitivity, reliably capturing all interpretable cases, even at the cost of moderate over-inclusiveness. In clinical contexts, this trade-off is acceptable and even preferable, as ensuring that no interpretable region is missed is often more critical than avoiding false positives.

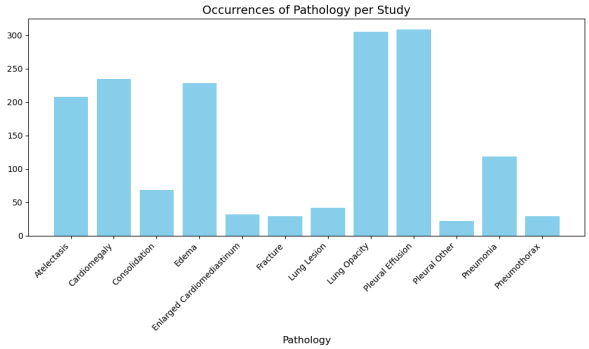


Figure 4.6: Occurrences of pathologies in the test dataset, including 2018 studies. The most frequent occurrences are observed for "Lung Opacity" and "Pleural Effusion," followed by "Cardiomegaly" and "Edema."

Finally, we present the results for each pathology class. Figure 4.6 illustrates the frequency of each pathology in our test dataset. The most frequent occurrences are observed for "Lung Opacity" and "Pleural Effusion," followed by "Cardiomegaly" and "Edema."

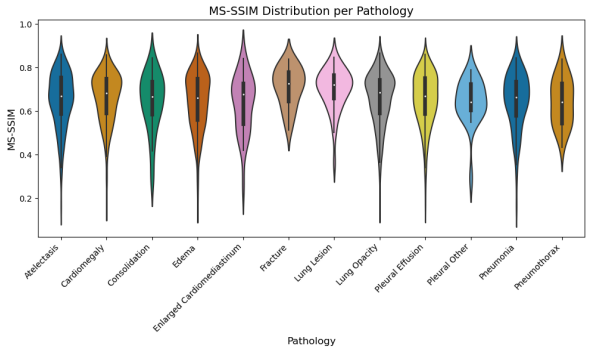


Figure 4.7: The MS-SSIM distribution per pathology in the test dataset.

For each pathology, we present the MS-SSIM score for each subject in Figure 4.7. The results show that the highest average scores are associated with "Lung Lesion" and "Fracture". Although these pathologies are less frequent in our test set, they are easier to interpret in a CXR image. Based on our findings, the most challenging pathologies to interpret in a CXR image are "Pneumonia", "Pneumothorax", and "Pleural others". Among these, "Pneumonia" occurs more frequently but remains the most difficult to assess. Additionally, our results

indicate that "Edema" is the most ambiguous class, as it shows the highest variability. This variability could be attributed to the subtle and often similarity in nature of the symptoms in CXR images, which may lead to difficulties in distinguishing them from each other. Our overall findings suggest that this interpretive study is crucial for enhancing the understanding and reliability of AI-generated reports. In the next section, a few case studies are presented to further demonstrate the effectiveness of our VICCA model.

4.5 Case Study

In this section, we present various scenarios and examples encountered during our study to illustrate the application and performance of our project. These include both successful cases and failed cases, which are included to provide deeper insights and facilitate a better understanding of the limitations and challenges of the approach.

4.5.1 Case 1: Multiple Pathologies

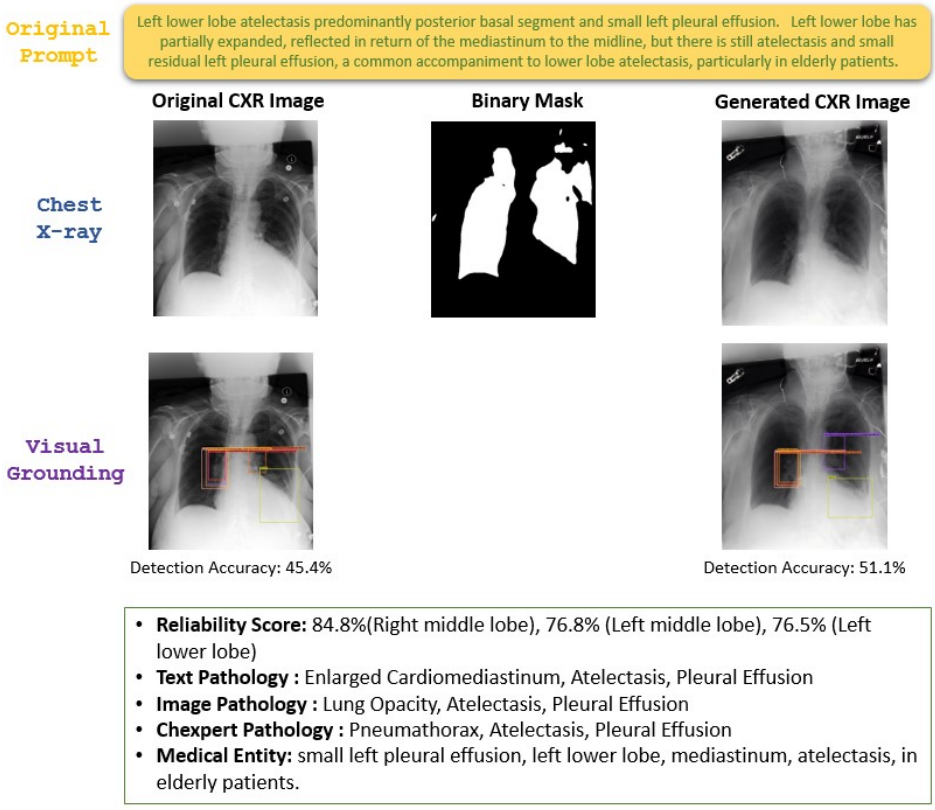


Figure 4.8: Case 1: Multiple Pathology

In the first scenario presented in figure 4.8, we examine a case involving multiple pathologies in both the text prompt and the image. Using our visual grounding visualization, we identify

three regions of interest (ROIs) in both the original and generated images. Although the detection probabilities are relatively low, the high similarity between the corresponding ROIs across both images indicates that the prompt can be considered trustworthy. Additionally, the pathologies identified through the prompt and the image using an AI model closely align with the Chexpert annotations derived from the MIMIC-CXR dataset [50].

4.5.2 Case 2: Single Pathology A Success Case

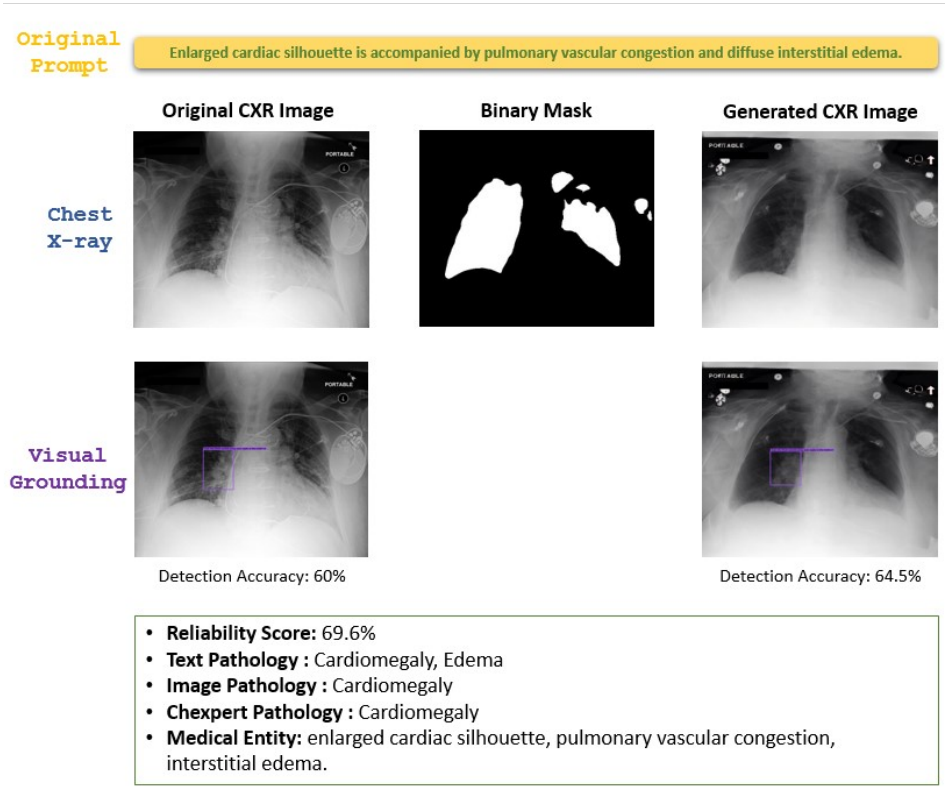


Figure 4.9: Case 2: Single Pathology A Success Case

In the second case, presented in Figure 4.9, we analyze a scenario involving a single pathology. While two pathologies are initially identified in the text prompt, both our image-based pathology detection and the Chexpert annotation confirm the presence of only one pathology, maintaining consistency across methods. Moreover, our visual grounding accurately detects a single pathology, which exhibits a high similarity with the pathology described in the prompt through the generated image. This alignment demonstrates that the prompt is consistent with the image and can be considered reliable.

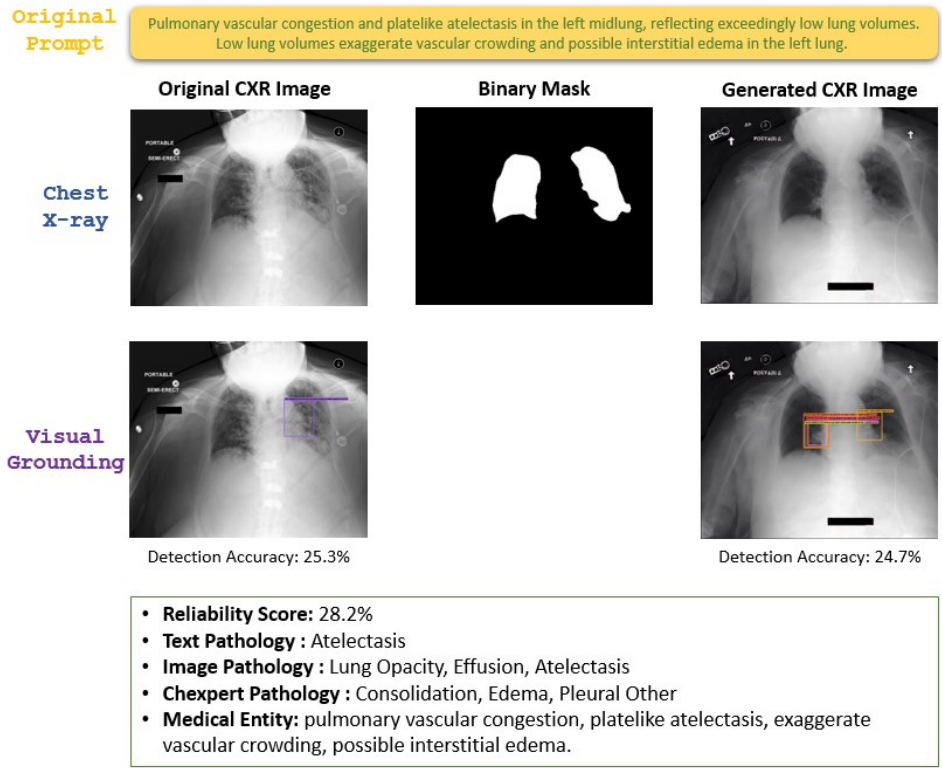


Figure 4.10: Case 3: Single Pathology A Failure Case

4.5.3 Case 3: Single Pathology A Failure Case

In contrast to the previous case, Figure 4.10 illustrates a failure to correctly align the original prompt with the corresponding image. Visual grounding on the original CXR image detects only one ROI associated with pathology, whereas pathology detection identifies three distinct pathologies in the image. Similarly, the generated image from the same prompt reveals two ROIs, but their similarity scores with the original image are notably low.

While the AI model identifies only one pathology from the text, this pathology corresponds to one of the detected pathologies in the image. However, these results do not align with the annotated pathologies in the dataset. Despite the prompt being originally written by a radiologist, the system fails to reliably associate it with the correct CXR image. This case highlights challenges in ensuring consistency and accuracy across modalities when relying solely on automated approaches without any interpretation.

4.5.4 Case 4: A Success Case of Using False Report

In this case which is illustrated in Figure 4.11, we examine a scenario where the prompt differs significantly from the original dataset report and shows no correlation with it. Although visual grounding on the original CXR image identifies a region of interest corresponding to

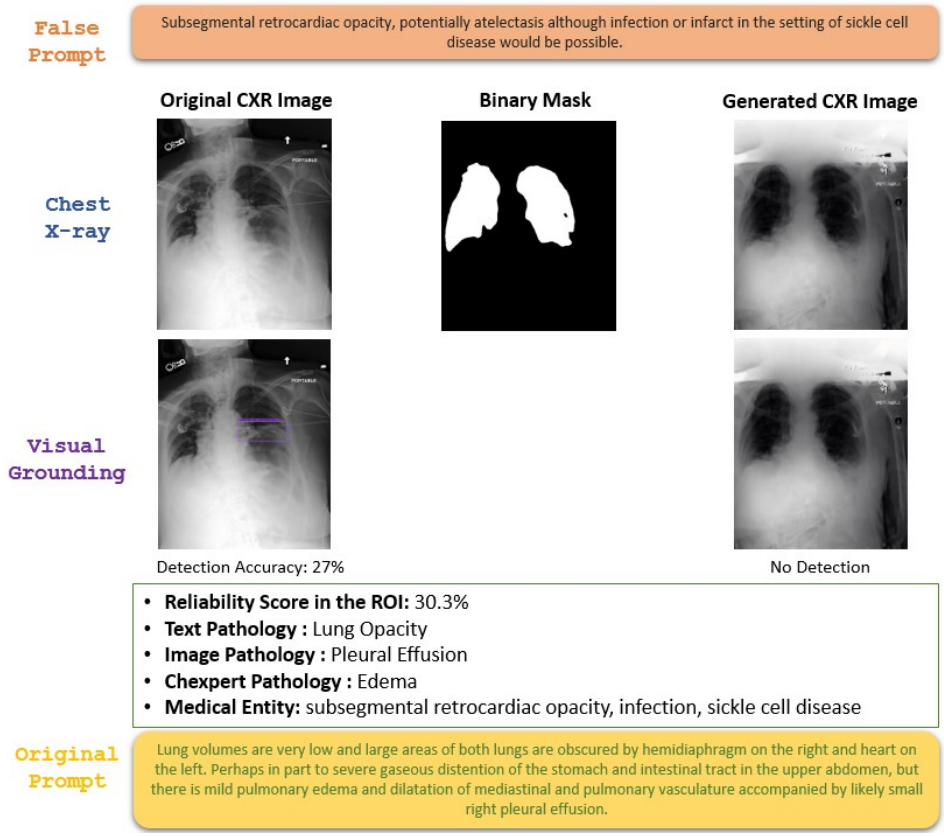


Figure 4.11: Case 4: A Success Case of Using False Report

the prompt, but no ROIs are detected when applying visual grounding on the generated image.

Pathology detection from the text is entirely misaligned with the pathologies identified in the image, and the similarity score between the detected ROI in the original image and the corresponding region in the generated image is notably low. Furthermore, the annotated pathology from Chexpert differs substantially. For instance, while the annotation does not include "pleural effusion" and instead notes only "edema," the original report clearly indicates the presence of "pleural effusion."

This case highlights two key insights: (1) even trusted annotations can occasionally lack accuracy or completeness, and (2) the model demonstrates robustness by not incorrectly aligning this mismatched prompt with the given CXR image. This underscores the importance of incorporating our VICCA approach to ensure reliability and trustworthiness in automated systems.

4.5.5 Case 5: No Pathology

In our final case in Figure 4.12, we examine an instance where the original annotation indicates no pathology. While the report describes an ROI in the image, our visual grounding model

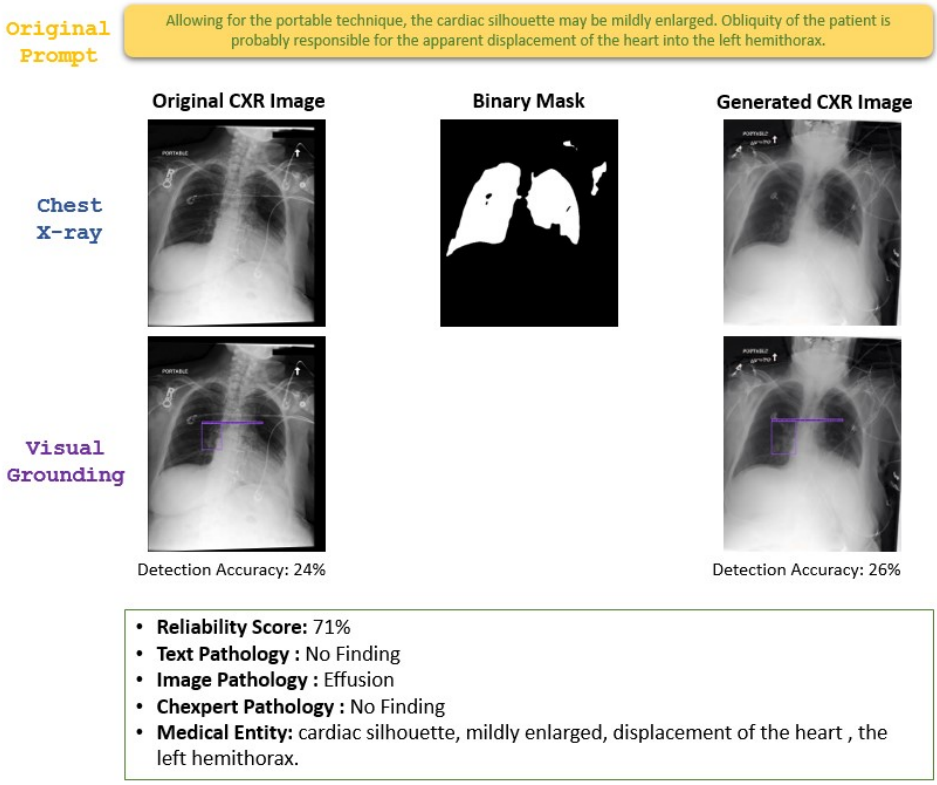


Figure 4.12: Case 5: No Pathology

detects an anomaly within the ROI in both the generated and original images. The reliability score indicates a high degree of similarity, but the detection accuracy score is notably low. Additionally, our image classification model identifies a pathology for this image.

Upon closer inspection of the text prompt, we observe a probable diagnosis linked to the ROI. Although the official annotation suggests no pathology, the findings from the VICCA model highlight a potential ambiguity. This case could be considered inconclusive, as the report itself is somewhat ambiguous, complicating the decision-making process.

4.5.6 Error Propagation and Score Interpretability

While the case studies above illustrate different scenarios of alignment between reports and images, it is also important to acknowledge potential error propagation within the VICCA pipeline. Since VICCA is a multi-stage framework, any failure at the module level can influence the final alignment score.

Error propagation across modules. If the visual grounding model fails to detect an existing pathology (false negative) or erroneously highlights an unrelated region (false positive), the resulting VICCA score may incorrectly penalize or validate the report. Likewise,

if the conditional diffusion model hallucinates structures not supported by the report, the comparison between original and synthetic images could wrongly suggest inconsistencies. In such cases, the error originates from the module rather than from the report itself. Although quantifying the exact frequency of these component-level errors was beyond the scope of this chapter, acknowledging their impact is essential. Future work could systematically analyze error propagation by testing VICCA under controlled perturbations of each module.

Interpretability of VICCA scores. Another limitation relates to the interpretability of VICCA's numerical scores. For example, a similarity score of 80% may suggest a good match between report and image, but in the absence of ground truth expert ratings it is difficult to calibrate what constitutes a "good" versus "poor" match. One possible direction is to leverage existing resources such as the CheXbert-annotated subset of MIMIC-CXR, which contains expert-labeled pathology mentions, or to design a smaller-scale radiologist evaluation study. By comparing VICCA scores with expert judgments, thresholds could be empirically calibrated, thereby improving the interpretability and clinical utility of the framework.

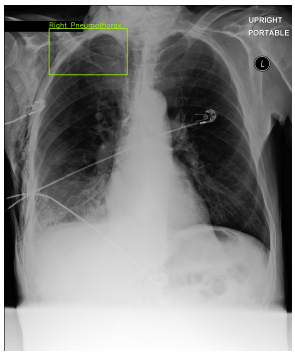
4.5.7 Anatomical Error Case Study

In clinical practice, errors in report writing may arise when the anatomical location of a pathology is incorrectly stated. For example, a radiologist or an automated system may mistakenly describe a finding in the "left lung" when it is actually located in the "right lung." To investigate whether VICCA can capture such inconsistencies, we systematically modified the anatomical references in a subset of reports (e.g., switching left $\leftrightarrow$ right or altering positional terms such as *basilar*, *apical*, or *middle*).

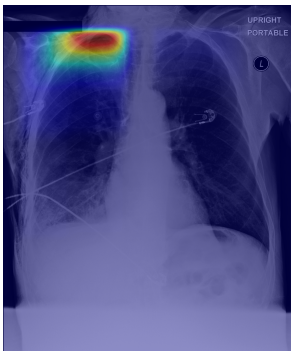
Figure 4.13 illustrates a representative case. The original report states: *"Right rib fractures with a small right apical pneumothorax. Right basilar opacity could represent an effusion or hemothorax."* The annotated ground truth confirms only the presence of a small right apical pneumothorax.

When the phrase was deliberately altered to *"Left rib fractures with a small left apical pneumothorax,"* the visual grounding model produced weak and uncertain detections (around 30%) in the left apical region, reflecting the uncertain existence of the pathology there. VICCA's evaluation gave a similarity score of 78% with the correct prompt and only 43.06% with the switched prompt, showing a clear drop in reliability.

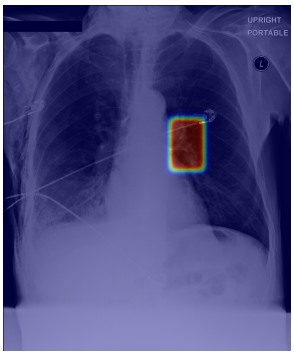
We extended the analysis to the second part of the same report (*"Right basilar opacity could represent an effusion or hemothorax"*), which did not have explicit annotations. With the correct phrase, the model focused on the right basilar region. When the phrase was altered to "left basilar opacity," the model shifted focus accordingly, though with reduced certainty. VICCA yielded 80.12% similarity with the correct phrasing versus 72.7% with the switched one. Additional experiments showed that modifying severity terms (e.g., "**right small basilar opacity**" versus "right basilar opacity") was reflected in the generation and evaluation,



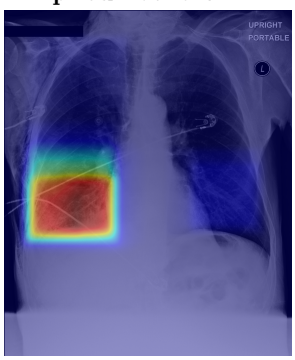
CXR Image with **Right** rib fractures and a small **right** apical **pneumothorax**.



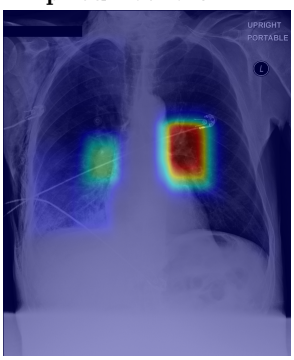
Visual Grounding with "**Right** rib fractures with a small **right** apical **pneumothorax**."



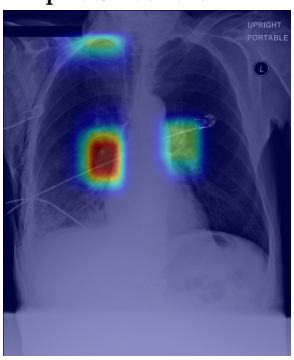
Visual Grounding with "**Left** rib fractures with a small **left** apical **pneumothorax**."



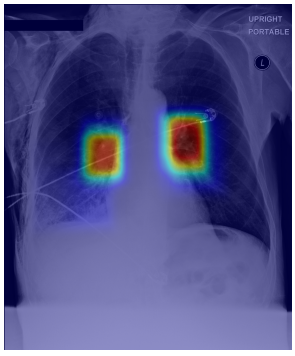
Visual Grounding with "**Right** basilar opacity could represent an effusion or hemothorax."



Visual Grounding with "**Left** basilar opacity could represent an effusion or hemothorax."



Visual Grounding with "**Right small** basilar opacity could represent an effusion or hemothorax."

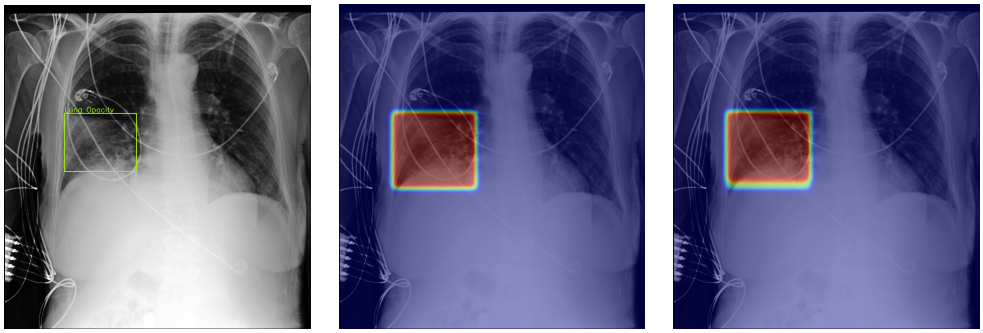


Visual Grounding with "**Middle** basilar opacity could represent an effusion or hemothorax."

Visual Grounding with "**Upper** basilar opacity could represent an effusion or hemothorax."

Figure 4.13: Case study of anatomical switching errors and their impact on grounding and VICCA reliability scores.

whereas removing location markers (e.g., using only "middle basilar opacity") caused ambiguous detections across both sides.



CXR Image with **Right** lung base opacity.

Visual Grounding with "opacity at the **right** lung base."

Visual Grounding with "opacity at the **left** lung base."

Figure 4.14: Case study of location-switching error for lung base opacity.

A second example is shown in Figure 4.14. The original report describes *"opacity at the right lung base."* When this was switched to "opacity at the left lung base," the grounding model still pointed to the correct region but lost specificity, labeling it generically as "opacity lung base." The localization accuracy dropped from 52.06% with the correct side to 38.55% with the switched side. VICCA similarly decreased from 71.28% to 50.44% under the incorrect phrasing.

These case studies demonstrate that VICCA, while not explicitly designed to detect linguistic errors in report phrasing, shows sensitivity to anatomical inconsistencies. By comparing generated versus original image evidence, VICCA reliability scores drop significantly when the location specified in the text is inconsistent with the actual pathology, suggesting its potential to highlight overlooked or erroneous statements in reports.

While the grounding model on its own can already act as a tool to flag such discrepancies, for example, by revealing low-confidence or misplaced detections when anatomical terms are switched, its integration within VICCA provides a more robust mechanism. The joint evaluation with the generative model amplifies these signals: if both grounding and generation follow the same incorrect phrasing, the final alignment score will still decrease when compared against the original image, thereby exposing inconsistencies that might otherwise go unnoticed.

This highlights an important extension of VICCA beyond its primary role as a reliability scoring framework. Not only can it quantify the consistency between reports and images, but it also offers a pathway for **error detection and quality control** in radiology reporting, whether the text originates from a human radiologist or an AI system.

4.6 Textual Complexity

An important clarification concerns the handling of textual complexity in radiology reports. As emphasized in Chapters 1 and 2, free-text reports often include modifiers such as severity ("mild," "moderate," "severe"), negations ("no evidence of pneumonia"), or uncertainty ("could represent"), all of which influence the interpretation of findings.

Within the scope of this chapter, VICCA's generation and grounding components primarily validate the **presence/absence** of pathologies and their spatial localization. Subtle linguistic attributes such as severity levels or negated expressions are not explicitly modeled here. Instead, we preprocess reports into structured entities that focus on the main pathology–region pairs, which are the essential inputs to both the grounding model and the conditional diffusion generator.

These additional attributes are not ignored, but are deferred to the semantic similarity metric introduced later in Chapter 5 (MCSE). That module is designed to capture semantic richness beyond binary presence/absence, including handling of negations and modifiers. In this way, VICCA integrates two complementary perspectives: (1) grounding and generation for spatial alignment, and (2) semantic similarity scoring for nuanced textual attributes.

Clarifying this distinction ensures that readers clearly understand the role of each module: the current chapter focuses on preserving spatial and anatomical fidelity, while subsequent chapters address the linguistic complexity of free-text reports. In particular, we later incorporate the entity extraction component of our MCSE model to preprocess reports, removing non-existent or negated pathologies. This step enriches localization quality and reduces the risk of hallucinated findings, thereby strengthening the overall reliability of the framework.

4.7 Generalizing VICCA Beyond Chest X-rays

While our work focuses on chest X-rays, the VICCA framework is not inherently limited to this modality. Its modular architecture enables potential adaptation to other medical domains, such as brain MRI for acute stroke assessment. For instance, a recently published dataset [70] provides brain MRI scans of stroke patients, including lesion volume data and detailed clinical annotations.

By fine-tuning VICCA on such MRI datasets, the visual grounding module could be trained using lesion segmentation maps aligned with corresponding clinical reports to enable visual interpretation of findings. Simultaneously, the text-to-image diffusion module could be adapted to synthesize pathology-aware MRI images, serving as a visual proxy for report validation. This would, however, require a dedicated text encoder trained on MRI-specific clinical language, as well as an image encoder suited for the MRI modality. Additionally, VICCA's diffusion component would need a pre-processing step to generate binary brain masks to preserve anatomical structure during image synthesis.

Together, these adaptations could enable VICCA to offer dual-score validation, quantifying both localization and semantic consistency, thereby advancing the transparency and interpretability of AI-generated reports in neuroimaging contexts [61]. Nonetheless, these extensions remain hypothetical and warrant further empirical validation.

4.8 Conclusion

Our VICCA framework demonstrates that integrating multiple AI models can substantially improve the interpretability of chest X-ray analysis by providing visual explanations paired with reliability scores and by introducing quantitative metrics to assess region-level alignment through generative models. Each component of the pipeline, as well as the overall system, has been validated using standard metrics.

Until now, the input text, whether authored by a radiologist or generated by an AI model, has been treated uniformly. However, one of the central motivations of this work is to improve the clarity and trustworthiness of AI-generated medical reports. Unlike expert-written reports, which inherently benefit from clinical authority, AI-generated reports must establish credibility through measurable and interpretable mechanisms. VICCA addresses this gap by offering a post-hoc validation tool to evaluate whether such reports are visually grounded and semantically aligned with the corresponding medical images.

In the following chapters, we shift our attention to the textual dimension of radiology data, with a particular focus on radiology reports. We first introduce our proposed semantic evaluation metric, MCSE, which is designed to overcome limitations in current report generation models, especially their lack of reliable mechanisms for capturing semantic fidelity. Building on this, we present a comparative study of state-of-the-art models for CXR report generation, highlighting their relative strengths and weaknesses. Finally, we evaluate these models using both VICCA and MCSE, providing a joint assessment that not only benchmarks their reliability but also demonstrates the complementary value of our framework in aligning textual and visual validation.

