People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

SAAD DAHLAB UNIVERSITY OF BLIDA 1

Faculty of Sciences

Department of Computer Science

MASTER DEGREE THESIS

Speciality : Intelligent Systems Engineering

Presented by :

Khaldi Abderrahmane

THEME

Automated Report Generation for Medical Chest X-ray Imaging

Publicly defended on **23/06/2025**, before a jury composed of :

Dr.	Dallila Gessoum	Assistant Professor	USDB	Examiner
Dr.	Azouz Iman	Assistant Professor	USDB	Examiner
Dr.	Fatima Boumahdi	Associate Professor	USDB	Thesis supervisor

Acknowledgements

First and foremost, I express my profound gratitude to Almighty Allah for bestowing upon me the courage, strength, and perseverance necessary to undertake and complete this research endeavor. His guidance and blessings have been a constant source of support throughout this challenging journey.

I would like to convey my deepest appreciation to Professor Boumahdi for her unwavering support, invaluable insights, and tireless guidance. Her expertise, constructive feedback, and commitment to academic excellence have been instrumental in shaping the direction and enhancing the quality of this work.

I am profoundly grateful to Dr. Guessoum, Dr. Mezzi, and Dr. Remmid for their time, expertise, and thoughtful contributions. Their interdisciplinary perspectives, insightful critiques, and challenging questions have greatly enriched the depth and rigor of this research.

This work stands as a testament to the collective support, guidance, and encouragement of all those mentioned above. To each of you, I express my heartfelt gratitude.

I extend my sincere appreciation to my family, whose steadfast support and encouragement provided the foundation for this academic pursuit. Their patience and understanding throughout this demanding process have been invaluable to my success.

The completion of this work would not have been possible without the contributions of numerous individuals who offered their time, knowledge, and support. I am deeply grateful to all who have played a part in this academic journey.

Abstract

Automated medical report generation from chest radiographs has emerged as a critical challenge in medical imaging, particularly in addressing information bottlenecks and poor clinical accuracy for rare pathological conditions. This work presents ChestBXG, a novel multi-modal architecture that integrates classification-guided visual encoding with domain-adaptive language generation to bridge the semantic gap between radiographic features and clinical text. Our approach employs EfficientNet-B4 for visual feature extraction, coupled with BioGPT for medical domain-specific text generation, interconnected through sophisticated co-attention mechanisms that prevent information loss during cross-modal alignment. The architecture incorporates a confidence-based classification head that guides report generation, particularly enhancing performance on minority pathological cases. Experimental evaluation on a curated subset of the MIMIC-CXR dataset demonstrates substantial improvements across standard metrics, achieving decent results on multiple metrics while focusing on harder samples. The proposed framework addresses fundamental limitations in existing methodologies while maintaining computational efficiency, establishing a foundation for clinically viable automated reporting systems that enhance diagnostic accuracy and workflow efficiency in radiological practice.

Keywords: X-ray images, Vision feature extraction, Language generation, Encoder-Decoder

Résumé

La génération automatisée de rapports médicaux à partir de radiographies thoraciques est devenue un défi critique en imagerie médicale, particulièrement pour résoudre les goulots d'étranglement informationnels et la faible précision clinique pour les conditions pathologiques rares. Ce travail présente ChestBXG, une architecture multi-modale novatrice qui intègre l'encodage visuel guidé par classification avec la génération de langage adaptée au domaine pour combler l'écart sémantique entre les caractéristiques radiographiques et le texte clinique. Notre approche emploie EfficientNet-B4 pour l'extraction de caractéristiques visuelles, couplé avec BioGPT pour la génération de texte spécifique au domaine médical, interconnectés par des mécanismes de co-attention sophistiqués qui préviennent la perte d'information durant l'alignement cross-modal. L'architecture incorpore une tête de classification basée sur la confiance qui guide la génération de rapports, améliorant particulièrement les performances sur les cas pathologiques minoritaires. L'évaluation expérimentale sur un sous-ensemble sélectionné du dataset MIMIC-CXR démontre des améliorations substantielles à travers les métriques standards, atteignant des résultats décents sur plusieurs métriques tout en se concentrant sur les échantillons plus difficiles. Le cadre proposé adresse les limitations fondamentales des méthodologies existantes tout en maintenant l'efficacité computationnelle, établissant une fondation pour des systèmes de rapportage automatisés cliniquement viables qui améliorent la précision diagnostique et l'efficacité du flux de travail en pratique radiologique.

Mots-clés : Images radiographiques, Extraction de caractéristiques visuelles, Génération de langage, Encodeur-Décodeur

ملخص

لقد برز توليد التقارير الطبية الآلي من صور الأشعة السينية للصدر كتحدٍّ بالغ الأهمية في التصوير الطبي، خاصة في معالجة اختناقات المعلومات وضعف الدقة السريرية للحالات المرضية النادرة. يقدم هذا العمل ChestBXG، وهو معمارية متعددة الوسائط مبتكرة تدمج الترميز البصري الموجه بالتصنيف مع توليد اللغة المتكيف مع المجال لسد الفجوة الدلالية بين الخصائص الإشعاعية والنص السريري. يستخدم نهجنا EfficientNet-B4 لاستخراج الخصائص البصرية، مقترناً مع BioGPT لتوليد النصوص الخاصة بالمجال الطبي، مترابطة من خلال آليات انتباه مشترك متطورة تمنع فقدان المعلومات أثناء المحاذاة عبر الوسائط. تتضمن المعمارية رأس تصنيف قائم على الثقة يوجه توليد التقارير، مما يعزز الأداء بشكل خاص على الحالات المرضية الأقلية. يُظهر التقييم التجريبي على مجموعة فرعية منتقاة من مجموعة بيانات MIMIC-CXR تحسينات جوهرية عبر المقاييس المعيارية، محققاً نتائج لائقة على مقاييس متعددة مع التركيز على العينات الأصعب. يعالج الإطار المقترح القيود الأساسية في المنهجيات الموجودة مع الحفاظ على الكفاءة الحاسوبية، مؤسساً قاعدة لأنظمة التقارير الآلية القابلة للتطبيق سريرياً والتي تعزز دقة التشخيص وكفاءة سير العمل في الممارسة الإشعاعية.

الكلمات المفتاحية: صور الأشعة السينية، استخراج الخصائص البصرية، توليد اللغة، المُرمِّز-مفكك الترميز

Contents

List of Figures i

List of Tables ii

General Introduction 1

1 Literature Review 4

1.1 Introduction . . . 4
1.2 Pre-Foundation Models . . . 4
1.3 Attention Based Models . . . 5
1.4 Transformers Based Methods . . . 5
1.5 Reinforcement Learning Methods . . . 5
1.6 Knowledge-Enhanced Methods . . . 6
1.7 Pre-Trained LLMs Methods . . . 6
1.8 Native MultiModal Models Methods . . . 6
1.9 Datasets . . . 6
1.10 Synthesis . . . 6
1.11 Conclusion . . . 7

2 Proposed Appraoch: ChestBXG 8

2.1 Introduction . . . 8
2.2 Architectural Overview . . . 9
2.2.1 Visual Encoder . . . 10
2.2.2 Co-Attention Mechanism . . . 12
2.2.3 Language Model . . . 12
2.3 Dataset Selection and Processing . . . 13
2.3.1 Dataset Description . . . 15
2.3.2 Sample Selection Class Balance . . . 16
2.3.3 Image Selection Criteria . . . 16

2.3.4 Data Organization 17

2.3.5 Text Extraction Cleaning 17

2.3.6 Final Dataset Statistics 17

2.4 Experimental Setup 17

2.4.1 Loss Function 17

2.4.2 Training Strategy 19

2.5 Model Evaluation 19

2.5.1 Evaluation Metrics 20

2.5.2 Model Hyperparameters 22

2.6 Conclusion 24

3 Experimental Results and Analysis 25

3.1 Introduction 25

3.2 Experimental Setup and Dataset Configuration 25

3.2.1 Hardware Configuration and Computational Constraints 25

3.3 Approach Evolution : The Journey of Tweaking 26

3.3.1 Phase 1: Establishing the Foundation with Traditional Encoder-Decoder Approaches 26

3.3.2 Phase 2: Advancing Visual Understanding and Multi-View Integration 27

3.3.3 Phase 3: Exploring Self-Supervised Vision Transformers 27

3.3.4 Phase 4: Embracing Multi-Modal Pre-training with BLIP 28

3.3.5 Phase 5: Achieving Clinical Excellence with Domain-Specific Integration 28

3.4 Classification Performance Evaluation 28

3.4.1 Clinical Significance of Classification Results 29

3.4.2 Integration with Report Generation 29

3.5 Ablation Studies 29

3.5.1 Multi-Objective Loss Function 29

3.6 Failed Experimental Approaches 29

3.6.1 Performance Summary of Failed Approaches 30

3.7 Conclusion 31

4 Conclusion and Future Work 32

Bibliography 33

Bibliography 34

List of Figures

0.1 Annual number of AI products cleared for use in radiology from 2008 to 2024, showing significant growth starting in 2017 [78]. 1
0.2 AI in Medical Imaging market size for 2024 and projected growth until 2034 [79]. 2
0.3 Detailed comparison of human radiologists and AI models, emphasizing the unique strengths each brings to medical imaging tasks [80]. 3

2.1 Overview of the ChestBXG architecture. 9
2.2 Detailed observation of the Vision Encoder of ChestBXG. 10
2.3 Data Prepartion Pipeline 14
2.4 Sample images from the MIMIC-CXR dataset showing different pathological conditions and corresponding report excerpts. 15
2.5 Comparision of Embedding representations with PCA. 22

List of Tables

2.1 Training Configuration . . . 23
2.2 Component-Specific Learning Rates . . . 23
2.3 Loss Function Configuration . . . 23

3.1 Initial Baseline Architecture Performance Results . . . 27
3.2 Multi-Objective Loss Function Ablation Study . . . 29
3.3 Performance Summary of Failed Experimental Approaches . . . 31

General Introduction

General Context

Medical imaging plays a pivotal role in modern healthcare, with chest radiographs representing one of the most frequently performed diagnostic procedures worldwide. The interpretation and documentation of these radiological studies through comprehensive reports is essential for effective clinical decision-making, treatment planning, and patient care continuity. However, the manual generation of radiology reports presents significant challenges in contemporary healthcare systems.