Chapter 5

Radiology Textual Report Generation Assessment

Contents

In this chapter, we critically examine the limitations of existing evaluation metrics for assessing AI-generated medical reports. Many of the commonly used metrics, originally developed for natural language generation tasks, are effective at measuring surface-level fluency and lexical similarity but fall short in clinical applications, where accuracy, semantic fidelity, factual correctness, and domain-specific relevance are essential. We highlight this gap through comparative examples involving both medical and non-medical texts, demonstrating how conventional metrics can produce misleading evaluations when applied to clinical narratives. Despite these shortcomings, such metrics remain the primary tools for validation in medical report generation research. To address this challenge, we conclude the chapter by introducing a new metric, specifically designed for radiology, that more reliably captures the quality, semantic accuracy, and clinical validity of generated content.

5.1 Introduction

Radiology reports are essential for clinical decision-making, serving as structured narratives that translate complex medical images into diagnostic insights. Traditionally authored by radiologists, these reports require careful interpretation of imaging data to identify pathologies and communicate findings using precise and domain-specific language. A comprehensive

report may include a patient's general health status, references to prior examinations, and the radiologist's assessment of anatomical and pathological observations.

However, producing such detailed reports is both time-consuming and highly dependent on expert knowledge. This creates a critical bottleneck in high-demand settings such as emergency departments and mass screening programs, where rapid yet accurate diagnostics are essential. These challenges have motivated the adoption of AI-based systems capable of automatically generating radiology and clinical text reports directly from medical images.

Evaluating such reports, however, is substantially more complex than evaluating natural image captions. In natural image domains, the goal is often to produce descriptions that demonstrate syntactic fluency (e.g., grammatically correct sentences such as "A man riding a bicycle") and semantic adequacy (e.g., capturing the gist of the image, such as identifying people, objects, or actions). These qualities can be reasonably assessed using automatic metrics such as BLEU, ROUGE, and CIDEr.

In contrast, radiology reports must adhere to domain-specific standards, such as the use of precise anatomical and pathological terminology, alignment with clinical reporting guidelines (e.g., sectioning into "Findings" and "Impression"), and the correct interpretation of subtle visual features that may have diagnostic implications. For example, the difference between "no acute cardiopulmonary abnormality" and "mild cardiomegaly with vascular congestion" is clinically significant, even if the former may seem more fluent or concise.

A generated radiology report may appear linguistically sound grammatically correct and well-structured yet still be misleading or factually incorrect. For instance, describing a chest X-ray as showing "clear lung fields" when there is subtle bilateral infiltrate could result in a missed diagnosis of pneumonia. Such discrepancies underscore the importance of evaluating medical report generation with metrics that prioritize clinical accuracy and semantic fidelity over surface-level language quality.

Existing evaluation metrics fall short in capturing the complexities of medical report generation. Lexical similarity metrics, such as BLEU, ROUGE, and CIDEr, measure word overlap between the generated and reference texts, rewarding exact n-gram matches while penalizing valid paraphrases. For instance, a model-generated sentence stating "The heart appears enlarged" may be penalized for differing from the reference "There is mild cardiomegaly," despite conveying the same clinical meaning. This rigid focus on surface-level wording fails to account for the diversity and nuance of medical expression.

To offer a more clinically informed view, domain-specific metrics like CheXpert F1 [47] and RadGraph F1 [48] attempt to evaluate reports based on extracted medical entities or structured graph representations of findings and anatomical locations. These metrics are designed to better reflect the factual correctness of a report from a clinical perspective. For example, RadGraph F1 compares whether both the generated and reference reports mention "consolidation" in the "right lower lobe," rather than whether the exact words match. However, these metrics still face limitations in granularity (e.g., failing to capture modifiers like "mild" vs. "severe"), dependency on high-quality structured annotations, and reduced generalizability

to datasets without such annotations.

Moreover, their reliance on structured ground truth significantly limits their scalability in real-world clinical settings, where such annotations are often unavailable or inconsistently labeled. A more robust solution must therefore move beyond both lexical overlap and annotation-heavy pipelines to assess the true semantic and clinical alignment of generated content.

In this next section, we explore conventional n-gram-based metrics such as BLEU, ROUGE, and CIDEr, which, although widely used in natural language generation, fail to capture the semantic and clinical nuances of radiology reports. We then review domain-specific metrics like CheXpert F1 and RadGraph F1 that introduce clinical knowledge into the evaluation process, while still facing challenges related to annotation dependence, generalizability, and granularity.

Building on this analysis, we motivate the need for more semantically grounded, entity-aware evaluation strategies tailored to the medical domain. This motivation led to our second core contribution, **C2** (Contribution 2): the development of *Medical Corpus Similarity Evaluation (MCSE)*, a novel evaluation metric that captures semantic similarity between generated and reference reports at the entity level, while preserving clinical relevance. MCSE bridges the gap between surface-level linguistic fidelity and deeper clinical correctness.

Through this chapter, we pose the following research question: **Q1.** How can we reliably evaluate the quality and clinical relevance of AI-generated radiology reports in the absence of structured annotations or expert feedback?

To answer this question, we propose MCSE as a robust, scalable alternative for assessing factual alignment and semantic accuracy. This work was published in our paper [86], and forms a central component of the evaluation pipeline proposed in this thesis. As such, this chapter links directly to the implementation and application of our third major contribution, **C3** (Contribution 3), which focuses on multimodal validation and integration within the VICCA reliability framework, presented in subsequent chapters.

5.2 Evaluation of Medical Report Generation

Evaluation of captioning models typically relies on n-gram-based similarity metrics such as BLEU [84], METEOR [7], ROUGE [66], CIDEr [106], and SPICE [4]. These metrics quantify overlap between generated and reference captions, either through direct n-gram matching (BLEU, ROUGE), semantic alignment using synonym sets (METEOR), or more sophisticated weighting of rare yet informative terms (CIDEr). SPICE, in particular, attempts to capture semantic content by comparing scene graphs, structured representations of objects and their relationships, between the reference and generated descriptions.

While these metrics provide convenient and automated tools for benchmarking model performance on large-scale natural image datasets such as MS COCO or nocaps, they are

not without limitations. One major drawback is their inability to account for the diversity of valid captions. A generated caption can be factually correct and semantically appropriate yet receive a low score if it diverges lexically from the reference annotations. For example, consider a reference caption that states: *"A man is riding a bicycle on a city street"*. A generated caption such as *"A cyclist moves through urban traffic"* expresses the same content with equivalent meaning, but due to differences in vocabulary and phrasing, traditional metrics like BLEU or ROUGE would assign a low similarity score. In medical contexts, this problem is exacerbated, as synonymous clinical terms (e.g., "pulmonary infiltrates" vs. "lung opacities") may differ lexically yet carrying identical diagnostic meaning. This reliance on surface-level overlap limits these metrics' ability to evaluate deeper semantic fidelity, especially in high-stakes domains like radiology. Moreover, these limitations become critical when evaluating AI-generated radiology reports, where factual correctness, clinical relevance, and contextual appropriateness are essential for safe and trustworthy deployment in clinical workflows.

Table 5.1: Evaluation of different metrics on natural and medical text examples.

Natural Text		BLEU	ROUGE	METEOR	CIDEr	SPICE
Reference	The cat sat on the mat.	0.00063	0.71	0.79	0.0	1.0
Candidate	The cat is sitting on the mat.					
Medical Text						
Reference	Pulmonary edema, cardiomegaly, likely pleural effusions.	0.00003	0.22	0.15	0.0	0.57
Candidate	Moderately severe bilateral pulmonary edema with no large pleural effusion.					

Table 5.1 illustrates how lexical rigidity in traditional evaluation metrics can lead to misleading assessments, both in natural and medical contexts. Take, for example, the reference caption *"The cat sat on the mat"* and the candidate caption *"The cat is sitting on the mat"*. Semantically, both sentences describe the same event, with only minor differences in tense and phrasing. Table 5.2 shows the n-gram tokenization for both texts. However, 3-gram and 4-gram-based metrics, such as BLEU and CIDEr, assign near-zero similarity because of the lack of exact n-gram matches, despite the fact that the core meaning remains intact. While METEOR and ROUGE-2 offer somewhat more flexible evaluations by considering synonymy and partial matches, SPICE provides the most accurate score, as it explicitly focuses on capturing the semantic structure of the captions.

This discrepancy becomes even more pronounced in medical text, where lex-

Table 5.2: Comparison of n-gram tokenization for a natural text and a medical text. The table shows unigrams, bigrams, trigrams, and 4-grams for both a reference and candidate caption. The first pair of texts corresponds to a natural description of a cat's action, while the second pair pertains to a medical diagnosis involving pulmonary edema and pleural effusion.

Reference: The cat sat on the mat. Candidate: The cat is sitting on the mat.	
Unigrams:	"the", "cat", "sat", "on", "the", "mat", "." "the", "cat", "is", "sitting", "on", "the", "mat", "."
Bigrams:	"the cat", "cat sat", "sat on", "on the", "the mat", "mat ." "the cat", "cat is", "is sitting", "sitting on", "on the", "the mat", "mat ."
Trigrams:	"the cat sat", "cat sat on", "sat on the", "on the mat", "the mat ." "the cat is", "cat is sitting", "is sitting on", "sitting on the", "on the mat", "the mat ."
4-grams:	"the cat sat on", "cat sat on the", "sat on the mat", "on the mat ." "the cat is sitting", "cat is sitting on", "is sitting on the", "sitting on the mat", "on the mat ."
Reference: Pulmonary edema, cardiomegaly, likely pleural effusions. Candidate: Moderately severe bilateral pulmonary edema with no large pleural effusion.	
Unigrams:	"pulmonary", "edema", "cardiomegaly", "likely", "pleural", "effusions." "moderately", "severe", "bilateral", "pulmonary", "edema", "with", "no", "large", "pleural", "effusion."
Bigrams:	('pulmonary', 'edema,'), ('edema,', 'cardiomegaly,'), ('cardiomegaly,', 'likely'), ('likely', 'pleural'), ('pleural', 'effusions.') ('moderately', 'severe'), ('severe', 'bilateral'), ('bilateral', 'pulmonary'), ('pulmonary', 'edema'), ('edema', 'with'), ('with', 'no'), ('no', 'large'), ('large', 'pleural'), ('pleural', 'effusion.')
Trigrams:	('pulmonary', 'edema,', 'cardiomegaly,'), ('edema,', 'cardiomegaly,', 'likely'), ('cardiomegaly,', 'likely', 'pleural'), ('likely', 'pleural', 'effusions.') ('moderately', 'severe', 'bilateral'), ('severe', 'bilateral', 'pulmonary'), ('bilateral', 'pulmonary', 'edema'), ('pulmonary', 'edema', 'with'), ('edema', 'with', 'no'), ('with', 'no', 'large'), ('no', 'large', 'pleural'), ('large', 'pleural', 'effusion.')
4-grams:	('pulmonary', 'edema,', 'cardiomegaly,', 'likely'), ('edema,', 'cardiomegaly,', 'likely', 'pleural'), ('cardiomegaly,', 'likely', 'pleural', 'effusions.') ('moderately', 'severe', 'bilateral', 'pulmonary'), ('severe', 'bilateral', 'pulmonary', 'edema'), ('bilateral', 'pulmonary', 'edema', 'with'), ('pulmonary', 'edema', 'with', 'no'), ('edema', 'with', 'no', 'large'), ('with', 'no', 'large', 'pleural'), ('no', 'large', 'pleural', 'effusion.')

ical variability is common and clinical nuance is critical. Consider the reference sentence *"Pulmonary edema, cardiomegaly, likely pleural effusions."* and the candidate

"Moderately severe bilateral pulmonary edema with no large pleural effusion." Although the two texts describe overlapping clinical findings, they exhibit minimal lexical overlap: only a few unigrams match, and there are virtually no matching bigrams or higher-order n-grams. As a result, BLEU, ROUGE, METEOR, and CIDEr all return very low scores, failing to acknowledge that the core pathology, pulmonary edema, is consistently identified. Furthermore, subtle variations like negations ("no large effusion") or modifiers ("moderately severe bilateral") can significantly alter n-gram-based evaluations without reflecting a loss in clinical accuracy. In contrast, SPICE, which parses the captions into scene graphs and evaluates semantic content, better captures this underlying alignment.

These examples highlight the limitations of traditional metrics when applied to domains that prioritize semantic accuracy over lexical similarity. In medical applications, where precise communication of findings is crucial, such metrics risk undervaluing clinically relevant predictions, potentially leading to misguided model development and deployment.

This issue is particularly pronounced in specialized domains like medical image captioning, where the language is highly technical, the vocabulary is domain-specific, and the visual features often contain subtle pathological cues that require expert interpretation. The goal in this context goes beyond merely describing visible elements; it is to accurately identify and communicate clinically significant findings. Consequently, standard captioning metrics may fall short in capturing the true diagnostic value or accuracy of generated reports.

This challenge is especially relevant to our research, where we work with medical text at **various stages** and frequently need to evaluate text similarity. Without reliable metrics that account for semantic nuances and subtle variations, such as modifiers or negations, evaluating feature extraction or ensuring the validity of downstream results becomes problematic. While conventional metrics may appear sufficient for comparative evaluations analysis across different radiology report generation models, we observed that in more fine-grained tasks, such as verifying semantic consistency between text and visual content or validating cross-modal alignment between independently trained models, these metrics often fail to provide meaningful or trustworthy signals. This limitation underscores the need for a more clinically grounded and semantically aware evaluation framework.

To address these challenges, recent research has explored alternative or complementary evaluation methods. These include expert-in-the-loop assessments, clinical correctness evaluations, and alignment-based techniques that assess the relationship between generated text and localized visual features in images. Additionally, model explainability and visual grounding, linking specific words in the caption to corresponding image regions, are increasingly integrated into both training and evaluation, particularly in safety-critical fields like healthcare, where transparency and trust are paramount.

While existing approaches have made progress, a reliable metric for capturing semantic similarity in medical text remains largely absent. To address this gap, we propose a more algorithmic solution designed for the nuances of medical language [86]. Our contribution builds on the core idea of SPICE by using a cosine-based similarity measure but extends it to better accommodate the structure of medical entities. Specifically, we decompose sentences into

clinically meaningful components (entities, modifiers, and negations) and compute similarity by weighing each entity according to its contextual importance and its relationship with other components. This allows us to more faithfully capture the semantic fidelity of generated reports. The following section introduces this metric in detail.

5.3 Semantic Textual Similarity Assessment in Chest X-ray Reports Using a Domain-Specific Cosine-Based Metric

In light of the existing challenges, we now introduce our proposed metric, **Medical Corpus Similarity Evaluation (MCSE)**, designed to provide a clinically meaningful, semantically grounded assessment of text similarity in radiology reports.

5.3.1 Overview of Prior Works

Several prior approaches have attempted to move beyond conventional n-gram-based metrics by incorporating domain-specific knowledge. For example, CheXbert vector similarity [31, 102] relies on binary disease labels extracted from chest X-ray reports but is limited by its fixed label space and dependency on structured ground truth. RadGraph F1 [120, 48] evaluates entity overlap using a predefined schema but does not account for semantic variation or detailed entity descriptions. More recent efforts like SciBERT [9] demonstrate the promise of transformer-based embeddings for semantic similarity, though they often require extensive post-processing and remain computationally expensive.

Table 5.3: Example of a report generated by the CXR-RePaiR model. Highlighted text represents inconsistencies and redundancies present in this particular output.

E.g: AP chest compared to —: Severe cardiomegaly and mediastinal widening due to vascular engorgement and fat deposition are longstanding. Moderate right pleural effusion is probably responsible for a right lower lobe collapse, although only slightly smaller left pleural effusion was shown to produce less atelectasis at the lung base on the chest CT, —: Pulmonary and mediastinal vascular congestion have increased since — and there is a new region of opacification in the right mid lung, which could represent early edema. Tracheostomy tube in standard placement. Right PIC line ends low in the SVC. No pneumothorax. AP chest compared to most recent prior chest radiograph, —: Allowing for differences in patient positioning, moderate to severe enlargement of the cardiac silhouette has not changed appreciably since —. New pericardial drainage catheter projects over the cardiac silhouette. Mediastinal veins are not distended. Radiographically at least there is no evidence of continued tamponade physiology. Left lower lobe atelectasis and small left pleural effusion are new. Right lung is clear, and there is no right pleural effusion or pneumothorax on either side.

Despite these innovations, a scalable, interpretable, and clinically sensitive similarity metric remains absent, particularly one that accounts for modifiers, negations, and contextual richness of medical entities.

5.3.2 Methodology

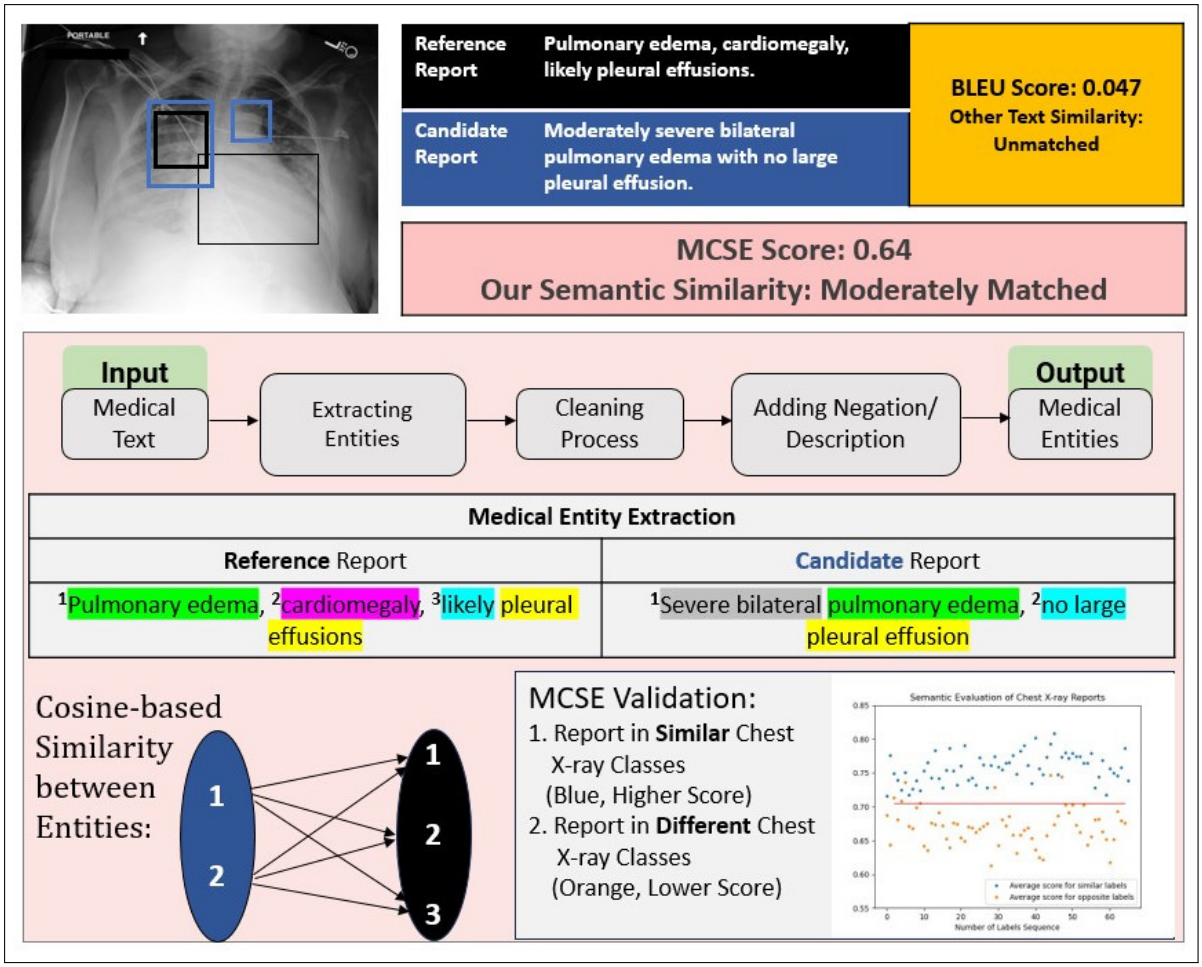


Figure 5.1: An overview of our Semantic Textual Similarity Assessment Evaluation.

We developed a novel metric for Medical Corpus Similarity Evaluation (MCSE), by exclusively extracting key medical entities and employing a pretrained BERT model to assess the semantic similarity of these entities within chest X-ray reports. Report generation methods often produce lengthy and redundant texts, such as the example in Table 5.3, making direct comparison with ground truth reports challenging when evaluating the entire text. This targeted approach allows BERT to concentrate solely on important information and reduces the computational load during comparison. Importantly, our methodology goes beyond extracting main entities, we also consider the negations and detailed descriptions associated with the primary medical entities in chest X-ray reports. Our MCSE metric consists of two essential steps as illustrated in Figure 5.1:

1. Clinical Entity Extraction.
2. Domain Similarity Evaluation.

5.3.2.1 Clinical Entity Extraction

The most important part of comprehending semantic similarity evaluation in text relies on identifying the key elements, often referred to as clinical entities, within medical texts. These entities typically fall into categories related to anatomical body parts, symptoms, laboratory equipment, and diagnoses. Each category is typically signaled by certain words within a sentence. However, there are additional words that precede or follow these main entities, offering descriptions.

Table 5.4: In the left column there is an example of medical text. In the right column, there are clinical entities extracted using the Scispacy model without any cleaning process, and in the middle column, there are clinical entities extracted using our method.

Medical Text	Extracted Entities using our method	Extracted Entities using Scispacy [79]
1. Interval clearance of left basilar consolidation. 2. Patchy right basilar opacities, which could be seen with minor atelectasis, but given the context clinical correlation is suggested regarding any possibility for recurrent or new aspiration pneumonitis at the right lung base. 3. Increased new interstitial abnormality, suggesting recurrence of fluid overload or mild-to-moderate pulmonary edema; aspiration could also be considered. Inflammation associated with atypical infectious process is probably less likely given the waxing and waning presentation.	fluid overload, inflammation, aspiration pneumonitis, minor atelectasis, mild to moderate pulmonary edema, left basilar consolidation, patchy right basilar opacities, interstitial abnormality	Interval, clearance, left basilar, consolidation, Patchy, right basilar, opacities, minor, atelectasis, clinical, recurrent, aspiration, pneumonitis, right lung base, Increased, interstitial abnormality, recurrence, fluid, overload, mild-to-moderate pulmonary edema, aspiration, Inflammation, associated with, atypical, infectious process, waxing, waning, presentation

To address these complexities, we employ the Scispacy model [79] for extracting primary clinical entities from medical text using the embedded clinical dictionary in this model (BC5CDR: a corpus comprising 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases, and 3116 chemical-disease interactions [64]). Subsequently, we automatically process

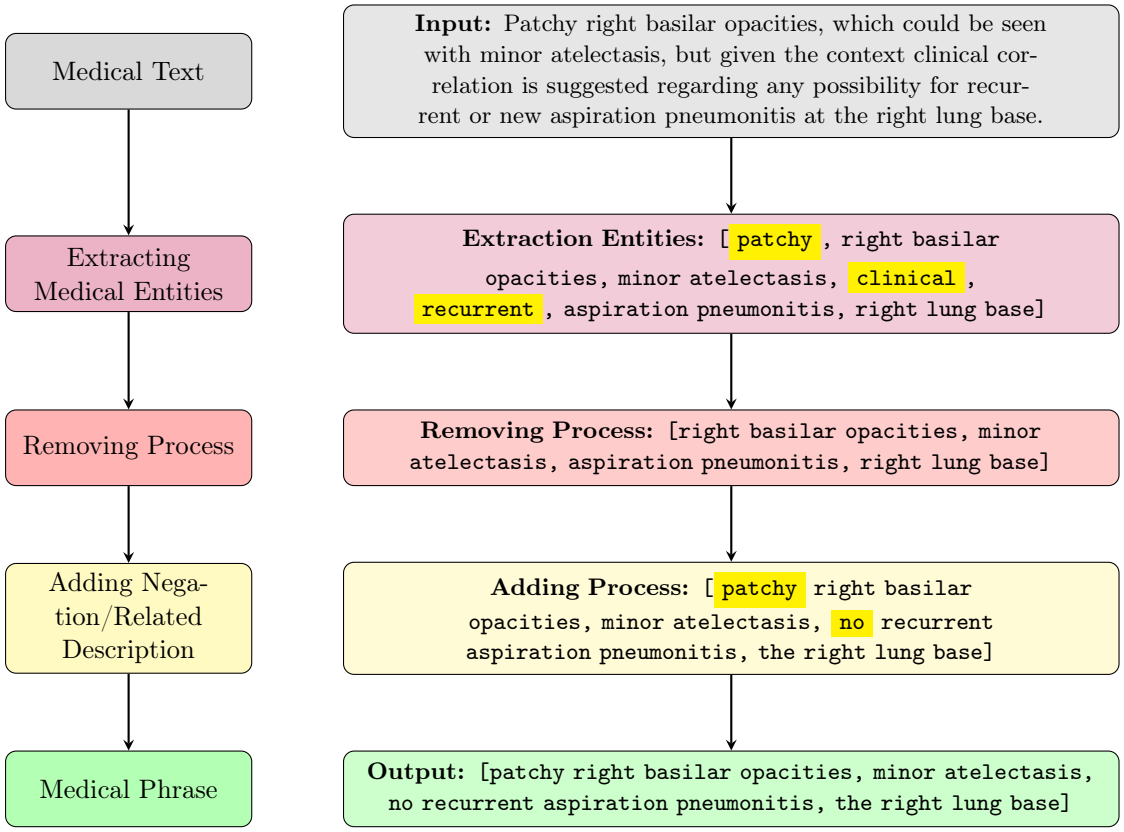


Figure 5.2: A step-by-step overview of the process of entity extraction using our MCSE method.

the entire text to identify associated negations and adjectives related to these key entities. These elements are then integrated to provide a comprehensive representation of the considered text. In the context of this research, the category of laboratory equipment is deliberately excluded, aligning with the specific focus of our application. Table 5.4 presents an example of medical text and the extracted entities using our method and the Scispacy method without any cleaning process. While we employ the Scispacy model for initial entity extraction, it is evident that this model alone may not suffice. An additional automated post-processing step is needed to refine and integrate related entities. The post-processing steps involve eliminating a single adjective or non-medical entities, excluding entities categorized as lab equipment, identifying and adding the relevant adjective to the remaining medical entities, including the existing negation into these primary entities, and screening out any reported diagnostic procedure terms. Figure 5.2 demonstrates the step-by-step process of our method. These processes are essential to ensure that the final output is presented as a cohesive set of primary medical entities, ready for practical use.

Table 5.5: An example of a medical similarity score between entities. Each score is calculated from equation (5.1), with the final row S_i being computed using equation (5.2). The scores highlighted in blue indicate the maximum value within each respective column.

Reference: 1. Interval clearance of left basilar consolidation. 2. Patchy right basilar opacities, which could be seen with minor atelectasis, but given the context clinical correlation is suggested regarding any possibility for recurrent or new aspiration pneumonitis at the right lung base. 3. Increased new interstitial abnormality, suggesting recurrence of fluid overload or mild-to-moderate pulmonary edema; aspiration could also be considered. Inflammation associated with atypical infectious process is probably less likely given the waxing and waning presentation.				
Candidate: Stable multiple bilateral pulmonary masses and right middle lobe collapse due to hilar adenopathy.				
		Candidate Medical Entities		
		pulmonary masses	right middle lobe	hilar adenopathy
Reference Medical Entities	fluid overload	0.61	0.49	0.45
	inflammation	0.64	0.48	0.55
	aspiration pneumonitis	0.65	0.39	0.50
	minor atelectasis	0.62	0.47	0.53
	mild to moderate pulmonary edema	**0.78**	0.31	0.51
	left basilar consolidation	0.52	**0.66**	0.32
	patchy right basilar opacities	0.64	**0.66**	0.49
	interstitial abnormality	0.69	0.63	**0.59**
S_i		0.548	0.563	0.545

5.3.2.2 Domain Similarity Evaluation

Having successfully extracted and shifted our focus to the primary entities within the medical corpus, the next step involves assessing their semantic similarities by assigning corresponding scores.

After processing entity extraction, we calculate a similarity score for the sequences of entities. Let $T = (t_1, \ldots, t_N)$ denote the entities from the reference text and $\hat{T} = (\hat{t}_1, \ldots, \hat{t}_M)$ the entities from the candidate (generated) text.

We first identify all entities that match exactly between the two corpora and collect them in the set

$$C = \{\, (i, j) \mid t_i = \hat{t}_j \,\},$$

so that $|C|$ is the number of exact matches.

The remaining (non-matching) entities are grouped into two reduced sequences

$$r_1, \ldots, r_{N'} \quad \text{and} \quad \hat{r}_1, \ldots, \hat{r}_{M'},$$

obtained by removing all entities that are part of C from T and $\hat{T}$ respectively. For these unmatched entities, we construct a similarity matrix $Y \in \mathbb{R}^{N' \times M'}$ with entries

$$y_{i,j} = \text{Similarity}(r_i, \hat{r}_j), \tag{5.1}$$

where Similarity$(\cdot, \cdot)$ is computed using spaCy [43] (a BERT-based model) and corresponds to a cosine similarity in a domain-specific embedding space.

For each candidate entity (i.e. for each column j of Y as illustrated in table 5.5), we define a normalized similarity score

$$S_j = \frac{\max_i y_{i,j}}{\max_i y_{i,j} + \overline{y_{i,j}}}, \tag{5.2}$$

where $\max_i y_{i,j}$ is the maximum similarity for column j and $\overline{y_{i,j}}$ is the average over all rows i in that column. Intuitively, S_j is high when there exists at least one reference entity that is strongly similar to the candidate entity $\hat{r}_j$, and low when the similarities are weak or diffuse.

The global similarity score between the two corpora, denoted MCSE, combines the exact matches and the soft matches encoded by S_j:

$$\text{MCSE} := \frac{|C| + \sum_{j=1}^{M'} S_j}{M}, \tag{5.3}$$

where M is the total number of candidate entities in $\hat{T}$. Here, $|C|$ accounts for entities that match exactly between reference and candidate texts, while the sum of S_j captures partial or paraphrastic matches for the remaining entities. Dividing by M normalizes the score to the range $[0, 1]$ and makes it comparable across examples.

For instance, Table 5.5 provides an example of the probable similarity score that two sets of entities can receive. These entities have been extracted using our medical entity extraction procedure.

In the table, the two corpora received a score of 0.55 according to our MCSE metric. However, the calculated BLEU, ROUGE, METEOR and even SPICE score for them is approximately zero. Upon analyzing the two medical texts, it becomes evident that although the candidate text does refer to the same side of the chest as in the reference text and that both texts indicate the presence of pulmonary edema and pulmonary masses, their overall similarity is relatively limited. The score of 0.55 carries a more meaningful value in this context compared to the nearly zero score generated by lexical similarity.

5.3.3 Validation

While the underlying logic of this metric is reasonable, it is imperative that we validate the results robustly. Given the use of chest X-ray reports for this particular application, we have conducted an extensive search within existing datasets to identify an appropriate validation method. After a comprehensive review of various datasets, we concluded that it would be more effective to conduct separate validations for the different steps of the proposed metric.

5.3.3.1 Clinical Entity Extraction Process

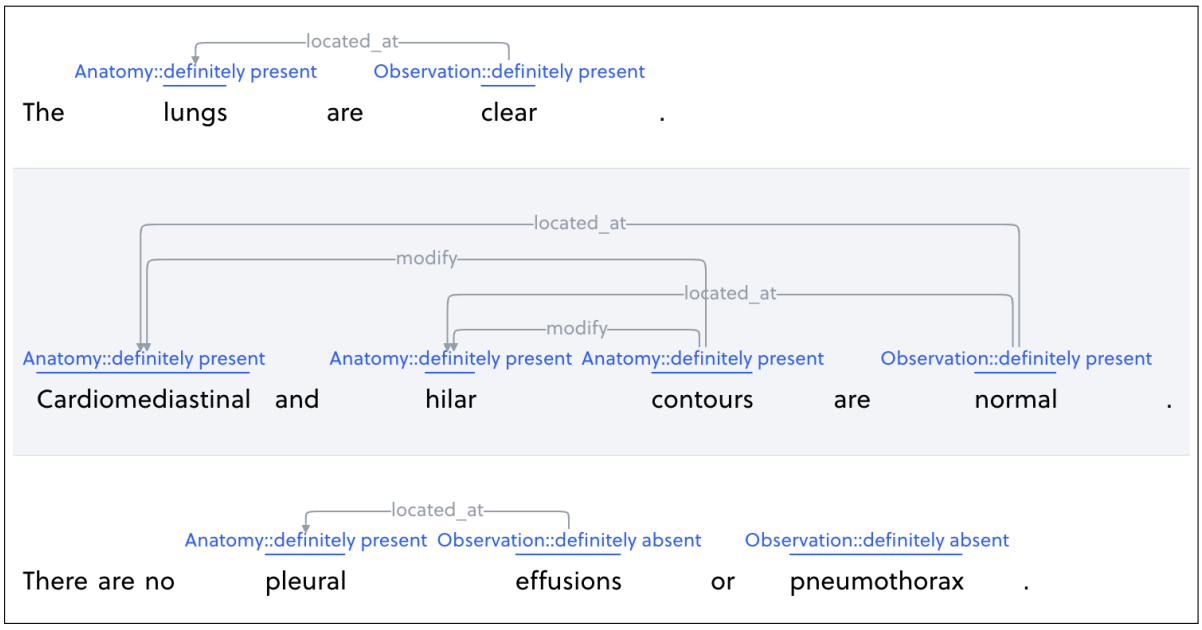


Figure 5.3: Sample report annotated according to the RadGraph schema [48].

In order to rigorously validate our clinical entity extraction process, we employ the RadGraph dataset [48]. This dataset is a valuable resource in which radiologists thoroughly annotated the primary clinical entities in chest X-ray reports as either "definitely present" within the report or "definitely absent". Figure 5.3 illustrates an example of an annotated report following the RadGraph schema. Importantly, in cases where a negation is associated with a particular entity, it is annotated as "definitely absent."

To achieve our validation objectives, we executed our entity extraction process on the reports within this dataset. Subsequently, we compare the number of similar entities extracted through our method with the annotations provided by radiologists, particularly focusing on the two categories of "definitely present" and "definitely absent". This systematic comparison allows us to assess the accuracy and effectiveness of our clinical entity extraction methodology in the context of chest X-ray reports, aligning with radiological standards. Throughout the validation process, covering all reports in our study, our method consistently achieves a high level of accuracy. On average, it accurately recognizes 75% of entities marked as "definitely present" and successfully identifies 76% of entities labeled as "definitely absent". In our entity extraction process, we deliberately omit anatomical entities like "chest" or "lung," as they are redundant to the chest X-ray application and do not contribute significantly to the process. This selective exclusion is one of the factors contributing to the approximately 75% accuracy in our results. Nevertheless, these results affirm the reliability and consistency of our methodology.

5.3.3.2 Domain Similarity Score

Table 5.6: A sample table featuring Chexpert labels (1. Atelectasis, 2. Cardiomegaly, 3. Consolidation, 4. Edema, 5. Enlarged Cardiomediastinum, 6. Fracture, 7. Lung Lesion, 8. Lung Opacity, 9. No FINDING, 10. Pleural Effusion, 11. Pleural Other, 12. Pneumonia, 13. Pneumothorax, 14. Support Devices) extracted from chest X-ray reports of five patients (Subject ##) from the MIMIC-CXR database [50].

Subject ##	Atelectasis	Cardiomegaly	Consolidation	Edema	Enlarged Cardiomediastinum	Fracture	Lung Lesion	Lung Opacity	No Finding	Pleural Effusion	Pleural Other	Pneumonia	Pneumothorax	Support Devices
01								0		1		1	-1	
02							1					1		
03									1			0		
04		1	0					-1					0	1
05	1									1				

In contrast to the initial phase of clinical entity extraction, validating the domain similarity score is more challenging. The scoring system itself is more controversial and subject to debate, and creating an automated validation method, free from reliance on radiologists, necessitates a creative and innovative approach. Nevertheless, through the available tools and databases, we establish a dedicated system for the validation of this scoring method for the application of chest X-rays.

In the chest X-ray application, the MIMIC-CXR dataset [50], is one of the biggest available databases for chest X-ray images and their corresponding reports. Notably, this dataset provides us with Chexpert labels (Medical Observation), including Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, and Support Devices labels [47]. The values of each label are 1 (definitely present), 0 (definitely absent), -1 (ambiguous), or it carries no value at all. Table 5.6 presents a sample of Chexpert labels ex-

Table 5.7: Reports corresponding to the subjects listed in Table 5.6 from the MIMIC-CXR dataset [50].

Subject_##	Report
01	Lung volumes remain low. There are innumerable bilateral scattered small pulmonary nodules which are better demonstrated on recent CT. Mild pulmonary vascular congestion is stable. The cardio mediastinal silhouette and hilar contours are unchanged. Small pleural effusion in the right middle fissure is new. There is no new focal opacity to suggest pneumonia. There is no pneumothorax.
02	A triangular opacity in the right lung apex is new from prior examination. There is also fullness of the right hilum, which is new. The remainder of the lungs are clear. Blunting of bilateral costophrenic angles, right greater than left, may be secondary to small effusions. The heart size is top normal.
03	Mild to moderate enlargement of the cardiac silhouette is unchanged. The aorta is calcified and diffusely tortuous. The mediastinal and hilar contours are otherwise similar in appearance. There is minimal upper zone vascular redistribution without overt pulmonary edema. No focal consolidation, pleural effusion or pneumothorax is present. The osseous structures are diffusely demineralized.
04	The endotracheal tube tip is 6 cm above the carina. Nasogastric tube tip is beyond the GE junction and off the edge of the film. A left central line is present in the tip is in the mid SVC. A pacemaker is noted on the right in the lead projects over the right ventricle. There is probable scarring in both lung apices. There are no new areas of consolidation. There is upper zone redistribution and cardiomegaly, suggesting pulmonary venous hypertension. There is no pneumothorax.
05	A moderate left pleural effusion is new. Associated left basilar opacity likely reflects compressive atelectasis. There is no pneumothorax. There are no new abnormal cardiac or mediastinal contours. Median sternotomy wires and mediastinal clips are in expected positions.

tracted from chest X-ray reports of five patients from the MIMIC-CXR database. The reports corresponding to these subjects are presented in Table 5.7.

Our approach involves two distinct strategies. Firstly, we seek to identify reports sharing the same sequence of labels and values. For instance, we search for reports from subjects with Chexpert label sequences similar to that of Subject_01 in Table 5.6. For these reports with matching label sequences, we proceed to similarity scores computation for each pair of reports. Simultaneously, we identify reports featuring only one or two labels and with a value of "definitely present" for these labels resembling Subject_02 in Table 5.6 and assess the similarity of these reports with the reports with different label sequences. As an example, we calculate the similarity between the reports of Subject_02 and Subject_05 from Table 5.6, given their entirely distinct label sequences. This two-fold method allows us to analyze the semantic similarity scores for both similar and contrasting reports in terms of their labels.

Figure 5.4 presents the results of the two-fold validation for our scoring method. Within the figure, blue dots represent the average scores for the semantic evaluation of reports with similar label sequences, while orange dots show the mean scores for reports with contrasting labels. The red horizontal line within the figure serves as the dividing line distinguishing between

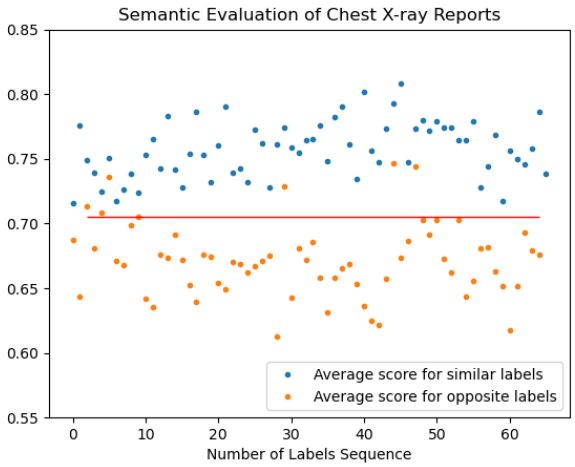


Figure 5.4: Semantic Evaluation of Chest X-ray reports. Each blue dot represents the mean score of semantic evaluation for reports with similar label sequences, while each orange dot signifies the mean score of semantic evaluation for reports with opposing labels. The red horizontal line represents the classification boundary.

similar and opposite evaluations. Upon reviewing these results, it becomes evident that a distinct boundary exists between reports sharing the same clinical diagnoses and those with entirely dissimilar diagnoses. Notably, there are no blue dots below a 70% similarity threshold, whereas six orange dots have scores above 70% across 70 label sequences, which is certainly not very high. Nevertheless, despite this differentiation between similar and opposite evaluations, some level of similarity, exceeding 50%, persists within the opposing category. This can be attributed to the implemented cosine similarity within the medical domain, which introduces a certain bias towards tokens in the same medical domain. Unfortunately, this bias cannot be entirely eliminated, as it plays a substantial role in the evaluation process. However, a clear boundary remains between similar and contrasting reports.

5.3.4 Results and Discussion

Table 5.8: The result of BLEU score of 2-gram for state-of-the-art models and the result of our novel metric on these models outcomes.

Models	BLEU	Our MCSE
R2Gen [20]	0.212	0.71
CXR-RePair [31]	0.069	0.64

In our original application of chest X-ray report generation, we incorporate our metric to assess the outputs of various models. We compare our results with the BLEU scores evaluated by these models, specifically, the CXR-RePaiR [31] and R2Gen [20] models, both being state-of-the-art models for generating chest X-ray reports. Our evaluation focuses on measuring the semantic similarity between the generated reports and the ground truth. Table 5.8 presents the BLEU scores obtained from these models and our metric's semantic evaluation. As anticipated, the BLEU scores are relatively low, signifying a substantial dissimilarity between the generated results and the ground truth for both the CXR-RePaiR and R2Gen models despite being regarded as state-of-the-art models for chest X-ray report generation. These models still employ the BLEU metric for evaluation, primarily due to the scarcity of more suitable metrics and the need for a standardized evaluation process for comparative purposes. Conversely, our metric produces more promising results for both of these models. While our metric's scores align with the BLEU scores, indicating higher scores for both BLEU and our MCSE metric in the case of the R2Gen model compared to the CXR-RePaiR, our metric provides a deeper evaluation. It suggests a degree of similarity to the ground truth rather than outright dissimilarity in BLEU, thus making the generated reports more reliable and trustworthy, which is a crucial advancement in the field.

Table 5.9: A comparative example of using the BLEU score and our adapted metric with medical reference and generated text.

	BLEU	MCSE
Reference Sentence: "Pulmonary edema, cardiomegaly, likely pleural effusions." **Generated Sentence:** "Moderately severe bilateral pulmonary edema with no large pleural effusion."	0.047	**0.64**

Table 5.9 provides an example of medical text generated and evaluated using both a BLEU score and our MCSE metric. It's evident that, according to the BLEU score, these two texts appear vastly different, even though they share the same primary medical entities. However, when we delve into the context, we can notice that "moderately severe" serves as a description for the main entity, "pulmonary edema", in the generated text. Similarly, in the second part of the text, the main medical entity is "pleural effusions", and terms like "likely" and "no large" are used to describe this entity, which may not be identical but share semantic similarities. This subtle context evaluation is precisely what our metric considers, yielding a similarity score of 0.64 for these texts, which we argue is a more accurate reflection compared to the BLEU score of 0.047.

Lastly, the significant benefit of employing this metric lies in its capacity for comparative analysis alongside other evaluation measures. For instance, when examining the outcomes of the BLEU score, with its word-by-word analysis, situations may arise where the results

are totally inaccurate, casting doubt on their reliability, despite the models performing well overall. Integrating the results of our novel MCSE metric into the evaluation process allows us to semantically analyze and ascertain the dependability of the models' textual outputs within the context of medical content.

5.4 Conclusion

In this chapter, we have explored the crucial task of automatic medical report generation, positioning it as a vital link between localized visual understanding and clinically meaningful communication within trustworthy, end-to-end diagnostic support systems. Drawing parallels with natural image captioning helped highlighting the unique challenges presented by the medical domain, where the language is highly technical and the goal extends beyond mere description to accurately identifying and articulating complex pathologies.

A significant challenge identified is the inadequacy of traditional image captioning evaluation metrics, such as BLEU, ROUGE, METEOR, and CIDEr, when applied to medical text. These metrics primarily rely on surface-level lexical matching and n-gram overlap, failing to capture crucial aspects like factual correctness, clinical relevance, and semantic fidelity. As demonstrated with examples, subtle variations in phrasing, modifiers, or negations can drastically reduce scores despite retaining core clinical meaning, while metrics like SPICE offer a better, though still imperfect, focus on semantic structure. This lexical rigidity risks undervaluing clinically appropriate predictions and can mislead model development in safety-critical fields like healthcare.

To address this critical gap, the absence of a comprehensive, general semantic similarity evaluation metric for medical content, we proposed a novel metric: the Medical Corpus Similarity Evaluation (MCSE). Our method builds on the idea of semantic evaluation by focusing on clinically meaningful components. The MCSE metric consists of two essential steps: Clinical Entity Extraction and Domain Similarity Evaluation. Using the Scispacy model and subsequent automated post-processing, our entity extraction method identifies primary medical entities, incorporating associated negations and detailed descriptions to create a comprehensive text representation. The domain similarity evaluation step then assesses the semantic similarities of these extracted entities using a domain-specific cosine-based measure derived from a pretrained BERT model. This process allows us to move beyond simple exact matches and weigh entities based on contextual importance and relationships.

Validation efforts, though challenging, provided support for our approach. Our clinical entity extraction process, evaluated against the RadGraph dataset, demonstrated a high level of accuracy in identifying entities marked as "definitely present" and "definitely absent" by radiologists. Validating the domain similarity score using the MIMIC-CXR dataset and Chexpert labels showed a distinct boundary between reports with similar and contrasting diagnoses, indicating that the metric captures meaningful semantic differences, despite a potential bias towards tokens within the medical domain.

Finally, applying our MCSE metric to evaluate state-of-the-art chest X-ray report generation models like CXR-RePaiR and R2Gen revealed a significant difference compared to traditional BLEU scores. While BLEU scores remained low, indicating substantial lexical dissimilarity, our MCSE metric yielded considerably higher scores, suggesting a greater degree of semantic similarity to the ground truth than previously indicated. This provides a deeper, more accurate reflection of the clinical correctness and dependability of the generated reports.

Integrating metrics like MCSE into the evaluation process allows for a more accurate assessment of semantic fidelity in medical text at each step of the project, which is crucial for building reliable and more interpretable AI systems for healthcare. While this textual semantic evaluation marks a significant step toward interpretable AI, true validation in medical imaging also requires bridging the gap between textual and visual modalities. Evaluating generation quality through the alignment of both textual and visual features can further enhance the reliability of these systems.

In the following chapters, we further develop this multimodal perspective, examining how medical texts can be cohesively interpreted alongside their corresponding images to provide a clinically grounded evaluation of model performance. Several state-of-the-art (SOTA) models for CXR report generation will be reviewed and assessed through the dual lens of our proposed framework: VICCA, which evaluates visual-textual consistency, and MCSE, which measures semantic similarity in clinical context.

Chapter 6

Radiology Report Generation: State of the Art and Comparisons

Contents

This chapter introduces the field of *Radiology Report Generation* (RRG), outlining several state-of-the-art approaches designed to automatically generate diagnostic reports from medical images. Our goal is to critically evaluate these models by highlighting their strengths, limitations, and the current trajectory of AI development in this area of clinical applications. In particular, we examine how these models handle the complex task of converting visual information into accurate and clinically relevant textual descriptions.

To ensure an objective and robust evaluation, we employ our VICCA framework, which uses a dual-scoring mechanism to assess both the reliability and alignment of model outputs. By comparing these models through the lens of VICCA, we aim to validate their performance beyond conventional metrics, focusing on interpretability and clinical trustworthiness.

This broader evaluation is structured across two chapters. The present chapter introduces the fundamental principles and representative models in RRG. And the next chapter applies the VICCA pipeline to objectively assess various RRG models, providing a comprehensive and reproducible framework for benchmarking medical report generation systems.

6.1 Introduction

Recent advances in artificial intelligence (AI), particularly in natural language processing and computer vision, have led to the development of Radiology Report Generation (RRG) systems, models that aim to automatically generate descriptive reports directly from medical images, including chest X-rays. These systems offer the potential to enhance clinical efficiency, reduce reporting delays, and alleviate the burden on medical professionals. While promising, automatic report generation remains a challenging task due to the complexity of medical language, long-form document structure, and the need for semantic accuracy.

Unlike general image captioning models, RRG systems must contend with subtle visual cues, overlapping pathologies, and the strict terminological conventions of clinical reporting. Trustworthy RRG models must not only produce linguistically coherent text, but also ensure that the content aligns with the visual evidence present in the image.

This chapter provides an overview of RRG methodologies. A more detailed comparative analysis, including model architectures, training strategies, and evaluation protocols, is deferred to Appendix D, allowing the main narrative to remain focused while still offering pedagogical depth for interested readers. To contextualize the evolution of these models, we begin with image captioning approaches in the domain of natural images. This comparison highlights both the shared methodological foundations and the distinctive challenges posed by medical report generation.

We conclude the chapter by focusing on radiology-specific adaptations, particularly for chest X-ray interpretation, and discuss how automatic report generation supports clinical workflows, diagnosis, and decision-making processes.

6.2 Image Captioning

Image captioning is a foundational task at the intersection of computer vision and natural language processing. The objective is to generate descriptive natural language sentences that accurately reflect the content of a given image. This task requires the model to identify objects, understand their spatial and semantic relationships, recognize actions or events, and compose a linguistically coherent and semantically faithful description [108].

The field gained attraction following the success of deep learning in vision and language tasks, particularly with the introduction of encoder-decoder architectures using *Convolution*

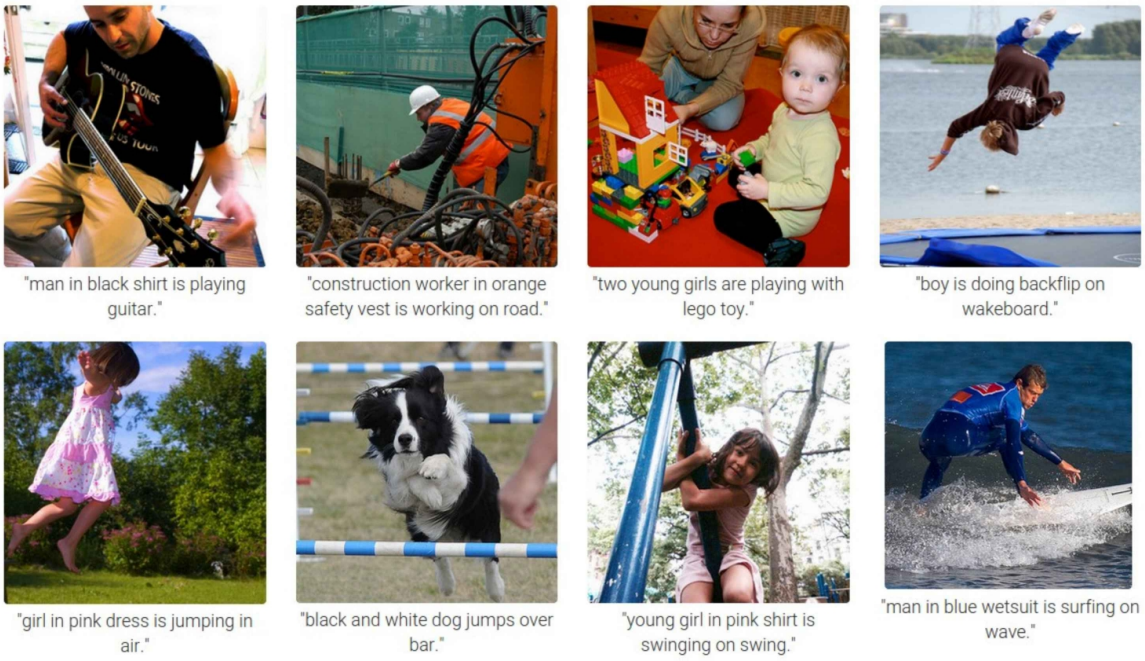


Figure 6.1: Image captioning examples on natural images. (image source [54])

*Neural Network*s (CNNs) for visual feature extraction and *Recurrent Neural Network*s (RNNs) or Transformer-based decoders for language generation [108, 116, 2]. Datasets such as MS COCO [68] and Flickr30k [119] played a pivotal role in advancing the field by providing large-scale, human-annotated image-caption pairs.

Image captioning has been successfully applied across a range of domains, including *Visual Question Answering* (VQA) [5], robotic perception and scene understanding [14], and accessibility tools for assisting visually impaired users [38]. These applications benefit from the ability of captioning systems to translate raw visual content into human-readable descriptions, enabling downstream tasks such as reasoning, navigation, and communication.

Despite these advancements, image captioning still faces notable challenges. Generated captions often lack explicit grounding, and tend to struggle with the composition and fine-grained visual detail [95, 3]. These limitations become more evident in specialized domains, where semantic precision and contextual understanding are critical.

Figure 6.1 illustrates examples of captioning in natural images, highlighting how models attempt to capture salient visual content in descriptive language.

6.2.1 Technical Foundations and Training Paradigm

The dominant architecture for image captioning follows an encoder-decoder framework. The encoder, typically a *Convolution Neural Network* (CNN) or more recently, a *Vision Transformer* (ViT), processes the image to extract high-level visual features. These features are

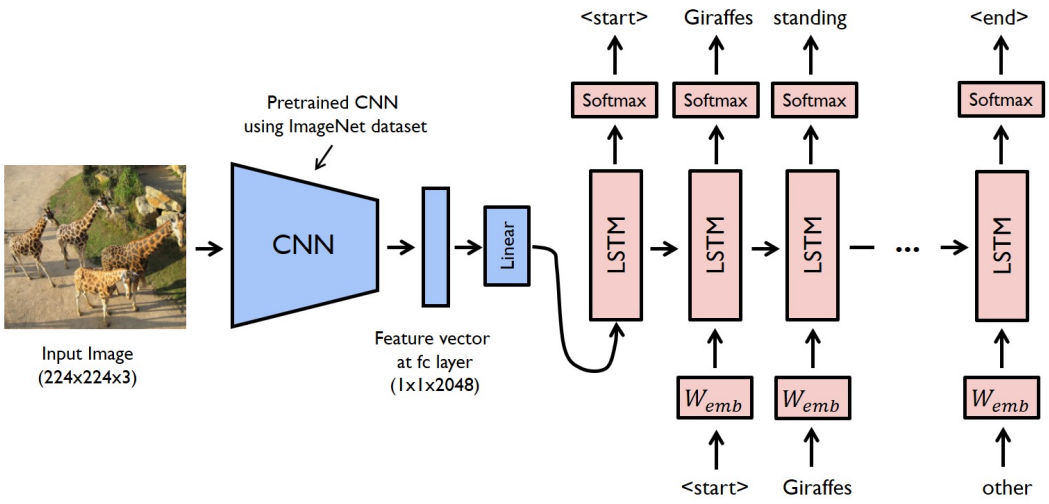


Figure 6.2: An encoder-decoder architecture for image captioning, where a CNN is used to extract and embed visual features into a vector representation. This feature vector is then passed to an LSTM-based decoder, which generates a descriptive sentence capturing the image context. Image adapted from Singh et al. [101].

then passed to a decoder, usually a *Recurrent Neural Network* (RNN), *Long Short-Term Memory* (LSTM), or transformer-based language model (e.g., GPT [89]), which generates a caption word by word. Figure 6.2 illustrates a typical image captioning pipeline, in which a CNN-based encoder extracts spatial and semantic features from the input image. These high-level visual representations are then transformed into a format compatible with the decoder, commonly a stack of LSTM layers, which sequentially generates a natural language description of the image content.

In some modern frameworks, *contrastive learning* is employed to align image and text features in a shared embedding space, significantly improving the model's ability to capture semantic relationships between modalities. Traditional encoder–decoder architectures often relied solely on supervised learning with cross-entropy losses, which focus on maximizing the likelihood of generating correct captions but lack explicit mechanisms for aligning modalities in feature space. In contrast, contrastive learning directly optimizes the embeddings: it pulls matched image–caption pairs closer together and pushes mismatched pairs farther apart. This training paradigm encourages the model to learn modality, invariant representations where semantic similarity is reflected by geometric closeness in the embedding space.

This encoder–decoder design, particularly when paired with contrastive loss and trained on large-scale paired datasets (e.g., COCO Captions [68]), has proven effective at modeling fine-grained visual–textual correlations. For example, a correctly matched pair such as an image showing “a red bus at a crosswalk" and its caption will have similar embeddings, while mismatched captions like “a man riding a horse" will be embedded further away. This alignment is essential for downstream tasks such as image retrieval, caption generation, and, in our

case, grounding clinical phrases in medical images.

A variety of vision-language datasets such as MS COCO [68], nocaps [1], and Flickr8k [42] have driven progress in natural image captioning. These benchmarks pair images with short, human-written descriptions and are typically learned using supervised objectives such as cross-entropy loss. Modern captioning systems, including BLIP [65], Show and Tell [108], and ViT–GPT2 [23], benefit from large-scale pretraining and transformer architectures, achieving state-of-the-art performance on natural image benchmarks.

6.3 Medical Report Generation

Medical report generation extends image captioning into a clinical setting, but differs fundamentally from natural-image captioning in scope, difficulty, and risk. Instead of producing brief descriptive sentences (e.g., "A dog playing with a ball"), a model must generate long, structured narratives that capture subtle anatomical cues, clinically significant abnormalities, and domain-specific terminology. Reports from datasets such as MIMIC-CXR [50] often include detailed statements such as: *"There is mild cardiomegaly with increased interstitial markings suggestive of early pulmonary edema."*

Several key challenges distinguish medical report generation from generic image captioning:

- **Linguistic and structural complexity.** Radiology reports are multi-sentence documents with implicit or explicit sections (e.g., Findings, Impression). Models must generate coherent, clinically appropriate text rather than short captions [20, 76].
- **Report length.** Medical reports are typically five to six times longer than captions for natural images and require fine-grained detail and contextual reasoning [109].
- **Clinical accuracy and safety.** Errors, such as missing a subtle opacity or incorrectly describing disease laterality, may directly impact patient care. Works such as [10, 63] emphasize the need for factual correctness, interpretability, and reduced hallucination.
- **Data scarcity and privacy.** Unlike public vision datasets, medical image–text pairs are limited by privacy regulations (e.g., *Health Insurance Portability and Accountability Act* (HIPAA)) and the need for expert annotation. Methods such as semi-supervised training and retrieval-based generation [31, 123] help mitigate this scarcity.
- **Semantic density and subtle findings.** CXRs often contain faint or overlapping abnormalities, mild interstitial edema, early consolidation, small nodules, that are easily overlooked by models trained on natural images. Fine-grained alignment strategies [103, 78] aim to improve detection of these nuanced cues.

These challenges have motivated the development of specialized architectures for clinical text generation. Early work adapted standard captioning models to IU-XRay [72] and MIMIC-CXR [50, 51], but subsequent research introduced mechanisms tailored to the medical domain.

Hierarchical decoders (e.g., HRGR-Agent [62]), memory-augmented transformers [75], and region-aware or multimodal alignment systems [103, 78] have all emerged to better model clinical structure, semantics, and visual–textual grounding.

In summary, although medical report generation builds on foundational ideas from image captioning, it requires substantial domain adaptation to meet the demands of clinical accuracy, interpretability, and safety. The following sections review these specialized radiology report generation models in more detail.

In the next section, we will focus more specifically on methods developed for chest X-ray report generation such as R2Gen [20], M2Trans [76], and CXR-RePair [31], exploring how state-of-the-art models address the unique demands of this task and how they are evaluated for both linguistic and clinical performance.

6.4 Chest X-Ray Radiology Report Generation Architectures

Chest X-rays (CXRs) are the most commonly performed imaging studies in clinical practice. Automating the generation of radiology reports from CXR images is a compelling application of AI, but one that comes with high clinical stakes. As a result, several specialized architectures have been proposed, each tailored to the unique challenges of medical language, image-text alignment, and diagnostic reliability. Below, we review some of the most prominent models developed for CXR report generation.

Chest X-rays (CXRs) are among the most frequently performed imaging studies in clinical settings due to their cost-effectiveness and diagnostic value. Automating the generation of radiology reports from CXRs is a promising application of AI that can alleviate clinical workload, reduce turnaround time, and support diagnostic workflows. However, given the high stakes involved in clinical decision-making, report generation systems must demonstrate not only linguistic fluency but also clinical accuracy, interpretability, and robustness.

In the context of this thesis, our multimodal reliability pipeline is designed to operate after the radiology report has been generated. Its objective is to assess the trustworthiness and enrichment of AI-generated reports by validating their alignment with the associated CXR images through visual grounding and cross-modal consistency. Consequently, understanding how these reports are produced, including the underlying model architectures, limitations, and training data is essential for contextualizing the output of our validation framework.

We therefore begin this section with a review of the most prominent CXR radiology report generation architectures. These models serve as upstream components whose outputs we evaluate using our pipeline. Rather than being dependent on a single report generation model, our pipeline functions as an objective assessment tool that measures interpretability, visual alignment, and clinical relevance across different RRG methods. By doing so, we aim to complement traditional evaluation metrics and offer a new perspective on model performance that reflects semantic and spatial consistency with medical imaging content.

6.4.1 R2Gen: Memory-driven Transformer

R2Gen [20] extends an encoder–decoder Transformer with *relational memory* and *memory-driven conditional layer normalization* to better handle long, structured radiology reports. A CNN (e.g., ResNet101) encodes the CXR; the Transformer decoder generates the report while conditioning normalization parameters on the evolving memory state. Report quality improves over vanilla Transformers on IU X-Ray and MIMIC-CXR using both NLG and CheXpert-based metrics.