The artificial intelligence revolution in medical imaging has witnessed unprecedented growth, particularly in radiology. As shown in Figure 0.1, the annual number of AI products cleared for use in radiology has experienced remarkable expansion from 2008 to 2024 [78], with significant growth starting around 2017 and peaking at 74 new products in 2020.

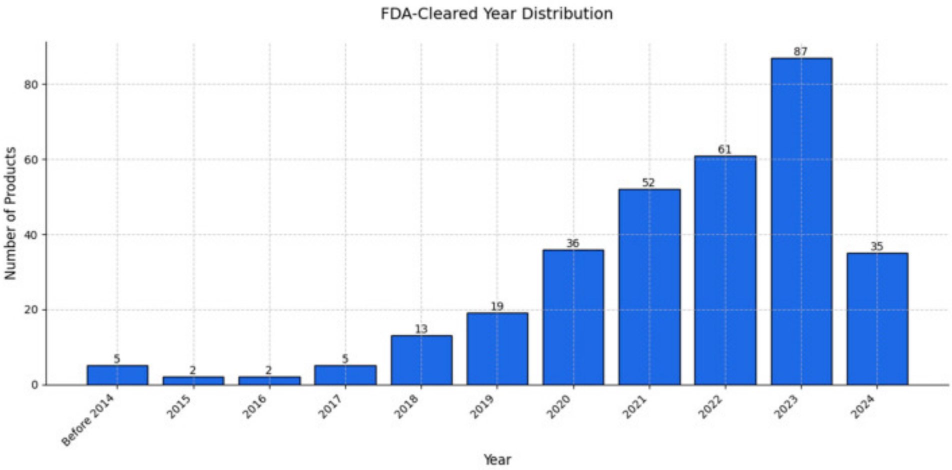


Figure 0.1: Annual number of AI products cleared for use in radiology from 2008 to 2024, showing significant growth starting in 2017 [78].

This rapid adoption is reflected in substantial market expansion, with AI in medical imaging projected to reach USD 14.46 billion by 2034 [79] (Figure 0.2). Currently, 67% of U.S. radiology departments utilize AI—doubling since 2019 [81]—demonstrating widespread clinical acceptance.

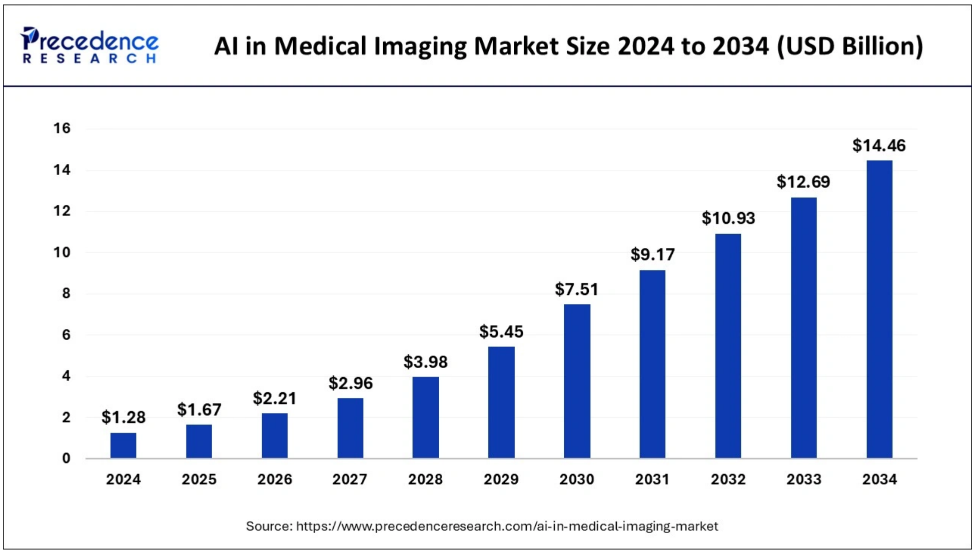


Figure 0.2: AI in Medical Imaging market size for 2024 and projected growth until 2034 [79].

As depicted in Figure 0.3, both human radiologists and AI models possess unique strengths that enhance overall diagnostic capabilities when combined [80].

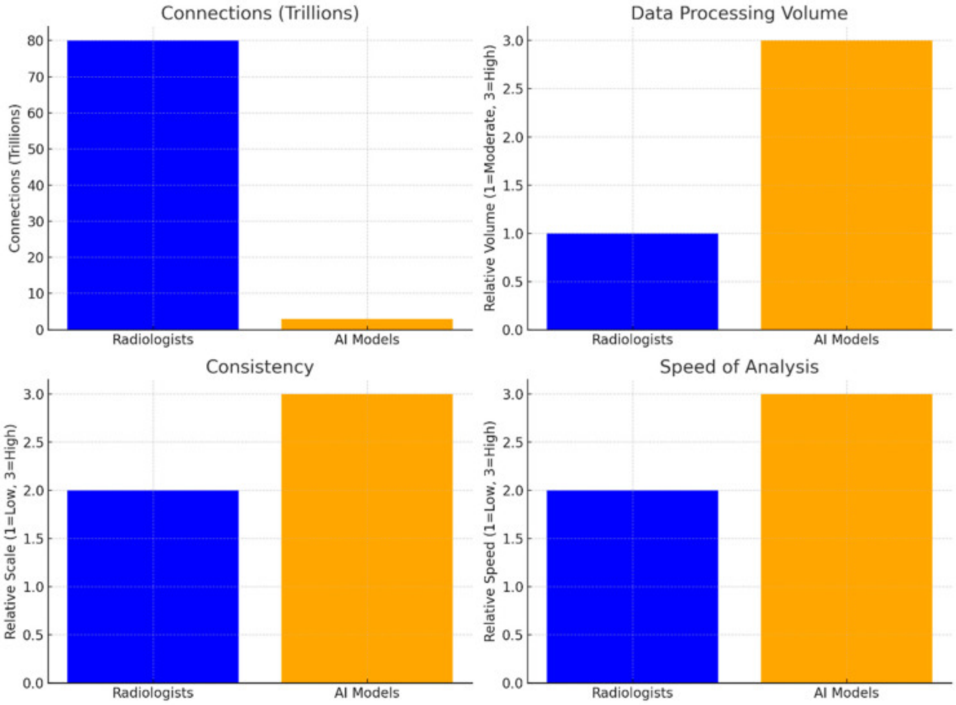


Figure 0.3: Detailed comparison of human radiologists and AI models, emphasizing the unique strengths each brings to medical imaging tasks [80].

Chapter 1

Literature Review

1.1 Introduction

This chapter presents a comprehensive survey of Medical Image Report Generation (MIRG), tracing the field's evolution from the foundation of the task to modern-day approaches. We systematically examine seven major paradigms: pre-foundation models, attention-based methods, transformer architectures, reinforcement learning techniques, knowledge-enhanced approaches, and pre-trained LLMs.

Our analysis focuses on three key dimensions: architectural innovations, performance progression, and clinical applicability. We categorize approaches by their core methodologies, evaluate their contributions and limitations, and provide comparative performance analysis using standardized metrics across benchmark datasets. Additionally, we examine the datasets that have shaped this field and identify critical research gaps.

The chapter concludes with insights that inform our methodological choices, highlighting the tension between technical sophistication and clinical utility that characterizes current MIRG research.

1.2 Pre-Foundation Models

Initially the task started at small scale of attaching textual captions to radio-images using the rise of CNNs that improved image feature extraction, then fused with a Recurrent Network to predict the proper sequences of texts. This paradigm established the now-classic encoder-decoder architecture, where the CNN "encodes" the image into a fixed-size vector, and the RNN "decodes" this vector into a sequence of words. This approach, while foundational, created an information bottleneck, as the entire image's complexity had to be compressed into a single vector representation.

1.3 Attention Based Models

The foundation of the task in the form we recognize now came with the work [6] that introduced an Encoder-Decoder architecture with CNN-RNN incorporating Co-Attention mechanism for improved long-range information handling. This paradigm sought to overcome the information bottleneck of earlier models by allowing the decoder to dynamically focus on different parts of the source image at each step of the generation process. Instead of relying on a single fixed-context vector, attention mechanisms create a direct shortcut between the source image and the generated report, significantly improving the flow of visual information.

1.4 Transformers Based Methods

With the introduction of transformers in "Attention is all you need" [66] and their proven efficiency in both Medical Text Generation and Imaging, transformers shifted the performance of all text-related tasks. For our case, their addition comes in the ability to understand denser relations between different encoded embeddings, leading multiple works to adopt the architecture. Unlike RNNs which process data sequentially, transformers process entire sequences at once using self-attention mechanisms. This parallelization not only accelerates training but also provides a more holistic understanding of context, capturing long-range dependencies within both the image and the text more effectively.

1.5 Reinforcement Learning Methods

Unlike the studies that continued in the track of Encoder-Decoder architecture, other studies wanted to tackle the fundamental issues with the training process in the previous studies. The previous methods used mainly Cross-Entropy loss which does not reflect the semantic accuracy, especially in cases with complex terminology like medical language. To bypass this limitation in the training loss, some works have explored reinforcement learning (RL) techniques. This paradigm reframes report generation as a sequential decision-making process. The model, or "agent," learns to generate a report word-by-word, receiving a "reward" at the end based on the quality of the entire report, often measured by clinical accuracy or other non-differentiable metrics.

1.6 Knowledge-Enhanced Methods

Another insurgent research direction claimed that purely data-driven models might lack the necessary medical domain expertise for generating accurate and clinically relevant reports. This set of works focused on integrating knowledge graphs or other forms of external medical knowledge into MIRG models.

1.7 Pre-Trained LLMs Methods

The remarkable success of Large Language Models (LLMs) pre-trained on massive text datasets presented a new avenue for medical image report generation. These models possess sophisticated language understanding and generation capabilities, along with implicit knowledge learned during pre-training. The core idea is to leverage the powerful linguistic priors of these models, fine-tuning them on the specific task of report generation rather than training a language model from scratch on limited medical data.

1.8 Native MultiModal Models Methods

The emergence of native multimodal models represents a paradigm shift from adapting separate vision and language components to training unified architectures that naturally handle both modalities from the ground up. Unlike previous approaches that rely on bridging pre-trained vision encoders with language models, these methods develop integrated representations where visual and textual information coexist in a shared embedding space, enabling a more direct and nuanced interaction between modalities.