6.4.2 M2Trans: Memory-Augmented Transformer

M2Trans [76] builds on a meshed-memory Transformer to aggregate multi-image features and optimize *factuality* via RL rewards: *factENT* (entity match) and *factENTNLI* (entity entailment). It yields higher clinical F1 and BERTScore than prior models on MIMIC-CXR/Open-i.

6.4.3 CXR-RePaiR: Retrieval-based Reporting

CXR-RePaiR [31] reframes report generation as retrieval with a CLIP-style dual encoder: select the report/sentences most similar to the image embedding. Variants retrieve full reports or compose top-k sentences, with corpus compression for efficiency. It matches/outperforms generative baselines on clinical F1 and semantic similarity.

6.4.4 RGRG: Region-guided Report Generation

RGRG [103] decomposes reporting by anatomical regions: detect 29 regions, select salient ones, classify abnormalities, and generate region-wise sentences with a GPT-2 decoder using pseudo self-attention. It improves anatomy-sensitivity and achieves competitive NLG/CE metrics on MIMIC-CXR.

6.4.5 MedGemma: Medical Vision-Language Model

MedGemma [100] is a multimodal extension of the Gemma 3 family, developed by Google DeepMind, tailored for medical image understanding and report generation. It is released in both 4B and 27B parameter variants, with multimodal models integrating a vision encoder and projection layers to map CXR features into the LLM space. MedGemma-4B is the model that can generate reports from radiology images. Training combines continued pretraining on medical corpora with reinforcement learning–based multimodal alignment, yielding strong performance on medical VQA, image classification, and radiology report generation benchmarks.

6.5 Comparative Evaluation of Report Generation Architectures

6.5.1 Performance: Textual Quality and Clinical Accuracy

The quality of generated reports is commonly evaluated using two categories of metrics. Standard NLG metrics such as BLEU, ROUGE, METEOR, and CIDEr measure lexical similarity, while clinically oriented metrics, including CheXpert F1 [47], RadGraph F1 [48], and RadCliQ [121], provide a more meaningful assessment of clinical content and alignment with radiologist judgment.

Table 6.1 reports the performance of the models considered in this thesis. Each metric value is taken directly from the original publications when available, or from the RadCliQ benchmark study, which re-evaluates several models in a standardized setting. For MedGemma, BLEU-4 and CIDEr were computed by us on the MIMIC-CXR test set, while RadGraph F1 is taken from the official paper.

Table 6.1: Reported performance of leading CXR report–generation models on MIMIC-CXR. BLEU-4 and CIDEr are taken from the original model papers when available; RadGraph F1 and RadCliQ values are taken from the RadCliQ benchmark [121], except for MedGemma where RadGraph F1 comes from the model paper.

Model	**BLEU-4**	**CIDEr**	**RadGraph F1**	**RadCliQ**
R2Gen	$0.103^{(\text{paper})}$	$0.406^{(\text{M2Trans})}$	$0.134^{(\text{RadCliQ})}$	$1.552^{(\text{RadCliQ})}$
M2Trans	$0.114^{(\text{paper})}$	$0.509^{(\text{paper})}$	$0.244^{(\text{RadCliQ})}$	$1.059^{(\text{RadCliQ})}$
CXR-RePaiR-Select	$0.050^{(\text{paper})}$	–	$0.091^{(\text{RadCliQ})}$	$1.642^{(\text{RadCliQ})}$
RGRG	$0.126^{(\text{paper})}$	$0.495^{(\text{paper})}$	$0.547^{(\text{paper})}$	–
MedGemma-4B	$0.140^{(\text{ours})}$	$0.289^{(\text{ours})}$	$0.295^{(\text{paper})}$	–

Models such as M2Trans and RGRG achieve stronger scores on linguistic metrics (BLEU and CIDEr), whereas MedGemma shows comparatively better clinical factuality as reflected by RadGraph F1. CXR-RePaiR, despite its low BLEU score, achieves high RadCliQ performance due to its retrieval-based nature, which reduces hallucinations and enforces clinical precision.

6.5.2 Interpretability and Transparency

Interpretability is a major concern for clinical deployment. Most models offer limited explainability, with varying degrees of attention visualization or component-level transparency.

Table 6.2: Comparison of interpretability strategies across models.

Model	**Mechanism**	**Limitations**
R2Gen	Transformer attention weights	No explicit visual grounding
M2Trans	Cross-modal attention + factual reward tuning	Diffuse attention, interpretability varies
CXR-RePaiR	Sentence-level scoring using CLIP	No internal reasoning transparency
RGRG	Anatomy-aware generation per region	Region-sensitive but coarse granularity
MedGemma	LLM attention with vision-token projection	Limited transparency in multimodal fusion

Among the models, RGRG offers anatomically localized sentence generation, enhancing transparency. MedGemma leverages vision–language fusion but inherits the opacity of large-scale LLM architectures. CXR-RePaiR provides post-hoc interpretability through sentence selection guided by semantic similarity. Table 6.2 summarizes these strategies, outlining how different architectures handle transparency and which limitations persist in their ability to provide clinically meaningful explanations.

6.5.3 Clinical Utility and Deployment Readiness

Clinical utility considers factual completeness, integration potential, and decision support capabilities. Models are compared below on how well they align with real-world clinical needs.

Table 6.3 compares the practical clinical utility of these models. It highlights how each approach balances factual consistency, reasoning ability, and readiness for real-world deployment within clinical workflows. CXR-RePaiR stands out for factual robustness, making it ideal for settings requiring safety-critical reporting. M2Trans and MedGemma are well-suited for decision support scenarios, thanks to their strengths in factual alignment and multimodal reasoning.

Table 6.3: Clinical utility comparison.

Model	**Factual Consistency**	**Supports Reasoning**	**Deployment Potential**
R2Gen	Medium	Low	Moderate
M2Trans	High	Moderate	Promising
CXR-RePaiR	Very High	High (via selection logic)	High (refinement use)
RGRG	Moderate	Moderate (region-specific)	Experimental (research-focused)
MedGemma	High	High (general LLM reasoning)	Promising (scalable multimodal foundation)

Table 6.4: Overall strengths and limitations of each model.

Model	**Strengths**	**Limitations**
R2Gen	Temporal coherence, easy to fine-tune	No explicit visual-textual alignment
M2Trans	Factuality via multimodal fusion and reward-based tuning	Training is resource-intensive
CXR-RePaiR	Reduces hallucination through retrieval, clinically accurate	Limited creativity in text generation
RGRG	Anatomy-sensitive generation, regional precision	Requires anatomy labels, limited scalability
MedGemma	Large-scale multimodal LLM, strong factuality	Limited transparency, high compute cost

6.5.4 Summary of Trade-offs

These trade-offs highlight the multidimensional nature of clinical report generation, where lexical performance, interpretability, and clinical utility must be carefully balanced. While models like R2Gen offer strong baselines with fluent generation, others like M2Trans and CXR-RePaiR enhance factual consistency and multimodal alignment at the cost of added

complexity or reduced transparency. As the field progresses, there is a growing consensus that future systems should adopt modular hybrid designs, integrating retrieval-based correction (as in CXR-RePaiR), anatomically guided attention (as in RGRG), and scalable vision–language integration (as in MedGemma). Such architectures can support more trustworthy, verifiable, and clinically actionable AI tools for radiology. Table 6.4 summarizes the main strengths and limitations of each model. It highlights how different architectural choices influence fluency, factuality, interpretability, and scalability, offering a concise overview of the trade-offs that shape current progress in radiology report generation.

These architectures reflect a broader trend in medical AI: moving beyond syntactic fluency toward clinically meaningful, interpretable, and verifiable generation. In the next section, we will explore how explainability, visual grounding, and semantic evaluation further enhance trust and reliability in report generation models.

Rather than selecting a single approach, this thesis evaluates five representative models, R2Gen, M2Trans, RGRG, and CXR-RePaiR, alongside MedGemma as a recent foundation model. Their diversity makes them well-suited as testbeds for assessing the VICCA framework, which aims to provide a systematic and comparable evaluation across different RRG paradigms in one single model.

Chapter 7

Objective Evaluation of Radiology Report Generation Models using VICCA

Contents

7.1 Introduction

In Chapter 6, we reviewed several *Radiology Report Generation* (RRG) models, R2Gen, M2Trans, CXR-RePaiR, RGRG, and MedGemma, highlighting their respective strengths, limitations, and comparative performance (see Appendix D for detailed descriptions). While these models are often evaluated with standard NLP metrics originally developed for machine

translation, such as BLEU and ROUGE, we have shown that these measures are poorly aligned with the requirements of clinical reporting, where semantic accuracy and factual validity are paramount. Nonetheless, due to their widespread adoption, these metrics remain a common baseline for benchmarking across RRG models.

In Chapter 5, we proposed a new evaluation metric, *Medical Corpus Similarity Evaluation* (MCSE), designed to assess the semantic similarity of medical texts with a focus on clinical accuracy. We demonstrated that MCSE provides a more reliable and domain-relevant evaluation framework than conventional lexical metrics. However, one limitation of MCSE is its reliance on the availability of a ground-truth or reference report, which may not be feasible in real-world clinical scenarios. In practice, RRG models are expected to assist clinicians by generating reports directly from radiological images, often without access to a predefined reference. This raises the critical question of how to assess the trustworthiness of AI-generated reports in the absence of ground-truth text.

To address this challenge, our VICCA framework offers an objective and reference-free evaluation strategy by assessing the visual-textual alignment and reliability of the generated report. VICCA leverages both visual grounding and a generative model to localize textual phrases within the chest X-ray and verify their correspondence using a diffusion-based auxiliary model. This process enables us to identify whether the generated report is aligned with the visual evidence, highlights missing pathologies, or includes inconsistencies, thus aiding clinical decision-making.

7.2 Evaluation Protocol

Building on the reviewed literature and the representative set of RRG models presented in Section 6, our evaluation protocol aims to provide a unified and clinically grounded framework for assessing the reliability of AI-generated chest X-ray reports. Unlike prior approaches that focus solely on lexical or semantic similarity, our goal is to integrate multiple complementary perspectives, textual, semantic, and visual, within a single standardized pipeline that reflects how radiologists assess report trustworthiness in practice.

Specifically, the proposed protocol combines:

1. **Pathology-level agreement**, which quantifies how well the generated reports capture medically relevant findings using the CheXbert classifier [102] to compute micro- and macro-F1 as well as Jaccard similarity.

2. **Semantic fidelity**, evaluated through the Medical Corpus Similarity Evaluation (MCSE) metric [86], which measures conceptual alignment between generated and reference reports by detecting clinical entities, modifiers, and negations rather than relying on lexical overlap.

3. **Visual-textual consistency**, assessed through the VICCA framework [34], which em-

ploys visual grounding to link report phrases to image regions and diffusion-based generation to test visual plausibility and detect potential hallucinations.

This multimodal evaluation strategy allows us to examine not only what each model says but also whether those statements are visually justified and clinically coherent. By analyzing these three dimensions jointly, we can uncover discrepancies that remain hidden in traditional text-only evaluations and gain a more holistic understanding of model behavior in realistic clinical scenarios.

7.3 Experimental Setup

To approximate a realistic clinical deployment scenario, we apply the proposed evaluation framework to the outputs of the five RRG models described in Section 6. All evaluations are conducted on a subset of 2,461 studies from the MIMIC-CXR test set [50, 51], ensuring that each case includes both a frontal chest X-ray and a corresponding reference report.

Test set filtering and leakage prevention The original MIMIC-CXR test split contains 3,269 studies. However, a subset of these studies is also included in the MS-CXR [10] annotations, which are used during training for visual grounding and visual alignment in VICCA. To prevent any form of data leakage between training and evaluation, we explicitly excluded all test studies that appear in the MS-CXR training set. After removing these overlapping cases, the final evaluation set consists of 2,461 studies. All reported results on MIMIC-CXR are computed exclusively on this filtered test set. No additional filtering or subsampling was applied beyond this exclusion. Table 7.1 summarizes the composition of the filtered MIMIC-CXR test set used throughout our experiments.

7.3.1 Text Processing Step

To generate the radiology reports for the test set, we used each model's publicly available pretrained weights. Reports were produced directly from the CXR images without any additional prompt engineering or text postprocessing, except in the case of the foundation model **MedGemma**. Consequently, the text inputs to the VICCA, MCSE, and pathology detection modules correspond to the raw outputs of the RRG models, with MedGemma being the only exception.

For MedGemma, we employed the 4B multimodal instruction-tuned variant (**MedGemma-4B-IT**) to produce radiology-style outputs, using a maximum of 1000 new tokens per generation. As a foundation model derived from Gemma and adapted for medical imaging, MedGemma often produces verbose, mixed-domain responses that extend beyond the typical radiology report format. To align its outputs with clinical reporting conventions,

Table 7.1: Statistics of the filtered MIMIC-CXR test set used for evaluation. Studies overlapping with the MS-CXR training set were excluded to prevent data leakage.

Statistic	**Value**
Original MIMIC-CXR test studies	3,269
Excluded due to MS-CXR overlap	808
Final evaluation studies	**2,461**
Studies labeled as *No Finding*	533
Studies with $\geq$1 pathology	1,563
Number of CheXbert pathology classes	14
Most frequent pathologies	Lung Opacity, Pleural Effusion, Cardiomegaly
Least frequent pathologies	Fracture, Lung Lesion, Pleural Other

an additional postprocessing step was applied to extract the most relevant *Findings* section, as illustrated in Table 7.2.

Table 7.2: Example of a MedGemma report before and after cleaning. The preprocessing step focuses on the clinically relevant *Findings* section to align outputs with radiology reporting conventions.

Original MedGemma Output	**Cleaned Report (Findings Only)**
"Okay, here's a description of the chest X-ray... **Overall Impression:** The heart appears enlarged, and there are areas of increased density... **Specific Findings:** Bones intact, mediastinum normal, diaphragms unremarkable... **Possible Considerations:** Cardiomegaly, infiltrates, need for CT... **Disclaimer:** Preliminary interpretation only..."	"The heart appears enlarged, with increased density in the lower lung fields suggesting possible consolidation or infiltrates. Bones are intact; mediastinum and diaphragm appear normal. Consider cardiomegaly and infiltrates; correlation with clinical context recommended."

Unlike the RRG models trained specifically for radiology report generation, MedGemma is primarily designed for general medical reasoning and non-expert interpretation of CXR images. As a result, its responses resemble those of conversational AI systems, frequently organized into sections such as "Overall Impression," "Specific Findings," "Differential Diagnosis," "Recommendations," or "Disclaimer," with varying structures across samples. This variability complicates the direct extraction of clinically focused content, particularly since the "Findings" or "Impression" sections are not consistently defined.

Moreover, due to the large number of generated tokens, MedGemma outputs occasionally contain hallucinated entities or redundant content. Although reducing the generation length could mitigate verbosity, this would compromise the accuracy and anatomical-pathological alignment of the output. Instead, we implemented a targeted summarization procedure that identifies and extracts the paragraph most relevant to clinical interpretation. This process relies on detecting the "Findings" keyword and analyzing the density of medical entities within each section to isolate the portion most suitable for evaluation. This postprocessing step affects only formatting and section selection and does not alter the semantic content of the generated reports.

After generation, we evaluate each model's output along the three complementary dimensions defined in our protocol 7.2. This end-to-end setup allows an integrated evaluation of each RRG model, simultaneously examining its textual coherence, semantic fidelity, and visual trustworthiness under clinically plausible conditions.

7.3.2 Evaluation Metrics

In this section, we describe the evaluation metrics used to assess the performance of radiology report generation (RRG) models that were referenced earlier but not yet formally defined. Table 7.3 summarizes the notation used throughout the evaluation framework.

Textual Pathology Label Accuracy (CheXbert)

- **Micro-F1** aggregates true positives, false positives, and false negatives across all pathology classes, and is therefore dominated by frequent conditions such as *No Finding* and *Pleural Effusion*.
- **Macro-F1** computes the F1-score independently for each pathology and then averages across classes, assigning equal weight to common and rare findings. This is particularly important in clinical settings, where infrequent but critical conditions (e.g., *Fracture*, *Lung Lesion*) should not be overshadowed by prevalent labels.

Pathology-set overlap (Jaccard similarity) To quantify overlap between pathology labels extracted from generated and reference reports, we compute the Jaccard similarity. Given

Table 7.3: Summary of notation used in the evaluation framework.

Symbol	Description
y	Reference radiology report associated with a chest X-ray image
$\hat{y}$	Report generated by a report generation model
$T = (t_1, \ldots, t_N)$	Sequence of clinical entities extracted from reference report y
$\hat{T} = (\hat{t}_1, \ldots, \hat{t}_M)$	Sequence of clinical entities extracted from generated report $\hat{y}$
C	Set of exact entity matches between T and $\hat{T}$
$r_i, \hat{r}_j$	Unmatched reference and generated entities after removing exact matches
$y_{i,j}$	Cosine similarity between entities r_i and $\hat{r}_j$ in clinical embedding space
S_j	Normalized semantic similarity score for generated entity $\hat{r}_j$
MCSE	Medical Corpus Similarity Evaluation score between y and $\hat{y}$
A, B	Sets of pathology labels extracted from generated and reference reports (CheXbert)
$J(A, B)$	Jaccard similarity between pathology sets A and B
τ	Confidence threshold used to accept a grounded bounding box
Coverage	Proportion of report phrases that yield at least one grounded region
Reliability	Mean MS-SSIM score between localized regions in original and generated images

the sets of pathologies A (generated) and B (reference), the Jaccard index is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{7.1}$$

This metric penalizes both missing findings and spurious mentions, providing a strict set-level agreement score.

Localization coverage Localization coverage measures the proportion of textual phrases in a report that can be visually grounded in the corresponding chest X-ray. For each extracted phrase, visual grounding is considered successful if the grounding model returns at least one bounding box with confidence greater than a fixed threshold τ.

Formally, coverage is defined as:

$$\text{Coverage} = \frac{\#\text{ grounded phrases}}{\#\text{ extracted phrases}}. \tag{7.2}$$

In all experiments, the confidence threshold was set to $\tau = 0.25$. This definition captures the presence of visually groundable content independently of localization precision, which is evaluated separately through reliability.

Region-level reliability (MS-SSIM) Reliability evaluates the consistency between localized regions in the original CXR and the corresponding regions in the text-conditioned generated image. For each successfully grounded phrase, MS-SSIM is computed between the cropped bounding box in the original image and the corresponding region in the generated image.

Reliability is first computed at the bounding-box level, then averaged across all grounded phrases within a study to obtain a per-study reliability score. Finally, model-level reliability is reported as the mean reliability across all studies with valid grounded regions.

Studies without any valid grounded regions are excluded from reliability aggregation, which may result in unequal sample sizes across models.

Together, these metrics allow us to disentangle pathology correctness, semantic fidelity, and visual grounding reliability in a unified evaluation framework.

The following sections detail the results of this multi-level evaluation, illustrating how VICCA can provide deeper insight into the strengths and limitations of state-of-the-art RRG models beyond traditional metrics.

7.4 Results and Analysis

7.4.1 Textual Pathology Label Accuracy

To evaluate whether the generated reports reflect clinically relevant findings, we first assess the accuracy of predicted pathologies using the CheXbert model. Trained on CheXpert classes, CheXbert classifies radiology reports into 14 thoracic disease categories. This comparison between generated and reference labels offers insight into each model's ability to capture medically meaningful content.

Given that pathology extraction is a multi-label classification task, we report both **micro-F1** and **macro-F1** scores. These scores range from 0 (worst) to 1 (best). Among the evaluated models, **R2Gen** shows the weakest alignment with reference findings (micro-F1 = 0.306, macro-F1 = 0.182), frequently producing vague descriptions and omitting key pathologies. In contrast, **MedGemma** (micro-F1 = 0.480, macro-F1 = 0.324) proves highly competitive, achieving the strongest balance across both common and rare classes, as reflected in its highest macro-F1 score. **M2Trans** reaches the second-highest pathology-level agreement (micro-F1 = 0.474), followed closely by **RGRG** (0.470) and **CXR-RePaiR** (0.452). These results are summarized in Figure 7.1. When no explicit pathologies are detected, models default to the "No Finding" label. This behavior affects performance on the dominant "No Finding" class.

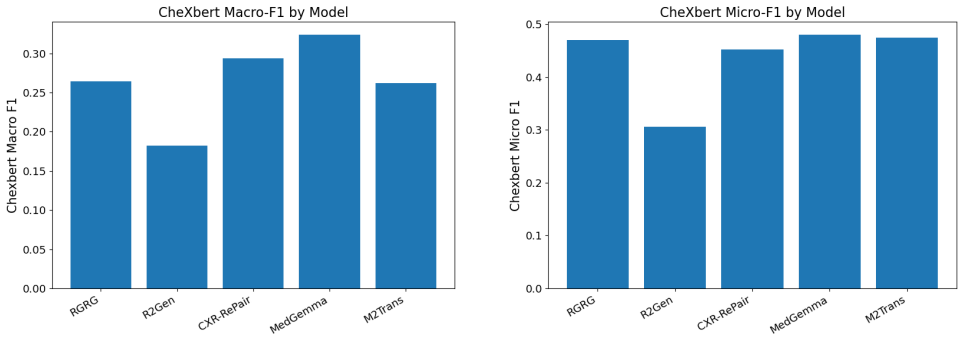


Figure 7.1: Comparison of CheXbert-based micro-F1 and macro-F1 scores for each RRG model.

In addition to global F1-scores, we also report performance for each pathology along with its support, i.e., the number of cases for that pathology in the test set. This breakdown is particularly relevant in medical evaluation, since certain findings (e.g., "Pleural Effusion," "Cardiomegaly") are much more prevalent than others (e.g., "Fracture," "Lung Lesion"). High-support classes tend to dominate micro-F1 scores, while low-support but clinically critical classes may be overlooked unless reported separately. By presenting pathology-by-support metrics in Figure 7.2, we highlight whether models generalize beyond frequent findings and remain clinically reliable for rare but important conditions.

- **High-support classes** (e.g., Pleural Effusion, Cardiomegaly, Lung Opacity): All models achieve relatively higher precision and recall since these findings dominate the dataset. M2Trans and RGRG capture them more consistently, while CXR-RePaiR shows balanced detection. R2Gen under-predicts them, often defaulting to "No Finding."

- **Low-support but clinically critical classes** (e.g., Fracture, Lung Lesion, Pleural Other): Performance is weak across all models, particularly in recall. This drags down macro-F1, which weights each pathology equally. MedGemma performs relatively better on these rare classes, as reflected in its higher macro-F1, though overall performance remains limited.

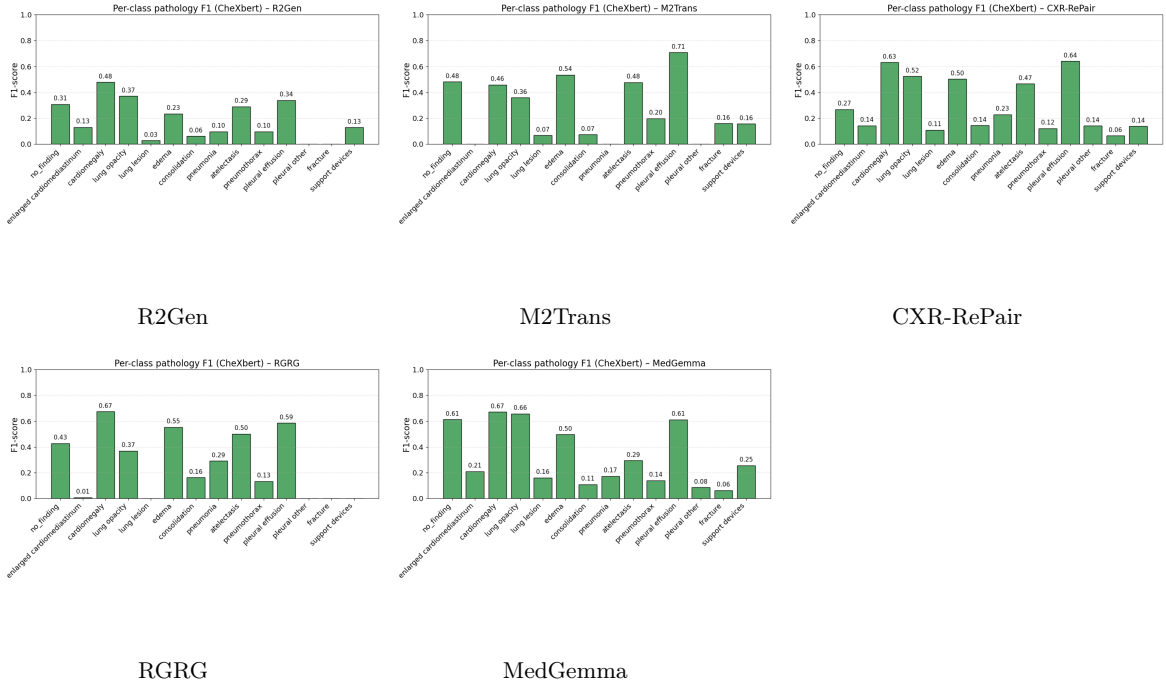


Figure 7.2: Per-class F1 score distributions for the top 15 most prevalent pathologies in the test set.

- **No Finding**: Predicted too frequently by R2Gen, inflating apparent accuracy in normal cases but masking pathology detection. CXR-RePaiR provides a better balance between "No Finding" and actual disease mentions, while MedGemma reduces this bias by predicting more diverse pathologies.

Key Takeaways

- **R2Gen:** Consistently underpredicts pathology classes and defaults heavily to "No Finding."
- **MedGemma:** Strongest overall balance, with the highest macro-F1; performs better than others on rare classes, though still limited in absolute terms.
- **M2Trans:** Achieves the second-best pathology alignment, especially strong on high-support classes, but still produces some empty predictions.
- **RGRG:** Balanced performance, capturing common pathologies well, though occasionally overpredicts compared to references.
- **CXR-RePaiR:** More diverse in its pathology predictions than R2Gen and maintains a better balance between "No Finding" and true disease mentions, but introduces some false positives.

Pathology-set overlap (Jaccard similarity) In addition to F1 scores, we report Jaccard similarity between the sets of pathologies extracted from generated and reference reports. Across all models, mean Jaccard values range from 0.25 to 0.38 (Table 7.4), with large interquartile ranges. Despite these numerical differences, Jaccard scores exhibit limited discriminative power, as many reports collapse to identical pathology sets dominated by the "No Finding" label. As a result, Jaccard similarity shows low variance and weak association with semantic fidelity, motivating the use of MCSE for narrative-level evaluation.

Table 7.4: Jaccard similarity between pathology sets extracted from generated and reference reports. Values are reported as mean ± standard deviation, along with median and interquartile range (IQR).

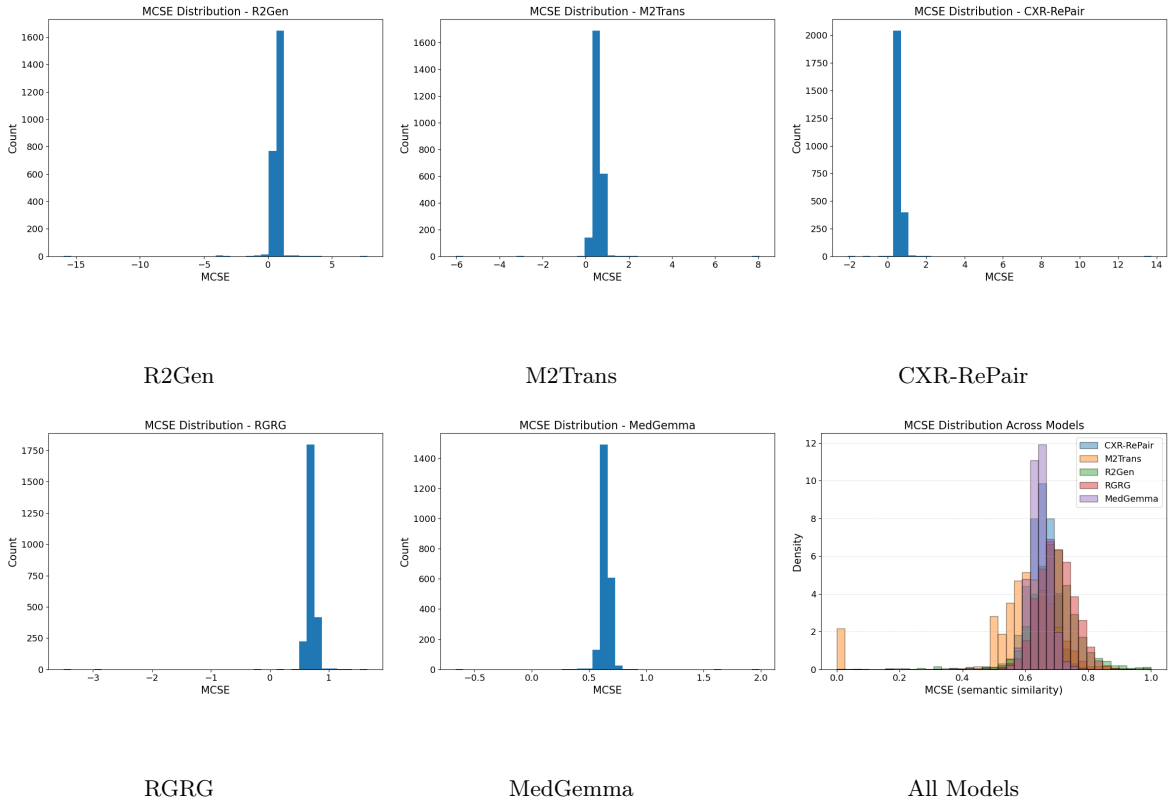
Model	Mean	Std	Median	IQR
RGRG	0.335	0.310	0.333	0.50
R2Gen	0.249	0.347	0.000	0.40
CXR-RePaiR	0.307	0.277	0.250	0.50
MedGemma	0.370	0.309	0.333	0.33
M2Trans	0.380	0.377	0.333	0.67

7.4.2 Semantic Similarity Evaluation using MCSE

To assess the semantic alignment between the generated and reference reports, we employ our proposed metric, **MCSE**, to evaluate conceptual consistency in clinical narratives illustrated in Figure 7.3.

Table 7.5 summarizes the distribution of MCSE scores across models, reporting mean, standard deviation, median, and interquartile range (IQR) over the 2,460 aligned studies.

Statistical significance of MCSE differences To evaluate whether observed differences in MCSE scores are statistically meaningful, we performed non-parametric testing on strictly paired samples using a Friedman test, followed by post-hoc paired Wilcoxon signed-rank tests with Holm correction. The Friedman test revealed a significant effect of model choice on semantic similarity ($\chi^2 = 2135.47$, $p < 10^{-16}$), indicating that MCSE scores differ substantially across models.



R2Gen M2Trans CXR-RePair

RGRG MedGemma All Models

Figure 7.3: Comparative distribution of MCSE scores across R2Gen, M2Trans, RGRG, and CXR-RePaiR. CXR-RePaiR and R2Gen cluster at intermediate values, while RGRG achieves higher semantic similarity on average.