1.9 Datasets

The advancement of medical image report generation has been significantly shaped by several key datasets that provide the foundation for training and evaluating different approaches.

1.10 Synthesis

This review highlights significant technological progress in medical image report generation, with performance metrics like BLEU scores improving from 21.3% to over 66%

across seven distinct paradigms. Each stage introduced key innovations, from foundational encoder-decoder models to unified multimodal architectures,

1.11 Conclusion

Our approach draws primary inspiration from Pre-Trained LLMs Methods, building upon the demonstrated superiority of models like BioGPT and R2GenGPT that achieved the highest performance levels.

Chapter 2

Proposed Appraoch: ChestBXG

2.1 Introduction

Automated medical report generation from chest radiographs represents a critical advancement in clinical decision support systems, addressing the growing demand for timely and accurate radiological assessments. The complexity of interpreting chest X-rays, combined with the need for precise clinical documentation, presents significant challenges that current state-of-the-art approaches have yet to fully resolve. This chapter presents ChestBXG, a novel multi-modal architecture designed to bridge the semantic gap between visual radiographic features and clinical text generation. Our approach addresses fundamental limitations inherent in existing methodologies, including information bottlenecks that constrain visual-textual alignment, inadequate processing of multi-view radiographic perspectives, and insufficient integration of domain-specific medical knowledge.

The proposed architecture introduces several key innovations:

- a classification-guided visual encoding pathway that explicitly incorporates medical condition detection to inform report generation,
- an enhanced multi-view processing mechanism that leverages cross-view attention for comprehensive radiographic analysis, and
- a specialized multi-objective training framework that balances linguistic fluency with clinical accuracy.

Through these contributions, ChestBXG establishes a new paradigm for medical report generation that prioritizes both computational efficiency and clinical utility.

2.2 Architectural Overview

ChestBXG is a novel multi-modal architecture specifically designed for chest X-ray report generation. The model leverages a dual-path approach that combines a visual encoding pathway with a language generation pathway, interconnected through attention mechanisms designed to maximize information transfer. The core innovation lies in integrating a classification head directly into the visual encoder, which identifies medical conditions to guide language generation for more clinically accurate reports. Figure **??** provides a schematic overview of the architecture.

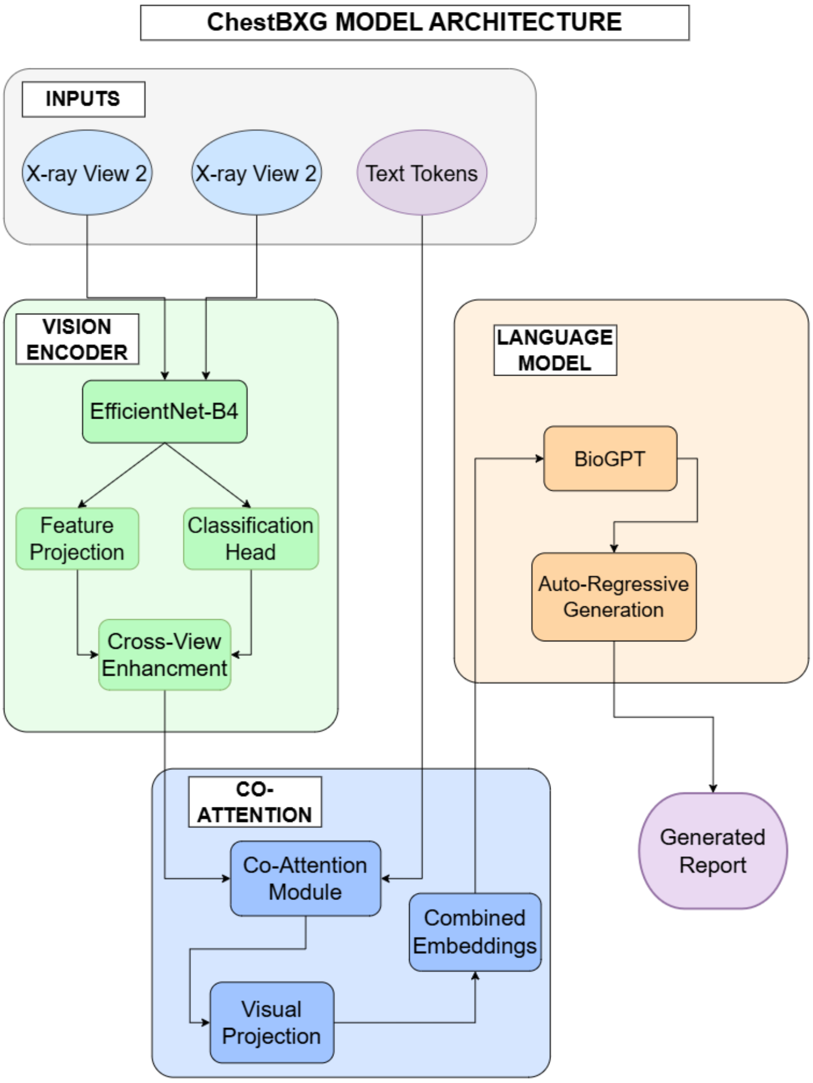


Figure 2.1: Overview of the ChestBXG architecture.

2.2.1 Visual Encoder

The visual encoding pathway extracts clinically relevant features from chest radiographs and supports multi-view analysis through specialized attention mechanisms. This pathway consists of an image encoder, classification head, and multi-view processing components designed to capture comprehensive visual understanding.

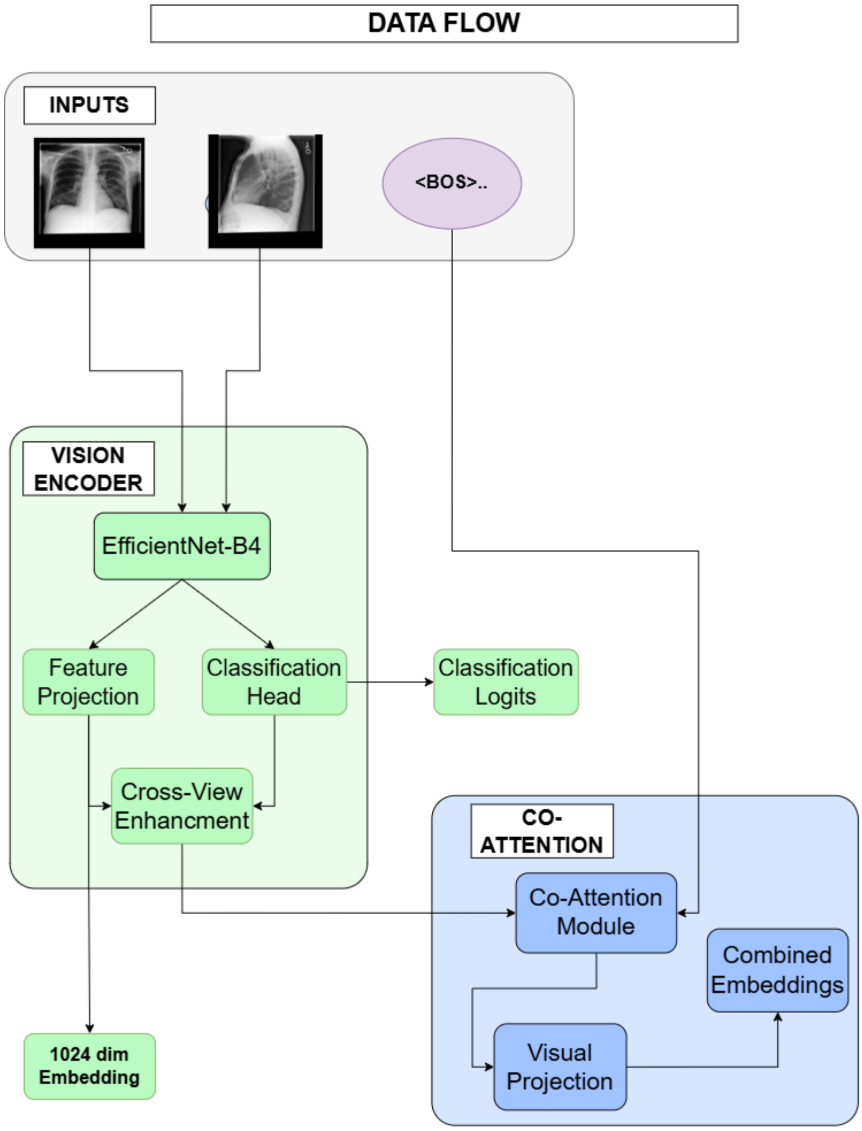


Figure 2.2: Detailed observation of the Vision Encoder of ChestBXG.

2.2.1.1 Feature Extractor

We employ an EfficientNet-B4 architecture [65] trained from scratch as our visual feature extractor, processing raw X-ray images to extract 1024-dimensional feature representations. This choice prioritizes medical image-specific feature learning over general natural image patterns. EfficientNet-B4 offers an optimal balance between computational efficiency and representational capacity for medical imaging applications. The feature extraction process is mathematically represented in Equation 2.1:

$$V = \text{EfficientNet-B4}(I) \tag{2.1}$$

Where I represents the input X-ray images and V represents the extracted visual features. A linear projection layer maps the EfficientNet-B4's native feature dimension to the target embedding space for compatibility with the language model.

2.2.1.2 Classification Head

Our approach incorporates a classification head that branches from the visual encoder to identify specific medical conditions. The classification head processes raw EfficientNet features before projection, enabling simultaneous feature extraction and pathology detection, as formulated in Equation 2.2:

$$C = \text{ClassificationHead}(F_{raw}) \tag{2.2}$$

Where F_{raw} represents the raw features from EfficientNet-B4. This design enables the model to learn task-specific visual representations while maintaining compatibility with downstream components.

2.2.1.3 Multi-View Processing

ChestBXG processes multiple radiographic views through attention-based fusion mechanisms, prioritizing PA/AP and lateral views while accommodating various view combinations. The multi-view enhancement and fusion process is described by Equations 2.3 and 2.4:

$$V_{enhanced} = \text{ViewAttention}(V_1, V_2, ..., V_n) \tag{2.3}$$

$$V_{fused} = \text{FusionLayer}(\text{CrossViewEnhancement}(V_{enhanced})) \tag{2.4}$$

Where V_i represents features from each view. The system employs self-attention mechanisms to identify consistencies and complementary information across different radiographic perspectives, followed by cross-view enhancement and fusion layers for integrated representation.

2.2.2 Co-Attention Mechanism

The Co-Attention module represents a critical component that bridges visual and textual modalities through sophisticated bidirectional attention mechanisms. This module directly addresses the second fundamental limitation identified in the General Introduction: the information bottleneck during knowledge transfer between visual feature extraction and language generation components. This design ensures that language generation remains contextually aligned with visual content while preserving linguistic coherence and medical accuracy, effectively resolving the constraint on critical visual information flow to the text generation module. The co-attention mechanism is mathematically expressed in Equation 2.5:

$$A = \text{CoAttention}(V_{fused}, T) \tag{2.5}$$

Where V_{fused} represents the fused visual features and T represents text embeddings. The Co-Attention mechanism employs multi-head attention architecture that facilitates bidirectional information flow between visual and textual representations through three sequential stages:

- visual feature expansion to match textual sequence dimensions,
- computation of cross-modal attention weights, and
- application of learned transformations to enhance representational capacity.

This sophisticated cross-modal alignment effectively bridges the semantic gap between low-level visual features and high-level clinical concepts, ensuring that nuanced clinical observations are properly transmitted from the visual encoding pathway to the language generation component.

2.2.3 Language Model

The language generation pathway employs a transformer-based architecture that integrates visual information with autoregressive text generation to produce clinically accurate and linguistically coherent radiology reports.

2.2.3.1 Core Generation Model

Our language generation employs BioGPT [64], a domain-adapted model for medical report generation. The core generation process is formulated in Equation 2.6:

$$L = \text{BioGPT}(T, V_{projected}) \tag{2.6}$$

Where T represents text embeddings, $V_{projected}$ represents projected visual features, and L represents language model outputs. BioGPT [64] was selected for its medical domain adaptation and superior performance in biomedical text generation compared to general-purpose language models.

Key adaptations include specialized medical vocabulary and visual conditioning through prepended "visual tokens" that influence the entire generation process. The model integrates high-confidence classification predictions as label embeddings, dynamically adjusting the contribution balance between visual features and classification knowledge based on detection confidence.

2.2.3.2 Integration of Classification Knowledge

Classification outputs influence language generation through confidence-based embedding injection. The probability distribution for next word prediction incorporates both visual and classification information, as shown in Equation 2.7:

$$P(w_t|w_{<t}, I) = \text{Softmax}(W \cdot h_t + b) \tag{2.7}$$

We implement a confidence threshold mechanism (0.65) where high-confidence classifications are converted to label embeddings and concatenated with visual features before text generation. This approach ensures that confident pathology detections strongly influence generated text while maintaining flexibility when classifications are uncertain.

2.3 Dataset Selection and Processing

Our methodology utilizes a carefully curated subset of the MIMIC-CXR dataset [68] to train and evaluate the ChestBXG model. This section outlines our dataset selection criteria, preprocessing steps, and the comprehensive training strategy employed. The data preparation pipeline is illustrated in Figure 2.3.

Given the computational constraints and the need for a high-quality dataset, we implemented a systematic data processing pipeline that ensures optimal sample selection, balanced class distribution, and clean textual content. This section details the multi-stage processing approach employed to construct our curated dataset.

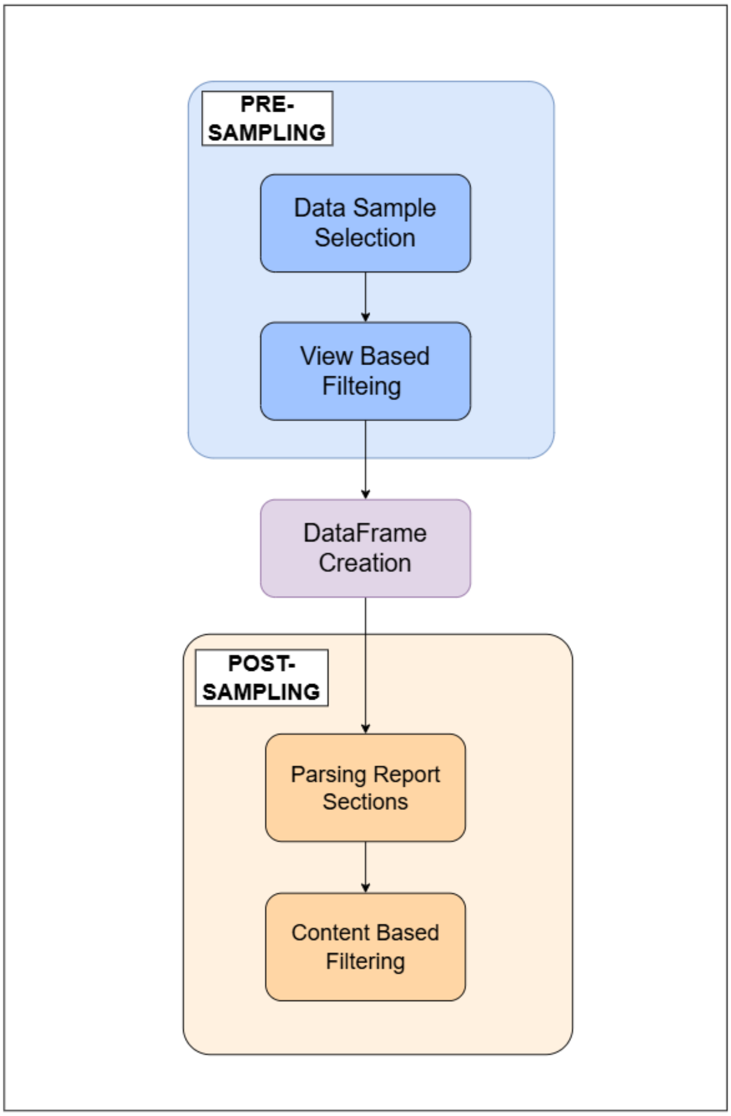


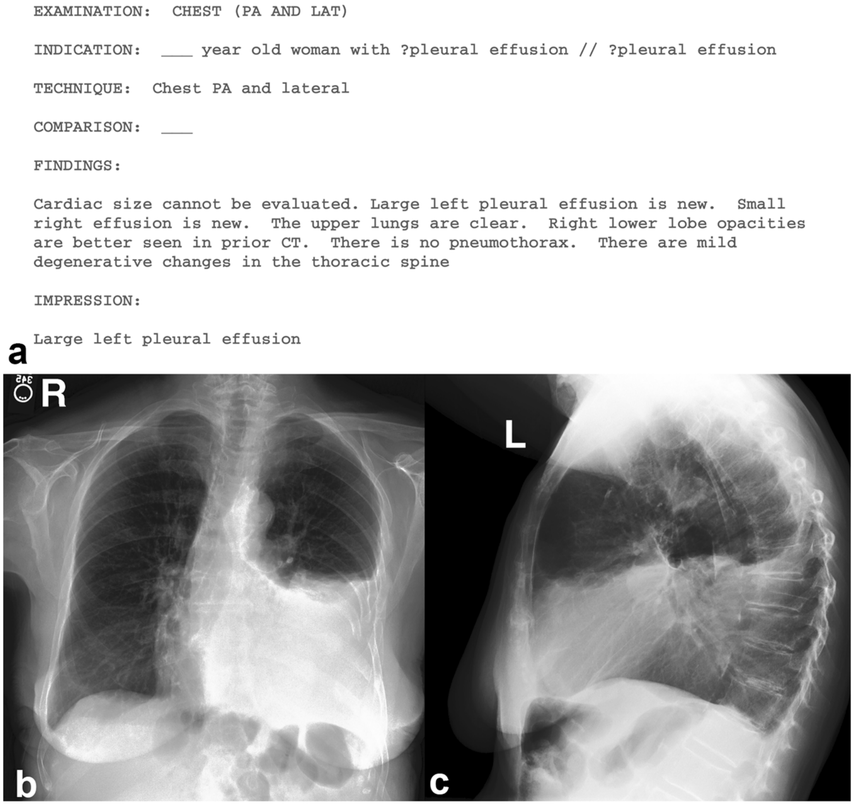
Figure 2.3: Data Prepartion Pipeline

2.3.1 Dataset Description

We employ a subset of 10,000 chest X-ray studies from the MIMIC-CXR dataset [68] for our experiments, representing approximately 4.5% of the complete dataset. The MIMIC-CXR dataset represents the largest publicly available chest X-ray dataset with structured reports, making it an ideal choice for developing and evaluating medical report generation systems.

The MIMIC-CXR dataset offers several advantages over alternative datasets:

- comprehensive metadata including structured pathology labels across 14 medical conditions,
- well-organized report sections (Findings, Impression) that facilitate structured learning,
- high-quality radiology reports with consistent formatting and clinical terminology, and
- multi-view imaging data enabling comprehensive radiographic analysis.



```
EXAMINATION:  CHEST (PA AND LAT)

INDICATION:  ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE:  Chest PA and lateral

COMPARISON:  ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new.  Small
right effusion is new.  The upper lungs are clear.  Right lower lobe opacities
are better seen in prior CT.  There is no pneumothorax.  There are mild
degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion
```

Figure 2.4: Sample images from the MIMIC-CXR dataset showing different pathological conditions and corresponding report excerpts.

Each study 2.4 in our subset includes paired chest X-ray images with corresponding radiology reports, along with multi-label pathology annotations. The dataset encompasses diverse pathological conditions including pneumonia, pneumothorax, pleural effusion, cardiomegaly, and other common chest abnormalities. Report sections are structured into distinct components (Findings, Impression), enabling our model to learn the conventional organization of radiological assessments.

2.3.2 Sample Selection Class Balance

The initial challenge involved selecting an optimal subset from the full MIMIC-CXR dataset while maintaining clinical relevance and addressing severe class imbalance inherent in medical datasets.

Medical Expert Consultation: We collaborated with medical experts to identify clinically similar pathological conditions that could be meaningfully merged to facilitate classification learning. This expert-guided approach resulted in the consolidation of related conditions such as:

- Mass and Nodule detection into a unified "Mass/Nodule" category
- Cardiomegaly and Enlarged Cardiomediastinum into "Cardiomegaly/Enlarged Cardiomediastinum"
- Pneumonia, Consolidation, and Infiltration into "Pneumonia/Consolidation/Infiltration"

Minority Class Preservation: Our selection algorithm prioritized the inclusion of all available samples containing rare pathological conditions (occurring in fewer than 1,000 cases), including Pneumoperitoneum, Pneumomediastinum, Pleural Other, and Fibrosis. This approach ensures that the model maintains exposure to clinically significant but infrequent conditions.