Post-hoc analysis confirmed that these differences are robust: **RGRG** significantly outperforms all other models, while **M2Trans** consistently underperforms. Notably, **R2Gen** also achieves significantly higher MCSE scores than M2Trans despite exhibiting greater variance in its outputs. Together, these results demonstrate that the observed differences in semantic similarity are not attributable to random variation.

As the reports are generated directly from true CXR images, we expect a meaningful degree of semantic similarity with their corresponding references. Among the models, **RGRG** achieved the highest average MCSE score (**0.696**), reflecting strong alignment with reference narratives. **R2Gen** followed closely (**0.676**), though with high variability across cases: some reports align well semantically, while others remain overly generic. **CXR-RePaiR** (**0.657**) and **MedGemma** (**0.645**) also performed competitively, maintaining reasonably coherent narratives. In contrast, **M2Trans** obtained the lowest MCSE (**0.587**), suggesting that while it identifies pathologies effectively, it struggles to embed them in cohesive and clinically meaningful text.

These findings highlight that a model may correctly detect clinical entities yet fail to present them in semantically rich, coherent narratives. Figure 7.3 illustrates the distribution

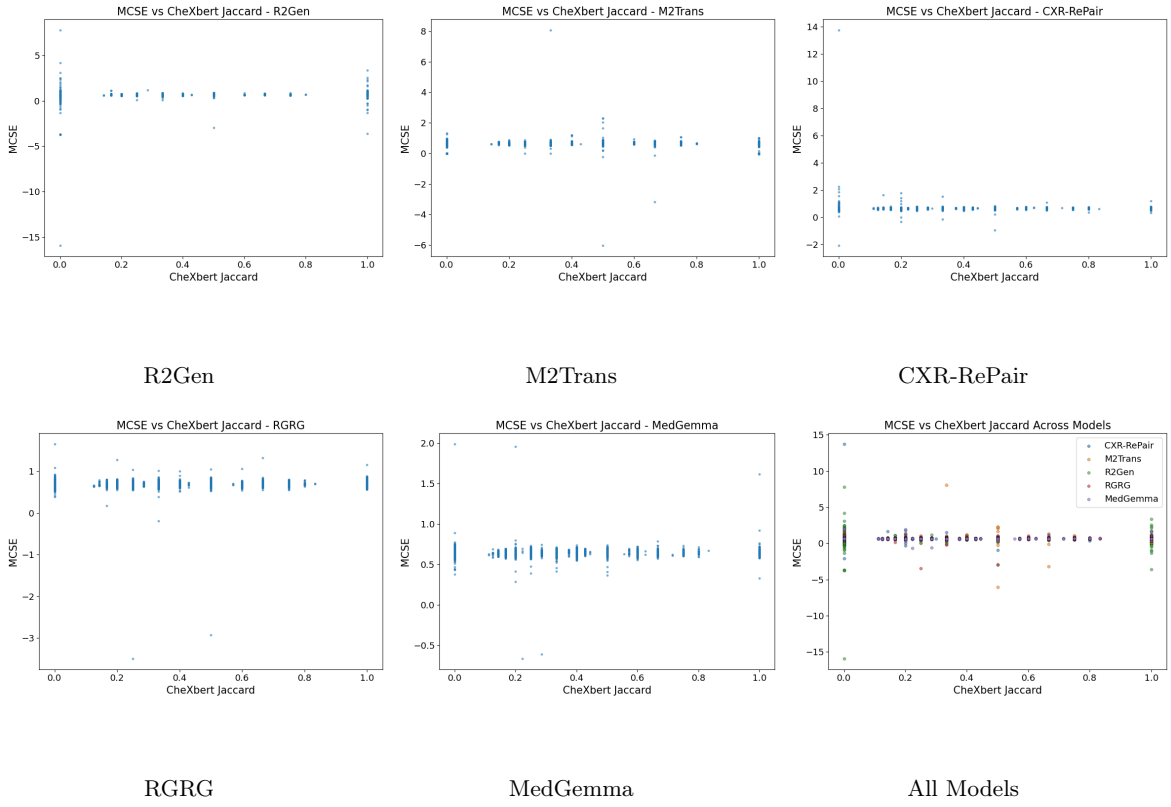
Table 7.5: Statistical summary of MCSE scores across RRG models. Higher mean MCSE indicates better semantic alignment with reference reports.

Model	Mean MCSE	Std	Median	IQR
RGRG	**0.696**	0.132	0.696	0.079
R2Gen	0.676	**0.468**	0.687	0.086
CXR-RePaiR	0.657	0.286	0.652	0.055
MedGemma	0.645	**0.070**	0.645	**0.042**
M2Trans	0.587	0.278	0.609	0.108

of MCSE scores across all models.

- **RGRG:** Strongest semantic alignment; produces structured, clinically coherent narratives.
- **R2Gen:** Moderate MCSE; variable performance, from semantically accurate to overly generic.
- **CXR-RePaiR:** Consistently coherent, reflecting retrieval-based grounding.
- **MedGemma:** Competitive semantics, slightly below CXR-RePaiR, with better rare-class coverage but weaker grounding.
- **M2Trans:** Lowest MCSE; lists pathologies correctly but struggles to form cohesive narratives.

Relation between MCSE and Jaccard similarity While Jaccard similarity measures set-level overlap of pathology labels, MCSE evaluates semantic fidelity of the full clinical narrative. As shown in Figure 7.4, these two metrics exhibit no meaningful association. In practice, CheXbert Jaccard similarity saturates at 1.0 for all models, yielding zero variance and rendering statistical testing and correlation analysis undefined. This saturation arises from coarse label extraction and the dominance of generic labels such as *No Finding*. Consequently, pathology-set overlap metrics fail to discriminate between report generators under this evaluation setup, indicating that pathology-set overlap alone does not discriminate between models under this evaluation setup.



R2Gen M2Trans CXR-RePair

RGRG MedGemma All Models

Figure 7.4: Comparative plot of MCSE vs chexbert Jaccard scores across R2Gen, M2Trans, RGRG, CXR-RePaiR, and MedGemma.

7.4.3 Visual Grounding and Localization

Beyond textual similarity, VICCA enables validation through visual grounding, associating textual phrases with corresponding regions on the chest X-ray. We evaluate each model's localization coverage, defined as the proportion of identified phrases successfully grounded with bounding boxes.

Localization coverage is defined as the proportion of report phrases that yield at least one valid grounded region (see Section 7.3.2), and should not be interpreted as a measure of localization accuracy or reliability. Figure 7.5 shows the coverage across all models. **RGRG** achieves the highest coverage (**0.844**), closely followed by **M2Trans** (**0.823**). **R2Gen** (**0.686**) and **CXR-RePaiR** (**0.677**) fall into the mid-range. By contrast, **MedGemma** records the lowest coverage (**0.297**), indicating that although it performs well textually, its grounding of findings to image regions remains limited.

Importantly, coverage alone does not guarantee clinical precision as grounding quantity does not directly reflect anatomical correctness. Models like M2Trans and RGRG produce more boxes overall, but the quality of alignment with meaningful anatomical structures varies.

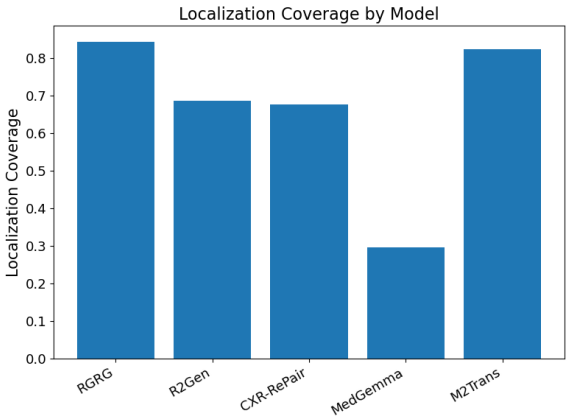


Figure 7.5: Comparative analysis of localization coverage histogram across SOTA models.

- **RGRG:** Strongest coverage and good phrase-to-region alignment.
- **M2Trans:** High coverage, though some boxes are noisy or weakly grounded.
- **R2Gen / CXR-RePaiR:** Moderate coverage; grounding consistency is weaker than RGRG.
- **MedGemma:** Struggles to produce bounding boxes, highlighting limited grounding capability.

This highlights the need to evaluate both the quantity and quality of grounding for clinically interpretable outputs.

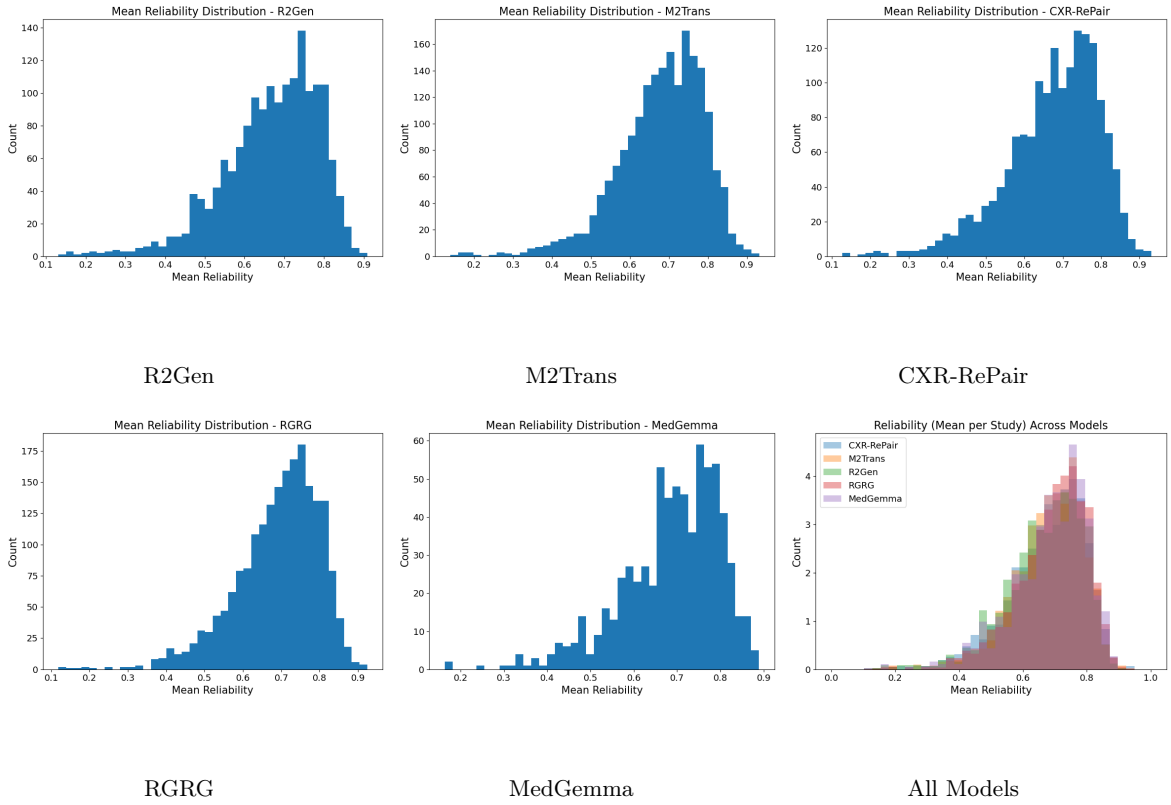
7.4.4 Reliability of Localized Regions

We further assess the reliability of grounded regions by evaluating the consistency between localized regions in the original and generated images using MS-SSIM. This metric captures structural similarity, providing insight into how well a given report preserves visual characteristics in the generated image.

Unlike localization coverage, reliability scores are very close across models, ranging from **0.666 to 0.690**. **RGRG** (0.690) and **MedGemma** (0.687) slightly outperform others, followed by **M2Trans** (0.680) and **CXR-RePaiR** (0.675). **R2Gen** lags marginally (0.666), though the difference is small.

These results in Figure 7.6 suggest that region-level structural consistency is relatively stable across models, even when textual semantics and pathology predictions differ significantly.

Region-level reliability scores are reported descriptively, as insufficient paired observations were available across all models to support non-parametric statistical testing.



R2Gen M2Trans CXR-RePair

RGRG MedGemma All Models

Figure 7.6: Comparative plot of distribution of mean reliability scores across models.

- **RGRG:** Most reliable overall; balances semantic quality with structural consistency.
- **MedGemma:** Strong reliability despite weak localization, suggesting coherent structural mapping when boxes exist.
- **M2Trans:** Reliable on average, though variability across samples is higher.
- **CXR-RePaiR:** Consistently reliable, though slightly lower than RGRG.
- **R2Gen:** Lowest reliability, consistent with weaker text-pathology alignment.

MS-SSIM thus captures a complementary dimension of model performance, structural plausibility, that is not apparent through text-based metrics alone.

7.4.5 Failure Mode Analysis (Qualitative Micro-Examples)

To complement aggregate quantitative metrics, we present representative qualitative failure cases that expose model-specific limitations. These micro-examples illustrate distinct failure modes that are not fully captured by label-based or reference-driven metrics, and highlight

the diagnostic value of jointly evaluating semantic consistency and visual grounding through MCSE and VICCA.

7.4.5.1 Case 1: CXR-RePaiR – Textually similar but clinically misleading retrieval

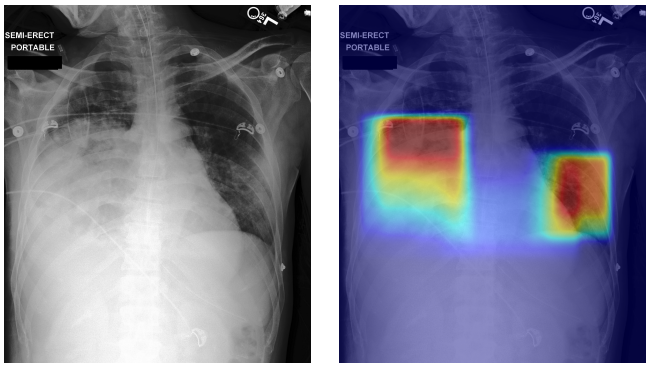
Table 7.6: Qualitative failure example for the CXR-RePaiR model. Despite high semantic similarity, the generated report introduces clinically unsupported findings.

Model: CXR-RePaiR	
Reference Report	**Generated Report**
There is unchanged evidence of the known massive irritated right lung process. Slightly progressive opacifications in the periphery of the left lung. Unchanged size of the cardiac silhouette.	Complete opacification of the right hemithorax is due to collapse of the small residual of aerated right lung compromised by the chronically elevated right hemidiaphragm and probable small right pleural effusion. Heterogeneous opacification in the left lung, particularly the suprahilar region, could be aspiration or early pneumonia. Previous pulmonary vascular congestion has improved, and moderate cardiomegaly is minimally smaller.
Reference Labels	**CheXbert Labels**
Lung Opacity, *Cardiomegaly*	*Lung Opacity*, *Pleural Effusion*, *Cardiomegaly*

Table 7.6 summarizes a representative failure case produced by the **CXR-RePaiR** model, comparing the reference report, generated report, and the corresponding pathology labels. Although the generated report remains semantically close to the reference narrative, it introduces additional findings not supported by the ground truth.

In this case, MCSE assigns a high semantic similarity score (**0.85**), reflecting strong overlap in disease progression, anatomical context, and descriptive structure. However, the pathology-set overlap measured by Jaccard similarity is low (**0.20**), as the generated report introduces spurious pathologies such as pleural effusion. A purely label-based evaluation would therefore suggest poor agreement, without revealing the underlying semantic alignment.

Figure 7.7 further illustrates this discrepancy through visual grounding. Although several phrases from the generated report can be localized, the predicted finding *Cardiomegaly* shows no corresponding visual support, as it is neither supported by the reference report nor visually evident in the image. This exposes a clinically misleading over-interpretation introduced by the model.



Original CXR

Visual grounding from CXR-RePaiR generated report

Figure 7.7: Example of a CXR-RePaiR failure case where high semantic similarity masks clinically unsupported findings.

This example highlights a characteristic failure mode of retrieval-augmented generation: reports that are fluent and semantically plausible yet introduce unsupported clinical hypotheses. Such behavior is difficult to identify using pathology-label metrics such as CheXbert or Jaccard similarity alone, but becomes apparent when semantic similarity and visual grounding are evaluated jointly.

7.4.5.2 Case 2: M2Trans – High Pathology Overlap but Low Visual Reliability

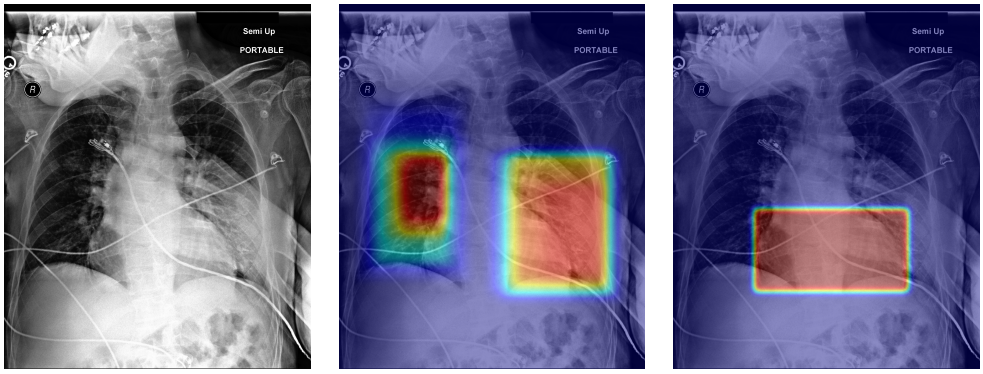
Figure 7.8 and Table 7.7 illustrates a case where pathology-level overlap appears deceptively strong, while visual reliability is poor. CheXbert extraction yields identical number of pathology sets for the reference and generated reports, resulting in a Jaccard similarity of 1.0. However, the generated report explicitly negates pulmonary findings and omits pneumonia, leading to a marked divergence in visual grounding.

As a result, the grounded regions derived from the generated report focus solely on the cardiac silhouette, whereas the reference grounding highlights basilar lung involvement for *Pneumonia*. This discrepancy yields a low region-level reliability score (**34.6%**), despite label overlap.

This example exposes a characteristic failure mode of transformer-based report generators: accurate repetition of dominant or frequent labels without preserving the spatial or clinical context in which those findings appear. Such failures are invisible to pathology-set overlap metrics but are revealed by joint evaluation of semantic consistency and visual reliability.

Table 7.7: Qualitative failure example for the **M2Trans** model. Although pathology-set overlap is high, the generated report fails to preserve the visual semantics of the reference findings.

Model: M2Trans	
Reference Report	**Generated Report**
The hand of the patient obscures the lower half of the left chest. There is enlargement of the cardiac silhouette with indistinctness of engorged pulmonary vessels, consistent with elevated pulmonary venous pressure. In the appropriate clinical setting, superimposed basilar pneumonia could be considered.	The cardiomediastinal and hilar contours are stable with moderate cardiomegaly. The heart is moderately enlarged. The lung fields are clear. There is no focal consolidation, pulmonary edema, pleural effusion, or pneumothorax.
Reference Labels	**CheXbert Labels (Generated)**
Pneumonia, *Cardiomegaly*	*Cardiomegaly*, *Cardiomegaly*



Original CXR | Grounding from reference report | Grounding from M2Trans generated report

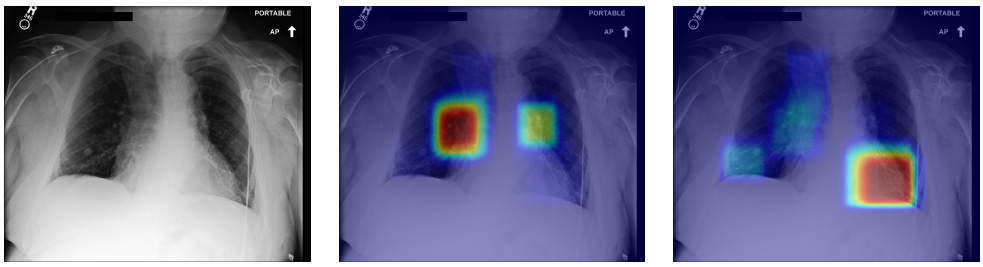
Figure 7.8: Failure case for M2Trans. The reference grounding highlights basilar pneumonia with partial overlap near the cardiac silhouette, whereas the generated report focuses exclusively on cardiomegaly, omitting the infectious component.

7.4.5.3 Case 3: MedGemma – Plausible Narrative with Fabricated Anatomical Detail

In this case as shown in Figure 7.9 and Table 7.8, MedGemma produces a fluent and clinically plausible narrative but introduces additional anatomical findings that are not supported by

Table 7.8: Qualitative failure example for the **MedGemma** model. The generated report introduces anatomically plausible but unsupported findings that are absent from the reference.

Model: MedGemma	
Reference Report	**Generated Report**
A frontal upright view of the chest was obtained portably. Since the prior study, mild interstitial edema persists but has improved. There is no focal consolidation, pleural effusion, or pneumothorax. Heart size is stable. Aortic tortuosity is unchanged. The left humeral head appears inferiorly subluxed with respect to the glenoid; however, it is not adequately visualized on this film and may be positional.	There are increased interstitial markings, which could be due to age-related changes, mild pulmonary vascular congestion, or underlying lung disease. The mediastinum appears within normal limits. **Bones: There is a fracture of the proximal humerus on the right side. Other: There is subcutaneous emphysema in the right upper chest.** Due to the portable nature of the image and patient positioning, further evaluation is recommended.
Reference Labels	**CheXbert Labels (Generated)**
Edema, no *Pneumonia*	*Lung Opacity*, *Fracture*



Original CXR | Grounding from reference report | Grounding from MedGemma generated report

Figure 7.9: Failure case for MedGemma. The reference grounding highlights diffuse interstitial edema, while the generated report induces grounding attempts for a fabricated humeral fracture and subcutaneous emphysema, resulting in low reliability.

the reference report or visual evidence. Although the reference mentions a possible positional subluxation of the "left" humeral head, the generated report instead describes a "right-sided" fracture, reflecting a semantic extrapolation beyond the available information.

During visual grounding, these fabricated entities trigger bounding-box predictions in anatomically inconsistent regions, yielding very low region-level reliability. At the same time,

the mild interstitial edema described in the reference is displaced spatially in the generated grounding, further reducing alignment between text and image.

This example highlights a characteristic behavior of foundation-scale instruction-tuned models: the generation of fluent, well-structured clinical language that may extrapolate beyond the available visual or textual evidence. Such hallucinations may remain undetected by surface-level semantic metrics or pathology extraction alone, but are effectively exposed by VICCA through joint evaluation of semantic consistency, anatomical localization, and visual reliability.

Taken together, these qualitative examples contextualize the aggregate results presented earlier. They demonstrate that strong pathology overlap, fluent language, or even high semantic similarity in isolation can mask clinically meaningful failure modes. By exposing discrepancies between textual claims and visual evidence, VICCA enables a more fine-grained diagnosis of model behavior that complements quantitative comparisons.

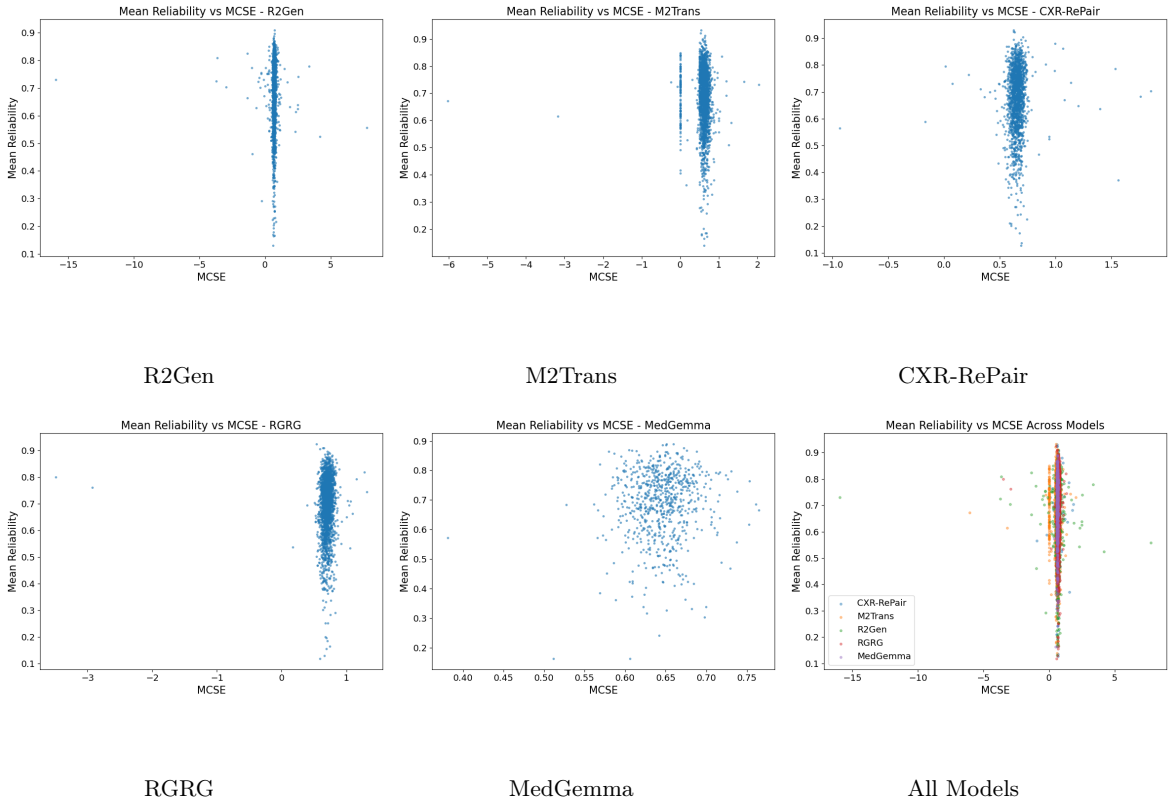
7.4.6 Integrated Comparison

Figure 7.10 illustrates the relationship between semantic similarity (MCSE) and region-level reliability. The observed positive yet shallow trend indicates that semantic alignment and visual reliability do not always evolve in parallel across models.

- **R2Gen:** Weak pathology alignment, limited semantics, and localization low coverage.
- **M2Trans:** Strongest in pathology-level prediction with high coverage, but narrative variability reduces semantic and reliability consistency.
- **RGRG:** Most balanced overall, combining strong semantic alignment with the highest localization coverage and reliable grounding.
- **CXR-RePaiR:** Clinically stable and consistent, providing moderate coverage with solid reliability across cases.
- **MedGemma:** Competitive in pathology accuracy and semantic similarity, with notably stable semantic outputs, but exhibits limited localization coverage, which constrains visual interpretability.

Correlation analysis between MCSE, CheXbert Jaccard similarity, and region-level reliability yielded undefined coefficients due to constant or missing inputs. This confirms that semantic similarity, pathology overlap, and visual reliability capture largely independent dimensions of report quality.

These findings also raise important ethical considerations regarding over-trust, accountability, and bias, which we discuss in Section 7.5.3.



R2Gen M2Trans CXR-RePair

RGRG MedGemma All Models

Figure 7.10: Scatter plot of mean reliability versus MCSE scores. Positive but shallow trends indicate that semantically similar reports do not always guarantee reliable localized grounding.

7.5 Discussion

While the Results section reports quantitative performance and statistical significance, this section focuses on interpreting the observed trade-offs and their implications for clinical deployment.

Table 7.9 summarizes the comparative evaluation. **RGRG** demonstrates the most balanced performance, excelling in semantic alignment, localization, and reliability. **M2Trans** provides the best pathology-level prediction but suffers from weaker semantic narratives. **MedGemma** performs strongly on rare pathologies and text metrics but underperforms in grounding. **CXR-RePaiR** offers stable and balanced performance across metrics. **R2Gen** remains the weakest overall, with limited pathology accuracy and grounding, though occasionally achieving competitive semantic similarity.

An important observation concerns the limited discriminative power of pathology-set overlap metrics. Although CheXbert-based Jaccard similarity reports non-zero mean values across models, these scores exhibit high variance, coarse medians, and large interquartile ranges, reflecting frequent collapse to identical or near-identical pathology sets dominated by generic labels such as “No Finding”. As a result, Jaccard similarity provides weak separation between

report generators and shows little association with semantic or visual fidelity. This limitation motivates the use of semantically informed metrics such as MCSE and image-aware evaluation through VICCA for meaningful comparison.

The inability to apply paired statistical testing to reliability scores reflects meaningful model behavior rather than missing methodology. Several models fail to produce grounded regions consistently, violating the assumptions required for paired analysis. This reinforces the role of grounding coverage itself as a discriminative signal and highlights the complementary nature of semantic and visual reliability metrics.

These findings suggest that future evaluation frameworks should explicitly combine semantic, visual, and uncertainty-aware signals rather than relying on any single metric.

Table 7.9: Comparison of RRG models evaluated with VICCA. RGRG provides the strongest balance across dimensions, while MedGemma is textually strong but visually weak.

Model	CheXbert Micro-F1	MCSE (mean)	Localization Coverage	Reliability (mean)
RGRG	0.470	**0.696**	**0.844**	**0.690**
R2Gen	0.306	0.676	0.686	0.666
CXR-RePaiR	0.452	0.657	0.677	0.675
M2Trans	0.474	0.587	0.823	0.680
MedGemma	**0.480**	0.653	0.297	0.687

7.5.1 Implications for deployment and clinical use

The proposed evaluation highlights that different report generation models may be preferable depending on the intended clinical context and risk tolerance. Importantly, these observations reflect relative behavior under the present evaluation framework rather than clinical endorsement of any system.

Models such as **RGRG**, which combine strong semantic fidelity (highest MCSE) with the highest localization coverage and reliable visual grounding, appear better suited for scenarios where interpretability and traceability of findings are critical, such as clinical triage, decision support, or auditing assistance. In such settings, the ability to verify whether textual statements are supported by image evidence is essential to mitigate hallucinations and omission errors.

Foundation-scale models such as **MedGemma** demonstrate stable semantic behavior and comparatively stronger performance on rare pathology classes, as reflected in macro-F1 and low MCSE variance. However, their limited grounding coverage suggests that human oversight remains necessary when anatomical localization or spatial verification is required. These models may therefore be more appropriate for exploratory analysis, educational support, or as assistive tools rather than autonomous reporting systems.

In contrast, **M2Trans** achieves strong pathology-level agreement but exhibits significantly weaker semantic coherence and higher variability. This indicates that accurate label prediction alone does not guarantee clinically meaningful narratives, limiting its suitability for settings where report readability and contextual reasoning are important.

Overall, these findings suggest that deployment decisions should consider not only pathology accuracy but also semantic robustness and visual grounding. Multidimensional evaluation frameworks such as VICCA can help practitioners and system designers identify model-specific strengths and failure modes before integration into clinical workflows.

7.5.2 Generalization under dataset shift: IU-Xray evaluation

To assess the robustness of the proposed evaluation framework under dataset shift, we additionally evaluated VICCA and MCSE on the IU-Xray test set [72] and results are presented in Table 7.10. Unlike MIMIC-CXR, IU-Xray contains shorter, more templated reports and follows different reporting conventions, providing a complementary evaluation setting.