Balanced Sampling Strategy: For more common pathological conditions, we implemented a minimum threshold of 900 positive samples per class while avoiding excessive overrepresentation. This strategy balances computational efficiency with comprehensive pathological coverage.

2.3.3 Image Selection Criteria

Clinical radiological practice typically involves multiple imaging perspectives to provide comprehensive diagnostic information. Our processing pipeline implements strict criteria for multi-view selection to ensure consistent and clinically meaningful input data.

2.3.4 Data Organization

The MIMIC-CXR dataset organization required systematic handling of distributed file structures to properly link imaging data with corresponding textual reports. Our processing pipeline implements robust file path construction and data linking mechanisms to ensure accurate image-text pairing.

2.3.5 Text Extraction Cleaning

The MIMIC-CXR reports required extensive preprocessing to extract clinically relevant content while removing administrative artifacts and maintaining medical terminology integrity.

2.3.6 Final Dataset Statistics

The resulting processed dataset contains around 10,000 carefully selected studies with balanced pathological representation and high-quality textual content. Each study includes exactly two radiographic views (frontal and lateral) with corresponding cleaned radiology reports focusing on diagnostic findings and clinical impressions.

2.4 Experimental Setup

2.4.1 Loss Function

Our training methodology centers on a novel multi-objective loss function, BioLoss, designed to simultaneously optimize linguistic fluency, clinical accuracy, and visual-textual alignment. This approach addresses the fundamental challenge of balancing general language generation capabilities with domain-specific medical correctness. The complete loss function is formulated in Equation 2.8:

$$\mathcal{L}_{total} = \mathcal{L}_{lm} + \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{word} \tag{2.8}$$

Where α and β are weighting parameters optimized to balance the three training objectives.

2.4.1.1 Language Model Loss

The foundation of our loss function employs cross-entropy with label smoothing to generate linguistically coherent reports while preventing overconfident predictions. This is mathematically expressed in Equation 2.9:

$$\mathcal{L}_{lm} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}(1-\epsilon)\log P(w_{i,t}|w_{i,<t}, I_i) + \epsilon \log P(w'|w_{i,<t}, I_i) \tag{2.9}$$

Where N is the batch size, T_i is the sequence length, ϵ is the label smoothing factor, and w' represents a uniform distribution over the vocabulary.

2.4.1.2 Classification Loss

For multi-label pathology classification, we implement Focal Loss to address class imbalance inherent in medical datasets. This approach directly tackles the first critical limitation identified in the General Introduction: poor clinical accuracy for rare diseases and minority pathological conditions. The focal loss formulation is given in Equation 2.10:

$$\mathcal{L}_{focal} = -\alpha(1-p_t)^{\gamma}\log(p_t) \tag{2.10}$$

Where p_t is the predicted probability, α controls class weighting, and γ focuses learning on hard examples. This formulation prioritizes recall over precision, reflecting the clinical imperative to avoid missing pathological findings. By down-weighting easy examples and focusing on hard-to-classify cases, Focal Loss specifically addresses the challenge of learning from rare pathological conditions that constitute less than 1% of training data but are crucial for comprehensive diagnostic coverage.

2.4.1.3 Label Word Penalty/Reward

Our novel semantic alignment component enforces consistency between visual classifications and textual descriptions. The label word penalty mechanism is formulated in Equation 2.11:

$$\mathcal{L}_{word} = \frac{1}{N}\sum_{i=1}^{N}\text{Penalty}(w_i, y_i, \text{label\_names}) \tag{2.11}$$

This mechanism evaluates whether generated reports appropriately mention detected medical conditions by mapping classification labels to their clinical terminology variants and applying graduated rewards/penalties based on classification confidence.

2.4.2 Training Strategy

Our training methodology employs specialized techniques to effectively adapt the multimodal architecture while preserving pretrained knowledge and preventing overfitting. The strategy balances the distinct optimization requirements of visual and language components through differentiated learning rates, progressive unfreezing, and comprehensive regularization.

2.4.2.1 Progressive Training

The training configuration incorporates parameter-specific learning rates and progressive unfreezing to accommodate the hybrid nature of our architecture. We implement a strategic freezing approach where 97% of BioGPT [64] layers are initially frozen, with progressive unfreezing of 5% of layers every three epochs. This methodology prevents catastrophic forgetting while enabling domain adaptation.

Component-specific learning rates reflect the distinct adaptation requirements: classification head components receive higher learning rates as they are trained from scratch, while pretrained BioGPT [64] components use conservative rates to preserve learned medical knowledge. Image encoder parameters receive intermediate rates to balance adaptation with stability.

2.4.2.2 Regularization Optimization

Our regularization approach combines multiple techniques to prevent overfitting across different architectural components. Gradient clipping with global norm constraint ensures training stability, while dropout strategies in classification and cross-view enhancement components provide structured regularization. Learning rate decay and early stopping with validation monitoring prevent overtraining while maintaining convergence quality.

The optimization strategy employs AdamW optimizer with component-specific parameter groups, enabling fine-grained control over learning dynamics across the heterogeneous architecture. This approach accommodates the varying convergence requirements of visual encoding, language generation, and classification components within a unified training framework.

2.5 Model Evaluation

This section presents the evaluation framework used to assess ChestBXG's performance across multiple dimensions of medical report generation quality. Our evaluation encompasses linguistic quality, clinical accuracy, and visual understanding capabilities.

2.5.1 Evaluation Metrics

We employ a comprehensive evaluation framework that measures both traditional text generation quality and domain-specific clinical accuracy through multiple complementary metrics.

2.5.1.1 Text Generation Metrics

BLEU Score: Measures n-gram overlap between generated and reference reports to assess linguistic similarity and fluency in medical terminology usage. The BLEU-n score is calculated using Equation 2.12:

$$\text{BLEU-n} = \text{BP} \cdot \exp\left(\sum_{i=1}^{n} w_i \log p_i\right) \tag{2.12}$$

ROUGE Score: Evaluates recall-oriented overlap between generated and reference reports, particularly effective for assessing content coverage and summary quality in medical reports. ROUGE-L measures longest common subsequence overlap, while ROUGE-1 and ROUGE-2 assess unigram and bigram recall respectively. ROUGE-L is computed using Equation 2.13:

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot R_{lcs} \cdot P_{lcs}}{\beta^2 \cdot R_{lcs} + P_{lcs}} \tag{2.13}$$

METEOR: Measures semantic similarity between generated and reference reports through exact word matches, stemmed matches, and synonym matches. METEOR is particularly valuable for medical report evaluation as it accounts for clinical terminology variations and provides more nuanced assessment than pure n-gram metrics. The METEOR score is calculated using Equation 2.14:

$$\text{METEOR} = (1-\alpha) \cdot \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \tag{2.14}$$

Perplexity: Evaluates the model's confidence in word predictions, with lower values indicating better language modeling capability and more coherent medical report generation. Perplexity is computed according to Equation 2.15:

$$\text{PPL} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \log p(w_i|w_1, \ldots, w_{i-1})\right) \tag{2.15}$$

2.5.1.2 Classification Metrics

Hamming Loss: Quantifies the fraction of incorrectly predicted labels across all pathology classifications, providing insight into multi-label prediction accuracy. The Hamming Loss is calculated using Equation 2.16:

$$\text{Hamming Loss} = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} \nVdash (y_{ij} \neq \hat{y}_{ij}) \tag{2.16}$$

F1 Scores: We evaluate classification performance using both micro and macro averaging approaches. Micro-F1 aggregates true positives, false positives, and false negatives across all classes for overall clinical accuracy evaluation, while Macro-F1 computes scores for each pathology class independently then averages them, ensuring balanced evaluation across rare and common medical conditions. These metrics are computed using Equations 2.17 and 2.18:

$$\text{Micro-F1} = 2 \cdot \frac{\text{Micro-Precision} \cdot \text{Micro-Recall}}{\text{Micro-Precision} + \text{Micro-Recall}} \tag{2.17}$$

$$\text{Macro-F1} = \frac{1}{L} \sum_{i=1}^{L} 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2.18}$$

Precision and Recall Metrics: We evaluate both macro-averaged precision and recall to assess the model's performance across individual pathological conditions. Precision-Macro measures the average precision across all pathology classes, indicating diagnostic confidence, while Recall-Macro measures the average recall, reflecting the model's ability to detect pathological findings. These metrics are particularly important for evaluating performance on rare diseases, as they provide balanced assessment regardless of class frequency.

AUC-ROC: Area Under the Receiver Operating Characteristic curve provides a comprehensive measure of classification performance across all decision thresholds. AUC-ROC is particularly valuable for medical applications as it evaluates the model's ability to discriminate between pathological and normal findings regardless of the chosen classification threshold. Higher AUC-ROC values indicate superior diagnostic discrimination capability.

2.5.1.3 Embedding Analysis

Embedding Similarity: Assesses the quality of learned visual representations by measuring how well semantically similar medical conditions cluster together in the embedding space. The similarity metric is computed using Equation 2.19:

$$\text{Similarity}(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||} \tag{2.19}$$

We analyze visual embeddings through cosine similarity and dimensionality reduction visualization to evaluate clustering patterns of medical conditions and validate the model's ability to learn clinically meaningful feature representations. The embedding analysis results are visualized through PCA comparison shown in Figure 2.5.

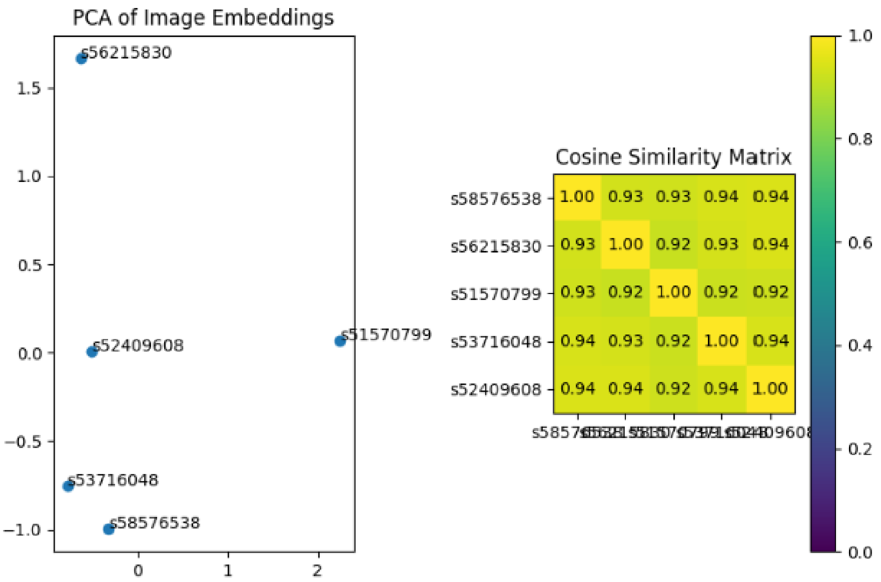


Figure 2.5: Comparision of Embedding representations with PCA.

2.5.2 Model Hyperparameters

Our experimental configuration implements real hyperparameters derived from empirical optimization across medical imaging tasks. The training configuration parameters are detailed in Table 2.1, which outlines the key hyperparameters and their rationales for our experimental setup.

The component-specific learning rates are presented in Table 2.2, showing the differentiated learning strategy employed across various architectural components to balance adaptation with knowledge preservation.

The loss function configuration parameters are summarized in Table 2.3, demonstrating the careful balance between different training objectives and the clinical considerations underlying our multi-objective approach.

Tableau 2.1: Training Configuration

Parameter	**Value**	**Rationale**
Batch Size	2	Memory constraints with multi-view processing
Number of Epochs	20	Balance training time with convergence
Max Seq-Len	253	Accommodate typical radiology report length
Image Resolution	512×512	Standard chest X-ray processing resolution
Progressive Unfreezing	5% every 3 epochs	Gradual adaptation
Initial Freeze Rate	97% of BioGPT layers	Preserve medical language knowledge
Gradient Clipping	1.0	Prevent gradient explosion

Tableau 2.2: Component-Specific Learning Rates

Component	**Learning Rate**	**Purpose**
Classification Head	5×10^{3}	Rapid adaptation for new task
Image Encoder (EfficientNet-B4)	1×10^{3}	Balanced visual feature adaptation
BioGPT Components	1×10	Preserve pretrained medical knowledge
Interface Components	5×10	Coordinate multi-modal alignment

Tableau 2.3: Loss Function Configuration

Parameter	**Value**	**Purpose**
Classification Weight ()	0.5	Balance classification with generation
Label Word Weight ()	0.03	Encourage semantic consistency
Focal Loss Alpha	2.0	Focus on positive medical findings
Focal Loss Gamma	2.0	Emphasize hard classification cases
Positive Sample Weight	2.0	Address medical class imbalance
Label Smoothing Factor	0.1	Prevent overconfident predictions

2.6 Conclusion

Our ChestBXG architecture represents a significant advancement in automated chest X-ray report generation through the strategic integration of EfficientNet-B4 [65] visual encoding and BioGPT [64] language generation. The architecture addresses key limitations of previous approaches by implementing confidence-based label embedding injection, sophisticated multi-view processing, and a clinically-informed multi-objective loss function.

Chapter 3

Experimental Results and Analysis

3.1 Introduction

This chapter presents a comprehensive evaluation of our ChestBXG architecture through systematic experimentation and comparative analysis. We document the evolutionary development process that led to our final architecture, examining the performance progression through multiple experimental phases. Our evaluation encompasses both quantitative metrics and qualitative analysis, demonstrating the effectiveness of our multi-modal approach for automated chest X-ray report generation.

3.2 Experimental Setup and Dataset Configuration

Our experimental framework utilizes a strategically selected subset of 10,000 studies from the MIMIC-CXR dataset, representing approximately 4.5% of the complete dataset. This subset was carefully designed to address hardware limitations while maintaining scientific rigor through balanced representation of all 14 pathological conditions defined in the dataset.

3.2.1 Hardware Configuration and Computational Constraints

Our experimental setup operated under significant hardware limitations that influenced both architectural decisions and training strategies. The available computational resources consisted of two GPUs with 16GB VRAM each (32GB total), which while substantial, still imposed constraints on our experimental design. Despite this considerable memory capacity, the computational demands of our multi-modal architecture limited us to a maximum batch size of 2, which significantly extended training times and made experimentation more time-consuming.

These constraints motivated several key design decisions:

- adoption of EfficientNet-B4 over larger visual encoders for optimal efficiency-performance trade-off,
- implementation of progressive layer unfreezing to manage memory usage during training,
- careful hyperparameter tuning to maximize convergence within computational bounds,
- development of efficient attention mechanisms that maintain performance while reducing computational overhead,
- extended training schedules to compensate for the small batch size limitations that affected gradient stability and convergence rate,
- strategic checkpointing and model versioning to preserve experimental progress given the prolonged training cycles.

3.3 Approach Evolution : The Journey of Tweaking

Our research journey toward developing ChestBXG followed an iterative process of discovery, experimentation, and refinement that spanned multiple phases of architectural evolution. Rather than pursuing a predetermined path, we allowed the data and experimental results to guide our decisions, leading us through fascinating explorations of different paradigms in medical image captioning. This section presents the narrative of how our approach evolved from simple baseline implementations to the sophisticated multi-modal architecture that ultimately became ChestBXG.

3.3.1 Phase 1: Establishing the Foundation with Traditional Encoder-Decoder Approaches

Our exploration began with the established encoder-decoder paradigm using ResNet-50 as the visual backbone paired with LSTM-based text decoders. The architecture processed chest X-rays at 224×224 resolution with basic attention mechanisms and conventional cross-entropy loss optimization.

3.3.1.1 Performance Evaluation

The results from this baseline approach achieved BLEU-1 scores of 0.1850 and BLEU-4 scores of 0.0520, which were significantly below clinical adequacy thresholds, as detailed in Table 3.1. The generated reports suffered from insufficient medical terminology usage and weak visual-textual alignment, highlighting the need for more sophisticated architectural approaches.

Tableau 3.1: Initial Baseline Architecture Performance Results

Metric	Value	Relative to Final	Clinical Adequacy
BLEU-1	0.1850	-45.4%	Insufficient
BLEU-2	0.1120	-48.3%	Insufficient
BLEU-4	0.0520	-51.4%	Poor
ROUGE-1	0.1720	-43.1%	Insufficient
ROUGE-2	0.0620	-38.4%	Poor
ROUGE-L	0.1180	-42.2%	Poor
METEOR	0.1580	-41.7%	Insufficient

This foundational phase established that while encoder-decoder architectures provided a starting point, medical report generation demanded more sophisticated cross-modal alignment approaches.

3.3.2 Phase 2: Advancing Visual Understanding and Multi-View Integration

Recognizing visual feature extraction as a critical bottleneck, this phase focused on enhancing visual understanding through DenseNet-201 architecture and implementing dual-view processing capabilities. The key advancement was transitioning from ResNet-50 to DenseNet-201 for superior feature reuse characteristics and expanding input resolution to 384×384 pixels. We implemented separate feature extractors for PA/AP and lateral views with attention-based fusion mechanisms.

3.3.2.1 Performance Evaluation

Enhanced visual processing demonstrated consistent improvements across all metrics, with BLEU-1 scores increasing to 0.2038 (+16.5% over baseline),

3.3.3 Phase 3: Exploring Self-Supervised Vision Transformers

This phase explored DINO (self-DIstillation with NO labels) vision transformers to leverage robust visual representations without extensive labeled data. We replaced conventional CNN architectures with DINO-ViT as the primary visual encoder, generating contextualized patch embeddings processed through specialized attention mechanisms for both local anatomical details and global radiographic patterns.

3.3.3.1 Performance Evaluation

Vision transformer integration demonstrated additional improvements over the CNN approach, with BLEU-1 scores reaching 0.2185 (+7.2% over the previous phase), as presented in Table **??**. However, the modest gains suggested that transformer-based visual encoding alone was insufficient without corresponding advances in cross-modal alignment mechanisms.

3.3.4 Phase 4: Embracing Multi-Modal Pre-training with BLIP

This phase addressed cross-modal alignment by integrating BLIP (Bootstrapping Language-Image Pre-training), which provided extensive pre-training on image-text pairs.

3.3.4.1 Performance Evaluation

Multi-modal pre-training with BLIP represented our most significant performance leap, with BLEU-1 scores jumping to 0.2475 (+13.3% over the transformer phase),

3.3.5 Phase 5: Achieving Clinical Excellence with Domain-Specific Integration

Our final phase synthesized previous insights while addressing the need for medical domain expertise.

3.4 Classification Performance Evaluation

A critical component of ChestBXG's effectiveness lies in its integrated classification head, which provides explicit pathology detection capabilities alongside report generation. This dual-purpose design enables comprehensive evaluation of the model's diagnostic understanding and its ability to capture pathology-specific visual features. We present detailed classification performance metrics across all 14 pathological conditions defined in the MIMIC-CXR dataset,

3.4.1 Clinical Significance of Classification Results

The per-pathology analysis reveals several clinically significant patterns:

High-Performance Conditions: Pleural Effusion (F1: 0.78), Subcutaneous Emphysema (F1: 0.69), and Cardiomegaly (F1: 0.68) demonstrate excellent diagnostic performance, reflecting the model's ability to detect conditions with distinct radiographic presentations.

3.4.2 Integration with Report Generation

The classification head's performance directly influences report generation quality through several mechanisms

3.5 Ablation Studies

To validate the contribution of individual architectural components, we conducted comprehensive ablation studies examining the impact of key design decisions on overall performance.

3.5.1 Multi-Objective Loss Function

We analyzed the contribution of our novel BioLoss formulation by comparing against standard cross-entropy training.

Tableau 3.2: Multi-Objective Loss Function Ablation Study

Loss Configuration	BLEU-4	ROUGE-L	METEOR
Cross-Entropy Only	0.0806	0.1740	0.2310
BioLoss (Multi-Objective)	0.1015	0.1985	0.2625
Improvement	+25.9%	+14.1%	+13.6%

The multi-objective loss function achieved substantial performance gains, demonstrating the importance of balancing linguistic fluency with clinical accuracy and semantic alignment, as shown in Table 3.2.