Table 7.10: Cross-dataset evaluation on IU-Xray test set using VICCA and MCSE.

Model	**CheXbert Micro-F1**	**CheXbert Macro-F1**	**MCSE (Mean)**	**Reliability (Mean)**
R2Gen	–	0.000	**0.708**	**0.518**
RGRG	0.245	0.134	0.681	0.516
CXR-RePaiR	0.233	0.148	0.612	0.507
MedGemma	0.218	0.094	0.654	0.499

Overall, pathology-level agreement measured by CheXbert F1 scores decreased substantially across all models compared to MIMIC-CXR, reflecting both domain shift and sensitivity of label extraction to reporting style. Despite this degradation, semantic similarity measured by MCSE remained relatively high, and visual grounding metrics remained stable across models. This indicates that semantic coherence and visual-textual alignment persist even when explicit pathology extraction becomes unreliable.

Notably, R2Gen achieved the highest MCSE score on IU-Xray despite near-zero macro-F1,

suggesting that its generated narratives remained semantically close to reference reports even when explicit pathology mentions diverged. RGRG maintained balanced performance across semantic similarity and visual grounding, while MedGemma produced fluent but occasionally less precise reports. CXR-RePaiR remained conservative, with lower semantic overlap and grounding coverage.

Localization coverage remained high for most models, confirming that VICCA successfully grounds the majority of generated phrases even under distribution shift. Region-level reliability scores were also consistent across models, suggesting that the diffusion-based visual reliability assessment captures stable structural alignment independent of dataset-specific reporting styles.

M2Trans was excluded from this evaluation, as its publicly available checkpoint failed to generate valid reports on IU-Xray. Adapting or retraining the model specifically for this dataset would introduce additional confounding factors and was therefore outside the scope of this analysis. Among the evaluated models, only R2Gen was trained on IU-Xray, while the remaining models operated in a zero-shot setting.

Taken together, these results demonstrate that while pathology-level metrics are highly sensitive to dataset-specific conventions, the proposed VICCA and MCSE framework provides a more robust and transferable assessment of report quality across datasets.

7.5.3 Ethical and Regulatory Considerations

The results of this study highlight several ethical risks associated with the deployment of radiology report generation systems. In particular, models that produce fluent and confident narratives may induce over-trust, even when their reports contain unsupported or weakly grounded findings. Our analysis demonstrates that strong textual coherence or pathology-level accuracy does not guarantee semantic fidelity or reliable alignment with image evidence, underscoring the risk of clinicians implicitly trusting well-formed text without sufficient verification.

Accountability and responsibility remain challenging in such settings. When errors arise from AI-generated reports, it is often unclear whether responsibility lies with the model developer, the deploying institution, or the clinician who relied on the output. By explicitly linking textual statements to visual evidence and quantifying semantic and grounding reliability, VICCA provides an audit-oriented framework that supports transparency and post-hoc analysis rather than autonomous decision-making. This distinction is particularly important for regulatory and clinical governance contexts, where explainability and traceability are prerequisites for responsible use.

Finally, performance disparities across datasets and pathology categories reveal potential sources of bias. Models may underperform on rare conditions, specific reporting styles, or patient subgroups, even when aggregate metrics appear favorable. Reference-free evaluation frameworks such as VICCA should therefore be viewed as tools for systematic auditing, model

comparison, and risk assessment, rather than as substitutes for expert judgment. Used appropriately, such frameworks can help identify failure modes, guide model selection, and inform regulatory evaluation without displacing clinical responsibility.

7.5.4 Conclusion

The comparative evaluation across five representative RRG models demonstrates that no single system excels uniformly across all dimensions of clinical assessment. These findings highlight an important insight: while individual models show strengths in specific areas, they also reveal weaknesses that would remain hidden if judged solely by conventional text-based metrics. By integrating pathology-level accuracy, semantic similarity (MCSE), visual grounding, and structural reliability, VICCA provides a multidimensional evaluation framework that exposes both strengths and blind spots of RRG systems.

Through this chapter, we demonstrated that VICCA is not only capable of benchmarking diverse RRG models under clinically meaningful criteria, but also crucial in moving beyond surface-level language metrics toward trustworthiness and interpretability. In practice, VICCA reveals that high pathology accuracy does not always imply coherent narratives, and semantically rich reports may still falter in visual grounding or structural consistency. This reinforces the necessity of multi step evaluation when deploying AI in sensitive domains like radiology, where clinical reliability depends on more than just textual overlap.

In summary, VICCA's approach closes the gap between linguistic evaluation and clinically grounded validation. It equips researchers and clinicians with a comprehensive tool for understanding how well AI-generated reports reflect medical reality, both textually and visually. By exposing subtle trade-offs across models, VICCA lays the groundwork for developing future RRG systems that are not only accurate but also trustworthy, interpretable, and clinically actionable. This contribution is essential for fostering safe and effective integration of AI into radiological workflows.

Chapter 8

Conclusion and Future Work

Contents

This concluding chapter summarizes the key contributions and findings of this thesis, followed by a discussion of its main limitations. Acknowledging these limitations offers a deeper perspective on the research topic, *visual interpretation and comprehension of chest X-ray anomalies in generated reports*, and helps identify promising directions for future investigation. The discussion of contributions and outcomes is organized in relation to the research questions introduced in Chapter 1. Finally, this chapter reflects on the broader impact of the work and addresses the ethical considerations associated with developing and deploying explainable AI in medical imaging.

8.1 Summary

This thesis proposed a multimodal framework aimed at enhancing the reliability, alignment, and expressive richness of radiology report generation, as well as its objective evaluation, with a particular focus on chest X-rays (CXRs). We introduced **VICCA**, a pipeline that (i) grounds report phrases in the original CXR, (ii) synthesizes an anatomically faithful, report-consistent image via conditional diffusion guided by a binary lung mask, and (iii) quantifies report-image consistency through a reliability score derived from cross-image alignment. In parallel, we introduced **MCSE**, a semantic evaluation metric tailored to clinical text that captures entity-level agreement, negations, and modifiers beyond surface lexical overlap.

Across the dissertation, Chapter 2 presented visual grounding for CXRs; Chapter 3 detailed guided image synthesis using diffusion and mask conditioning; Chapter 5 motivated

and defined MCSE for semantic evaluation of radiology reports; and the subsequent chapters benchmarked representative report-generation models (e.g., R2Gen, M2Trans, CXR-RePaiR, RGRG, and MedGemma), applying both VICCA and MCSE for a joint visual-textual assessment.

8.2 Contributions

1. **A reliability framework (VICCA) for report-image alignment.** We formulate a cross-modal validation loop that compares grounded regions in the original CXR with regions synthesized from the report, yielding a reliability score that reflects visual support for textual claims.

2. **Automatic Anatomical Region Detection for Chest X-rays.** To improve the anatomical interpretability and evaluation capabilities of VICCA, we trained a DETR-style model to automatically detect 36 anatomical regions in CXR images. This detector was later integrated to augment training data for the visual grounding task. Since many public datasets lack paired image–report annotations, our trained detector was used to infer the most relevant anatomical region associated with each pathology and to generate custom region-level descriptors, thereby enriching weakly supervised datasets with structured spatial information.

3. **Visual Grounding for Chest X-rays Using a Customized Text Encoder.** We fine-tuned a visual grounding model specifically for CXR data by replacing the generic text encoder with a domain-adapted CXR-specific encoder and subsequently adjusting the image encoder for improved feature alignment. This adaptation enhances the model's ability to accurately localize key entities within the report, such as anatomical regions and pathological findings, resulting in more precise and clinically reliable grounding.

4. **Guided CXR synthesis via conditional diffusion with mask control.** We adapt Stable Diffusion with ControlNet and a medical text encoder to generate CXRs from reports while preserving anatomical structure using a binary lung mask. This makes synthesis spatially faithful to parenchymal regions and robust for downstream visual alignment.

5. **A clinical semantic metric (MCSE) for report evaluation.** We introduce an entity and relation aware similarity measure that better captures clinical semantics (e.g., presence/absence, negation, and modifiers) than generic NLG metrics, improving textual assessment for medical reports.

6. **A unified evaluation of SOTA report generators.** We conduct a comparative study of state-of-the-art models for CXR report generation (including the foundation model MedGemma), demonstrating how *VICCA + MCSE* together reveal strengths and weaknesses not visible to BLEU/ROUGE/CIDEr alone (e.g., visual grounding mismatches, semantic omissions).

8.3 Outcomes

- **Enhanced visual grounding accuracy.** The customized CXR-specific text encoder improved phrase-level localization, enabling the model to more precisely link textual entities, such as anatomical structures and pathologies, to their visual counterparts. This adaptation increased the grounding reliability and interpretability of the VICCA pipeline.

- **Automated anatomical region detection.** The DETR-based anatomical region detector achieved robust identification of 36 key thoracic structures, providing the foundation for evaluating anatomical fidelity in generated images. Its integration further allowed augmentation of weakly annotated datasets, strengthening both the grounding and evaluation components of VICCA.

- **Visual fidelity and structural consistency in generation.** The guided diffusion process successfully preserved lung morphology and localized findings under binary mask constraints. Quantitative measures (MS-SSIM, Dice, FID) and IoU scores from the anatomical detector validated the structural realism of the generated CXRs.

- **Semantic adequacy beyond lexical similarity.** The MCSE metric demonstrated stronger alignment with clinically meaningful variations, such as negations, severity modifiers, and entity relations, than traditional NLP metrics, effectively distinguishing fluent yet clinically inaccurate outputs from truly reliable ones.

- **Integrated evaluation for reliability assessment.** Combining VICCA and MCSE provided a dual-layer validation of report quality. Reliability scores decreased when text–image misalignments occurred (e.g., incorrect laterality or misplaced findings), underscoring the framework's sensitivity to subtle inconsistencies often overlooked by lexical metrics.

- **Refined evaluation of foundation models.** Applying the pipeline to MedGemma revealed the importance of domain-specific preprocessing. Focusing on the clinically relevant Findings section improved grounding accuracy and report–image coherence, illustrating the adaptability of VICCA and MCSE to modern multimodal LLMs.

8.4 Broader Impact

The proposed framework advances multimodal verification for clinical AI by shifting emphasis from generic fluency to clinically grounded agreement between text and image. Beyond evaluation, guided synthesis can support data augmentation for rare findings (e.g., inserting a small pulmonary nodule in the left lower lung to address class imbalance). The general idea, closing the loop between language and vision with alignment scores, could extend to other modalities (e.g., ultrasound, CT projections) and to education (e.g., training clinicians using pairs of real/synthetic examples that illustrate specific findings).

8.5 Limitations

- **Mask granularity and out-of-mask control.** Spatial control relies on a binary lung mask. Structures outside the lungs (e.g., ribs, devices) are modeled implicitly by the text encoder and learned priors. More granular, multi-class masks could provide finer spatial control.
- **Module error propagation.** The pipeline is multi-stage. Errors in grounding (false negatives/positives) or synthesis (hallucinations) can affect the final reliability score. While joint evaluation mitigates some issues, systematic calibration of module-level uncertainties is needed.
- **Score calibration with expert judgments.** VICCA and MCSE produce continuous scores, but clinical thresholds (e.g., what constitutes a "good" alignment) require calibration against radiologist ratings on a representative subset.
- **Data and domain shift.** Results depend on training corpora (e.g., MIMIC-CXR). Generalization to other institutions, devices, and populations requires further validation and domain adaptation.

8.6 Ethical Considerations

- **Privacy and governance.** Use of clinical data must follow strict de-identification and governance. Synthetic images should be labeled as such to prevent misuse or data leakage.
- **Bias and fairness.** Dataset imbalances can induce biased decisions. Augmentation with guided synthesis should be monitored to avoid reinforcing spurious correlations or masking minority patterns.
- **Transparency and human oversight.** VICCA/MCSE are assistive tools, not diagnostic authorities. Outputs should be accompanied by uncertainty and alignment indicators, and remain under clinician oversight.
- **Responsible augmentation.** Synthetic data should be used with caution: its distribution differs from real images and may influence downstream models in unintended ways if overused or misapplied.

8.7 Future Directions

1. **From phrases to full reports.** Extend grounding to report-level graphs that encode entities, relations, negations, and temporal references; integrate MCSE signals directly into the visual alignment objective.

2. **Richer spatial control.** Replace binary masks with multi-class anatomical segmentation (e.g., lobes, mediastinum, skeletal landmarks), and explore soft attention maps learned from detectors for fine-grained placement.

3. **Template-Guided Visual Encoding.** Incorporate a new Medical Template-Guided Visual Encoder (TGVE), which uses structural priors and learned anatomical templates to guide feature extraction and localization. By embedding spatial templates derived from anatomical atlases or prior detector outputs, TGVE can constrain visual attention to medically relevant regions, reducing grounding drift and improving cross-domain consistency.

4. **Uncertainty and calibration.** Attach calibrated uncertainty to VICCA and MCSE scores; learn thresholds from radiologist panels, and quantify module-level error propagation via controlled perturbations.

5. **Generalization and robustness.** Evaluate across multi-center datasets and domain shifts; incorporate test-time adaptation and robustness checks (e.g., device artifacts, post-operative changes, OOD conditions).

6. **Human-in-the-loop workflows.** Prototype interactive tools where VICCA flags low-alignment regions, MCSE highlights semantic conflicts (e.g., negations vs. findings), and clinicians can correct or approve with minimal friction.

7. **Augmentation for rare conditions.** Systematically study guided synthesis for targeted rebalancing (e.g., small nodules, pneumothorax at apices), with safeguards against overfitting to synthetic patterns.

Closing Remarks

This thesis advances the clinically grounded evaluation of radiology report generation by coupling **what the text says** with **what the image shows**. By unifying VICCA's visual alignment and MCSE's semantic fidelity, we provide complementary lenses on correctness that go beyond generic NLG metrics. We believe these tools can help the field transition from fluency-centric benchmarking toward verifiable, trustworthy systems that better reflect clinical reality.

Appendix

Appendix A

Database Description

This appendix provides a detailed description of the datasets referenced in the thesis. For each dataset, we outline the available modalities, associated metadata, annotation formats, and potential challenges for visual-textual tasks such as report generation and grounding.

A.1 MIMIC-CXR

(see page 24, 103)

The MIMIC-CXR dataset [50, 51] is one of the largest publicly available chest radiograph datasets, released by the MIT Lab for Computational Physiology in collaboration with Beth Israel Deaconess Medical Center. It contains both imaging data and associated free-text radiology reports.

General Statistics

- **Number of Studies**: ∼377,000
- **Number of Images**: ∼227,000
- **Number of Patients**: ∼65,000
- **Image Views**: Frontal (AP/PA), Lateral
- **Image Format**: DICOM (.dcm), with de-identified metadata

Modalities

- **Chest X-ray Images**: Frontal and lateral views in grayscale DICOM format.
- **Radiology Reports**: Unstructured free-text reports, usually composed of three sections, *Findings*, *Impression*, and *Indication*.
- **DICOM Metadata**: Includes patient age, sex, acquisition parameters, view position, and study timestamps.

Annotations and Derived Labels

While the raw dataset does not contain structured labels or bounding boxes, several derived label sets exist:

- **CheXpert Labels**: Automatically generated using the CheXpert labeler [47], assigning 14 pathology labels (e.g., consolidation, edema).
- **Negation/Uncertainty Tags**: Labeler output includes negation and uncertainty for better clinical fidelity.

Use Cases in This Thesis

MIMIC-CXR is used throughout this work for:

- Training and evaluating report generation models (e.g., R2Gen, CXR-RePaiR).
- Visual grounding tasks where reports are aligned with localized regions.
- Phrase extraction for semantic similarity metrics like MCSE.

Challenges

- Lack of ground-truth bounding box annotations for most studies.
- Highly variable report length and linguistic style.
- Imbalanced label distributions across findings.

A.2 MS-CXR Dataset

(Mentioned on page 26)

The MS-CXR dataset [10] is a domain-specific benchmark tailored for the task of phrase grounding in chest radiography. Derived from the larger MIMIC-CXR corpus, it comprises 1,162 chest X-ray images, each annotated with one or more bounding boxes associated with short textual descriptions extracted from the corresponding radiology reports from MIMIC-CXR. These annotations represent a fine-grained mapping between linguistic entities, such as “opacity in the right upper lobe” or “enlarged cardiac silhouette”—and the image regions where they are observed.

Each image is accompanied by a set of text phrases and their corresponding spatial annotations, enabling supervised training and evaluation of visual grounding models in the medical

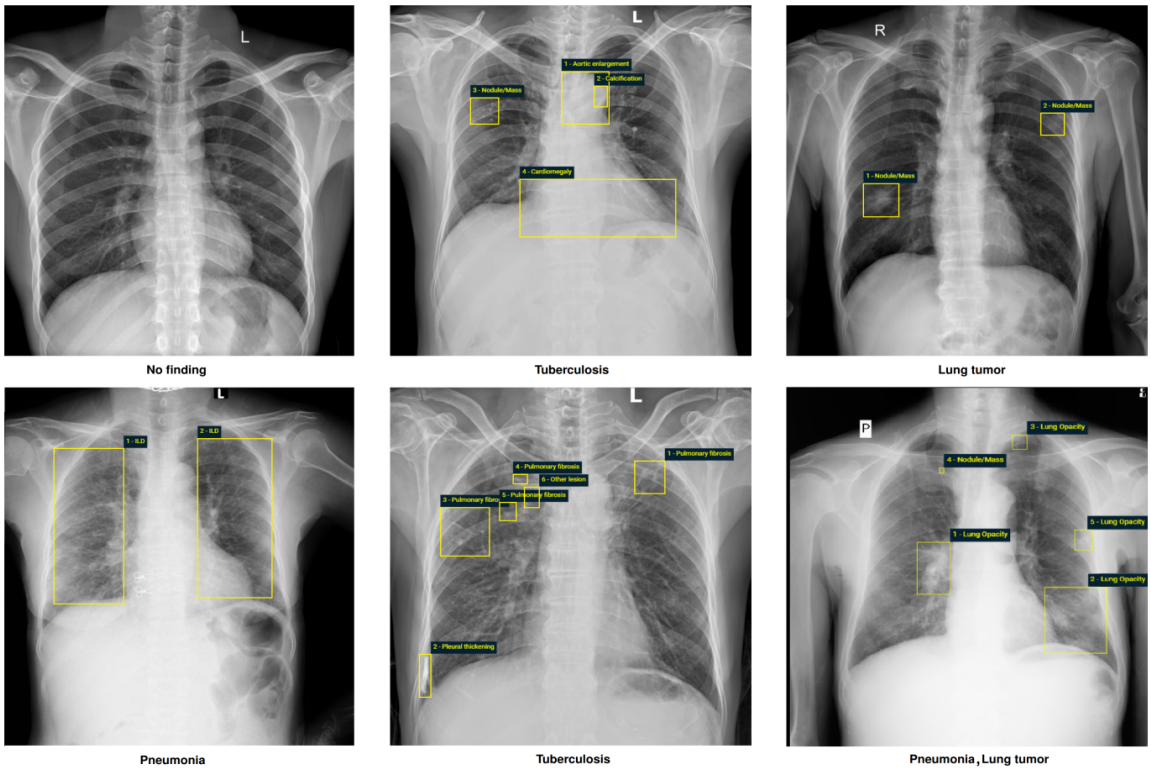


Figure A.1: Examples of chest X-rays with expert radiologist annotations. Local findings are highlighted with bounding boxes overlaid on the original images for visualization. Global diagnostic labels are shown in bold below each example that come from [81].

context. The dataset supports the alignment of natural language queries with radiological findings, making it particularly valuable for developing explainable AI systems that integrate vision and language in diagnostic workflows.

While MS-CXR provides high-quality annotations, its relatively small size limits its use for training deep learning models from scratch. As such, it is most effective when used for fine-tuning or evaluation, often in conjunction with larger datasets that provide complementary information (e.g., anatomical locations or disease presence). Despite its scale limitations, MS-CXR remains a foundational resource for phrase-level localization tasks in chest X-ray interpretation.

A.3 VinDr-CXR Dataset

(Mentioned on page: 26)

VinDr-CXR is a large-scale, publicly available chest X-ray dataset that offers rich expert-annotated data, making it highly suitable for medical vision-language research, Figure A.1. The dataset comprises 18,000 postero-anterior (PA) chest radiographs collected retrospectively

from two major hospitals in Vietnam between 2018 and 2020 [81].

Each image in VinDr-CXR is annotated with both local and global labels: 22 local radiographic findings (e.g., *pleural effusion*, *nodule*) are provided as bounding boxes, while 6 global diagnoses represent impression-level conclusions. The annotations were generated through a rigorous labeling process using VinDr Lab, a web-based annotation tool built on a PACS (Picture Archiving and Communication System) infrastructure.

- **Modality:** Chest X-ray (PA view)
- **Annotations:** 22 findings and 6 diseases, including bounding boxes for findings: (1) Aortic enlargement, (2) Atelectasis, (3) Cardiomegaly, (4) Calcification, (5) Clavicle fracture, (6) Consolidation, (7) Edema, (8) Emphysema, (9) Enlarged PA, (10) Interstitial lung disease (ILD), (11) Infiltration, (12) Lung cavity, (13) Lung cyst, (14) Lung opacity, (15) Mediastinal shift, (16) Nodule/Mass, (17) Pulmonary fibrosis, (18) Pneumothorax, (19) Pleural thickening, (20) Pleural effusion, (21) Rib fracture, (22) Other lesion, (23) Lung tumor, (24) Pneumonia, (25) Tuberculosis, (26) Other diseases, (27) Chronic obstructive pulmonary disease (COPD), and (28) No finding.
- **Metadata:** Image-level labels, radiologist IDs, acquisition parameters
- **Use case:** Phrase-level localization, object detection in CXR, weakly-supervised learning

The dataset is divided into two main parts:

- **Training set:** 15,000 images annotated independently by three experienced radiologists.
- **Test set:** 3,000 images annotated by a panel of five senior radiologists using a majority-vote consensus.

VinDr-CXR provides high-resolution DICOM images, preserving clinical quality standards. The combination of detailed spatial annotations and impression-level diagnoses makes this dataset particularly valuable for phrase grounding tasks in medical imaging. It enables training of models that align textual phrases from radiology reports with specific regions in chest X-ray images, an essential step toward interpretable and trustworthy AI systems in healthcare.

A.4 Chest ImaGenome Dataset

(Mentioned on pages: 29)

The **Chest ImaGenome** dataset is a richly annotated extension of the MIMIC-CXR dataset, designed to enable fine-grained visual grounding tasks in chest radiographs. It provides **comprehensive region-level annotations** across 10,000 frontal-view chest X-ray im-

ages, making it one of the first large-scale medical datasets to support detailed phrase grounding.

Each image is annotated with both:

- **Anatomy and pathology labels** (ranging from 36 anatomical regions and 26 condition types), and
- **Structured visual descriptors** in natural language, totaling over **700,000 phrase-region associations**.

The dataset structure follows a hierarchical schema where each annotated region includes:

- A bounding box,
- A label (e.g., "right lower lung zone"),
- A natural language descriptor (e.g., "there is a subtle opacity"), and
- The associated radiology report from the original MIMIC-CXR study.

In addition to facilitating supervised learning for phrase grounding and region tagging, the dataset supports interpretability studies and model benchmarking for clinical report understanding.

A.5 Visual Genome Dataset

(see page 20)

The **Visual Genome** dataset [58] is a large-scale multimodal dataset designed to bridge the gap between computer vision and natural language processing. It serves as a foundational resource for tasks involving visual grounding, scene understanding, and image captioning. Figure A.2 presents a sample from this dataset with the graph of text features and their cnnection.

Description

Visual Genome contains over 108,000 images sourced primarily from the MS COCO dataset [68]. Each image is richly annotated with region descriptions, object labels, attributes, relationships, and question–answer pairs. Unlike many image datasets that offer only image-level labels, Visual Genome provides dense annotations at the region level, making it ideal for fine-grained visual-textual alignment tasks.

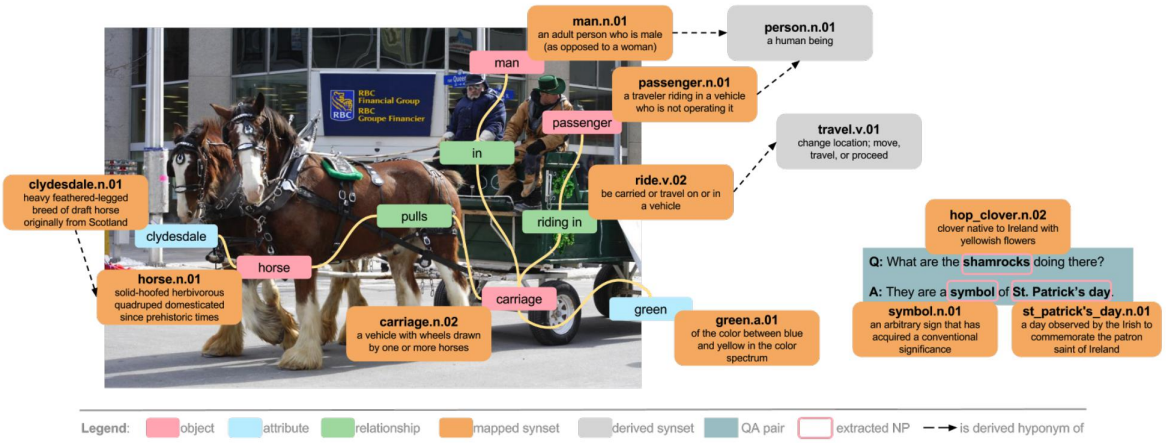


Figure A.2: An example of Visual Genome dataset showing the graph of text features and their cnnection.

Modalities and Annotations

The dataset provides multimodal annotations across several structured layers:

- **Region Descriptions**: Each image is segmented into multiple subregions, each associated with a natural language phrase that describes the content of the region.
- **Objects**: Over 21,000 object categories are annotated with bounding boxes, allowing for detailed spatial localization.
- **Attributes**: Descriptive properties (e.g., "red", "large") are associated with objects.
- **Relationships**: Triplets such as `<man, holding, umbrella>` are annotated to capture inter-object relations.
- **Scene Graphs**: Graph-based representations of object–attribute–relationship combinations for structured scene understanding.
- **Question-Answer Pairs**: Around 1.7 million Q&A pairs for visual question answering (VQA) tasks.

Metadata Overview

- **Image Count**: ∼108,000
- **Average Regions per Image**: 42
- **Total Region Descriptions**: 5.4 million
- **Total Objects**: ∼2.3 million

- **Total Relationships**: ∼1.5 million
- **Image Source**: MS COCO
- **Modality**: RGB images + Natural language text

Relevance to This Work

Visual Genome is mentioned in this thesis as a pretraining and evaluation resource for visual grounding models, especially in the context of grounding free-form text phrases to image regions. Its detailed region–text associations help build models that can later be transferred or adapted to medical datasets where similar grounding between radiological findings and anatomical regions is desired.

A.5.1 CheXpert Dataset

CheXpert [47] is a large chest X-ray dataset consisting of 224,316 radiographs from 65,240 patients collected at Stanford Hospital. Each study is associated with an automatically extracted label set covering 14 thoracic conditions, including *Atelectasis*, *Cardiomegaly*, *Edema*, *Pleural Effusion*, and *Consolidation*. Labels take one of four states: **positive**, **negative**, **uncertain**, or **not mentioned**, assigned via a rule-based labeler designed for high coverage and consistency.

Although CheXpert does not include free-text radiology reports, it serves as an important benchmark for evaluating report-generation models through clinical label agreement metrics such as CheXpert F1 (see Appendix C.1.1). Its well-defined diagnostic categories and uncertainty modeling make it a widely used reference for assessing the clinical validity of generated reports.

Appendix B

Large Language Models

B.1 BERT: Bidirectional Encoder Representations from Transformers

(Mentioned on page 23)

BERT (Bidirectional Encoder Representations from Transformers) is a groundbreaking language representation model introduced by Devlin et al. [28]. Unlike previous language models that are unidirectional (e.g., left-to-right in GPT), BERT employs a deep bidirectional Transformer architecture that jointly conditions on both left and right contexts at all layers, enabling a more holistic understanding of language.

BERT is pre-trained on large corpora using two self-supervised tasks:

- **Masked Language Modeling (MLM)**: Randomly masks a subset of input tokens and predicts them using surrounding context, encouraging bidirectional encoding.

- **Next Sentence Prediction (NSP)**: Learns sentence-pair relationships by predicting whether one sentence logically follows another.

Once pre-trained, BERT can be fine-tuned with minimal architectural changes on a wide range of NLP tasks such as question answering (SQuAD), natural language inference (MNLI), and sentiment analysis (SST-2).

The model achieved state-of-the-art results across multiple benchmarks at the time of publication, with the $\texttt{BERT}_{\texttt{LARGE}}$ variant surpassing previous records on GLUE, SQuAD, and SWAG tasks. The architecture's core strength lies in its ability to leverage deep bidirectional representations, setting a new standard in NLP model pre-training.

B.2 CLIP: Learning Transferable Visual Models From Natural Language Supervision

(Mentioned on page 23)

The CLIP (Contrastive Language–Image Pretraining) model, introduced by Radford et al. [90], presents a scalable approach to learning vision-language representations by training on a vast dataset of 400 million image-text pairs collected from the internet. Instead of relying on supervised labels, CLIP leverages natural language supervision using a contrastive learning objective.

CLIP trains two separate encoders: a Vision Transformer (ViT) or ResNet-based image encoder and a Transformer-based text encoder. Both encoders map their respective modalities into a shared embedding space. During training, the model learns to maximize the similarity between corresponding image and text embeddings (positive pairs) while minimizing it for mismatched pairs (negative examples). This contrastive loss enables CLIP to align visual and textual semantics in a highly generalizable way.

A key strength of CLIP lies in its zero-shot transfer capabilities. Without any additional fine-tuning, the model can perform various vision tasks by simply providing text prompts (e.g., "a photo of a cat") as classifier surrogates. This approach allows CLIP to rival or outperform fully supervised models on over 30 datasets, demonstrating strong generalization and flexibility. CLIP's pretraining method unlocks rich cross-modal understanding, paving the way for downstream applications such as image captioning, visual question answering, and open-vocabulary detection.

B.3 BioViL-T: BiomedVLP-CXR-BERT for Vision-Language Pretraining

(Mentioned on page 33)

BioViL-T BiomedVLP-CXR-BERT, also known as **CXR-BERT**, is a transformer-based vision-language model specifically designed for chest X-ray (CXR) interpretation. Introduced in the MS-CXR benchmark paper [10], BioViL-T extends the BERT architecture [28] with radiology-specific pretraining using paired image-report data. The architecture is presented in Figure B.1, It aligns visual and textual modalities by jointly learning from CXR images and corresponding radiology reports in MIMIC-CXR, enabling the model to understand clinically relevant associations.

BioViL-T shows superior performance in various CXR downstream tasks, including zero-shot classification and grounding of findings. The model is pretrained using image-text match-

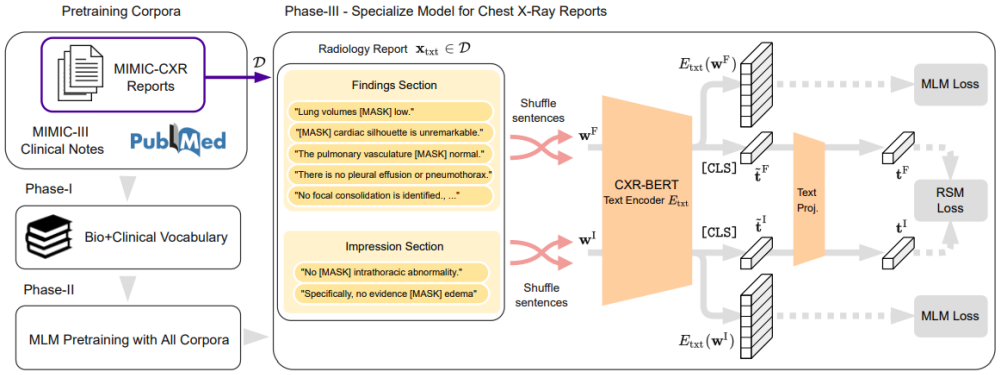


Figure B.1: Overview of the CXR-BERT text encoder architecture. The model undergoes a three-phase pretraining strategy, leveraging a domain-specific vocabulary alongside masked language modeling (MLM) and radiology section matching (RSM) objectives. The training is further enhanced by regularization techniques and radiology-specific text augmentations.

ing (ITM) and image-text contrastive (ITC) objectives, and uses a ResNet-50 as the image encoder and a BERT-style text encoder initialized with ClinicalBERT [44].

Table B.1: Comparison of Language Models in Clinical Contexts

Model	**Training Data**	**Domain Specialization**	**Example Use Case**
BERT [28]	General-domain corpora (Wikipedia, BooksCorpus)	None (general-purpose NLP)	Sentence classification (e.g., "The heart is enlarged." → positive abnormality)
ClinicalBERT [44]	MIMIC-III clinical notes	Clinical terminology and abbreviations	Named Entity Recognition (NER) in discharge summaries (e.g., detect "congestive heart failure")
BiomedVLP-CXR-BERT [10]	MIMIC-CXR reports (paired with images)	Chest X-ray radiology reports	Align image regions to phrases (e.g., "right basilar opacity") or generate CXR captions

Compared to prior models:

- **BERT** is a general-purpose language model that lacks domain-specific medical knowledge.
- **ClinicalBERT** is fine-tuned on clinical notes (e.g., MIMIC-III [52]), improving contextual understanding in healthcare narratives, but it lacks visual modality integration that associate textual descriptions with corresponding medical images.
- **BioViL-T (CXR-BERT)** is trained on paired chest X-ray and report data, allowing it to bridge visual and textual modalities for clinically meaningful image interpretation.

Table B.1 compares further these three models.

B.3.1 Comparison of CXR-BERT with ClinicalBERT and PubMedBERT

Table B.2: Vocabulary comparison for domain-specific terminology.

Term	**ClinicalBERT**	**PubMedBERT** [37]	**CXR-BERT**
pneumonia	✓	✓	✓
opacity	op-acity	✓	✓
effusion	e-ff-usion	✓	✓
pneumothorax	p-ne-um-oth-orax	✓	✓
atelectasis	ate-lect-asis	ate-le-ct-asis	✓
cardiomegaly	card-io-me-gal-y	cardio-me-gal-y	✓
bibasilar	bi-bas-ila-r	bib-asi-la-r	✓

CXR-BERT uses a 30k WordPiece vocabulary derived from PubMed abstracts, MIMIC-III clinical notes [52], and MIMIC-CXR reports [50]. This tailored vocabulary significantly reduces subword fragmentation for domain-relevant terms compared to its predecessors.

Table B.2 compares how different biomedical language models tokenize domain-specific medical terms. It highlights the advantage of CXR-BERT, whose specialized vocabulary, built from clinical and radiology corpora—preserves key medical entities more effectively, reducing subword fragmentation and improving linguistic precision for chest X-ray reporting tasks.

Intrinsic Evaluation Results

Table B.3 summarizes performance on masked token prediction and natural language inference (NLI) using RadNLI.

Table B.3 summarizes the performance of the evaluated models on two intrinsic language understanding tasks: masked token prediction and natural language inference (NLI). The NLI task is based on the RadNLI dataset [77], which contains radiology-specific sentence pairs labeled as entailment, contradiction, or neutral. This dataset is designed to assess a model's ability to understand and reason over clinical language nuances, making it works as a benchmark for evaluating domain-specific language representations in radiology.

Table B.3: Intrinsic evaluation of language models on MIMIC-CXR.

Model	**RadNLI Accuracy**	**Mask Acc.**	**Avg. Tokens**
ClinicalBERT	47.67%	39.84%	78.98
PubMedBERT	57.71%	35.24%	63.55
CXR-BERT (Phase III)	60.46%	77.72%	58.07
CXR-BERT + Joint Training	**65.21%**	**81.58%**	58.07

Table B.4: Contrast-to-noise ratio (CNR) on MS-CXR dataset.

Text Encoder	**Avg. CNR**	**Training Objective**
ClinicalBERT	0.769	Global
PubMedBERT	0.773	Global
CXR-BERT	1.027	Global
CXR-BERT (BioViL-L)	**1.142**	Global & Local

Phrase Grounding Performance

When used in the BioViL framework for vision-language tasks, CXR-BERT outperforms both PubMedBERT and ClinicalBERT on the MS-CXR dataset, as shown in Table B.4. These results confirm that domain-specific adaptation significantly enhances both linguistic and multimodal grounding performance in chest X-ray applications.

Appendix C

Language Assessment Metrics

C.1 Domain-Specific Evaluation Metrics

To address the limitations of general-purpose metrics in radiology report generation, several medically informed metrics have been proposed. Among them, **CheXpert F1** and **RadGraph F1** are two of the most widely used for evaluating clinical accuracy.

C.1.1 CheXpert F1

(Mentioned on page 80)

CheXpert F1 is a pathology-centric metric that evaluates whether a generated report correctly identifies the presence or absence of specific thoracic diseases. It uses the CheXbert labeler [102], a rule-based or transformer-based tool trained to classify reports according to the 14 pathologies defined in the CheXpert dataset [47] (e.g., Atelectasis, Edema, Cardiomegaly).

Both the reference and generated reports are passed through the CheXbert labeler, which extracts labels (positive, negative, uncertain) for each condition. The F1 score is then computed based on the overlap of these predicted labels.

Strengths:

- Offers interpretable, label-based evaluation of clinical content.
- Considers uncertainty explicitly.

Limitations:

- Limited to 14 predefined labels.
- Ignores semantic richness and syntactic structure.

Table C.1: Comparison between RadGraph F1 and CheXpert F1 metrics

Metric	What it Evaluates	Limitations
CheXpert F1	Presence or absence of 14 thoracic diseases using rule-based/ML-based labeler	Limited to predefined labels; ignores phrasing or contextual differences
RadGraph F1	Overlap of entities and relations in a structured graph (observations and anatomical locations)	Requires trained NER and RE models; limited generalizability outside CXR domain

C.1.2 RadGraph F1

(Mentioned on page 80)

RadGraph F1 [48] evaluates the clinical correctness of a report by converting its content into a structured graph of medical entities and relations. The metric relies on a dedicated **Named Entity Recognition (NER)** and **Relation Extraction (RE)** model trained specifically on radiologist-annotated chest X-ray reports from the MIMIC-CXR dataset [50].

NER Model. The RadGraph NER component identifies and classifies two categories of clinically meaningful entities:

- **Observations:** pathology terms and findings (e.g., "consolidation", "effusion", "opacity").
- **Anatomical Locations:** spatial descriptors describing the body region (e.g., "right lower lobe", "left costophrenic angle").

These entities are extracted using a transformer-based architecture (a Bio+ClinicalBERT encoder) fine-tuned on the RadGraph training corpus. The model jointly predicts:

1. entity spans (start/end tokens),
2. entity types (observation vs. anatomical location),
3. entity attributes (e.g., uncertainty labels).

Relation Extraction. Once entities are identified, the RE model predicts semantic relations between them, most notably:

- **located_at**: linking an observation to a specific anatomical location;
- **suggestive_of**: linking a finding to a possible diagnosis.

This structured representation allows evaluation beyond surface text similarity by checking whether the generated report conveys the same factual medical meaning as the reference.

RadGraph F1 Computation. The RadGraph F1 score is calculated by comparing the overlap of:

- extracted entities (type + textual span),
- extracted relations between entities,

between the generated and reference report. The final score is the harmonic mean of precision and recall over the combined entity+relation set.

Strengths:

- Evaluates structured clinical meaning rather than lexical similarity.
- Captures both what is reported and where it is located.

Limitations:

- Performance depends on NER/RE accuracy; errors propagate to the metric.
- Does not capture semantic near-matches or paraphrases.
- Limited to the ontology and relation set defined in the RadGraph annotation scheme.

C.1.3 RadCliQ

(Mentioned on page 106)

RadCliQ [121] is a clinically oriented evaluation metric designed to better reflect radiologist preferences when assessing the quality of automatically generated radiology reports. Unlike lexical metrics (e.g., BLEU, ROUGE) or entity-based metrics (e.g., CheXpert F1, RadGraph F1), RadCliQ operates as a **learned metric**: a regression model trained to predict how closely a generated report aligns with expert judgment.

Method Overview

RadCliQ is trained by comparing model-generated reports to radiologist-labeled ground truth reports using several input features:

- RadGraph entity and relation overlaps (observation and anatomy consistency);
- Report-level semantic features (e.g., entity density, relation coverage);
- Textual characteristics (e.g., sentence length, structural patterns).

A human-labeled dataset is used to regress the perceived report quality onto these features, producing a score that acts as a surrogate for radiologist evaluation.

Interpretation

Higher RadCliQ scores indicate better agreement with radiologist judgment. Because it incorporates structured clinical information (entities, relations) rather than purely lexical similarity, RadCliQ is more robust to paraphrasing and better aligned with diagnostic correctness.

Strengths

- Predicts radiologist preference rather than surface text similarity.
- Leverages RadGraph features, improving clinical sensitivity.
- Less penalized by stylistic differences or paraphrasing.

Limitations

- Requires a trained regression model, making it dataset-dependent.
- Cannot be directly interpreted without calibration (scores are relative).
- Sensitive to RadGraph extraction errors.

RadCliQ provides a complementary perspective to entity-based and lexical metrics, offering a more clinically aligned measure of report quality. It is particularly useful when evaluating large models (e.g., MedGemma) or architectures that produce fluent text with potential factual inconsistencies.

Appendix D

Radiology Report Generation Comprehensive Study

D.1 R2Gen: Generating Radiology Reports via Memory-driven Transformer

(Mentioned on page 105)

R2Gen [20] is a neural architecture designed specifically for generating radiology reports from CXR images. It extends the conventional encoder-decoder Transformer framework by incorporating two key components: Relational Memory (RM) and Memory-driven Conditional Layer Normalization (MCLN). These components are introduced to address the challenges of generating long, structured, and clinically coherent reports by modeling temporal dependencies and enhancing contextual awareness during decoding.

Model Architecture

The R2Gen architecture consists of three primary components as illustrated in Figure D.1: a visual encoder, a Transformer encoder, and a Transformer decoder augmented with relational memory and memory-conditioned normalization.

A pretrained CNN, typically ResNet101, is used to extract visual features from the input CXR image; this network is pretrained on ImageNet [26] to leverage general visual representations before fine-tuning on the medical domain. These features are tokenized and passed into a standard Transformer encoder to generate high-level visual representations. The decoder then takes these representations and sequentially generates tokens of the report, conditioned on the encoded visual information and the previously generated tokens.

Relational Memory

The relational memory module maintains a learnable memory matrix M_t that evolves throughout the generation process. At each time step t, the memory interacts with the previously

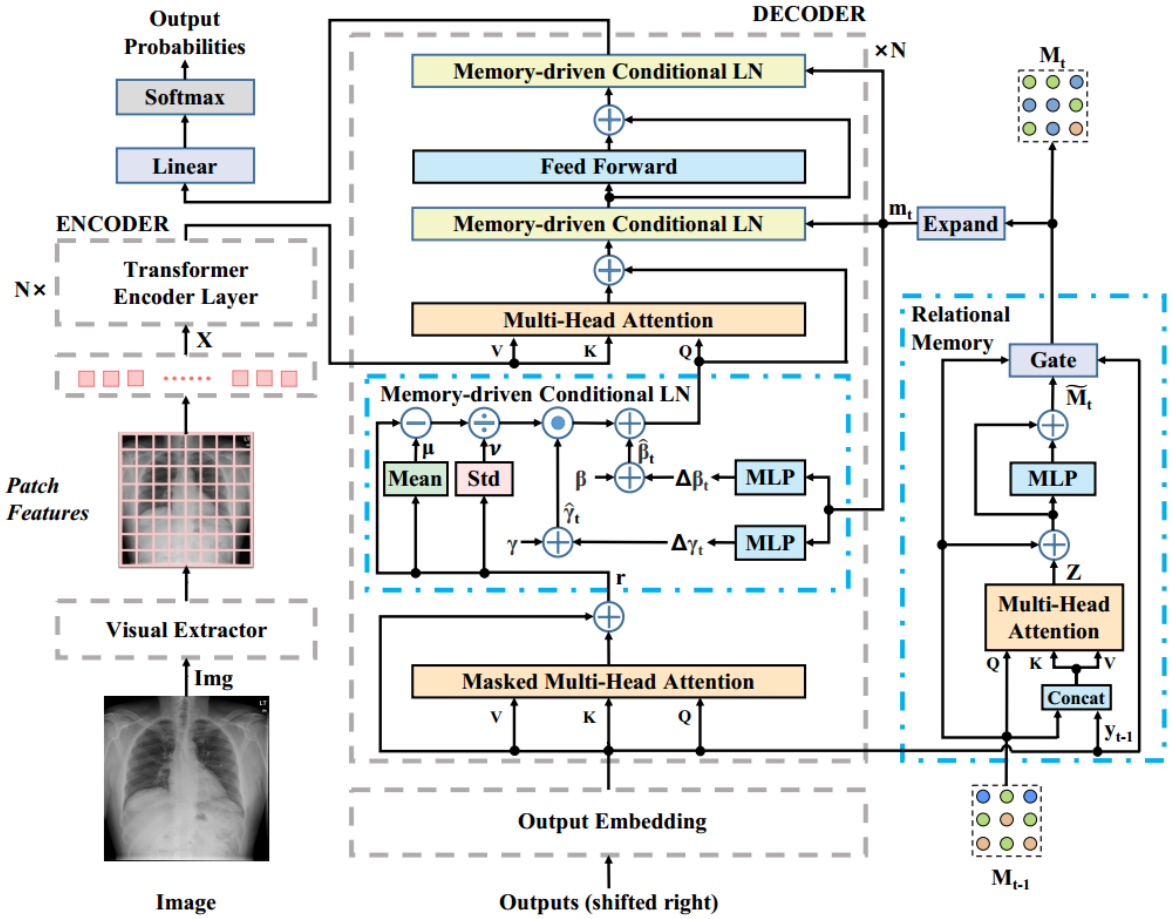


Figure D.1: The overall architecture of R2Gen model, where the visual extractor, encoder and decoder are shown in gray dash boxes and the details of the visual extractor and encoder are omitted. The relational memory and memory conditional layer-normalization are illustrated in grey solid boxes with blue dash lines.

generated token embedding y_{t-1} to update its state. This interaction is governed by a multi-head attention mechanism, where M_{t-1} is used as the query and $[M_{t-1}; y_{t-1}]$ serves as the key and value.

The updated memory is then modulated using gating mechanisms, including an input gate and a forget gate presented in Figure D.2, which regulate the contribution of the new input versus the retained memory. A residual connection and a multi-layer perceptron (MLP) further refine the memory update, allowing it to capture long-range dependencies and recurring clinical phrases.

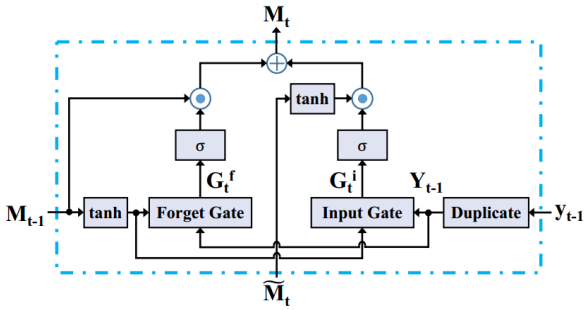


Figure D.2: The illustration of the gate mechanism of R2Gen model.

Memory-Driven Conditional Layer Normalization

To enhance the integration of contextual information, R2Gen introduces a dynamic normalization strategy where the parameters of the layer normalization are conditioned on the current memory state. Specifically, the memory output is flattened and passed through an MLP to generate offset parameters $\Delta\gamma_t$ and $\Delta\beta_t$, which are then added to the base normalization parameters γ and β:

$$\gamma_t = \gamma + \Delta\gamma_t, \quad \beta_t = \beta + \Delta\beta_t.$$

This allows the normalization layer to adapt based on the evolving semantic context stored in the relational memory, improving the fluency and coherence of the generated reports.

Training Objective

R2Gen is trained using the negative log-likelihood loss over the report tokens:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \text{Image}),$$

where y_t is the token at step t and the model conditions on all previous tokens and the encoded image. Beam search with a width of 3 is used during inference to enhance the quality of generation.

Experimental Results

R2Gen is evaluated on two public datasets: IU X-Ray [72] and MIMIC-CXR [50, 51]. Performance is measured using both traditional natural language generation metrics (BLEU [84], METEOR [7], ROUGE-L [66]) and clinical efficacy metrics based on CheXpert label classification [47] (precision, recall, F1 score). The results demonstrate that the inclusion of relational memory and conditional layer normalization improves both linguistic and clinical accuracy.

For instance, on the MIMIC-CXR dataset, R2Gen achieves a BLEU-4 score of 0.103 and a clinical F1 score of 0.276, outperforming baseline models such as standard Transformers and previous medical captioning systems.

Relevance to VICCA Framework. While R2Gen demonstrates improved linguistic fluency and clinical coherence through memory-enhanced generation, it does not explicitly validate the factual alignment between generated text and corresponding image regions. As such, errors may remain undetected if they are semantically plausible but visually unsupported. Our VICCA framework addresses this limitation by assessing the visual consistency of the generated report via visual grounding and CXR reconstruction. Additionally, MCSE enables a more nuanced evaluation of R2Gen's text output by capturing semantic relationships beyond surface-level token similarity.

D.2 M2Trans: Memory-Augmented Transformer for Factual Report Generation

(Mentioned on page 105)

The M2Trans model [76] builds upon the Meshed-Memory Transformer architecture illustrated in Figure D.3 to improve the factual consistency and completeness of image-to-text radiology report generation. The method is designed to address limitations in conventional report generators, which often score well on natural language generation (NLG) metrics like BLEU [84] or CIDEr [106] but produce factually inaccurate or inconsistent content.

Architecture

Given K chest X-ray images $\{x_1, \ldots, x_K\}$ for a patient, M2Trans aims to generate a textual report $\hat{y} = \{y_1, \ldots, y_T\}$ describing the relevant clinical findings. Each image x_k is passed through a CNN (e.g., DenseNet-121) to extract regional features X_k. These features are encoded using a *memory-augmented attention* mechanism:

$$M_{\text{mem}}(X) = \text{Att}(W_q X, [W_k X; M_k], [W_v X; M_v])$$

where W_q, W_k, W_v are projection weights, and M_k, M_v are learnable memory matrices. The attention mechanism helps encoding both spatial and prior knowledge, which is useful in medical image interpretation.

The decoder uses a *meshed attention* mechanism to aggregate information across encoder layers and multiple images:

$$M_{\text{mesh}}(\tilde{X}^{N,K}, \ddot{Y}) = \sum_n \alpha_n \odot C(\tilde{X}^{n,K}, \ddot{Y})$$

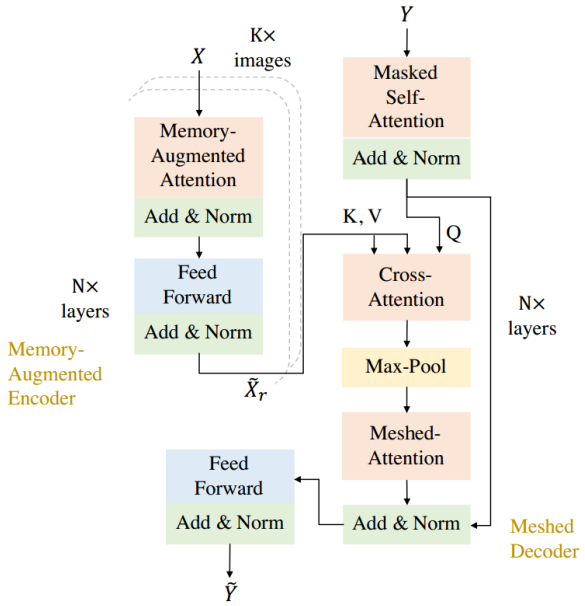


Figure D.3: An overview of Meshed-Memory Transformer extended to multiple images in M2Trans model.

where $\ddot{Y}$ is the intermediate textual representation, and α_n is a learned gating weight that modulates contributions from different encoder layers. This meshed attention fuses low- and high-level visual features with linguistic tokens during generation.

Optimizing for Factual Completeness and Consistency

To mitigate factual errors, M2Trans incorporates two domain-specific reward signals into reinforcement learning:

factENT (Exact Entity Match Reward): This reward encourages the generator to produce clinical entities that match those in the reference. Entities are extracted using Stanza's clinical NER models. Precision and recall are computed on the set of extracted entities, and their harmonic mean is used as the reward:

$$\text{factENT} = \frac{2 \cdot \text{pr}_{\text{ENT}} \cdot \text{rc}_{\text{ENT}}}{\text{pr}_{\text{ENT}} + \text{rc}_{\text{ENT}}}$$

factENTNLI (Entailing Entity Match Reward): To enforce logical consistency, factENTNLI extends factENT using a domain-specific natural language inference (NLI) model. If an entity in the generated text contradicts the reference, it is penalized. The NLI model is trained via weak supervision using entity overlap and semantic similarity heuristics applied to MIMIC-CXR reports.

Joint Loss Function

The full training objective combines negative log-likelihood (NLL) loss with reinforcement learning (RL) losses based on NLG and factual rewards:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{NLL}} + \lambda_2 \mathcal{L}_{\text{RL\_NLG}} + \lambda_3 \mathcal{L}_{\text{RL\_FACT}}$$

where $\mathcal{L}_{\text{RL\_FACT}}$ is computed using either factENT or factENTNLI, and $\mathcal{L}_{\text{RL\_NLG}}$ may use CIDEr [106] or BERTScore [85].

Results

On the MIMIC-CXR and Open-i [25] datasets, M2Trans achieves significantly improved clinical F1 scores compared to previous models such as R2Gen. For example, integrating factENT boosts the clinical F1 by up to 63.9%. The method also exhibits higher BERTScore and stronger alignment with reference entities, with human evaluations by board-certified radiologists preferring M2Trans outputs over R2Gen.

In summary, M2Trans enhances radiology report generation by explicitly optimizing for factual alignment between generated reports and ground-truth entities. This focus on factual consistency is particularly relevant to the objectives of our multimodal reliability framework, which independently evaluates whether generated reports accurately reflect corresponding visual features in chest X-rays. As such, M2Trans serves as a strong candidate for assessing whether improvements in textual factuality also translate into improved cross-modal consistency as measured by our pipeline.

Relevance to This Thesis. M2Trans directly addresses factual alignment using entity-level reinforcement learning, a step forward compared to traditional generative models. However, its alignment validation is constrained to textual entities and does not incorporate cross-modal verification against the image. This creates a gap in assessing whether generated clinical facts are visually grounded. Our VICCA framework complements M2Trans by validating the semantic and visual reliability of these generated entities through text-to-image synthesis and phrase grounding. Additionally, our MCSE metric provides an entity-aware similarity evaluation that captures semantic nuance beyond exact textual matches, which aligns well with M2Trans's factual consistency goals.

D.3 CXR-RePaiR: Retrieval-Based Chest X-ray Report Generation

(Mentioned on page 105)

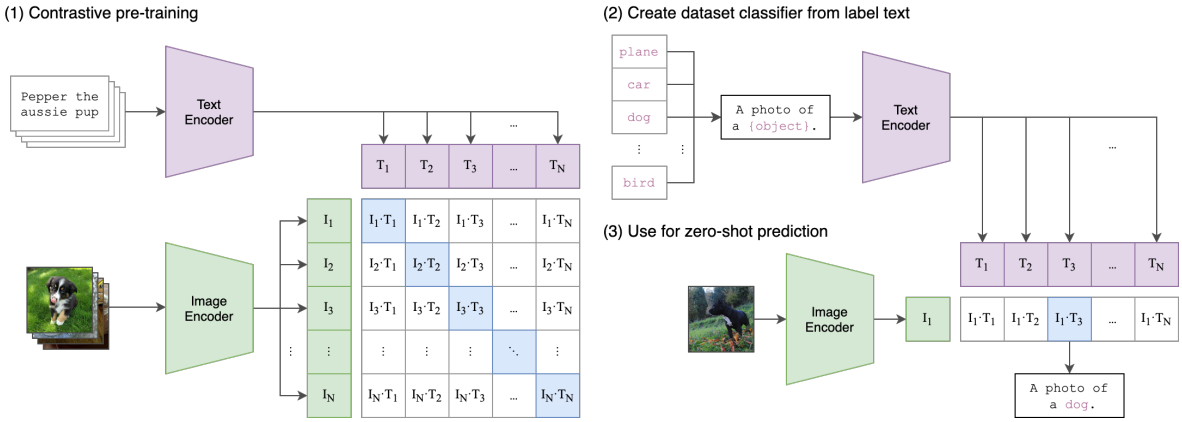


Figure D.4: Summary of CLIP approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some labels, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

CXR-RePaiR [31] is a retrieval-based radiology report generation method that reframes report synthesis as a contrastive matching task rather than a language generation problem. Instead of generating text from scratch, CXR-RePaiR selects clinically relevant sentences or full reports from a large corpus by leveraging the semantic alignment between radiology images and textual descriptions learned through contrastive pretraining.

Architecture Overview

CXR-RePaiR is built on top of the CLIP framework [90] presented in Figure D.4, consisting of a dual-encoder architecture with an image encoder $h(\cdot)$ and a text encoder $g(\cdot)$ as illustrated in Figure D.5. Given an input chest X-ray image x, the model encodes it into a visual embedding $I = h(x)$. A large corpus of candidate reports or report sentences $\mathcal{R} = \{r_1, ..., r_N\}$ is encoded using the text encoder to obtain text embeddings $T_i = g(r_i)$. The report (or sentence) r_i that maximizes the cosine similarity with the image embedding is selected as the generated report:

$$\hat{r} = \arg\max_{r_i \in \mathcal{R}} \langle h(x), g(r_i) \rangle.$$

Variants and Adaptations

Three retrieval variants are proposed:

- **CXR-RePaiR-R**: Retrieves entire reports from the corpus.
- **CXR-RePaiR-k**: Selects the top-k matching sentences from the corpus and concatenates them.

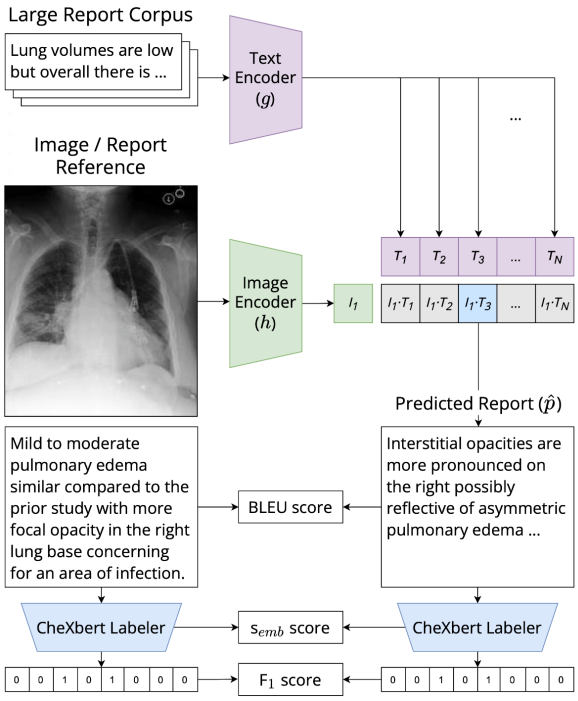


Figure D.5: CXR-RePaiR approach. Reports or report sentences from a large corpus are passed through a pre-trained text encoder, and the input chest X-ray is similarly passed through a pre-trained image encoder. The prediction is generated by selecting the report that maximizes the similarity between the text and image embeddings. The predicted and ground truth reports are then passed through a labeler and performance scores are computed.

- **CXR-RePaiR-Select**: Dynamically determines the number of retrieved sentences based on diagnostic value, using CheXbert labels to adaptively select k.

To improve runtime, a corpus compression method based on CheXbert-based [102] clustering is introduced. This reduces the retrieval space while preserving semantic diversity, enabling up to 65% faster inference with minimal accuracy drop.

Training and Evaluation

The CLIP model is pretrained first on natural image-text pairs and subsequently fine-tuned on radiology image-report pairs. Performance is evaluated on both the MIMIC-CXR and CheXpert [47] datasets using three metrics:

1. **F1 Score (macro-average)** based on CheXbert diagnostic labels.
2. **BLEU-2 Score** for n-gram overlap.
3. **Semantic Similarity (semb)**, measured via cosine similarity of CheXbert-encoded report embeddings.

Results and Observations

CXR-RePaiR-Select outperforms or matches SOTA generative models like R2Gen and M2Trans on clinical F1 and semantic similarity, especially on the out-of-distribution CheXpert dataset. The architecture benefits from the bounded vocabulary and diagnostic structure of medical reports, leveraging retrieval as a robust alternative to generative methods, particularly when combined with strong pretrained image-text encoders.

Moreover, retrieving individual sentences rather than entire reports increases the expressive power of the model, allowing it to compose tailored outputs from multiple partial matches. This sentence-level retrieval strategy outperforms report-level retrieval in both F1 and semantic similarity, especially when k is adaptively chosen.

Limitations

Despite strong performance, CXR-RePaiR depends on the diversity and completeness of its retrieval corpus. Rare pathologies not represented in the corpus cannot be inferred. The sentence selection approach may also introduce redundant or inconsistent information if not carefully constrained.

Relevance to This Thesis. CXR-RePaiR bypasses generation entirely by retrieving semantically similar sentences or reports, thereby reducing grammatical errors and hallucinations. Yet, this approach is inherently limited by the diversity and coverage of its retrieval corpus. It cannot adapt well to rare findings or unseen report variations, and it lacks mechanisms for evaluating how well the retrieved text corresponds to the actual visual content. Our VICCA framework offers an orthogonal perspective by testing whether the selected sentences are visually consistent with the CXR via grounding. The MCSE metric adds value by scoring semantic alignment when textual variation is present, ensuring reliable evaluation even in retrieval-based contexts.