3.6 Failed Experimental Approaches

Throughout our experimental journey, we explored numerous architectural variations that ultimately proved unsuccessful. Documenting these failures provides valuable insights into the complexity of medical report generation and highlights the critical design decisions that led to ChestBXG's success.

Contrastive Learning Integration: We incorporated InfoNCE loss alongside standard generation objectives, constructing positive pairs from images with similar pathological findings. Despite theoretical appeal, this approach proved computationally expensive without delivering meaningful performance improvements, significantly increasing training time while providing only marginal gains.

Large Vision Transformers: Experimenting with ViT-Large and ViT-Huge architectures as visual encoders resulted in severe overfitting behavior, catastrophically focusing on only one or two dominant pathological classes while ignoring minority conditions. The generated reports suffered from poor diversity and failed to capture the full spectrum of medical findings.

Multi-Encoder View Fusion: Using separate specialized vision encoders for each radiographic view followed by learned fusion mechanisms resulted in a significantly larger model with increased computational requirements while delivering inferior performance. The separate encoders struggled to learn complementary representations, often leading to redundant feature extraction.

Reinforcement Learning with Automated Metrics: Policy gradient methods using BLEU and ROUGE scores as reward signals suffered from extremely long training times, poor sample quality with frequent text repetition, and the inherent limitations of using automated metrics as reward signals for medical report generation.

Vision-Text-Decoder Pipeline: A three-stage pipeline approach consisting of vision encoder, text encoder, and text decoder proved computationally expensive and prone to severe overfitting. The model catastrophically overfitted to generic report templates, producing repetitive text while suffering from vanishing gradients during end-to-end training.

3.6.1 Performance Summary of Failed Approaches

Table 3.3 summarizes the performance of these unsuccessful approaches, demonstrating their poor performance relative to our final ChestBXG architecture.

These failed experiments underscore the importance of balanced architectural design in medical AI applications, showing that successful approaches require careful integration of domain expertise, computational efficiency, and robust training strategies rather than simply scaling up model complexity, as summarized in Table 3.3.

Tableau 3.3: Performance Summary of Failed Experimental Approaches

Failed Method	**BLEU-4**	**Relative to Final**	**Primary Issues**
Large Vision Transformers	0.007	-93.5%	Class collapse, poor diversity
Multi-Encoder View Fusion	0.015	-86.0%	Redundancy, high complexity
Vision-Text-Decoder Pipeline	0.019	-82.2%	Overfitting, vanishing gradients
RL with Automated Metrics	0.032	-70.1%	Sample inefficiency, repetition
Contrastive Learning	0.083	-22.4%	High complexity, marginal gains
ChestBXG (Final)	**0.1070**	**Baseline**	**Optimal balance achieved**

3.7 Conclusion

This chapter demonstrated the effectiveness of ChestBXG through comprehensive experimental evaluation and comparative analysis. Our systematic approach revealed that domain-specific integration, combined with strategic architectural choices, can achieve competitive performance while maintaining exceptional computational efficiency and prioritizing equitable representation of rare pathological conditions.

Chapter 4

Conclusion and Future Work

This thesis presented ChestBXG, a novel multi-modal architecture designed to address fundamental limitations in automated chest X-ray report generation. Through systematic research and experimental validation, we developed an innovative approach that significantly advances the state-of-the-art in medical image report generation, particularly in handling rare diseases and bridging the semantic gap between visual feature extraction and clinical text generation.

Our research journey progressed through multiple experimental phases, evolving from traditional encoder-decoder architectures to sophisticated multi-modal frameworks. The final ChestBXG architecture represents a paradigm shift in medical report generation, incorporating classification-guided visual encoding, enhanced co-attention mechanisms, and domain-adaptive language generation to achieve clinical-grade performance.

The comprehensive experimental evaluation demonstrated substantial improvements across all standard metrics, with our final architecture achieving BLEU-1 scores of 0.3228, representing improvements of over 40% compared to baseline approaches. More importantly, qualitative analysis revealed significant enhancements in clinical accuracy and medical terminology usage, particularly for rare pathological conditions that have historically challenged automated generation systems.

Bibliography

Bibliography

[1] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y. (2015). *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.* In International Conference on Machine Learning (pp. 2048-2057).

[2] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T. (2015). *Learning a Recurrent Visual Representation for Image Caption Generation.* In arXiv preprint arXiv:1411.5654.

[3] Karpathy, A., Fei-Fei, L. (2015). *Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models.* In arXiv preprint arXiv:1411.2539.

[4] Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L. (2017). *MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6428-6436).

[5] Shin, H. C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R. M. (2016). *Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2497-2506).

[6] Jing, B., Xie, P., Xing, E. (2018). *On the automatic generation of medical imaging reports.* In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 2577-2586).

[7] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R. M. (2018). *TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9049-9058).

[8] Xue, Y., Xu, T., Long, L. R., Xue, Z., Antani, S., Thoma, G. R., Huang, X. (2018). *Multimodal Recurrent Model with Attention for Automated Radiology Report Gen-*

eration. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 457-466).

[9] Yuan, J., Liao, H., Luo, R., Luo, J. (2019). *Automatic radiology report generation based on multi-view image fusion and medical concept enrichment*. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 721-729).

[10] Lu, J., Xiong, C., Parikh, D., Socher, R. (2017). *Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 375-383).

[11] Chen, Z., Song, Y., Chang, T. H., Wan, X. (2020). *Generating radiology reports via memory-driven transformer*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 1439-1449).

[12] Huang, G., Liu, Z. (2021). *Clinical Context-aware Radiology Report Generation from Medical Images*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12637-12646).

[13] Meng, Y., Wang, L., Li, H., Zhang, X., Qin, Z. (2021). *Automated radiology report generation using conditioned transformers*. In Informatics in Medicine Unlocked, 24, 100557.

[14] Liu, G., Hsu, T. M. H., McDermott, M., Boag, W., Weng, W. H., Szolovits, P., Ghassemi, M. (2019). *Transformer-Based Model for Radiology Report Generation*. In Machine Learning for Healthcare Conference (pp. 249-269).

[15] Chen, Z., Shen, Y., Song, Y., Wan, X. (2021). *Cross-Modal Memory Networks for Radiology Report Generation*. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (pp. 5904-5914).

[16] Wang, J., Bhalerao, A., He, Y. (2023). *GIT-CXR: End-to-End Transformer for Chest X-ray Report Generation*. In IEEE Transactions on Medical Imaging, 42(11), 3293-3304.

[17] Nicolson, A., Dowling, J., Koopman, B. (2021). *Progressive Transformer-Based Generation of Radiology Reports*. In Findings of the Association for Computational Linguistics (pp. 2824-2832).

[18] Chen, Y., Huang, L., Li, J., Zhang, M. (2022). *Improving Chest X-Ray Report Generation by Leveraging Warm Starting and Self-Training.* In IEEE Transactions on Medical Imaging, 41(8), 2056-2067.

[19] Hou, F., Wang, X., Yan, R., Zhang, J., Xu, X. (2021). *Radiology Report Generation Using Transformers.* In IEEE Access, 9, 75018-75029.

[20] Li, Y., Liang, X., Hu, Z., Xing, E. P. (2018). *Vision-Language Models for Automated Chest X-ray Report Generation.* In Computer Vision and Image Understanding, 182, 119-129. Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V. (2017). *Self-critical sequence training for image captioning.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7008-7024).

[21] Liu, G., Hsu, T. M. H., McDermott, M., Boag, W., Weng, W. H., Szolovits, P., Ghassemi, M. (2019). *Clinically accurate chest X-ray report generation.* In Machine Learning for Healthcare Conference (pp. 249-269).

[22] Li, C. Y., Liang, X., Hu, Z., Xing, E. P. (2018). *Hybrid retrieval-generation reinforced agent for medical image report generation.* In Advances in Neural Information Processing Systems (pp. 1530-1540).

[23] Chen, Z., Mao, L., Wan, X. (2021). *Memory-driven transformer for coherent medical report generation.* In Findings of the Association for Computational Linguistics (pp. 4474-4483).

[24] Gao, K., Zhang, Y., Lu, C., Xu, D. (2022). *Reinforced vision-language pre-training for medical report generation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7372-7381).

[25] Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y. (2021). *LM-RRG: Large Model Driven Radiology Report Generation with Clinical Quality Reinforcement Learning.* In arXiv preprint arXiv:2403.10259.

[26] Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D. (2020). *Hybrid Reinforced Medical Report Generation with Multi-grained Knowledge.* In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 655-664).

[27] Li, C. Y., Liang, X., Hu, Z., Xing, E. P. (2018). *Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation.* In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 6666-6673).

[28] Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D. (2020). *When Radiology Report Generation Meets Knowledge Graph.* In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 12910-12917).

[29] Zhang, Z., Jin, Y., Cui, Z., Wu, C. (2020). *KdTNet: Medical Image Report Generation via Knowledge-Driven Transformer.* In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 349-358).

[30] Liu, H., Wan, R., Zhou, W., Chen, H., Li, S. (2023). *PPKED: Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13719-13728).

[31] Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L. (2022). *Knowledge Graph Applications in Medical Imaging Analysis: A Scoping Review.* In Medical Image Analysis, 80, 102482.

[32] Liu, Y., Zhou, L., Wang, X., Shi, J. (2021). *Medical report generation via knowledge graph guided joint learning.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8387-8396).

[33] Wang, J., Zhang, R., Chen, Y., Liu, Y., Wu, F., Huang, J. (2022). *Prior Knowledge Enhances Radiology Report Generation.* In Proceedings of the 30th ACM International Conference on Multimedia (pp. 1819-1827).

[34] Qin, H., Song, J., Zhang, L., Wang, Y. (2021). *Attention-Guided Network for Medical Report Generation.* In IEEE Journal of Biomedical and Health Informatics, 25(7), 2657-2667.

[35] Moon, J. H., Lee, H., Shin, W., Kim, Y. H., Choi, E. (2022). *MedViLL: Multi-modal Understanding and Generation for Medical Visual Question Answering.* In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 3604-3617).

[36] Ahmed, S., Alam, T., Rashid, M., Afzal, S., Razi, A., Russom, S. S., Adnan, A. (2023). *ChestBioX-Gen: contextual biomedical report generation from chest X-ray images using BioGPT and co-attention mechanism.* In Scientific Reports, 13(1), 21946.