D.4 RGRG: Interactive and Explainable Region-guided Radiology Report Generation

(Mentioned on page 105)

RGRG [103] is a novel framework that aims to enhance factual accuracy, explainability, and interactivity in automatic chest X-ray report generation. Unlike previous methods that rely on global image-level features, RGRG decomposes the report generation task by focusing on localized anatomical regions. Each anatomical region is processed individually, and sentences are generated specifically for regions deemed salient or abnormal.

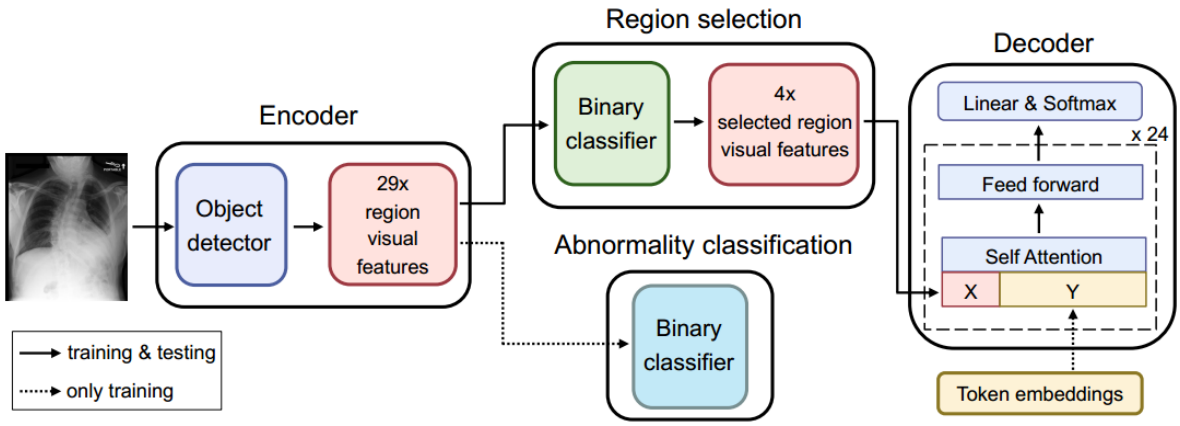


Figure D.6: Region-Guided Radiology Report Generation (RGRG): the object detector extracts visual features for 29 unique anatomical regions of the chest. Two subsequent binary classifiers select salient region features for the final report and encode strong abnormal information in the features, respectively. The language model generates sentences for each of the selected regions (in this example 4), forming the final report.

Model Architecture

The RGRG framework is built around four core modules: an object detector, a region selection module, an abnormality classification module, and a transformer-based language model as in Figure D.6.

Object Detection and Region Feature Extraction. The model employs Faster R-CNN with a ResNet-50 backbone to detect 29 unique anatomical regions. From the region proposals, region of interest (RoI) features are extracted and downsampled via average pooling followed by a linear projection to yield 29×1024 dimensional visual features.

Region Selection and Abnormality Classification. Two parallel multilayer perceptron classifiers determine whether each region should be included in the report and whether it exhibits abnormal findings. These classifiers are trained jointly with the object detector and language model, with binary cross-entropy losses.

Sentence Generation. A GPT-2 Medium model (355M parameters) [89], fine-tuned on PubMed abstracts, acts as the language decoder. Sentence generation for each region is conditioned on the region's visual features via a *pseudo self-attention* (PSA) mechanism. This

injects region-specific context directly into the self-attention layers of the language model:

$$\mathrm{PSA}(X, Y) = \mathrm{softmax}\left((YW_q) \begin{bmatrix} XU_k \\ \\ YW_k \end{bmatrix}^\top \right) \begin{bmatrix} XU_v \\ \\ YW_v \end{bmatrix},$$

where X are the region visual features and Y the token embeddings.

Post-processing. To mitigate redundancy, similar generated sentences are compared using BERTScore. The shorter sentence is discarded in favor of the longer, more informative version.

Training Strategy

Training is performed in three stages: (1) training the object detector independently, (2) training it together with the classifiers, and (3) end-to-end fine-tuning with the language model, where only the PSA projection layers are updated. The total loss is:

$$\mathcal{L} = \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{select}}\mathcal{L}_{\text{select}} + \lambda_{\text{abnormal}}\mathcal{L}_{\text{abnormal}} + \lambda_{\text{lang}}\mathcal{L}_{\text{language}}.$$

Empirically chosen weights balance the components.

Evaluation

RGRG is evaluated on the MIMIC-CXR dataset using both natural language generation (NLG) metrics (BLEU, ROUGE, METEOR, CIDEr) and clinical efficacy (CE) metrics. RGRG achieves state-of-the-art results on METEOR and is competitive in BLEU and CIDEr. Furthermore, it demonstrates high anatomy-sensitivity and robustness to bounding box variations, validating its applicability in clinical settings.

Relevance to This Thesis. RGRG provides localized sentence generation and enables interpretability through region-wise analysis. Despite its anatomical grounding, the model does not assess whether generated sentences semantically match the visual abnormalities within each region. There is also no mechanism to evaluate overall report reliability or inter-region consistency. VICCA addresses this by comparing localized predictions across original and synthetic CXRs, offering an external verification of region-sentence alignment. Additionally, MCSE supports the fine-grained evaluation of sentence-level entity correctness, providing structured insight into how accurately the report conveys meaningful clinical findings.

D.5 MedGemma: Medical Vision–Language Models Based on Gemma 3

(Mentioned on page 105)

MedGemma [100] is a collection of medically tuned vision–language foundation models built on the Gemma 3 family (4B and 27B), released by Google DeepMind. The suite includes both text-only and multimodal variants, with the multimodal models designed for medical image understanding and free-text generation (e.g., radiology reporting, VQA). Relative to similar-sized generative models, MedGemma reports strong performance on a wide range of medical tasks while retaining general capabilities of the base Gemma models. Use cases emphasized by the developers include medical image report generation, image interpretation, and text-based medical QA. [100]

Overview of Architecture

MedGemma adopts the Gemma 3 backbone and augments it with a multimodal interface for 2D medical images (e.g., CXR, 2D CT/MRI slices). In the multimodal variants, images are encoded into visual tokens and fused with text tokens inside the LLM through cross-attention, enabling grounded generation (e.g., narrative reports) conditioned on image evidence. The collection is released in three primary variants: a 4B multimodal model and two 27B models (text-only and multimodal).

- **Backbone:** Gemma 3 LLM (4B/27B) with instruction-following capabilities; the multimodal path adds a vision encoder and projection layers to map image features into the LLM token space. [100]
- **Vision Modality:** Focused on 2D medical imaging; the technical report explicitly notes that 3D volumetric inputs are out of scope for this release (unlike Med-Gemini's 3D). [104]
- **Interfaces/Deployment:** Distributed via Model Garden / developer resources with guidance for report-generation and VQA applications. [100]

Training and Post-training Alignment

The technical report describes a multi-stage medical adaptation pipeline: continued pretraining and post-training on curated medical corpora and image–text data, followed by reinforcement learning (RL)–based multimodal post-training. Notably, the authors report that RL yields better generalization than purely supervised fine-tuning for the multimodal setup, and they removed lower-quality VQA datasets during curation. [100, 104]

Inference and Prompting

MedGemma is positioned for (i) medical image report generation (free-text summaries grounded in an input image), (ii) visual question answering over medical images, and (iii) text-only medical QA. The developer documentation highlights these as primary, supported workflows.

Evaluation Summary

Across the technical report and blog/model-card materials, MedGemma is evaluated on multiple categories (medical text QA, image classification, VQA, and report-like generation), often approaching task-specific baselines and surpassing similarly sized open models. Public posts also emphasize that MedGemma is suited for free-text medical image tasks, while complementary models (e.g., MedSigLIP) are suggested for structured imaging outputs (classification/retrieval).

Relevance to This Thesis (VICCA). MedGemma's multimodal generation makes it a natural candidate for report-to-image and image-to-report consistency checks. In our evaluation setting, MedGemma-generated reports can be grounded and compared against synthetic CXRs to compute reliability scores, revealing visual–textual mismatches. And MedGemma's free-text outputs can be semantically scored with entity-aware similarity (negations, modifiers) to complement surface NLG metrics; this helps quantify semantic fidelity in clinically salient terms. Together, VICCA and MCSE provide an external cross-modal and semantic lens on MedGemma's reported capabilities. (See Chapters 3 and 5.)

Bibliography

[1] Harsh Agrawal et al. "nocaps: novel object captioning at scale". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 8948–8957 (cit. on p. 103).

[2] Peter Anderson et al. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6077–6086 (cit. on p. 101).

[3] Peter Anderson et al. "Bottom-up and top-down attention for image captioning and vqa". In: *arXiv preprint arXiv:1707.07998* 2.4 (2017), p. 8 (cit. on p. 101).

[4] Peter Anderson et al. *SPICE: Semantic Propositional Image Caption Evaluation*. 2016. arXiv: 1607.08822 [cs.CV] (cit. on p. 81).

[5] Stanislaw Antol et al. "VQA: Visual Question Answering". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2425–2433 (cit. on p. 101).

[6] Yunfei Bai et al. "Misalignment-resistant domain adaptive learning for one-stage object detection". In: *Knowledge-Based Systems* 305 (2024), p. 112605 (cit. on p. 19).

[7] 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*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72 (cit. on pp. 81, 163).

[8] Shruthi Bannur et al. *MAIRA-2: Grounded Radiology Report Generation*. Tech. rep. MSR-TR-2024-18. Microsoft, 2024 (cit. on pp. 2, 3).

[9] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620 (cit. on p. 85).

[10] Benedikt Boecking et al. "Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing". In: *The European Conference on Computer Vision (ECCV)*. 2022 (cit. on pp. 2, 3, 8, 25–27, 33, 36, 38, 103, 113, 144, 152, 153).

[11] Nicolas Carion et al. "End-to-End Object Detection with Transformers". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 213–229 (cit. on pp. 8, 16, 19, 20, 29, 55).

[12] Pierre Chambon et al. *RoentGen: Vision-Language Foundation Model for Chest X-ray Generation*. 2022 (cit. on pp. 46, 54).

[13] Pierre Joseph Marcel Chambon et al. "Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains". In: *NeurIPS 2022 Foundation Models for Decision Making Workshop*. 2022 (cit. on p. 46).

[14] David Chen and Raymond Mooney. "Learning to interpret natural language navigation instructions from observations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 25. 1. 2011, pp. 859–865 (cit. on p. 101).

[15] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020 (cit. on p. 15).

[16] Wenting Chen et al. "Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9494–9509 (cit. on p. 15).

[17] Xieling Chen et al. "Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics". In: *Artificial Intelligence Review* 57.4 (2024), p. 91 (cit. on p. 2).

[18] Y. Chen et al. "MIMO: A Medical Vision Language Model with Visual Referring and Pixel Grounding". In: *CVPR 2025*. 2025 (cit. on p. 2).

[19] Zhihao Chen et al. "Medical Phrase Grounding with Region-Phrase Context Contrastive Alignment". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan et al. Cham: Springer Nature Switzerland, 2023, pp. 371–381 (cit. on p. 25).

[20] Zhihong Chen et al. "Generating Radiology Reports via Memory-driven Transformer". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1439–1449 (cit. on pp. 1, 94, 95, 103–105, 161).

[21] Joseph Paul Cohen et al. "TorchXRayVision: A library of chest X-ray datasets and models". In: *Medical Imaging with Deep Learning*. 2022 (cit. on pp. 8, 56, 63).

[22] Olivier Colliot, Elina Thibeau-Sutre, and Ninon Burgos. "Reproducibility in Machine Learning for Medical Imaging". In: *Machine Learning for Brain Disorders*. Vol. 197. Neuromethods. Springer, 2023, 631–653 (cit. on p. 4).

[23] NLP Connect. *vit-gpt2-image-captioning (Revision 0e334c7)*. 2022 (cit. on p. 103).

[24] Yonghao Dang et al. "DBNet: A New Generalized Structure Efficient for Classification". In: *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 2019, pp. 1–6 (cit. on p. 20).

[25] Dina Demner-Fushman et al. "Preparing a collection of radiology examinations for distribution and retrieval". In: *Journal of the American Medical Informatics Association* 23.2 (2016), pp. 304–310 (cit. on p. 166).

[26] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255 (cit. on p. 161).

[27] Jiajun Deng et al. "TransVG: End-to-End Visual Grounding with Transformers". In: *arXiv preprint arXiv:2104.08541* (2021) (cit. on p. 38).

[28] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *North American Chapter of the Association for Computational Linguistics*. 2019 (cit. on pp. 8, 23, 33, 151–153).

[29] Prafulla Dhariwal and Alexander Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794 (cit. on p. 8).

[30] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021 (cit. on p. 16).

[31] Mark Endo et al. "Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model". In: *Proceedings of Machine Learning for Health*. Vol. 158. Proceedings of Machine Learning Research. 2021, pp. 209–219 (cit. on pp. 1, 3, 85, 94, 95, 103–105, 167).

[32] Fabio Ferreira et al. "Beyond Random Augmentations: Pretraining with Hard Views". In: *The Thirteenth International Conference on Learning Representations*. 2025 (cit. on p. 34).

[33] Sayeh Gholipour Picha., Dawood Al Chanti., and Alice Caplier. "How far Generated Data Can Impact Neural Networks Performance?" In: *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,* INSTICC. SciTePress, 2023, pp. 472–479 (cit. on p. 46).

[34] Sayeh Gholipour Picha, Dawood Al Chanti, and Alice Caplier. "VICCA: Visual interpretation and comprehension of chest X-ray anomalies in generated report without human feedback". In: *Machine Learning with Applications* 21 (2025), p. 100684 (cit. on pp. 8, 9, 112).

[35] A L Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". en. In: *Circulation* 101.23 (June 2000), E215–20 (cit. on pp. 50, 55).

[36] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014 (cit. on p. 43).

[37] Yu Gu et al. *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. 2020. eprint: `arXiv:2007.15779` (cit. on p. 154).

[38] Danna Gurari et al. "VizWiz Grand Challenge: Answering Visual Questions from Blind People". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3608–3617 (cit. on p. 101).

[39] Anees Ur Rehman Hashmi et al. "XReal: Realistic Anatomy and Pathology-Aware X-ray Generation via Controllable Diffusion Model". In: *arXiv preprint arXiv:2403.09240* (2024) (cit. on pp. 46, 54).

[40] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778 (cit. on pp. 20, 23, 35).

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG] (cit. on p. 42).

[42] M Hodosh, P Young, and J Hockenmaier. "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics". In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 853–899 (cit. on p. 103).

[43] Matthew Honnibal et al. "spaCy: Industrial-strength Natural Language Processing in Python". In: (2020) (cit. on p. 90).

[44] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". In: *arXiv:1904.05342* (2019) (cit. on pp. 33, 153).

[45] Shih-Cheng Huang et al. "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 3922–3931 (cit. on pp. 2, 3).

[46] Akimichi Ichinose et al. "Visual Grounding ofWhole Radiology Reports for3D CT Images". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan et al. Cham: Springer Nature Switzerland, 2023, pp. 611–621 (cit. on p. 25).

[47] Jeremy Irvin et al. *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*. 2019. arXiv: 1901.07031 [cs.CV] (cit. on pp. 63, 80, 92, 106, 144, 149, 157, 163, 168).

[48] Saahil Jain et al. "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021 (cit. on pp. 80, 85, 91, 106, 158).

[49] Haseeb Javed, Shaker El-Sappagh, and Tamer Abuhmed. "Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications". In: *Artificial Intelligence Review* 58.12 (2024), pp. 1–30 (cit. on p. 4).

[50] Alistair E. W. Johnson et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific Data* 6.1 (2019), p. 317 (cit. on pp. 16, 50, 69, 92, 93, 103, 113, 143, 154, 158, 163).

[51] Alistair E. W. Johnson et al. *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. 2019. arXiv: 1901.07042 (cit. on pp. 53, 103, 113, 143, 163).

[52] Alistair E.W. Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3.1 (2016), p. 160035 (cit. on p. 154).

[53] Aishwarya Kamath et al. "MDETR - Modulated Detection for End-to-End Multi-Modal Understanding". In: Oct. 2021, pp. 1760–1770 (cit. on pp. 17, 19, 20).

[54] Andrej Karpathy and Li Fei-Fei. "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4 (Apr. 2017), 664–676 (cit. on p. 101).

[55] Sahar Kazemzadeh et al. "ReferItGame: Referring to Objects in Photographs of Natural Scenes". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 787–798 (cit. on pp. 17, 24).

[56] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *CoRR* abs/1312.6114 (2013) (cit. on p. 42).

[57] Ivan Krasin et al. "OpenImages: A public dataset for large-scale multi-label and multi-class image classification." In: *Dataset available from https://github.com/openimages* (2016) (cit. on p. 22).

[58] Ranjay Krishna et al. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". In: *Int. J. Comput. Vision* 123.1 (May 2017), 32–73 (cit. on pp. 17, 20, 22, 24, 147).

[59] Harold W. Kuhn. "The Hungarian Method for the assignment problem". In: *Naval Research Logistics Quarterly* 2 (1955), pp. 83–97 (cit. on p. 16).

[60] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. "Zero-data learning of new tasks". In: *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*. AAAI'08. Chicago, Illinois: AAAI Press, 2008, 646–651 (cit. on p. 17).

[61] Jiayu Lei et al. *AutoRG-Brain: Grounded Report Generation for Brain MRI*. 2024. arXiv: 2407.16684 [eess.IV] (cit. on p. 77).

[62] Christy Y. Li et al. "Hybrid retrieval-generation reinforced agent for medical image report generation". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, 1537–1547 (cit. on p. 104).

[63] Chunyuan Li et al. "Llava-med: Training a large language-and-vision assistant for biomedicine in one day". In: *arXiv preprint arXiv:2306.00890* (2023) (cit. on p. 103).

[64] Jiao Li et al. "BioCreative V CDR task corpus: a resource for chemical disease relation extraction". In: *Database* 2016 (May 2016), baw068. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baw068/8224483/baw068.pdf (cit. on p. 87).

[65] Junnan Li et al. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". In: *ICML*. 2022 (cit. on p. 103).

[66] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81 (cit. on pp. 81, 163).

[67] Tsung-Yi Lin et al. "Focal Loss for Dense Object Detection". In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2999–3007 (cit. on p. 35).

[68] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 740–755 (cit. on pp. 20, 22, 24, 101–103, 147).

[69] Geert Litjens et al. "A survey on deep learning in medical image analysis". en. In: *Med. Image Anal.* 42 (Dec. 2017), pp. 60–88 (cit. on p. 1).

[70] Chin-Fu Liu et al. "A large public dataset of annotated clinical MRIs and metadata of patients with acute stroke". In: *Scientific Data* 10.1 (2023), p. 548 (cit. on p. 76).

[71] Shilong Liu et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection". In: *arXiv preprint arXiv:2303.05499* (2023) (cit. on pp. 17, 20, 22, 23, 25).

[72] Weihua Liu et al. *IU X-ray dataset*. 2025 (cit. on pp. 103, 133, 163).

[73] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692 (cit. on pp. 20, 21).

[74] Ze Liu et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 (cit. on pp. 16, 35).

[75] Xin Mei et al. "PhraseAug: An Augmented Medical Report Generation Model With Phrasebook". In: *IEEE Transactions on Medical Imaging* 43.12 (2024), pp. 4211–4223 (cit. on p. 104).

[76] Yasuhide Miura et al. "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*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 5288–5304 (cit. on pp. 1, 103–105, 164).

[77] Yasuhide Miura et al. "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*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 5288–5304 (cit. on p. 155).

[78] Philip Müller, Georgios Kaissis, and Daniel Rueckert. *ChEX: Interactive Localization and Region Description in Chest X-rays*. 2024. arXiv: 2404.15770 (cit. on pp. 2, 3, 25, 38, 103, 104).

[79] Mark Neumann et al. "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing". In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. eprint: arXiv:1902.07669 (cit. on p. 87).

[80] Chee Ng, Liliang Sun, and Shaoqing Tang. "X-Ray-CoT: Interpretable Chest X-ray Diagnosis with Vision-Language Models via Chain-of-Thought Reasoning". In: *arXiv preprint arXiv:2508.12455* (2025) (cit. on p. 2).

[81] Ha Q. Nguyen et al. *VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations*. 2022. arXiv: 2012.15029 [eess.IV] (cit. on pp. 26, 27, 30, 36, 145, 146).

[82] Felix Nützel, Mischa Dombrowski, and Bernhard Kainz. "Generate to Ground: Multimodal Text Conditioning Boosts Phrase Grounding in Medical Vision-Language Models". In: *Proceedings of MIDL 2025*. 2025 (cit. on p. 2).

[83] Kai Packhäuser et al. "Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data". In: *Scientific Reports* 12.1 (2022), p. 14851 (cit. on p. 46).

[84] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318 (cit. on pp. 81, 163, 164).

[85] Jessica Patricoski et al. "An evaluation of pretrained BERT models for comparing semantic similarity across unstructured clinical trial texts". In: *Stud Health Technol Inform* 289 (2022), pp. 18–21 (cit. on p. 166).

[86] Sayeh Gholipour Picha, Dawood Al Chanti, and Alice Caplier. "Semantic Textual Similarity Assessment in Chest X-ray Reports Using a Domain-Specific Cosine-Based Metric". In: *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 1: BIOINFORMATICS*. INSTICC. SciTePress, 2024, pp. 487–494 (cit. on pp. 9, 63, 81, 84, 112).

[87] Walter H. L. Pinaya et al. *Brain Imaging Generation with Latent Diffusion Models*. 2022. arXiv: 2209.07162 [eess.IV] (cit. on p. 46).

[88] Dustin Podell et al. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis". In: *The Twelfth International Conference on Learning Representations*. 2024 (cit. on p. 42).

[89] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019) (cit. on pp. 102, 170).

[90] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV] (cit. on pp. 23, 152, 167).

[91] Vishwanatha M Rao et al. "Multimodal generative AI for medical image interpretation". en. In: *Nature* 639.8056 (Mar. 2025), pp. 888–896 (cit. on p. 46).

[92] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788 (cit. on p. 19).

[93] Shaoqing Ren et al. "Faster R-CNN: towards real-time object detection with region proposal networks". In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, 91–99 (cit. on p. 19).

[94] Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 1530–1538 (cit. on p. 42).

[95] Anna Rohrbach et al. “Grounding of textual phrases in images by reconstruction”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 817–834 (cit. on p. 101).

[96] Beddiar Romaissa et al. “ACapMed: Automatic Captioning for Medical Imaging”. In: *Applied Sciences* 12 (Nov. 2022) (cit. on pp. 2, 3).

[97] Robin Rombach et al. “High-Resolution Image Synthesis With Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684–10695 (cit. on pp. 46, 49).

[98] Abhilasha Sancheti, Koustava Goswami, and Balaji Srinivasan. “Post-Hoc Answer Attribution for Grounded and Trustworthy Long Document Comprehension: Task, Insights, and Challenges”. In: *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*. Ed. by Danushka Bollegala and Vered Shwartz. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 49–57 (cit. on p. 16).

[99] Alexander Selivanov et al. “Medical image captioning via generative pretrained transformers”. In: *Scientific Reports* 13.1 (2023), p. 4171 (cit. on p. 2).

[100] Andrew Sellergren et al. “MedGemma Technical Report”. In: *arXiv preprint arXiv:2507.05201* (2025) (cit. on pp. 105, 172).

[101] Yajush Pratap Singh et al. “Image Captioning using Artificial Intelligence”. In: *Journal of Physics: Conference Series* 1854.1 (2021), p. 012048 (cit. on p. 102).

[102] Akshay Smit et al. “CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT”. In: *Conference on Empirical Methods in Natural Language Processing*. 2020 (cit. on pp. 85, 112, 157, 168).

[103] Tim Tanida et al. “Interactive and Explainable Region-guided Radiology Report Generation”. In: *CVPR*. 2023 (cit. on pp. 3, 25, 103–105, 169).

[104] Gemma Team et al. *Gemma 3 Technical Report*. 2025. arXiv: 2503.19786 [cs.CL] (cit. on p. 172).

[105] Ashish Vaswani et al. “Attention is all you need”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, 6000–6010 (cit. on pp. 1, 15).

[106] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. “CIDEr: Consensus-based image description evaluation”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 4566–4575 (cit. on pp. 81, 164, 166).

[107] Konstantinos Vilouras et al. *Zero-Shot Medical Phrase Grounding with Off-the-shelf Diffusion Models*. 2024. arXiv: 2404.12920 [cs.CV] (cit. on p. 25).

[108] Oriol Vinyals et al. “Show and tell: A neural image caption generator”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3156–3164 (cit. on pp. 100, 101, 103).

[109] Jun Wang et al. “CAMANet: Class Activation Map Guided Attention Network for Radiology Report Generation”. In: *IEEE Journal of Biomedical and Health Informatics* 28.4 (2024), pp. 2199–2210 (cit. on p. 103).

[110] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612 (cit. on p. 62).

[111] Zifeng Wang et al. "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text". en. In: *Proc Conf Empir Methods Nat Lang Process* 2022 (Dec. 2022), pp. 3876–3887 (cit. on p. 18).

[112] Tobias Weber et al. "Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis". In: *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference, PAKDD 2023*. Springer. 2023 (cit. on pp. 46, 54).

[113] Lilian Weng. "Flow-based Deep Generative Models". In: *lilianweng.github.io* (2018) (cit. on p. 42).

[114] Lilian Weng. "What are diffusion models?" In: *lilianweng.github.io* (2021) (cit. on p. 43).

[115] Joy T. Wu et al. "Chest ImaGenome Dataset for Clinical Reasoning". In: *ArXiv* abs/2108.00316 (2021) (cit. on pp. 29, 36, 55).

[116] Kelvin Xu et al. "Show, attend and tell: neural image caption generation with visual attention". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, 2048–2057 (cit. on p. 101).

[117] Nur Yildirim et al. "Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024) (cit. on pp. 2, 3).

[118] W J Youden. "Index for rating diagnostic tests". en. In: *Cancer* 3.1 (Jan. 1950), pp. 32–35 (cit. on p. 66).

[119] Peter Young et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". In: *TACL* 2 (2014), pp. 67–78 (cit. on p. 101).

[120] Feiyang Yu et al. *Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation*. en. preprint. Radiology and Imaging, Aug. 2022 (cit. on p. 85).

[121] Feiyang Yu et al. "Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation". In: *medRxiv* (2022). eprint: https://www.medrxiv.org/content/early/2022/08/31/2022.08.30.22279318.full.pdf (cit. on pp. 106, 159).

[122] Licheng Yu et al. "MAttNet: Modular Attention Network for Referring Expression Comprehension". In: *CVPR*. 2018 (cit. on p. 20).

[123] Jianbo Yuan et al. "Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Cham: Springer International Publishing, 2019, pp. 721–729 (cit. on p. 103).

[124] Haotian Zhang et al. "GLIPv2: Unifying Localization and Vision-Language Understanding". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022 (cit. on pp. 17, 20, 21, 35).

[125] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023 (cit. on pp. 8, 49).

[126] Ke Zou et al. *MedRG: Medical Report Grounding with Multi-modal Large Language Model*. 2024. arXiv: 2404.06798 (cit. on pp. 2, 25).

Résumé — Cette thèse explore des stratégies multimodales visant à améliorer la fiabilité, l'alignement et l'interprétabilité des rapports radiologiques générés par l'intelligence artificielle à partir de radiographies thoraciques (CXR). Bien que les modèles de vision-langage modernes puissent produire des rapports fluides et riches en contexte, ils manquent souvent de mécanismes permettant de vérifier si les observations décrites sont réellement soutenues par l'image ou de détecter d'éventuelles hallucinations. Pour remédier à cette limitation, nous proposons **VICCA** (Visual Interpretation and Comprehension of Chest X-ray Anomalies), un cadre de validation multimodal qui évalue la cohérence entre le rapport et l'image au moyen de deux composantes complémentaires : (i) un *modèle de localisation visuelle* reliant les descriptions textuelles à leurs régions correspondantes dans la radiographie, et (ii) un *modèle de diffusion conditionnelle* guidé par des masques pulmonaires binaires afin de générer des images anatomiquement cohérentes et conformes au rapport. Un score de fiabilité est ensuite calculé à partir de l'alignement visuel entre les images réelles et synthétisées. En parallèle, nous introduisons **MCSE** (Medical Corpus Similarity Evaluation), une métrique de similarité sémantique conçue pour les textes cliniques. MCSE capture les relations entre entités, les négations et les modificateurs au-delà de la simple similarité lexicale, offrant ainsi une évaluation plus pertinente de la précision des rapports. Ensemble, VICCA et MCSE forment un système d'évaluation unifié permettant de vérifier à la fois la validité visuelle et sémantique des modèles de génération de rapports radiologiques. Ce cadre a été appliqué à plusieurs modèles de pointe, dont le nouveau modèle fondationnel de grand langage **MedGemma**, démontrant sa capacité à révéler des incohérences et à améliorer l'interprétabilité dans des contextes multimodaux. Dans l'ensemble, ce travail contribue au développement d'une intelligence artificielle explicable et vérifiable en imagerie médicale, établissant de nouvelles bases pour la confiance, la transparence et la fiabilité clinique dans la génération automatisée de rapports.

Mots clés: Radiographies thoraciques; Intelligence artificielle explicable; Apprentissage multimodal; Localisation sémantique; Modèles de diffusion; Similarité sémantique; Alignement image–texte médical; Génération de rapports radiologiques; Fiabilité clinique; Cohérence visuo-textuelle.

GIPSA-LAB, 11 rue des Mathématiques
38402 Saint Martin d'Hères, France

Abstract — This thesis explores multimodal strategies for improving the reliability, alignment, and interpretability of AI-generated radiology reports from chest X-rays (CXRs). While modern vision-language models can generate fluent and contextually rich reports, they often lack mechanisms for verifying whether the described findings are truly supported by the image or to detect potential hallucinations. To address this limitation, we propose **VICCA** (Visual Interpretation and Comprehension of Chest X-ray Anomalies), a multimodal validation framework that assesses report-image consistency through two complementary components: (i) a *visual grounding model* that links textual descriptions to their visual regions in the CXR, and (ii) a *conditional diffusion model* guided by binary lung masks to synthesize anatomically faithful and report-consistent images. A reliability score is then derived from the visual alignment between real and generated CXRs. In parallel, we introduce **MCSE** (Medical Corpus Similarity Evaluation), a semantic similarity metric designed for clinical text. MCSE captures entity-level relations, negations, and modifiers beyond surface lexical overlap, offering a more meaningful assessment of report accuracy. Together, VICCA and MCSE form a unified evaluation system for verifying both the visual and semantic validity of radiology report generation models. The framework was applied to several state-of-the-art models, including the new foundation large language model MedGemma, demonstrating its ability to reveal inconsistencies and improve interpretability across multimodal settings. Overall, this work contributes to the development of explainable and verifiable AI in medical imaging, establishing new pathways for trust, transparency, and clinical reliability in automated report generation.

Keywords: Chest X-rays; Explainable Artificial Intelligence; Multimodal Learning; Visual Grounding; Diffusion Models; Semantic Similarity; Medical Image–Text Alignment; Radiology Report Generation; Clinical Reliability; Visual–Textual Consistency.

GIPSA-LAB, 11 rue des Mathématiques
38402 Saint Martin d'Hères, France