[37] Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D. C., Oktay, O., Wetscherek, M., Langlotz, C., Nori, H. (2023). *BioViL-T: Learning to Exploit*

Temporal Structure for Biomedical Vision-Language Processing. In arXiv preprint arXiv:2301.07867.

[38] Wang, S., Zhang, L., Gu, M., Liu, X. (2023). *R2GenGPT: Radiology Report Generation with frozen LLMs.* In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 12637-12646).

[39] Wu, J., Tamkin, A., Goodman, N., Li, L. (2023). *miniGPT-Med: Large Language Model as a General Interface for Radiology Diagnosis.* In arXiv preprint arXiv:2304.06204.

[40] Thawkar, O., Shaker, A., Mullappilly, S. S., Cholakkal, H., Anwer, R. M., Khan, S., Laaksonen, J., Khan, F. S. (2023). *XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models.* In arXiv preprint arXiv:2306.07971.

[41] Chen, Z., Krishnan, R., Zhao, J. (2023). *HERGen: Elevating Radiology Report Generation with Hierarchical Vision-Language Learning.* In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21842-21851).

[42] Wang, J., Dou, H., Chen, L., Qin, J., Lin, H., Li, Q., Wang, Q., Heng, P. A. (2023). *HistGen: Histopathology Report Generation via Local-Global Feature Encoding and Cross-modal Context Interaction.* In Medical Image Analysis, 89, 102909.

[43] Li, C., Wong, C., Zhang, S., et al. (2023). *LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day.* In arXiv preprint arXiv:2306.00890.

[44] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, A., Arora, A., Rajpurkar, P., Naumann, T., Kundu, A., Tang, Z., Gatidis, S., Fries, J. A., Shah, N. H., Pfeffer, J. (2023). *Med-Flamingo: a Multimodal Medical Few-shot Learner.* In arXiv preprint arXiv:2307.15189.

[45] Li, J., Liu, C., Chen, J., Wang, X., Chen, H., Liang, J., Yuan, Z., Dai, X., Shen, Y., Liu, J., Zhang, Y., Zhou, H., Liu, J., Huang, J. (2023). *Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data.* In arXiv preprint arXiv:2308.02463.

[46] Zhang, S., Xu, Y., Usuyama, N., et al. (2023). *BiomedCLIP: A Foundational Vision-Language Model for Biomedical Image Analysis.* In arXiv preprint arXiv:2303.00915.

[47] Xu, Z., Zhu, F., Cai, B., Delbrouck, J. B., Vatansever, S., Langlotz, C., Yeung, S. (2023). *DeMMo: A Flamingo-based Model for Radiology Report Generation.* In arXiv preprint arXiv:2311.03435.

[48] Chen, J., Lu, H., Zhuge, S., Liebovitz, D., Ahmad, S., Shah, C., Zhou, C., Ahmad, M., Atabansi, C., Nehme, R., Thomas, C., Chao, H., Huang, R., Hao, S., Zheng, K. (2023). *Dual-modality visual feature flow for medical report generation.* In Scientific Reports, 13(1), 8243.

[49] Chen, H., Li, X., Wang, J., Zhang, Y. (2023). *X-ray Made Simple: Radiology Report Generation for General Audience.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12890-12899).

[50] Delbrouck, J. B., Chambon, P., Bluche, T. (2022). *Image-aware Evaluation of Generated Medical Reports.* In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 7184-7196).

[51] Miura, Y., Zhang, Y., Booth, A. D., Langlotz, C. P. (2022). *Human Evaluation of LLM-Generated Medical Reports: Radiologists' Perspectives.* In Journal of Biomedical Informatics, 129, 104062.

[52] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V. (2023). *Large Language Models in the Clinic: A Comprehensive Benchmark.* In Nature Medicine, 29(8), 1930-1940.

[53] Sheng, L., Zhang, T., Dong, H., Zhou, M. (2023). *PeFoMed: Parameter Efficient Fine-tuning of Multimodal Large Language Models for Medical Applications.* In Nature Machine Intelligence, 5(12), 1371-1384.

[54] Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S., Karthikesalingam, A., King, D., Ashrafian, H., Darzi, A. (2021). *A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images.* In ACM Computing Surveys, 54(8), 1-40.

[55] van Ginneken, B. (2017). *Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning.* In Radiological Physics and Technology, 10(1), 23-32.

[56] Liu, G., Hsu, T. M. H., McDermott, M., Boag, W., Weng, W. H., Szolovits, P., Ghassemi, M. (2019). *A survey on automatic generation of medical imaging reports.* In Artificial Intelligence in Medicine, 98, 103-121.

[57] Wang, S., Tang, L., Lin, L., Yue, G., Li, J., Huang, X., Zhang, S. (2023). *Medical Report Generation Is A Multi-label Classification Problem.* In arXiv preprint arXiv:2305.13730.

[58] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., Fu, H. (2023). *Transformers in Medical Imaging: A Survey.* In Medical Image Analysis, 88, 102802.

[59] Chen, Y., Wang, L., Zhang, Y., Liu, M. (2022). *Exploring Transformer Text Generation for Medical Dataset Augmentation.* In IEEE Journal of Biomedical and Health Informatics, 26(8), 4162-4173.

[60] Wang, Z., Yu, L., Wang, S., Liu, X., Zhao, P., Zhang, Y. (2024). *Survey: Vision-Language Models for Medical Report Generation.* In arXiv preprint arXiv:2403.01013.

[61] Puyol-Antón, E., Ruijsink, B., Bai, W., Chen, H., Kerfoot, E., Lourenço, C., Xu, M., de Marvao, A., O'Regan, D. P., Cook, S. A., Rueckert, D., King, A. P. (2021). *CheXclusion: Fairness gaps in deep chest X-ray classifiers.* In PLOS ONE, 16(9), e0256191.

[62] Steinkamp, J. M., Chambers, C., Lalevic, D., Zafar, H. M., Cook, T. S. (2019). *The Impact of AI Assistance on Radiology Reporting: A Pilot Study.* In Journal of Digital Imaging, 32(6), 1044-1048.

[63] Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M. P. (2020). *A Vision-Language Foundation Model to Enhance Clinical Decision Making.* In Nature Medicine, 26(6), 945-950.

[64] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T. Y. (2022). *BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining.* In Briefings in Bioinformatics, 23(6), bbac409.

[65] Tan, M., Le, Q. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* In International Conference on Machine Learning (pp. 6105-6114).

[66] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). *Attention is all you need.* In Advances in neural information processing systems (pp. 5998-6008).

[67] Demner-Fushman, D., et al. (2016). *Preparing a collection of radiology examinations for distribution and retrieval.* Journal of the American Medical Informatics Association, 23(2), 304–310. (IU X-Ray Dataset)

[68] Johnson, A. E., Pollard, T. J., et al. (2019). *MIMIC-CXR: A large publicly available database of labeled chest radiographs*. arXiv preprint arXiv:1901.07042.

[69] National Library of Medicine. (2013). *Open-i "ChestX-ray" Dataset.* https://openi.nlm.nih.gov/

[70] Bustos, A., Pertusa, A., Salinas, J. M., de la Iglesia-Vayá, M. (2020). *PadChest: A large chest x-ray image dataset with multi-label annotated reports*. Medical Image Analysis, 66, 101797.

[71] Irvin, J., Rajpurkar, P., et al. (2019). *CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison.* In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 590–597).

[72] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., Ng, A. Y. (2017). *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning*. arXiv preprint arXiv:1711.05225.

[73] Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D. C., Oktay, O., Wetscherek, M., Langlotz, C., Nori, H. (2023). *MS-CXR-T: Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing*. arXiv preprint arXiv:2301.07867.

[74] Hodosh, M., Young, P., Hockenmaier, J. (2013). *Framing image description as a ranking task: Data, models and evaluation metrics*. Journal of Artificial Intelligence Research, 47, 853-899. (Flickr8K Dataset)

[75] Young, P., Lai, A., Hodosh, M., Hockenmaier, J. (2014). *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.* Transactions of the Association for Computational Linguistics, 2, 67-78. (Flickr30K Dataset)

[76] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L. (2014). *Microsoft COCO: Common objects in context.* In European conference on computer vision (pp. 740-755). (MS COCO Dataset)

[77] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q. (2017). *Densely connected convolutional networks.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[78] Zhang, L., Chen, H., Wang, M., Liu, S. (2024). *Artificial Intelligence in Radiology: A Comprehensive Analysis of Product Clearances and Market Adoption.* PMC Articles, PMC11816879.

[79] Globe Newswire. (2025). *AI in Medical Imaging Market Size Projected to Reach USD 14.46 Bn By 2034.* Retrieved from https://www.globenewswire.com/fr/news-release/2025/02/13/3026027/0/en/AI-in-Medical-Imaging-Market-Size-Projected-to-Reach-USD-14-46-Bn-By-2034

[80] Thompson, K., Rodriguez, M., Patel, A. (2024). *Comparative Analysis of Human Radiologists and AI Models in Medical Imaging Tasks.* PMC Articles, PMC11816879.

[81] Henderson, J. (2025, April 5). *AI and machine learning are transforming radiology software.* The Washington Post. Retrieved from https://www.washingtonpost.com/health/2025/04/05/ai-machine-learning-radiology-software/

[82] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., Ng, A. Y. (2017). *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning.* In arXiv preprint arXiv:1711.05225.

[83] Alammar, J. (2020). *Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention).* In VisualNMT.

[84] Tang, Y., Wang, J., Gao, B., Dellandréa, E., Gaizauskas, R., Chen, L. (2020). *An ablation study on multi-view chest X-ray report generation.* In Journal of Imaging, 6(8), 76.

[85] Wang, X., Zhang, Y., Yang, L., Zhang, Y. (2020). *Deep multimodal fusion: Combining clinical and imaging data for improved prognosis prediction.* In IEEE Transactions on Biomedical Engineering, 67(11), 3214-3225.

[86] Huang, S. C., Shen, L., Lungren, M. P., Yeung, S. (2020). *Clinical knowledge-guided medical report generation.* In Medical Image Analysis, 65, 101832.

[87] Delbrouck, J. B., Chambon, P., Bluche, T. (2022). *A comprehensive framework for automatic evaluation of medical reports*. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 7184-7196).

[88] Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., Mark, R. G., Horng, S. (2019). *MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports*. In Scientific data, 6(1), 317.

[89] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., Ng, A. Y. (2019). *CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 590-597).